Research Article

International Journal of Advanced Technology and Engineering Exploration, Vol 11(113)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2023.10102095

Credibility assessment of social media images shared during disasters

Saima Saleem[1*], Akash Shah[1] and Monica Mehrotra[2]
Senior Research Fellow, Department of Computer Science, Jamia Millia Islamia, New Delhi, India[1]
Professor, Department of Computer Science, Jamia Millia Islamia, New Delhi, India[2]

Received: 03-June-2023; Revised: 15-April-2024; Accepted: 18-April-2024

Abstract

Social media (SM) has emerged as a critical tool in disaster response, offering real-time visual insights through image sharing. This visual information aids responders in assessing the severity of situation and formulating effective strategies. However, the prevalence of forged images on SM poses a significant challenge, potentially misleading responders and hindering the humanitarian efforts. Therefore, it's crucial to verify the credibility of information sourced from SM images before incorporating it into any crucial decision-making process. However, detecting forged disaster images uploaded to SM platforms presents additional challenges. These images undergo various post-processing operations including compression, which introduces additional noise and degrades image quality, thereby complicates forgery detection. This study is the first to focus on SM disaster image Forgery detection. A novel dataset named Forge Disaster is presented, comprising both authentic and forged SM images with copy-move and splicing forgeries. The primary objective of this dataset is to serve as a benchmark for evaluating novel techniques and methodologies in the domain. Additionally, this paper presents a unified approach for robust detection of both copy-move and spliced disaster images on SM. Leveraging image enhancement filters, local binary pattern (LBP) combined with discrete fourier transform (DFT), and support vector machine (SVM), the proposed approach achieved an impressive detection accuracy of 91%, outperforming existing forgery detection methods. These contributions address the growing concern of misinformation through forged images on SM platforms during disaster situations, enhancing the reliability of disaster-related information for effective response.

Keywords

Disaster response, Forgery detection, Social media, Machine learning.

1.Introduction

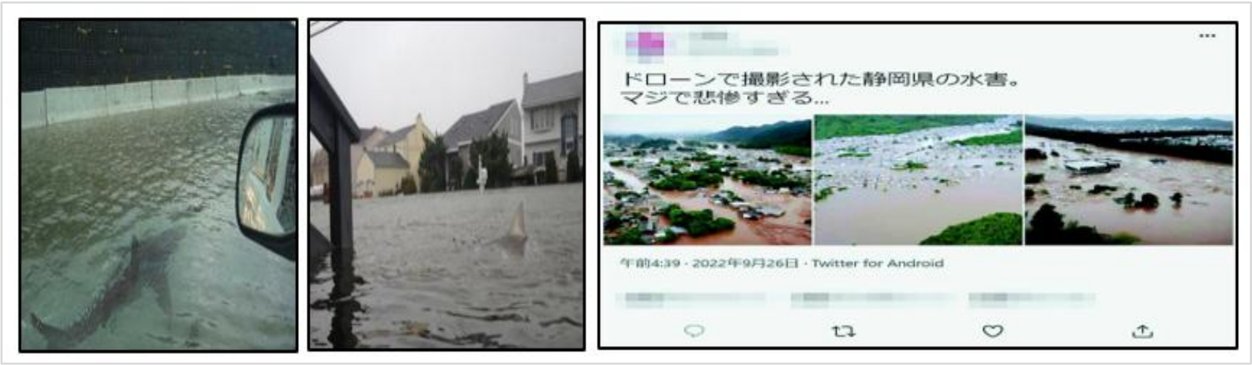
As per the Global Assessment Report (GAR 2022) published by the "United Nations Office for Disaster Risk Reduction" [1], the frequency of disasters has increased significantly over the years and is projected to surge by 40 percent by the year 2030. This increased prevalence of disasters has also led to a rise in both human fatalities and economic ramifications across the globe. Addressing such ecological disruptions necessitates the acquisition of human-centric information for disaster response organizations [2]. In recent years, social media (SM) has emerged as a pertinent alternative information conduit alongside traditional media during disaster events such as floods, earthquakes, hurricanes, etc. [3].Throughout the past decade, various SM platforms, particularly Twitter, have played a significant role in humanitarian response tasks due to their widespread utilization in disseminating information and gathering valuable insights [4].

Recent studies have highlighted the significant role of SM images in the aftermath of disasters. They serve to convey accurate details regarding the extent of devastation [5], aid humanitarian groups in assessing infrastructural harm [6], and expedite the identification and rescue of missing or injured persons [7].

However, despite the potential benefits of integrating SM as an additional resource within established disaster response frameworks, humanitarian authorities raise concerns about the veracity of the data disseminated on such platforms. The credibility of SM-driven information remains a significant issue for both the general public and disaster response agencies [4]. SM platforms have become significant contributors to the propagation of fake news [8, 9] and numerous instances have occurred where forged images have gone viral on SM platforms, especially Twitter, during disasters, creating panic among the population. For instance, during Hurricane Sandy in 2012, Twitter played a crucial role in keeping people

[*] Author for correspondence

informed, however, malevolent users also extensively utilized SM to propagate misinformation and misleading images in real-time [10]. Similarly, during the recent Japan Typhoon, a series of forged images shared on Twitter, purportedly showing flooded homes, quickly gained widespread attention, as reported by the Shizuoka prefectural government (GOV) [11]. The forged images are depicted in *Figure 1 (a), (b), (c),* which rapidly went viral on Twitter during disasters.



ドローンで撮影された静岡県の水害。
マジで悲惨すぎる…

午前4:39 · 2022年9月26日 · Twitter for Android

(a) (b) (c)

Figure 1 Viral forged pictures shared during disasters on Twitter (a) Picture of shark pasted into the flooded street during Hurricane Sandy (b) Another forged picture of a shark on the flooded street during Hurricane Sandy, and (c) Forged photos of flooding during Japan Typhoon on Twitter

With the widespread availability of image-altering software and advancements in artificial intelligence, images can be easily manipulated. To detect image manipulations, image forensics methods are broadly categorized into active and passive, as illustrated in *Figure* 2. Active methods require prior knowledge regarding the evaluated image to function effectively. In these methods, an original image is implanted with a digital signature or digital watermark, which is subsequently removed at the receiver end and compared to the original. Conversely, passive methods [12] are employed in the absence of prior knowledge and rely on the inherent characteristics of the image. Often referred to as blind techniques for detecting forgery, passive methods are more practical and viable. These methods include the detection of copy-move, splicing, and other forgeries. Copy-move forgery is a deceptive technique where a section of an image is duplicated and then inserted into another area of the same image, either to hide or duplicate specific elements. While image splicing involves generating a composite image by combining parts of different images.

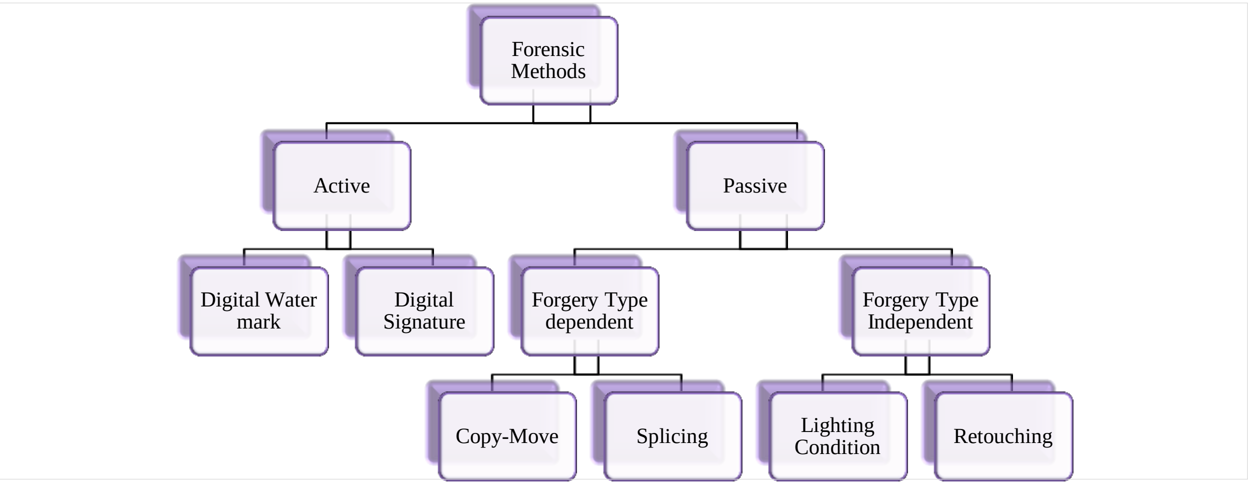


Figure 2 Forensic methods

Copy-move and splicing are recognized as the most prevalent and harmful digital image forgeries, often created for spreading disinformation [13, 14]. Evaluating the credibility of a digital image is currently one of the most pressing issues. Forged images shared in disasters can provide inaccurate depictions of the disaster's impact, hindering effective decision-making, misguided resource allocation, delayed responses, and compromised situational awareness. This can result in increased risks for both responders and affected communities. Therefore, assessing the credibility of information gleaned from SM images is crucial, yet remains largely underexplored in the realm of disaster response.

Moreover, the detection of forged images over SM platforms presents additional challenges. As images traverse these platforms, they undergo compression to accommodate server limitations, both during upload and download processes. Additionally, SM platforms often employ multilevel quality compression techniques customized to their needs, enabling rapid data transmission for swift image sharing. However, compression introduces blurring, noise and can mask manipulation artifacts, thus hindering accurate detection [15–17]. Therefore, it is necessary to build specific methods to detect altered images shared in disasters over SM platforms. As far as current knowledge suggests, there hasn't been any research conducted specifically to detect forged disaster images on SM. There is no benchmark dataset for forged disaster image detection either. Therefore, the objective of this study is twofold: first, to introduce a novel ForgeDisaster SM disaster image forgery dataset, and second, to present a unified approach for detecting both copy-move and splicing forgeries on SM disaster images.

The dataset is designed to serve as a benchmark for evaluating novel techniques and methodologies in the domain. It comprises balanced sets of authentic and forged SM disaster images with copy-move and splicing forgeries. Furthermore, the approach offers a promising solution to the challenges posed by manipulated content on SM platforms during disaster situations. The proposed approach encompasses three key phases: pre-processing, feature extraction, and classification. Firstly, the pre-processing stage aims to enhance the quality of low-quality images by leveraging enhancement filters within the YCbCr color space. This step not only mitigates the effects of compression but also facilitates the identification of key features by reducing noise and blurring. Subsequently, in the feature extraction stage, a robust combination of local binary pattern (LBP) and discrete fourier transform (DFT) is employed. LBP excels at capturing local texture patterns, making tampering artifacts more discernible, while DFT transforms the LBP image from the spatial domain to the frequency domain. This transformation enhances the detection of frequency fluctuations caused by these artifacts, leveraging standard deviation (STD) to further refine the analysis. Lastly, the classification stage utilizes support vector machine (SVM) to classify the STD based features, enabling accurate identification of forged and authentic images on SM during disasters. This comprehensive approach aims to establish a reliable framework for detecting forged disaster images, thus enhancing the integrity of information dissemination during critical situations on SM.

The following list outlines the contributions of the proposed work:

- A novel ForgeDisaster dataset for SM image forgery detection is developed using copy-move and splicing forgery methods, ForgeDisaster dataset addresses the lack of a benchmark dataset for disaster domain SM image forgery detection research.
- This paper provides benchmark results for SM image forgery detection in the disaster domain by presenting a new approach for identifying both copy-move and splicing forgeries, utilizing image enhancement filters in YCbCr color space for pre-processing, LBP and DFT for feature extraction and SVM for classification.
- The proposed approach is juxtaposed against existing general domain forgery detection methods, demonstrating superior performance when evaluated using the ForgeDisaster dataset.
- In addition, various pre-trained deep learning models are employed and evaluated using Forge Disaster dataset to provide results that can be used as a baseline for future deep learning solutions.

