Hydrology Research

Hydrology Research Vol 53 No 6, 795 doi: 10.2166/nh.2022.093

The need for training and benchmark datasets for convolutional neural networks in flood applications

Abdou Khouakhi *, Joanna Zawadzka and Ian Truckell

School of Water, Energy and Environment, Centre for Environmental and Agricultural Informatics, Cranfield University, Cranfield, UK
*Corresponding author. E-mail: a.khouakhi@cranfield.ac.uk

AK, 0000-0002-1224-3208; JZ, 0000-0002-6172-901X

ABSTRACT

Flood-related image datasets from social media, smartphones, CCTV cameras, and unmanned aerial vehicles (UAVs) present valuable data for the management of flood risk, and particularly for the application of modern convolutional neural networks (CNNs) to specific flood-related problems such as flood extent detection and flood depth estimation. This review discusses the increasing role of CNNs in flood research with a growing number of published datasets, particularly since 2018. We note the lack of open and labelled flood image datasets and the growing need for an open, benchmark data library for image classification, object detection, and segmentation relevant to flood management. Such a library would provide benchmark datasets to advance CNN flood applications in general and serve as a resource, providing data scientists and the flood research community with the necessary data for model training and validation.

Key words: convolutional neural networks, flood images, floods

HIGHLIGHTS

- There is a growing need for an open, benchmark image library for CNNs relevant to flood applications.
- The majority of papers reviewed used relatively small sets of data often manually annotated for training which leads to weak models, reproducibility, and replicability issues.
- Building an open dataset is a community effort and requires close collaboration with researchers and organizations that generate the data.

1. INTRODUCTION

Recent developments in cloud computing, artificial intelligence (AI) algorithms, and image processing have impacted many research fields, such as flood detection and prediction. Technologies such as high-resolution smartphone cameras, connectivity, social media, unmanned aerial vehicle (UAV), and satellite imagery have led to valuable and new types of visual data sources relevant to flood applications (Fohringer *et al.* 2015; Tkachenko *et al.* 2017; Hashemi-Beni & Gebrehiwot 2021). The use of these data to develop predictive models, however, remains an ongoing challenge, due to their diverse and disparate nature.

Visual data (i.e., images), image processing, and computer vision algorithms in flood application have been used for decades. For example, image processing and edge detection techniques have been used to estimate information from images, including water level (Park *et al.* 2009; Yang *et al.* 2014), flow velocity (Le Coz *et al.* 2010), and flood inundation extent (Horritt *et al.* 2001). Recently, convolutional neural networks (CNNs) have demonstrated unprecedented performance for image detection and segmentation. CNNs are a type of deep neural network which is in turn a specific type of machine learning belonging to the broader field of AI. Like classic machine learning algorithms, CNNs learn features from labelled training data to predict an accurate output from an unknown input. CNNs are formed of a network of stacked layers, which are used to extract richer features from the input data. The model becomes deeper as more layers are stacked leading to the increasing complexity of the network allowing for features to be learned (Krizhevsky *et al.* 2012). Therefore, the collection and labelling of training data are crucial components of the CNN training process. For an in-depth review on recent advances in deep learning including CNNs and their architectures, we refer the reader to the literature (Hoeser & Kuenzer 2020; Alzubaidi *et al.* 2021).

Historically, satellite and aerial images are two of the most extensively used visual data types for flood-related applications. Landsat MultiSpectral Scanner was one of the first sources of imagery used for flood delineation, for instance, to delineate flood-affected areas in regions of the United States (Deutsch & Ruggles 1974; Rango & Anderson 1974). Many other satellite missions followed, such as Landsat thematic mapper, SPOT (Satellite Pour l'Observation de la Terre) imagery, and Very High-Resolution Radiometer (AVHRR), were used extensively for flood management (Brown *et al.* 1987; Profeti & Macintosh 1997; Jain *et al.* 2006; Qi *et al.* 2009; Inman & Lyons 2020). Recent missions, such as Sentinel-1 and -2, provide open and even higher spatial resolution imagery opening many opportunities to apply machine learning and CNNs for improved flood-related research. More recently, the UAV technology is being used to generate on-demand and high-resolution imagery at lower costs for rapid flood assessment. UAVs can acquire high-resolution data (compared to satellite imagery) for the fast and accurate detection of inundated areas under complex landscapes and inaccessible flooded areas.