Following is the arrangement of the remaining portions of the paper. The related literature is presented in section 2. Section 3 covers the materials and methods. Section 4 presents the results of various experiments. The analysis of results is discussed in section 5. Lastly, section 6 serves to conclude the paper by outlining future directions for further exploration.

2.Literature review

As the current study focuses on detecting SM image forgery within the context of disasters, the related literature is presented in two distinct sections: (i) SM disaster image analysis, and (ii) Image forgery detection methods.

2.1SM disaster image analysis

During disasters, SM content has been shown to be beneficial in assisting various stakeholders, particularly humanitarian organizations [18]. There has also been a rise in efforts to create automated methods and systems for SM content analysis and the extraction of useful insights, mostly driven by Twitter and Facebook content. The use of Twitter content has continued to be more popular than Facebook because of its rapid access to timely multi-modal content, which is critical for GOV and non-government organizations (NGOs) [19]. The majority of earlier studies in the disaster domain have centered on the analysis of textual content [5]. However, there has recently been a surge in interest in visual data analysis due to the greater role that images play in disaster management tasks as per various studies [20, 21]. A number of disaster-related SM image datasets such as DAD [5], CrisisMMD [22], and MEDIC [19] are publicly available and used for developing automated methods for various disaster management tasks.

Nguyen et al. [5] conducted an analysis of images posted on SM platforms during natural disasters to gauge the severity of damage caused. They employed a fine-tuned deep convolutional neural network (CNN) model to classify the images into three categories: severe damage, mild damage, and no damage. It achieved an accuracy of up to 90%. Similarly, in [23] damage severity assessment was conducted from SM images during disasters using visual geometry group (VGG) model. However, the credibility of the images was not verified before assessing the damage depicted in the disaster images. Ning et al. [24] proposed a flood detection system based on SM images. They iteratively trained a CNN model and achieved a 93% accuracy in detecting flood-related images. However, the credibility of the SM flood images was not assessed in their study.

Hassan et al. [25] presented a deep visual sentiment analyzer for disaster-related images, employing CNN and transfer learning techniques. While their work could potentially aid responders in analyzing the emotions of affected individuals based on image content, a significant limitation arises from the lack of consideration for image authenticity.

Kotha et al. [26] proposed a model based on RegNetY320 for categorizing SM images into humanitarian information classes. Their model attained an accuracy of 80.20%. These works hold significant potential for humanitarian aid workers, but the authenticity of images has not been verified in these studies which can have serious repercussions on response efforts, public trust, and the well-being of affected communities.

2.2Image forgery detection methods

For detecting forgery in images, numerous passive methods have been developed over the years by researchers utilizing different image forgery datasets such as CASIAv1.0 and CASIAv2.0 [27], COLUMNBIA [28], MICC-F220, MICC-F2000 [29], and CoMoFod [30]. An approach for detecting forgery based on the fusion of LBP, discrete wavelet transforms (DWT), and principal component analysis has been proposed in [31]. The integration of LBP and DWT yielded a notable enhancement in accuracy. It achieved an accuracy rate of 97.21% on CASIAv1.0. and 95.13% on Columbia dataset. However, this method was limited to detecting splicing forgery exclusively.

Another method [32] developed for tamper detection, utilized DWT with histograms of LBP to detect spliced images. The feature vector was formed by combining LBP histogram from the four wavelet sub bands. To determine the accuracy, SVM was used on 10 folds which attained an accuracy of 96.62% on CASIAv1.0, 94.04% on CASIAv2.0 and 87.05% on Columbia datasets. However, the performance was impacted by the small size of images, and this approach may not work well for detecting copy-move forgery.

Hayat and Qazi [33] presented an approach centered on the fusion of DWT and discrete cosine transform (DCT) for copy-move forgery detection. They started by obtaining the DWT approximation sub band, and then they applied DCT to the overlapping picture blocks. In order to make a more accurate comparison of the blocks, additional correlation coefficients were utilized. Their approach achieved an accuracy of 73.62% for detecting images tampered with copy-move. In addition to having low accuracy, this method may underperform in scenarios involving occlusion and images with repeated patterns and spliced regions.

In [34], a multi-scale LBP and DCT was utilized for tamper detection. This method involved computing multi-scale-LBP, where each pixel had multiple LBP codes and passing these multi-scale-LBP representations to DCT to extract coefficients. To construct a feature set for the image, STD was computed with respect to the coefficients. This approach achieved an accuracy of up to 97.3%. However, it was limited to detecting splicing and may not be effective for other types of tampering.

In a separate study, Parnak et al. [35] focused on splicing detection. They proposed a novel feature set derived from the mantissa distribution of DCT coefficients in images, aiming to enhance detection performance. The approach demonstrated exceptional accuracy, achieving a remarkable 99.78% accuracy when evaluated on the CASIAv1.0 dataset. However, their approach was specific to detecting splicing forgery only.

Copy-move detection was addressed in [36], wherein DCT coefficients were utilized as features for blocks of various sizes. The process involved transforming the images from red green blue (RGB) to grayscale and dividing them into overlapping blocks, from which DCT coefficients were computed. These 2-dimensional (2D) coefficients were then rearranged into a feature vector using zig-zag scanning. Subsequently, all blocks were sorted using lexicographic order, facilitating the identification of duplicated blocks through Euclidean Distance computation. The study demonstrated that employing 8 × 8 overlapping blocks yielded superior performance in terms of precision and recall for forged detection. However, performing post-processing procedures on tampered images resulted in significant computational complexity and inaccuracies in detecting tampered regions.

For detecting both copy-move and splicing image forgery, Alahmadi et al. [14] achieved promising outcomes by employing a combination of LBP and DCT. Initially, LBP was applied to blocks extracted from the picture chroma channel, followed by the application of DCT on these blocks. They then evaluated a feature for each DCT coefficient, calculated as the STD of DCT coefficients contained within each block. This method yielded impressive accuracy rates, with 97% on CASIAv1.0, 97.50% on CASIAv2.0, and 97.77% on the Columbia dataset. Conversely, in [37], Islam et al. developed a forgery detection system that initially applied DCT followed by LBP. They computed the mean value of all LBP blocks to obtain a fixed number of features. Their method demonstrated promising results for detecting both copy-move and splicing forgeries, achieving an accuracy of 99.55% on CASIAv1.0, 99.88% on CASIAv2.0, and 98.20% on the Columbia dataset. Additionally, they conducted experiments with internet of things data, attaining good performance in this domain as well. However, these methods were not tested with SM disaster images.

Similarly, Dua et al. [38] presented a forgery detection system to detect both forgeries. They took advantage of the diversity in statistical features of the overall image's AC coefficients by calculating the STD and count of non-zero DCT coefficients with respect to every AC frequency component individually. The proposed features were examined for the test image and its cropped counterpart. The retrieved features then utilized in conjunction with the SVM to distinguish between tempered and authentic photos with an accuracy of 93.2% on CASIAv2.0 and 98.3% on CASIAv2.0 datasets.

Recently, various advanced models built on deep-learning methods that automatically learn features have been investigated. More specifically, CNN is being used increasingly in recent methods for feature extraction and categorization. Xiao et al. [39] provided a forgery detection method combining a "coarse-to-refined CNN" and "diluted adaptive clustering". Their approach used CNNs to obtain differences in image properties between tampered and untampered regions across varying scales. After identifying suspicious regions, forgery regions were generated using adaptive clustering. The method achieved an F1 score of up to 69.5%. However, it's important to note that the approach was specific to detecting splicing forgery and may exhibit longer runtime due to its sophisticated detection mechanisms.

In another work, Qazi et al. [40] introduced an approach for splicing detection based on the ResNet50v2 architecture. In this approach, image batches were inputted, and the weights of a "you only look once" CNN were utilized through the architecture of ResNet50v2. This method achieved a high accuracy of 99.3% on the CASIAv2.0 dataset. However, it's important to note that this method was specifically tailored for detecting splicing forgery and may not be suitable for other types of forgery detection.

Abdalla et al. [41] utilized a fusion processing model for forgery detection, which combined a deep convolutional model with an adversarial model. Operating on a two-branch architecture alongside a fusion module, the method employed CNN and generative adversarial network to pinpoint and characterize copy-move regions. Performance results fall within the range of 93% to 97%. However, the effectiveness of the method was contingent upon the configuration of its parameters.

Goel et al. [42] presented an algorithm for copy-move detection employing a novel dual-branch CNN. This CNN architecture obtained multi-scale features through varying kernel sizes in each branch. Subsequently, the multi-scale features were fused to enhance accuracy. Remarkably, the proposed approach achieved a commendable accuracy of 96% on the MICC F-2000 dataset. However, the robustness of the model wasn't tested on other types of forgeries.

In [43], an end-to-end dual-channel U-Net model was introduced for detecting and locating splicing forgery. High-pass filters were used to get residual information, capturing tampered area edges. The model then fused deep features from original and residual images, then extracted tampered features with varying granularity and performs secondary fusion. This approach achieved 97.93% accuracy on CASIAv2.0 and 97.27% on Columbia datasets, however, it was limited to fixed-size images and specialized for splicing forgery detection only.

Walia et al. [44] introduced a feature fusion approach to detect forgeries in digital images. In this approach handcrafted features and deep high-level features were amalgamated to form a comprehensive feature vector. This vector was then employed for classification using a shallow neural network. With a high accuracy of 99.3% on CASIAv1.0 and 97.94% on CASIAv2.0 datasets, this proposed approach demonstrated promise for offline forensic digital image analysis. However, the high dimensionality of the fused features posed a challenge for real-time analysis, serving as a potential bottleneck in practical implementation.

Sabeena and Abraham [45] presented a forgery detection approach centered on a convolutional attention-based model. This study employed the "convolutional block attention module" for feature extraction, image segmentation, and forgery localization. By leveraging spatial and channel attention features fused through the convolutional block attention mechanism, comprehensive context information was captured, enriching feature representation. The approach achieved an impressive 99.6% accuracy for copy-move detection on the ComoFod dataset. However, the method lacks in generalizability to other types of forgeries.

Vijayalakshmi et al. [46] presented a copy-paste detection approach, comprising three key operations: "pre-processing", "image augmentation", and "classification". In the pre-processing phase, tasks such as image normalization, rescaling, and error level analysis (ELA) were conducted to enhance accuracy performance. Image augmentation techniques were then employed to expand the dataset size. Finally, a convolutional autoencoder-based technique was provided for forgery classification. The method achieved an impressive accuracy of 99.2%. However, this approach was specifically designed for detecting copy-paste image forgeries and may experience reduced performance when images are subjected to high noise.

Ali et al. [47] presented a custom model based on CNN consisting of three convolutional layers and a dense fully connected layer. They used ELA as an input to CNN for detecting both splicing and copy-move manipulations and achieved a validation accuracy of 92.23% on CASIA.2.0 dataset.

The existing literature on disaster management utilizing SM has explored various facets, yet a significant gap persists: the absence of systematic assessment of image credibility from SM platforms before their integration into disaster-related tasks. Despite the availability of numerous SM image datasets for disaster management, the lack of ground truth labels regarding image credibility impedes their utility in assessing image authenticity. Based on the analyzed literature on image forgery detection, numerous image forgery detection methods exist, however, none have been evaluated on SM disaster images. The unique challenges posed by SM platforms, such as compression-induced noise and degradation, complicate SM image forgery detection, necessitating specific approaches. Moreover, majority of the approaches focus on either copy-move or splicing forgery types.

In response to these challenges, this study addresses critical gaps by systematically evaluating the credibility of images sourced from SM platforms before their integration into disaster-related tasks. A labeled dataset addressing the credibility of SM disaster images is provided, overcoming limitations in existing SM image datasets. Moreover, the challenges posed by forged images on SM platforms are recognized, and a specialized approach is offered, tailored to accommodate these limitations. Unlike existing forgery detection methods that typically focus on either copy-move or splicing forgery, the proposed approach provides comprehensive coverage for both types of forgery, aiming to enhance the credibility of SM images in disaster response contexts.

3.Materials and methods

3.1Experimental dataset

This section introduces the novel ForgeDisaster dataset. The creation of this dataset involved three stages described below. The overall dataset creation process is shown in *Figure* 3.

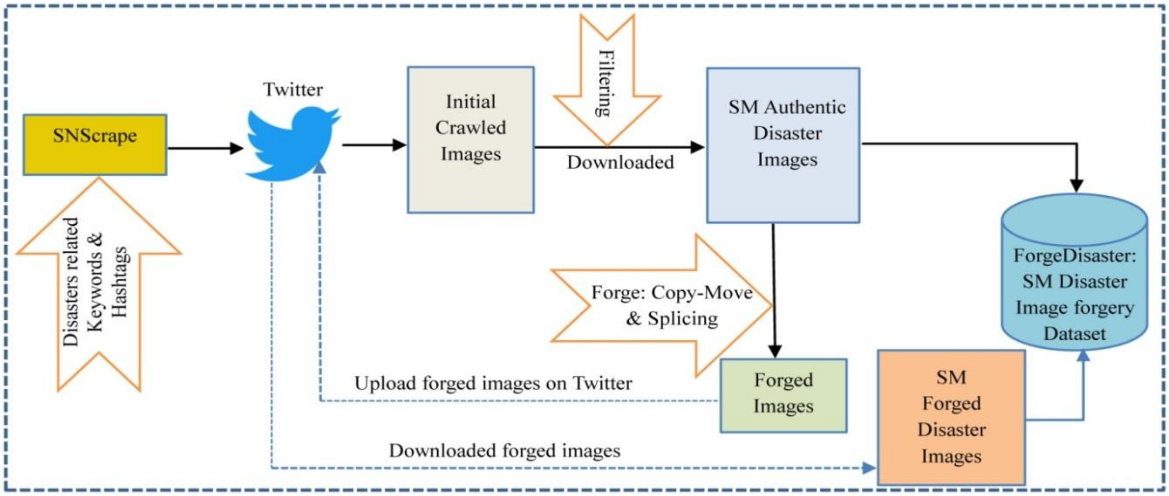

SNScrape
Disasters related Keywords & Hashtags
Twitter
Initial Crawled Images
Filtering
Downloaded
SM Authentic Disaster Images
ForgeDisaster: SM Disaster Image forgery Dataset
Forge: Copy-Move & Splicing
Upload forged images on Twitter
Forged Images
SM Forged Disaster Images
Downloaded forged images

Figure 3 ForgeDisaster dataset creation process

3.1.1Dataset collection

The images were collected from the Twitter SM platform that were posted during different natural disasters like floods, earthquakes, hurricanes, droughts, tsunamis, wildfires, cyclones, and tornados occurred in different places of the world using the SnScrape Twitter scrapping tool. The images were extracted based on various hashtags and keywords that involved words/phrases about disasters.

3.1.2Dataset filtering

The quality of the training data has a substantial effect on how well any detection system can classify forged content. For this reason, images from reliable resources “from the official Twitter accounts of GOV, NGOs, public figures and news channels” were deemed authentic/genuine, and the rest of the images were discarded. The rationale behind this selection is that such organizations publish content that either combats fake news or is inherently genuine. To further ensure the veracity of the collected images, additional validation measures were implemented. Specifically, cross-referencing was conducted using various fact-checking sites such as Boomlive[1], Snopes[2], and PolitiFact[3]. This thorough approach aimed to verify the credibility of the images, thereby further strengthening the reliability of the dataset.

Besides, duplicate images and those unrelated to disasters despite the posts containing disaster-related hashtags were also discarded at this stage. Furthermore, formats of the images were also standardized to a single JPEG format as few images were in PNG format. Finally, a set of 740 genuine JPEG images were obtained, having a varying resolution in the range of 235 × 156 to 4096 × 3139 as shown in *Figure 4*. The majority of images were related to flood disaster, sourced from NGOs (such as UNOCHA, UNICEF, Red Cross, UNDP, etc.) followed by news channels. The number of authentic images according to image source, and type of disaster is depicted in *Table 1* and *Figure* 5 *(a),* and *(b)* show their distribution using charts.

[1] https://www.boomlive.in/
[2] https://www.snopes.com/
[3] https://www.politifact.com/

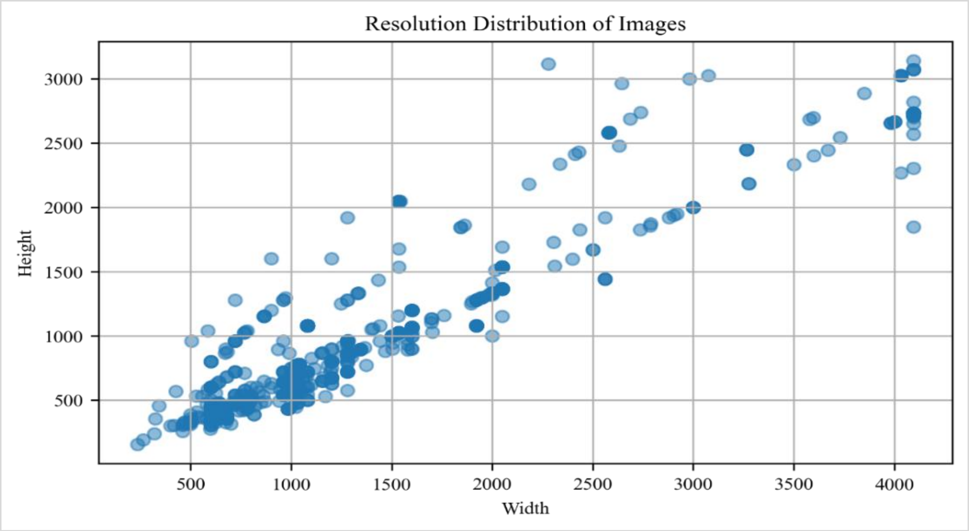
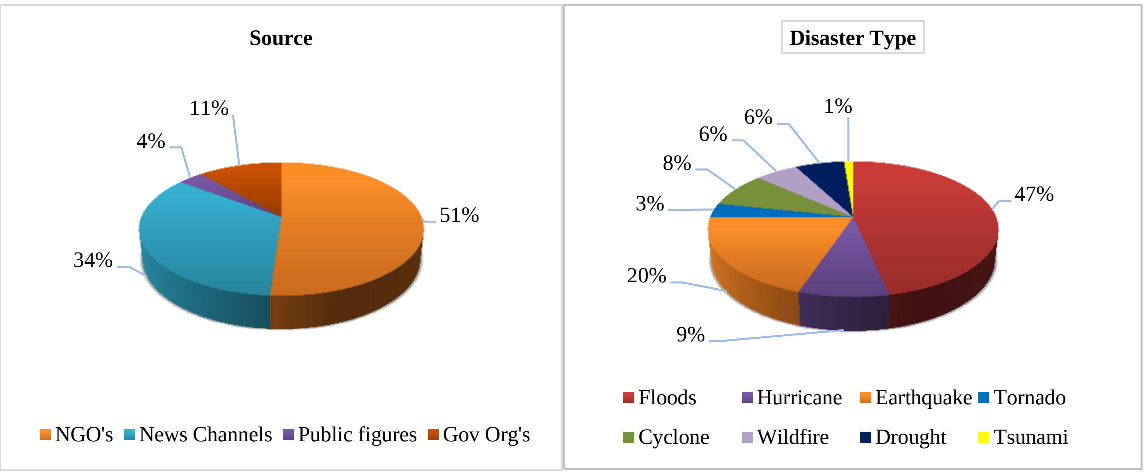


Figure 4 Resolution distribution of authentic images

Table 1 Number of authentic images as per source, and disaster type

	Type	**#images**
Source	NGO’s	378
	News Channel	253
	Gov Org	84
	Public figure	25
Disaster	Floods	345
	Earthquake	147
	Hurricane	63
	Cyclone	61
	Drought	48
	Wildfire	42
	Tornado	25
	Tsunami	9

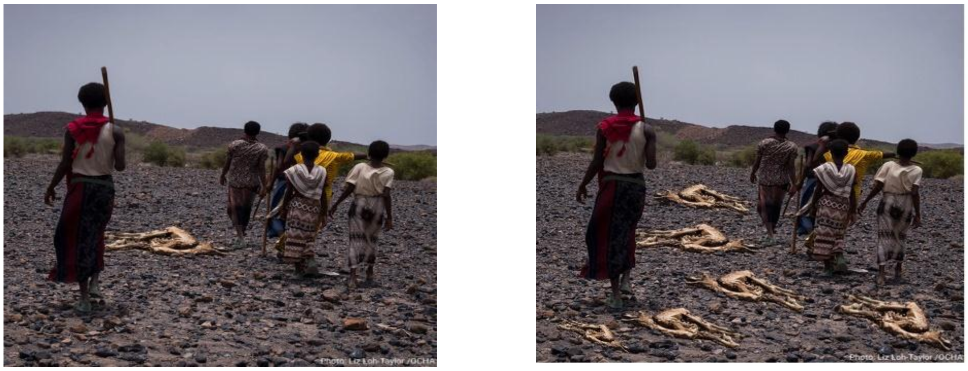


(a) (b)

Figure 5 Percentage of authentic images according to (a) Image Source, and (b) Disaster type

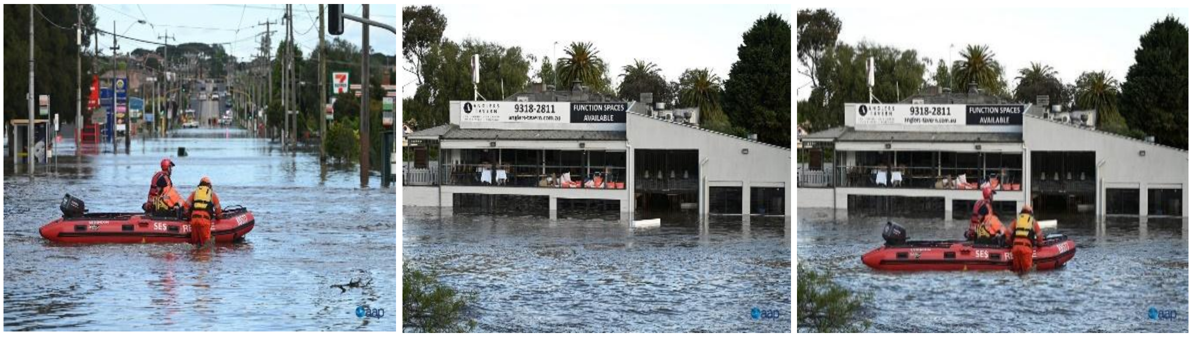
3.1.3 Dataset preparation

The set of genuine images were forged using two forgery methods: copy-move and splicing. Copy-move forgery entails duplicating a portion of an image and inserting it elsewhere within the same image either to conceal or replicate certain elements. An example illustrating a copy-move forgery is depicted in *Figure* 6. The original image in *Figure 6 (a)* depicts the Ethiopia drought, capturing the landscape and relevant elements within the scene. The copy-move version in *Figure 6 (b)* is created by selecting a specific region of the original image, such as an animal corpse, and duplicating it by pasting it onto another area within the same image. This results in the appearance of multiple instances of the selected object within the same scene. Conversely, splicing forgery involves combining parts of multiple images to create a composite image. An example illustrating a splicing forgery is depicted in *Figure 7*. The splicing forgery was executed through a multi-step process aimed at seamlessly integrating elements from one image into another. Initially, image (a) depicting the Victoria floods was selected as the source for the manipulated section. A portion of this image, specifically the rescue boat, was carefully extracted using image editing software. Subsequently, this extracted segment was superimposed onto the image (b). To seamlessly integrate the pasted regions with the background, operations such as cropping, resizing, rotating, brightness adjustment, color enhancement, etc., were applied achieving an illusion of a cohesive and natural scene. The forging process utilized various image editing software, including professional tools such as Adobe Photoshop and basic ones like Photpea and MS Paint. The majority of images underwent tampering via Photoshop, followed by Photopea. The images were then uploaded and downloaded on Twitter. The class distribution of ForgeDisaster is shown in *Figure 8*.