Despite efforts in recent years to exploit the power of CNNs, machine learning, and visual data for flood risk management, there are still many challenges, such as the limited availability of labelled images to train and validate CNN models (Iqbal *et al.* 2021). Labelling images is a tedious, manual process in most cases without collaboration and sharing of annotated images, and the creation of benchmark datasets within the flood application community, a significant amount of time will be spent by individual research groups in the labelling process. For example, approximately 13 h was required to label just 100 UAV images as water, buildings, road, and vegetation (Gebrehiwot *et al.* 2019a).

Here, we present a review of different initiatives using CNNs for flood applications; we highlight the recent advances and provide benefits of increasing the availability of annotated visual datasets for deep CNNs applied to flood research. The main research questions that we address are: (1) what is the current usage of CNNs for flood research? (2) what are the training datasets and techniques used to address flood-related problems as related to CNNs? and (3) what is the benefit of open training data and data annotation for flood applications?

2. FLOOD APPLICATIONS AND DEEP CNNS

To review existing approaches using CNNs for flood application, we first searched the Scopus database using the search terms '*deep learning*' *OR* '*fully convolutional*' *OR* '*convolutional neural network*' *OR* '*CNN*' *OR* '*convolutional network*' *AND* '*flood*'. We performed the search on titles, abstracts, and keywords and retained only journal articles. The results (236 journal papers) were filtered to retain only journal articles published after 2010 (222 journal papers), given that advances in CNN algorithm development have only occurred since 2010. Figure 1 shows the number of articles indexed in Scopus published from 2010 and reveals that most papers have been published in the 3 years only. We, therefore, focus the review only on papers published since 2018 (190 journal articles). Several papers were omitted from the list after the first examination of abstracts.

A preliminary examination of the papers confirms that the role of CNNs in flood research has been growing steadily, particularly from 2018 with the number of Scopus journal papers almost doubling every year. For the purpose of this work and given the importance of the datasets in training CNN models, we structured this review in the following four main aspects: (1) sources of visual data (images) used for model training; (2) labelling strategy; (3) CNN algorithms employed; and (4) types of flooding problems solved (e.g., detection and extent mapping).

2.1. Sources of visual data

Satellite imagery, airborne imagery, CCTV footage, ground camera photographs and images shared on social media, and the internet are the main sources of data used across the reviewed papers. We found about 43% of the 70 reviewed papers used ground camera images, whereas 37% of papers used satellite images such as Landsat and Sentinel imagery. Other authors used a combination of both satellite and camera images. Only ~7% of reviewed papers used UAV imagery (Figure 2). To provide further details, we clustered the reviewed papers into studies that used ground RGB (red, green, blue) camera images (i.e., ground camera photos including social media and internet images) and those used satellite and airborne images including multispectral imagery. We also highlighted studies that used both datasets.

2.1.1. RGB ground images

RGB ground camera images shared on social media have been widely retrieved and used to aid with flood detection and delineation. For example, Sazara *et al.* (2019) used 253 images of size 385×512 pixels with a flood and 238 without floods, and shuffled and split the data into training and testing subsets to train and evaluate a CNN classifier of images containing a scene