(a) (b)

Figure 6 Copy-move forgery example: (a) Original image depicting Ethiopia drought, b) Copy-move version created by pasting an animal corpse within the image on to the same image



(a) (b) (c)

Figure 7 Splicing forgery example: (a) and (b) represent original images of Victoria floods, while (c) depicts the spliced image, created by pasting the rescue boat part from image (a) on to image (b)

3.2 Proposed approach

The proposed approach is structured into three main phases: pre-processing, feature extraction, and classification as shown in *Figure 9*. In the subsections that follow, an in-depth explanation of each one of these phases is given.

3.2.1 Pre-processing

The primary goal of the pre-processing phase was to improve the quality of SM disaster images. Initially, all images underwent a color space transformation into the YCbCr color system. The YCbCr representation was obtained from RGB colorspace, as depicted in the Equations 1, 2, and 3. This transformation separated the image into its luminance (Y) and chrominance (Cb and Cr) channels. By isolating luminance from chrominance, a more precise analysis of the image can be performed. Moreover, working in the YCbCr color space reduced computational requirements, facilitating efficient image processing.

$$Y = 16 + \frac{65.738\ R}{256} + \frac{129.057\ G}{256} + \frac{25.064\ B}{256} \quad (1)$$

$$Cb = 128 - \frac{37.945\ R}{256} - \frac{74.494\ G}{256} + \frac{112.439\ B}{256} \quad (2)$$

$$Cr = 128 + \frac{112.439\ R}{256} - \frac{94.154\ G}{256} - \frac{18.285\ B}{256} \quad (3)$$

Given that images on SM are typically compressed to varying degrees depending on the platform's requirements, this compression can introduce unwanted noise and blurring by discarding portions of the original image data to reduce file size, consequently leading to a degradation in image quality. To mitigate this issue, the pre-processing phase employed enhancement filters such as non-local means (NLM) and bilateral filtering for low quality images with a compression quality of less than 90%. Images having higher compression quality exhibited satisfactory clarity and detail, reducing the necessity for additional enhancement. The filters effectively reduced noise and blur thereby improved the clarity of key features in the images. The NLM filter was chosen as it excels in suppressing noise while preserving textures and details, while the bilateral filter is known for its ability to retain image edges while smoothing overall image appearance. Following noise reduction, subsequent processing was carried out on the chrominance channels, particularly (Cr) component, as it was found to give maximum accuracy after experimental verification.

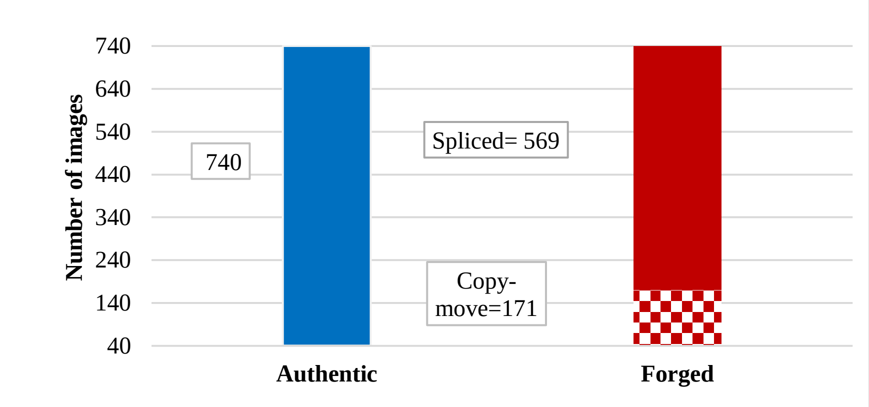


Figure 8 Class distribution of authentic and forged images: 740 authentic and 740 forged JPEG images

3.2.2 Feature extraction

In this phase, the motive was to extract features that were sensitive enough to spot image manipulations like copy-move and splicing. Splicing and copy-move operations disrupt the fine boundary of tampered regions, leading to structural changes in the image. Consequently, the local frequency distribution within the forged region is altered, and the correlation among pixels is disturbed. These operations specifically affect the continuity of host image pixels, particularly around the edges of the forged region. Therefore, it's important to record any structural alterations that may have occurred in an image due to tampering. LBP and DFT techniques were employed to model the tampering traces in images.

LBP

LBP was used to accentuate the subtle alterations introduced by tampering in images. The Cr components extracted in the previous step were divided into 16×16 non-overlapping blocks. Breaking down the image into smaller blocks allows for a more focused analysis of specific regions and is computationally less expensive compared to processing the entire image at once. LBP values were then computed for each block to effectively highlight the tempering artifacts. LBP is a powerful technique for extracting texture information, revealing patterns that may indicate manipulation. It achieved this by analyzing the intensity values of each pixel in relation to those of its neighbouring pixels, generating a binary value for each pixel based on these intensity variations. This binary code is then transformed into a decimal value, which represents the texture of that pixel's neighbourhood. Equation 4 and 5 are utilized in the calculation of the LBP.

$$\mathrm{LBP}_{p,r} = \sum_{n=0}^{p-1} S(g_n - g_c) \times 2^n \qquad (4)$$

Where g_c is the central pixel value, g_n denotes the values of the nearby pixels, p represents the neighbourhood pixel count, and r represents the distance from the central pixel to the surrounding pixels. The threshold function S(z), where z is $g_n - g_c$is defined as:

$$S(Z) = \begin{cases} 0, & z < 0 \\ 1, & z \geq 0 \end{cases} \qquad (5)$$

The number of neighbours was set to eight (p= 8) with radius, r = 1 to compute LBP in the proposed method. If the intensity of the central pixel g_c was higher than that of its neighboring pixels, a value of 0 was assigned; otherwise, a value of 1 was assigned. Therefore, each central pixel value was assigned an 8-bit binary code, which was further converted into its respective decimal value. This resulting decimal value served as the LBP code for the central pixel, and this computation was repeated for every pixel.

DFT

To track the shifts in the LBP blocks local frequency distribution, DFT was employed to convert the LBP blocks from spatial domain to the frequency domain, and then statistical measures of the individual DFT coefficients for each block were computed. To efficiently compute the DFT, fast fourier transform (FFT) was utilized. In terms of calculation time, FFT is more efficient; it reduces the number of operations for a problem of size N from O(N^2) to (NlogN), which makes it more practical for real-time applications. In the spatial domain, the images are processed in their raw form. The pixel values change depending on the scene in the image. In contrast, in the frequency domain, the rate of change of spatial pixel values is examined. Tamper detection relies heavily on the rate of change of spatial pixels, and the frequency domain is useful for addressing the issue related to this. The output of the DFT algorithm was a complex number array, which was extremely challenging to directly visualize. For visualization purposes, it was transformed into a 2D space called a spectrum image. *Figure 10* shows an example of spectrum image of a forged image from the dataset. Each point in the spectrum image denotes a specific frequency in the spatial domain image. The central values around the origin (middle) of the spectrum are the DC components. These coefficients describe the regions of the image that have low frequencies or are smooth. Components with higher frequencies that are dispersed across the spectrum fill in the detail and edges. Each LBP block was converted into the frequency domain with 2D DFT in order to capture shifts in the local frequency distribution. The 2D DFT of an input block was transmuted to discrete Fourier coefficients by formulation in Equation 6. This formula defines the DFT of an n×n matrix.

$$F(x,y) = \sum_{u=0}^{n-1} \sum_{v=0}^{n-1} f(u,v) e^{-i2\pi\left(\frac{xu}{n}\right)\left(\frac{yv}{n}\right)} \quad (6)$$

An image is broken down into its sine and cosine components, as demonstrated by the Equation 7.

$$F(x,y) = \sum_{u=0}^{n-1} \sum_{v=0}^{n-1} f(u,v) \left[\cos\frac{2\pi xu}{n} - i\sin\frac{2\pi yv}{n}\right] \qquad (7)$$

Where u and v are pixel locations in the spatial domain, x, and y are pixel locations in the frequency domain. In Fourier space, the exponential term corresponds to the basis function for every point F(x,y). The equation is interpreted as follows: for each point F(x,y), the value is derived by multiplying the spatial image by the associated base function and adding the resulting product. The base functions are sine and cosine waves with increasing frequencies. The last thing to do was to compute the STD of every individual DFT coefficients of all blocks and input these features to the classifier for classification. The STD provides a measure of the variability or spread of pixel intensities within each block in the frequency domain. By computing the STD across DFT blocks, regions where manipulation has occurred can be identified. For example, areas of the image that have been tampered with may exhibit higher variability in pixel intensities compared to the surrounding authentic regions. This increased variability is indicative of the alterations introduced by the manipulation process, which may include the addition or removal of content, resulting in changes to the frequency distribution.

3.2.3 Classification

Detecting forged images is a binary classification problem that falls into two main categories (i.e., authentic vs. forged). After the features were acquired, they were input to an SVM classifier, which subsequently categorized the features as Authentic or Forged. The popular radial basis function was used with SVM for the classification. Algorithm 1 describes the proposed method in pseudo-code form.

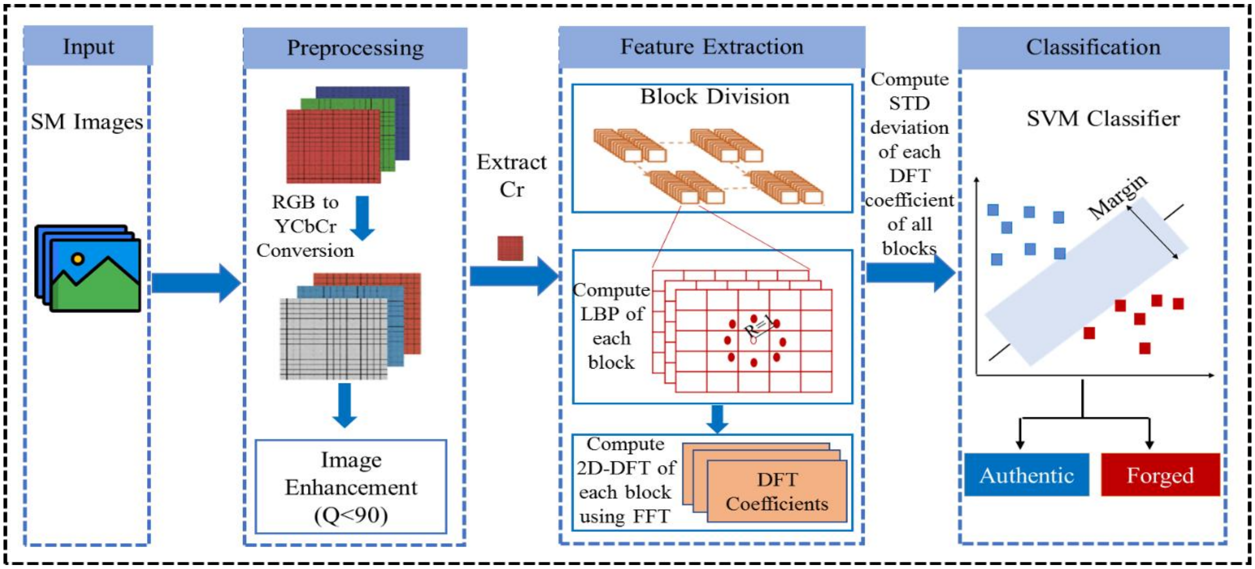


Figure 9 Proposed Approach: it begins with a pre-processing phase, wherein all images are converted into the YCbCr color space. Subsequently, images with compression quality below 90% (Q<90%) undergo enhancement. Then feature extraction is done using the Cr component images, which are divided into blocks. For each block, LBP computation is performed followed by 2D-DFT. The STD of each frequency coefficient in the DFT is calculated which are subsequently utilized for classification by the SVM

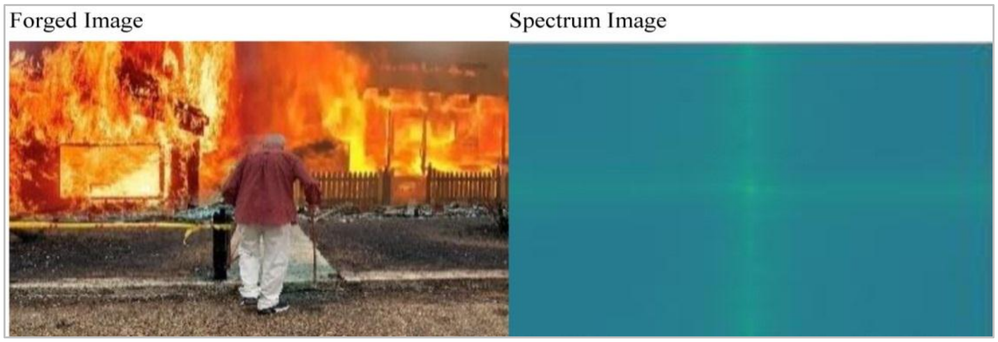


Figure 10 Example of spectrum image in frequency domain

Algorithm 1: Classification of authentic and forged disaster images

```
Input:
    Images: A set of N images ($M_1$, $M_2$, …$M_N$)
    Parameters: Block_Size = (16,16), LBP parameters P= 8, R=1
Output: Classification result (Authentic or Forged)
Procedure:
  for each image $M_i$ ($i$: $1 \rightarrow N$):
      Pre-process image $M_i$
      Divide the image $M_i$ into non-overlapping blocks of Block_Size
      for each block $b$ in image $M_i$:
          Compute LBP of block b: $lbp_b = LBP(b, P, R)$
          Compute DFT of LBP block: $dft = FFT(lbp_b)$
          Append the DFT block to an array $dft\_blocks \leftarrow dft$
      end for
```

```
        Compute the STD of each DFT coefficient of all blocks:
                $std = STD(dft\_blocks)$
        Append the computed standard deviations as a feature vector to an array:
                $X\_feature \leftarrow std$
    end for
    Use the final $X\_feature$ array as input to SVM for detection of authentic or forged images:
        $X\_feature = \begin{bmatrix} [0][0] & . \; . & [0][256] \\ . & . \; . & . \\ . & . \; . & . \\ . & . \; . & . \\ [N][0] & . \; . & [N][256] \end{bmatrix}$
        $predictions = predict(SVM, X\_feature)$
    end procedure
```

Hyperparameter setting

To discover best values for hyperparameters of the SVM model, Bayesian optimization [48] was employed. Bayesian optimization is an informed search method. It has the ability to intelligently navigate the hyperparameter space, leveraging information from previous iterations to guide the search towards optimal values. For the SVM model, the two crucial hyperparameters considered are C and gamma. The parameter C regulates the balance between achieving a smooth decision boundary and accurately classifying training points, whereas gamma impacts the shape of the decision boundary. To determine the best values for C and gamma, a search space was specified encompassing various candidate values. Specifically, for C, a range from 0.1 to 50 was explored, and for gamma, values ranging from 0.01 to 9 were considered. Throughout the Bayesian Optimization process, the algorithm iteratively evaluated different combinations of hyperparameters, guided by the goal of maximizing the model's performance on the given task. By leveraging this informed search strategy, the combination of C=40 and gamma=0.05 were identified as yielding the best performance for the current task.

3.3 Experimental setup

This section provides details about the hardware and software requirements and metrics used for evaluation.

3.3.1Hardware and software requirements

All experiments were carried out on a computing environment equipped with an Intel Xeon processor and 32GB of random-access memory utilizing Python programming language (version 3.7.13) within the Jupyter Notebook environment. Anaconda software was also employed for managing the experimental setup. The experiments made use of various libraries including Pandas, Matplotlib, NumPy, os, OpenCV, PIL, scikit-learn, skimage, and SciPy. For implementing machine learning models such as random forest (RF), naïve bayes (NB), decision tree (DT), eXtreme gradient boosting (XGB), and SVM, the popular scikit-learn library was utilized. In addition, for implementing deep learning models, libraries such as Keras and TensorFlow were employed. These libraries provided the necessary tools and frameworks for developing and training neural network architectures.

3.3.2 Evaluation metrics

The description of the evaluation metrics used is provided below.

Accuracy: is the proportion of accurately predicted images to all images and is defined by Equation 8.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \qquad (8)$$

Sensitivity: is also called as recall is the ratio of accurately classified forged images to all forged images and is defined by Equation 9.

$$Sensitivity = \frac{TP}{TP+FN} \qquad (9)$$

Specificity: is the proportion of accurately classified authentic images to all authentic images and is defined by Equation 10.

$$Specificity = \frac{TN}{TN+FP} \qquad (10)$$

False negative rate (FNR): is the proportion of forged images being incorrectly predicted as authentic and is defined by Equation 11.

$$FNR = \frac{FN}{TP+FN} \qquad (11)$$

False positive rate (FPR): is the proportion of authentic images being incorrectly predicted as forged and is defined by Equation 12.

$$FPR = \frac{FP}{FP+TN} \qquad (12)$$

Area under the curve of "receiver operating characteristic" curve (AUC-ROC): illustrates the efficacy of a classifier by comparing the "true positive rate" with the "FPR" at different thresholds of the classifier outcome. An AUC-ROC=1 represents the perfect model performance.

4.Results

The results of all the experiments conducted in this study along with a comparison of results with the existing methods are provided in this section.

4.1Performance of the proposed approach

Different classification models including SVM, RF, DT, XGB, and NB were evaluated using the obtained feature set, to get the best classifier for the current task on the balanced ForgeDisaster dataset comprising 740 authentic and 740 forged (mixed set of copy-move and spliced) images. The LBP and DFT features were applied in two ways: (1) applying LBP before DFT (LBP-DFT) and (2) applying LBP after DFT (DFT-LBP). *Table 2* reports the results of all experiments. Among all classifiers, SVM provided the best results using both LBP-DFT and DFT-LBP features. Notably, the proposed LBP-DFT feature arrangement along with SVM outperformed DFT-LBP feature arrangement with SVM in detection accuracy. The proposed method exhibited high accuracy of 0.91, sensitivity value of 0.95, and a specificity value of 0.89. Moreover, lowest FNR of 0.05 and FPR of 0.11 was obtained. These metrics hold significant implications in disaster scenarios, where a failure to detect actual forgery can result in the spread of misinformation and rumors, exacerbating panic among affected populations and potentially diverting resources from actual areas of need. On the other hand, false positives can lead to the overlooking of genuinely affected areas, leading to delays in assistance and support to those in need. So, achieving a low FNR and FPR indicates a reliable detection approach.

To further validate the effectiveness of the proposed approach, AUC-ROC analysis was performed. The AUC-ROC metric provides a comprehensive measure of the model's ability to discriminate between forged and authentic instances across different threshold settings. A higher mean AUC-ROC of 0.92 was obtained on ten folds as shown in *Figure 11*, which is good and validates the model’s correctness. Furthermore, the CPU-processing time required for extracting features, training and prediction was measured. The process of extracting features from a single image required approximately 0.952 seconds. The training phase, which involved training the model with the features of images, lasted about 0.067 seconds. Finally, the prediction process, where the trained model was used to predict authentic or forged image, took approximately 0.0001 seconds.

Table 2 Results of the proposed method and different models

Model name	Accuracy	Sensitivity	Specificity	FNR	FPR
LBP-DFT-RF	0.86	0.84	0.87	0.16	0.13
LBP-DFT-DT	0.83	0.81	0.83	0.19	0.17
LBP-DFT-XGB	0.83	0.79	0.87	0.21	0.13
LBP-DFT-NB	0.60	0.39	0.81	0.61	0.19
LBP-DFT-SVM (Proposed Method)	**0.91**	**0.95**	**0.89**	**0.05**	**0.11**
DFT-LBP-RF	0.81	0.84	0.77	0.16	0.23
DFT-LBP-DT	0.79	0.75	0.83	0.25	0.17
DFT-LBP-XGB	0.80	0.80	0.80	0.20	0.20
DFT-LBP-NB	0.72	0.61	0.83	0.39	0.17
DFT-LBP-SVM	0.89	0.91	0.87	0.09	0.13

4.1.1Performance of the proposed approach for each forgery type

The proposed method was evaluated for copy-move and splicing forgery type individually. Each type was assessed individually to gain insights into the method's effectiveness across different manipulation scenarios. The results are reported in *Table 3*. For the detection of copy-move forgery, a dataset comprising 171 copy-move images and 740 authentic images was formed, resulting in an imbalanced dataset. Similarly, for splicing detection, a dataset containing 569 spliced images and 740 authentic images was used. The proposed method exhibited robust performance for both copy-move and splicing detection. For copy-move forgery, an accuracy of 0.90, sensitivity of 0.93, and specificity of 0.87, with an FNR of 0.07 and FPR of 0.13 was achieved. This indicates that the proposed method accurately identifies 93% of copy-move manipulated images while maintaining a low rate of false negatives and false positives. Similarly, for splicing forgery detection, the proposed method achieved an accuracy of 0.92, sensitivity of 0.94, and specificity of 0.92, with an FNR of 0.06 and FPR of 0.08. The AUC-ROC analysis further validated the robustness of the proposed method, with values of 0.90 and 0.93 obtained for copy-move and splicing over ten folds, respectively as shown on the AUC-ROC graphs in *Figure 12 (a)* and *(b).* It can be observed that the proposed method performs well in detecting both types of forgeries even with imbalanced samples. This shows the robustness of the proposed method across imbalanced samples and forgery types.

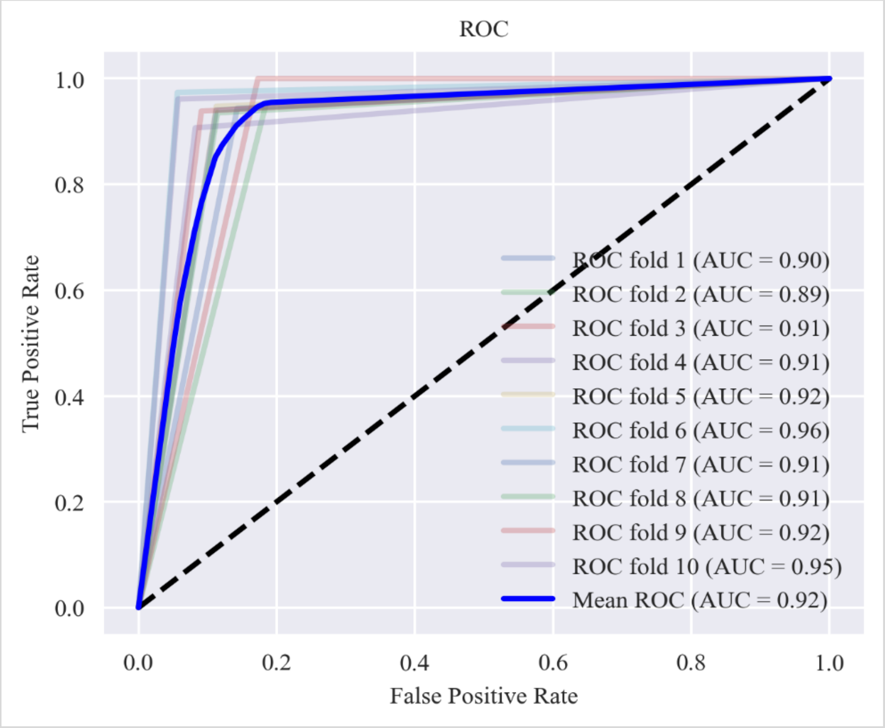


Figure 11 AUC-ROC of the proposed approach over 10 folds

Table 3 Results of the proposed method for each forgery type

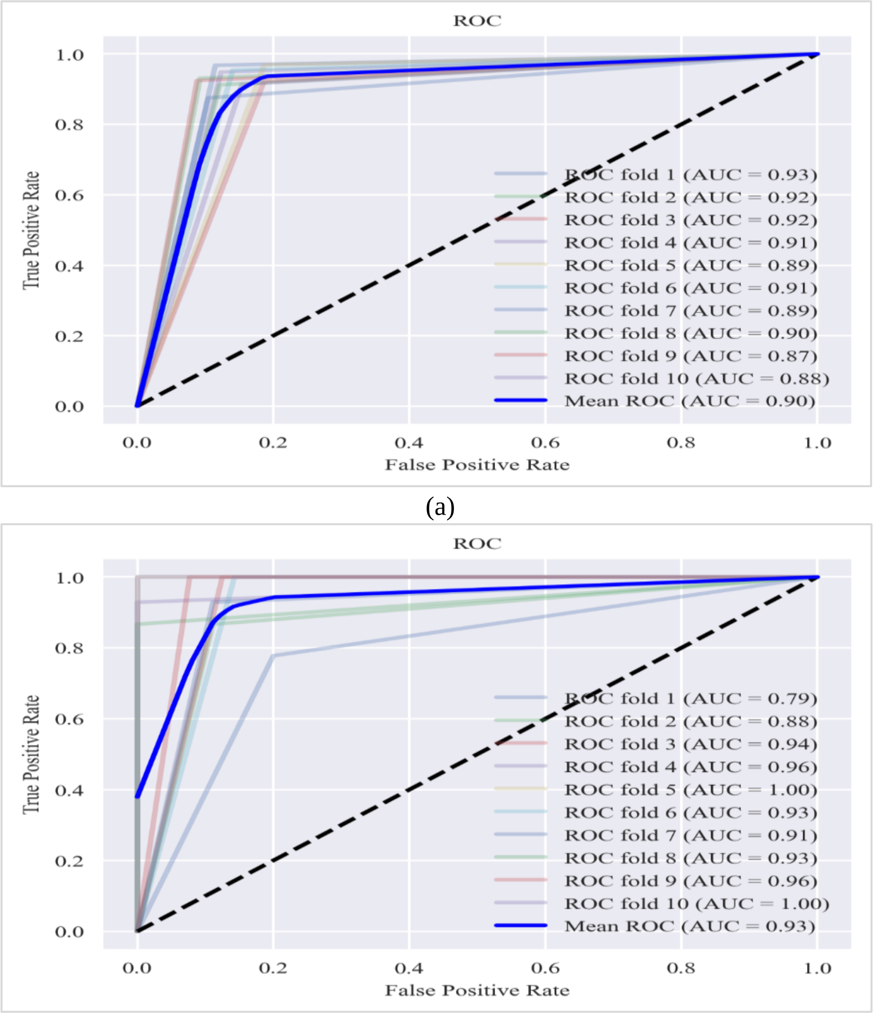
Metric/forgery type	Copy-move	Splicing
Accuracy	0.90	0.92
Sensitivity	0.93	0.94
Specificity	0.87	0.92
FNR	0.07	0.06
FPR	0.13	0.08

4.1.2 Ablation study

In this section, analysis of ablation experiments is presented to thoroughly investigate the impact of pre-processing steps and feature extraction on the proposed method. The pre-processing steps, specifically image enhancement filtering was excluded from the proposed approach to discern its effect. *Figure 13 (a)* presents the results of the proposed method with and without pre-processing. Without pre-processing, utilizing LBP with DFT achieved an accuracy of 88%. Conversely, when incorporating image enhancement filters with LBP and DFT achieved higher accuracy of 91%. It demonstrates that the pre-processing significantly improves the obtained results. Additionally, the performance of the proposed method is compared using Cr channel against utilizing Y and Cb channels, as depicted in *Figure 13 (b).* Notably, the experiments revealed that utilizing the Cr channel yields better results than the other channels. Moreover, the efficacy of feature extraction techniques was explored, examining the performance when employing LBP alone, DFT alone, and a combination of both. The results are illustrated in *Figure 13 (c),* showcasing that the proposed approach, which combines LBP and DFT features, outperformes the utilization of these features individually. Overall, through systematic ablation studies, the critical importance of pre-processing steps and the synergistic effects of feature combination is demonstrated, leading to superior performance.

4.1.3 Error analysis

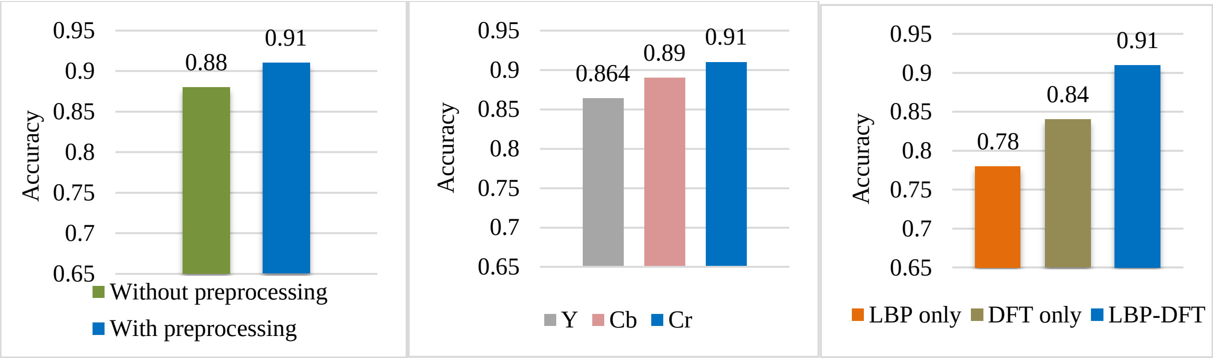
This subsection delves into error analysis aimed at examining cases where the model's predictions diverge from the actual labels. *Figure 14* displays instances of misclassified cases. It was observed that the proposed method tended to misclassify images, where only a small region has been manipulated. For instance, *Figure 14 (a)*, shows an image of an earthquake scene, the yellow circled region encompassing a distressed dog indicates a forged area. Similarly, *Figure 14 (b)*, depicting a wildfire scenario, the yellow circled animal corpse indicates a forged area. The actual label of both the images is forged, however, the proposed method predicted them as authentic. This observation suggests that the proposed method may not accurately detect subtle alterations in the image, especially when they are localized to a small area and are blended well in the target image. The feature representation might not capture these nuanced differences between authentic and manipulated regions leading to misclassification of forged images as authentic. Additionally, in *Figure 14 (c)* and *(d)*, the actual label of the images is authentic, however, the proposed method predicted them as forged. The proposed method misclassified those images which have blurred background with sharp objects in the foreground, capturing clear difference between the sharp objects and the blurred background and detected the image as forged. The method's sensitivity to sharp edges and contrast between foreground and background elements resulted in erroneous classification of images.



(a)

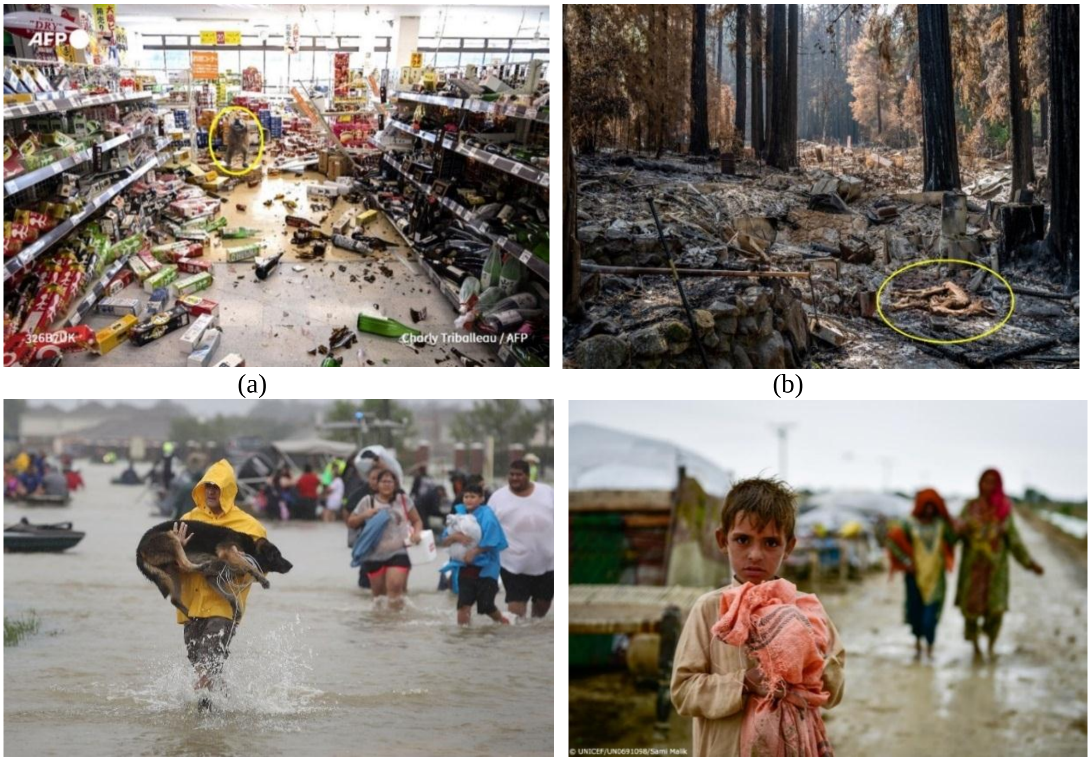
(b)

Figure 12 AUC-ROC of the proposed approach over 10 folds for (a) Copy-move forgery (b) Splicing forgery



(a) (b) (c)

Figure 13 Comparison of accuracy of the proposed method (a) without and with pre-processing (b) with Y, Cb, and Cr channels, and (c) with LBP only, DFT only, and LBP combined with DFT



(a) (b)

(c) (d)

Figure 14 Misclassified cases: (a) and (b) are forged images falsely predicted as authentic, (c) and (d) are authentic images falsely predicted as forged

4.2Evaluating pre-trained deep learning models using the ForgeDisaster dataset

The evaluation of various pre-trained CNN models on the ForgeDisaster dataset was conducted to assess their performance in detecting manipulated images. The models evaluated include InceptionV3, VGG16, EfficientNet, and ResNet50, which are widely recognized architectures in the field of computer vision. These models are pretrained on ImageNet dataset. *Table 4* presents a comparative analysis of the performance metrics achieved by each model, with the proposed method's results highlighted for comparison. The highest accuracy, sensitivity, and specificity achieved by these deep learning models was 0.83, 0.91, and 0.84, respectively. While the lowest FNR and FPR achieved was 0.09 and 0.16, respectively. From the table, it can be observed that these deep learning models does not seem to provide improved performance and the proposed method consistently outperforms them across all the metrics.

Table 4 Results of the proposed method and pre-trained deep learning models

Model Name	Accuracy	Sensitivity	Specificity	FNR	FPR
InceptionV3	0.81	0.78	0.84	0.22	0.16
VGG16	0.83	0.88	0.78	0.12	0.22
EfficientNet	0.81	0.91	0.72	0.09	0.28
ResNet50	0.79	0.70	0.77	0.30	0.23
Proposed Method	**0.91**	**0.95**	**0.89**	**0.05**	**0.11**

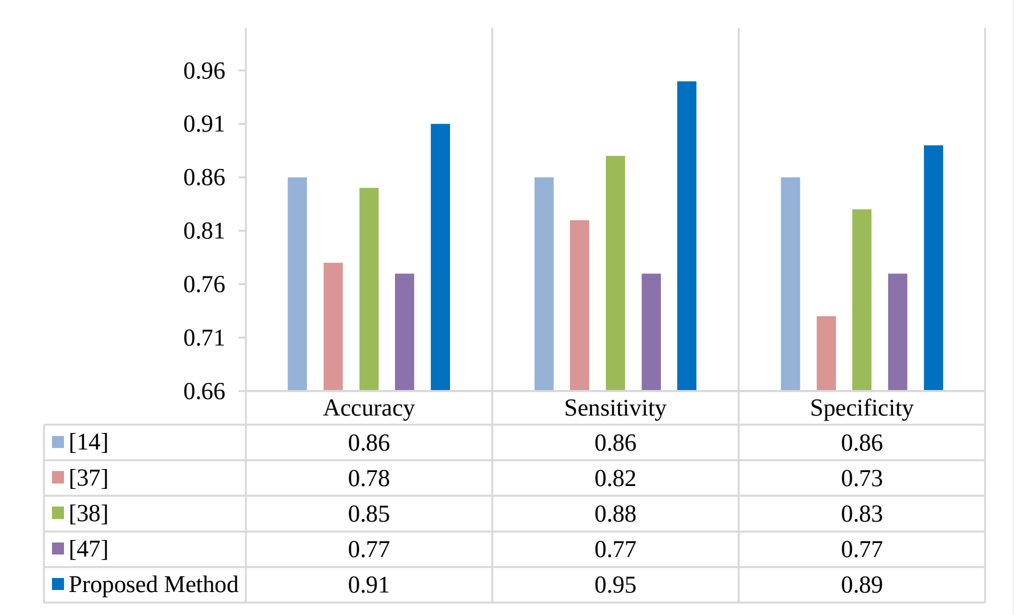
4.3Comparative analysis

As discussed in section 2.2, literature presented various methods for detecting splicing and copy-move attacks. Researchers, including Alahmadi et al. [14], Islam et al. [37], Dua et al. [38], and Ali et al. [47], have evaluated these methods on mixed collections of both copy-move and splicing forged images, showcasing robust performance on general forgery datasets. Experiments were conducted with these methods on the ForgeDisaster dataset, composed of real-world SM disaster-related forged images. In this section, a comparative analysis of the proposed approach against these established methods for SM disaster image forgery detection is provided.

Alahmadi et al. [14], presented a frequency-based method that combined LBP with DCT, reporting a high detection performance on general forgery datasets. On the ForgeDisaster dataset, they achieved an accuracy, sensitivity, and specificity of 86%. While Islam et al. [37], also utilized LBP and DCT, applying DCT first followed by computing LBP and applying a mean operation on the feature set, attained an accuracy score of 78%, with sensitivity of 82% and specificity of 73% on ForgeDisaster dataset.

Dua et al. [38], applied DCT and computed STD and count of ones in DCT coefficients corresponding to each AC frequency component independently to capture the variation in statistical properties of AC coefficients of an entire image. When tested on ForgeDisaster, they obtained an accuracy of 85%, sensitivity of 88%, and specificity of 83%. Ali et al. [47] employed deep learning with ELA features for detecting copy-move and spliced images, achieved an accuracy, sensitivity, and specificity of 77%.

The comparative performance of the proposed method against these referenced existing methods is illustrated in *Figure 15*. The results clearly indicate that the proposed approach consistently outperforms all the existing methods achieving an accuracy of 91%, sensitivity of 95% and specificity of 89%.

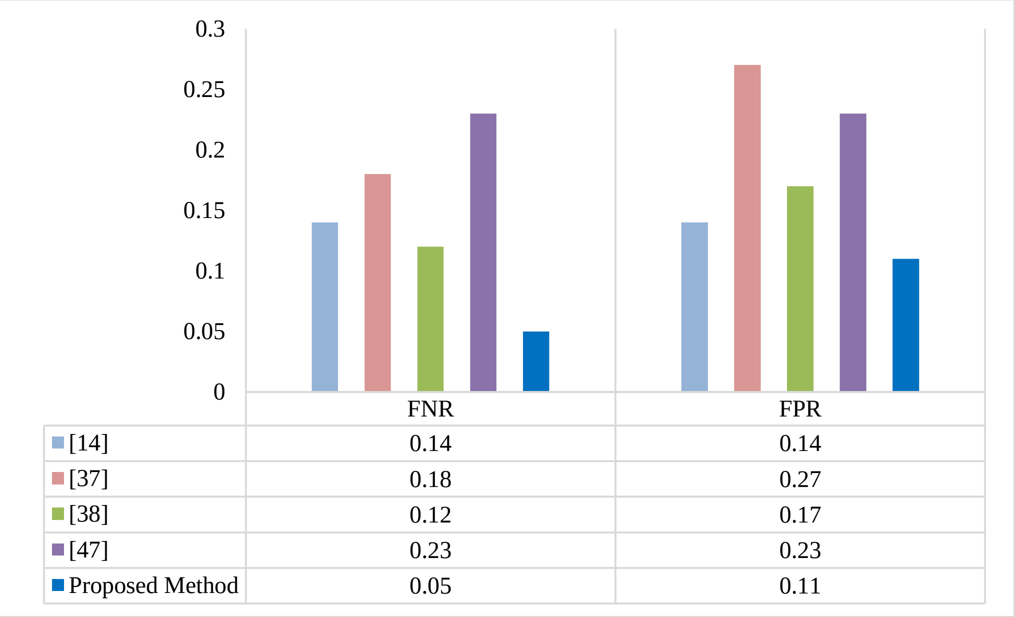


	Accuracy	Sensitivity	Specificity
■ [14]	0.86	0.86	0.86
■ [37]	0.78	0.82	0.73
■ [38]	0.85	0.88	0.83
■ [47]	0.77	0.77	0.77
■ Proposed Method	0.91	0.95	0.89

Figure 15 Comparison of results between the proposed method and existing methods

Furthermore, when comparing the FNR and FPR of the proposed method with the existing methods, Alahmadi et al. [14] provided an FNR and FPR of 14%, Islam et al. [37] provided 18% and 27% of

FNR and FPR, respectively, Dua et al. [38] provided 12% and 17% FNR and FPR, and Ali et al. [47] provided the highest FNR and FPR of 23%. *Figure 16* highlights the FNR and FPR comparison between the proposed method and the existing techniques. Notably, the proposed method demonstrated the lowest FNR of 5% and FPR of 11% compared to the other methods, indicating its capability to minimize the instances of incorrectly identifying authentic images as forged and forged images as authentic.



	FNR	FPR
🟦 [14]	0.14	0.14
🟥 [37]	0.18	0.27
🟩 [38]	0.12	0.17
🟪 [47]	0.23	0.23
🟦 Proposed Method	0.05	0.11

Figure 16 FNR and FPR of the proposed method and existing methods

5.Discussion

Upon thorough evaluation, it was observed that SVM classifier exhibited the best results using the proposed LBP-DFT feature arrangement as compared to other classifiers and feature arrangement. The combination of LBP before DFT for feature extraction enhanced the discriminative power of STD-based features, making them more effective in detecting and highlighting subtle alterations introduced by tampering in images. By achieving high performance in terms of accuracy, specificity, sensitivity, and AUC-ROC, the proposed approach demonstrated superior efficacy, offering a promising solution for disaster authorities seeking credible information from SM platforms. Also, the low FNR and FPR produced by the method indicated its robustness in distinguishing between authentic and forged images. Moreover, the proposed approach exhibited robustness in handling both balanced and imbalanced datasets, as well as different types of forgeries including copy-move and splicing.

Through ablation experiments, the importance of pre-processing steps, such as image enhancement through NLM and bilateral filtering, is elucidated. NLM and bilateral filtering proved to be effective in reducing noise while preserving important image details, thus improving feature extraction from low-quality SM images. Also, experimenting with different color channels revealed that utilizing the Cr channel yields the best results. This shows that Chrominance channels do a better job than any other channel type for encoding tampering traces. These pre-processing steps not only mitigate the impact of noise and compression artifacts but also enable the extraction of relevant forgery indicators, thereby enhancing the overall effectiveness of the detection process.

As per the ablation experimental results, the combination of LBP and DFT features proved to be particularly effective for SM disaster images. As discussed, LBP excels in capturing local texture patterns, and enhances tampering artifacts crucial for identifying subtle irregularities introduced by image manipulations. While DFT analyses the frequency content of images to capture the variations in the local frequency distributions which can reveal inconsistencies introduced by forgery. It provides a comprehensive representation of image features in the frequency domain, complementing the spatial information captured by LBP. By utilizing both LBP and DFT, the proposed approach captures a diverse range of features from SM images, including both texture and frequency characteristics. This comprehensive feature representation enhances the discriminative power of the detection system.

The proposed approach stands out from existing forgery detection methods by specifically addressing the challenges associated with SM disaster images. Unlike existing methods, the proposed approach incorporated image enhancement filtering. By utilizing this pre-processing step, the low-quality SM disaster images are enhanced by reducing noise while preserving crucial image details. This enhancement significantly improves the clarity of key features in the images, thereby enhancing the accuracy of the analysis. Furthermore, the proposed approach utilizes a combination of LBP and DFT for extracting features. Even without pre-processing, this combination achieves an impressive accuracy as detailed in section 4.1.2. This underscores the robustness of the LBP and DFT combination in extracting features, making them particularly well-suited for analyzing disaster images on Twitter SM platform.

In summary, the combination of pre-processing techniques, robust feature extraction methods, and SVM classification collectively enhances the robustness of forgery detection for disaster images on the Twitter SM platform.

5.1Limitations

One notable limitation is that the proposed method is evaluated using a specific ForgeDisaster dataset which includes images from Twitter SM only. This may introduce bias and limit the generalizability of the findings to other SM platforms. Additionally, the error analysis revealed misclassifications, suggesting the need for further improvements. It is recommended to augment the training dataset with a diverse range of forged images that encompass various degrees of manipulation and visual complexity, may improve the method's ability to generalize across different scenarios.

5.2Practical implications

The proposed approach offers disaster response authorities a reliable method to verify the authenticity of SM images before using them in decision-making processes during disasters. By integrating the proposed methodology into their workflows, organizations can mitigate risks posed by forged content, enhancing the efficiency of rescue missions and optimizing resource allocation. Additionally, the availability of the ForgeDisaster dataset provides a valuable resource for researchers and practitioners, facilitating advancements in SM disaster forgery detection. *Appendix I* contains an exhaustive compilation of all abbreviations referenced throughout this paper for easy reference.

6.Conclusion and future work

The potential dissemination of manipulated or fake images can impede rescue missions, prolong recovery efforts, and even endanger lives. To address this critical issue, effective detection of forged SM disaster images is imperative. However, the post-processing operations like compression commonly performed on SM platforms further complicate the detection process by introducing noise and degrading image quality, necessitating specific approaches. To this end, a new dataset ForgeDisaster was introduced in this study which consisted of authentic and forged images collected from Twitter SM related to various natural disasters that occurred worldwide. Additionally, a new passive approach was presented specific to SM disaster image forgery detection. The approach included pre-processing phase that enhanced the low image quality on SM by additional filters. Feature extraction phase that included LBP and DFT for extracting robust features and classification phase including SVM for classification. The combination of all three phases in the proposed approach collectively contributed to the effective detection of SM disaster image forgery detection, resulting an accuracy of 91% and demonstrating superior efficacy compared to existing approaches. Thus, it presented a promising solution for disaster authorities seeking credible information from SM platforms during disasters.

Moreover, while the proposed dataset was utilized to evaluate various pre-trained deep learning architectures, the obtained results did not exhibit significant improvements. However, the obtained results can be used as a baseline for future deep learning solutions. This observation suggests that while deep learning models offer the potential for automatic feature learning, the basic models employed in this study may not be sufficiently advanced to address the complexities of forgery detection. A logical step to enhance detection performance could involve augmenting the dataset. Augmentation would enable the training of more advanced deep learning models, which is crucial given that these models often struggle with limited data availability. In addition to this, future research should consider expanding the dataset to encompass images from a broader spectrum of SM platforms. This expansion would ensure the robustness and applicability of the proposed approach across diverse SM environments. To further enhance the proposed forgery detection method, it's imperative to incorporate forgery localization techniques. This addition would provide finer-grained insights into the specific regions of an image affected by tampering, thereby improving the overall accuracy and efficacy of the proposed approach. Furthermore, in the current study, the assessment of the credibility of SM disaster images is limited to copy-move and splicing forgeries. However, it's noteworthy that in contemporary times, there is a rising trend in circulating AI-generated fake images or the proliferation of deepfakes during disasters on SM. Hence, future research efforts could extend to encompass the detection of such AI-generated fake images or deepfakes in disaster contexts.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data availability

The ForgeDisaster dataset utilized in this study is available at https://github.com/saimasaleem/ForgeDisaster

Author’s contribution statement

Saima Saleem: Data collection, dataset preparation, conceptualization, investigation, implementation, writing, analysis and interpretation of results. **Akash Shah:** Dataset preparation, conceptualization, investigation and analysis. **Monica Mehrotra:** Supervision, investigation, reviewing and editing.

References

[1] United nations office for disaster risk reduction. Global assessment report on disaster risk reduction 2022: our world at risk: transforming governance for a resilient future. UN; 2022.

[2] Saleem S, Mehrotra M. Emergent use of artificial intelligence and social media for disaster management. In proceedings of international conference on data science and applications 2022 (pp. 195-210). Springer Singapore.

[3] Saleem S, Mehrotra M. Context-aware transfer learning approach to detect informative social media content for disaster management. International Journal of Advanced Computer Science and Applications. 2024; 15(1):680-9.

[4] Imran M, Ofli F, Caragea D, Torralba A. Using AI and social media multimodal content for disaster response and management: opportunities, challenges, and future directions. Information Processing & Management. 2020; 57(5):102261.

[5] Nguyen DT, Ofli F, Imran M, Mitra P. Damage assessment from social media imagery data during disasters. In proceedings of the international conference on advances in social networks analysis and mining 2017 (pp. 569-76). IEEE.

[6] Mouzannar H, Rizk Y, Awad M. Damage identification in social media posts using multimodal deep learning. In proceedings of the 15th ISCRAM conference, Rochester, NY, USA 2018 (pp. 1-15).

[7] Kalliatakis G, Ehsan S, Fasli M, DMconald-maier K. Displacenet: recognising displaced people from images by exploiting dominance level. In proceedings of the conference on computer vision and pattern recognition workshops 2019 (pp. 33-8). IEEE.

[8] Shah A, Varshney S, Mehrotra M. DeepMUI: a novel method to identify malicious users on online social network platforms. Concurrency and Computation: Practice and Experience. 2024; 36(3):e7917.

[9] Gupta A, Lamba H, Kumaraguru P, Joshi A. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In proceedings of the 22nd international conference on world wide web 2013 (pp. 729-36). IEEE.

[10] https://www.bbc.com/future/article/ 20121031-how-to-spot-a-fake-sandy-photo. Accessed 02 June 2023.

[11] https://japannews.yomiuri.co.jp/society/ general-news/20220929-61020/ Accessed 02 June 2023.

[12] Walia S, Kumar K. Digital image forgery detection: a systematic scrutiny. Australian Journal of Forensic Sciences. 2019; 51(5):488-526.

[13] Redi JA, Taktak W, Dugelay JL. Digital image forensics: a booklet for beginners. Multimedia Tools and Applications. 2011; 51:133-62.

[14] Alahmadi A, Hussain M, Aboalsamh H, Muhammad G, Bebis G, Mathkour H. Passive detection of image forgery using DCT and local binary pattern. Signal, Image and Video Processing. 2017; 11:81-8.

[15] Sun W, Zhou J, Lyu R, Zhu S. Processing-aware privacy-preserving photo sharing over online social networks. In proceedings of the 24th international conference on multimedia 2016 (pp. 581-85). ACM.

[16] Mitra A, Mohanty SP, Corcoran P, Kougianos E. A novel machine learning based method for deepfake video detection in social media. In international symposium on smart electronic systems 2020 (pp. 91-96). IEEE.

[17] Sun W, Zhou J, Li Y, Cheung M, She J. Robust high-capacity watermarking over online social network shared images. IEEE Transactions on Circuits and Systems for Video Technology. 2020; 31(3):1208-21.

[18] Imran M, Qazi U, Ofli F, Peterson S, Alam F. AI for disaster rapid damage assessment from microblogs. In proceedings of the AAAI conference on artificial intelligence 2022 (pp. 12517-23).

[19] Alam F, Alam T, Hasan MA, Hasnat A, Imran M, Ofli F. MEDIC: a multi-task learning dataset for disaster image classification. Neural Computing and Applications. 2023; 35(3):2609-32.

[20] Saleem S, Mehrotra M. An analytical framework for analyzing tweets for disaster management: case study of turkey earthquake. In 14th international conference on computing communication and networking technologies 2023 (pp. 1-7). IEEE.

[21] Alam F, Ofli F, Imran M. Processing social media images by combining human and machine computing during crises. International Journal of Human–Computer Interaction. 2018; 34(4):311-27.

[22] Alam F, Ofli F, Imran M. Crisismmd: multimodal twitter datasets from natural disasters. In proceedings of the international AAAI conference on web and social media 2018 (pp.465-73).

[23] Li X, Caragea D, Caragea C, Imran M, Ofli F. Identifying disaster damage images using a domain adaptation approach. In proceedings of the 16th international conference on information systems for crisis response and management 2019 (pp. 633-45).

[24] Ning H, Li Z, Hodgson ME, Wang C. Prototyping a social media flooding photo screening system based on deep learning. ISPRS International Journal of Geo-Information. 2020; 9(2):104.

[25] Hassan SZ, Ahmad K, Hicks S, Halvorsen P, Al-fuqaha A, Conci N, et al. Visual sentiment analysis from disaster images in social media. Sensors. 2022; 22(10):1-18.

[26] Kotha S, Haridasan S, Rattani A, Bowen A, Rimmington G, Dutta A. Multimodal combination of text and image tweets for disaster response assessment. International Workshop on Data-driven Resilience Research 2022 (pp. 1-10).

[27] Dong J, Wang W, Tan T. Casia image tampering detection evaluation database. In China summit and international conference on signal and information processing 2013 (pp. 422-6). IEEE.

[28] Hsu YF, Chang SF. Detecting image splicing using geometry invariants and camera characteristics consistency. In international conference on multimedia and explore 2006 (pp. 549-52). IEEE.

[29] Amerini I, Ballan L, Caldelli R, Del BA, Del TL, Serra G. Copy-move forgery detection and localization by means of robust clustering with J-linkage. Signal Processing: Image Communication. 2013; 28(6):659-69.

[30] Tralic D, Zupancic I, Grgic S, Grgic M. CoMoFoD-new database for copy-move forgery detection. In proceedings ELMAR 2013 (pp. 49-54). IEEE.

[31] Hakimi F, Hariri M, Gharehbaghi F. Image splicing forgery detection using local binary pattern and discrete wavelet transform. In 2nd international conference on knowledge-based engineering and innovation 2015 (pp. 1074-7). IEEE.

[32] Kaur M, Gupta S. A passive blind approach for image splicing detection based on DWT and LBP histograms. In 4th international symposium security in computing and communications 2016 (pp. 318-27). Springer Singapore.

[33] Hayat K, Qazi T. Forgery detection in digital images via discrete wavelet and discrete cosine transforms. Computers & Electrical Engineering. 2017; 62:448-58.

[34] Shah A, El-alfy ES. Image splicing forgery detection using DCT coefficients with multi-scale LBP. In international conference on computing sciences and engineering 2018 (pp. 1-6). IEEE.

[35] Parnak A, Baleghi Y, Kazemitabar J. A novel forgery detection algorithm based on mantissa distribution in digital images. In 6th Iranian conference on signal processing and intelligent systems 2020 (pp. 1-4). IEEE.

[36] Alkawaz MH, Sulong G, Saba T, Rehman A. Detection of copy-move image forgery based on discrete cosine transform. Neural Computing and Applications. 2018; 30:183-92.

[37] Islam MM, Karmakar G, Kamruzzaman J, Murshed M. A robust forgery detection method for copy–move and splicing attacks in images. Electronics. 2020; 9(9):1-22.

[38] Dua S, Singh J, Parthasarathy H. Image forgery detection based on statistical features of block DCT coefficients. Procedia Computer Science. 2020; 171:369-78.

[39] Xiao B, Wei Y, Bi X, Li W, Ma J. Image splicing forgery detection combining coarse to refined convolutional neural network and adaptive clustering. Information Sciences. 2020; 511:172-91.

[40] Qazi EU, Zia T, Almorjan A. Deep learning-based digital image forgery detection system. Applied Sciences. 2022; 12(6):1-17.

[41] Abdalla Y, Iqbal MT, Shehata M. Copy-move forgery detection and localization using a generative adversarial network and convolutional neural-network. Information. 2019; 10(9):1-26.

[42] Goel N, Kaur S, Bala R. Dual branch convolutional neural network for copy move forgery detection. IET Image Processing. 2021; 15(3):656-65.

[43] Ding H, Chen L, Tao Q, Fu Z, Dong L, Cui X. DCU-Net: a dual-channel U-shaped network for image splicing forgery detection. Neural Computing and Applications. 2023; 35(7):5015-31.

[44] Walia S, Kumar K, Kumar M, Gao XZ. Fusion of handcrafted and deep features for forgery detection in digital images. IEEE Access. 2021; 9:99742-55.

[45] Sabeena M, Abraham L. Convolutional block attention based network for copy-move image forgery detection. Multimedia Tools and Applications. 2024; 83(1):2383-405.

[46] Vijayalakshmi KNV, Sasikala J, Shanmuganathan C. Copy-paste forgery detection using deep learning with error level analysis. Multimedia Tools and Applications. 2024; 83(2):3425-49.

[47] Ali SS, Ganapathi II, Vu NS, Ali SD, Saxena N, Werghi N. Image forgery detection using deep learning by recompressing images. Electronics. 2022; 11(3):1-17.

[48] Martinez-cantin R. Bayesopt: a bayesian optimization library for nonlinear optimization, experimental design and bandits. Journal of Machine Learning Research. 2014; 15(1):3735-9.

Saima Saleem is a Ph.D candidate at the Departemt of Computer Science, Jamia Millia Islamia, New Delhi, India. She has published several articles in international conferences, journals, and books. Her research area of interest include Natural Language Processing, Machine Learning, Deep Learning, and Social Media Analysis.
Email: saimak6.sk@gmail.com

Akash Shah is a research scholar pursuing Ph.D in the Department of Computer Science at Jamia Millia Islamia, New Delhi, India. He has contributed to numerous international journals and conferences through his published articles. The scope of his research encompasses the detection of Malicious Profiles, Identification of Inference Attacks, and Social Media Anlaysis.
Email: akashshah.dsc@gmail.com

Monica Mehrotra is a professor and Head of the department at the department of computer science, Jamia Millia Islamia, New Delhi, India. She has over twenty-five years of teaching experience. She has received the Ph.D. degree in computer science from Jamia Millia Islamia University. She has published over 70 papers in International conferences & Journals of repute. She has won 'Excellent Researcher Award (female)' in 2nd International academic and research excellence awards (IARE – 2020) ceremony organized by GISR foundation. She is currently a member of the Institute of Electrical and Electronics Engineers (IEEE). Her research interests include Data Mining, Information Retrieval and Social Network Analysis, and Machine Learning.
Email: mmehrotra@jmi.ac.in

Appendix I

S. No.	Abbreviation	Description
1	2D	2-Dimensional
2	AUC-ROC	Area Under the Curve of Receiver Operating Characteristic Curve
3	CNN	Convolutional Neural Network
4	DT	Decision Tree
5	DC	Direct Component
6	DCT	Discrete Cosine Transform
7	DFT	Discrete Fourier Transform
8	DWT	Discrete Wavelet Transform
9	ELA	Error Level Analysis
10	XGB	eXtreme Gradient Boosting
11	FNR	False Negative Rate
12	FPR	False Positive Rate
13	FFT	Fast Fourier Transform
14	GOV	Government
15	LBP	Local Binary Pattern
16	NB	Naïve Bayes
17	NGO	Non-Government Organization
18	NLM	Non-Local Means
19	RF	Random Forest
20	RGB	Red Green Blue
21	SM	Social Media
22	STD	Standard Deviation
23	SVM	Support Vector Machine
24	VGG	Visual Geometry Group