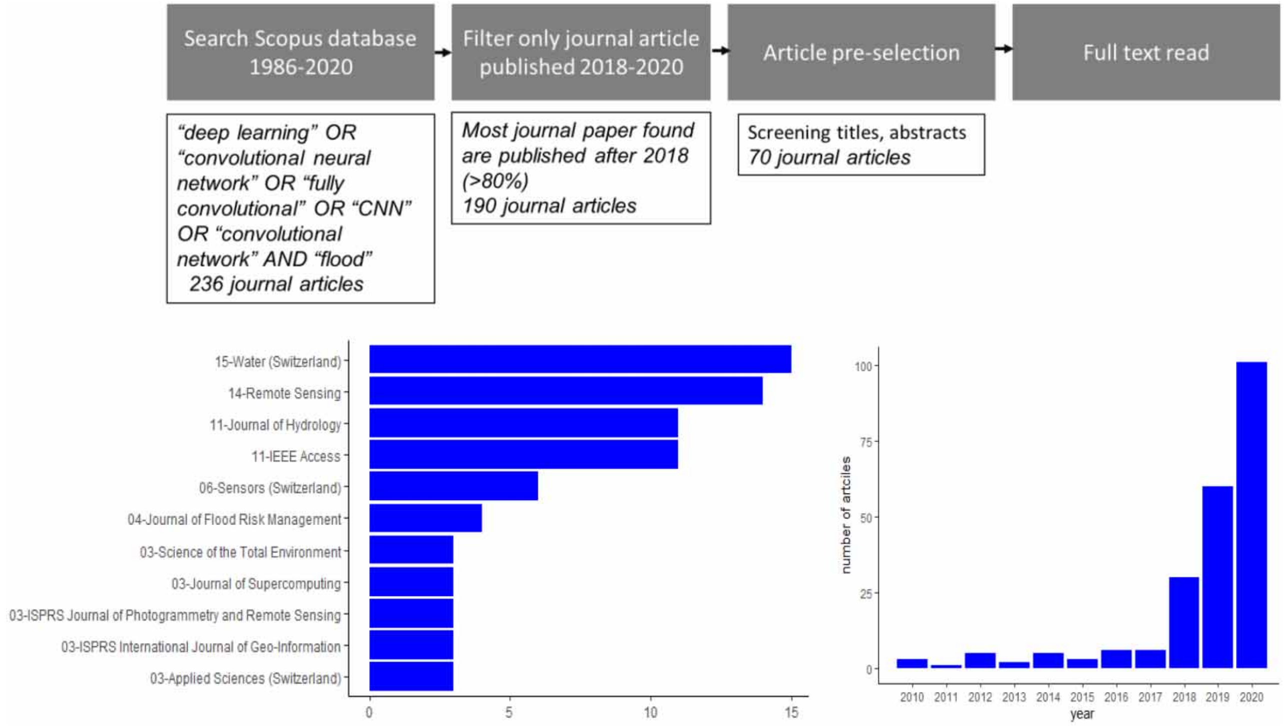

Figure 1 | Summary of the review process (top), number of published papers on deep CNNs and floods across (right panel) different journals and peer-reviewed papers and their corresponding journals (left panel). Numbers to the left of each journal indicate the number of published papers in each journal.

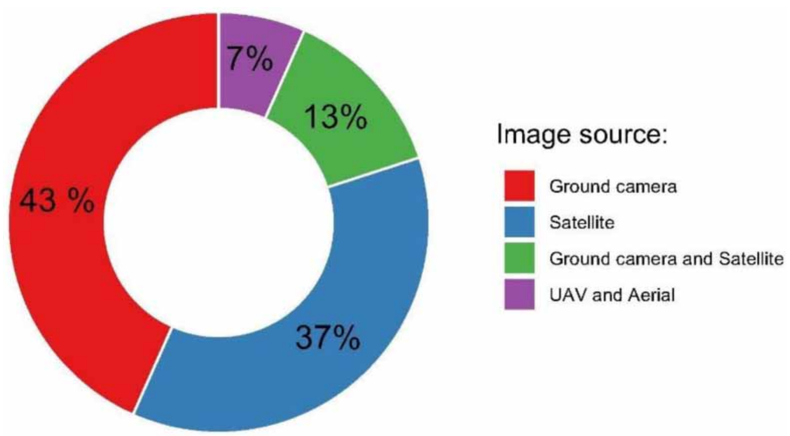

Figure 2 | Sources of visual data (Image source) used in the 70 reviewed papers.

with floodwater, before segmenting the flooded area. Moy De Vitry *et al.* (2019) used existing CCTV footage to provide qualitative flood level information over time. The authors first delineated flooded areas using CNNs and developed a proxy named 'static observer flooding index' indicating water level fluctuations visible from a surveillance camera's viewpoint. Photos of floods from social media typically contain geolocation, date-time, and other metadata about floods relevant for model training. Crowdsourced photos from predefined online applications and from Twitter have also been used to locate the flooded area and can be used to validate numerical models (Wang *et al.* 2018). Another example where images of drains and gullies from Google Images and YouTube were employed to train a CNN model to detect and classify drainage blockages (Mishra *et al.* 2020). The images were classified according to different blockage severity classes. The authors suggested that the trained model could then be used with an Internet of things (IoT)-enabled camera to monitor gullies and drainage. Finally, Chaudhary *et al.* (2020) gathered images containing flooded objects to estimate the flood level based on features partially submerged in water. They employed a CNN model on images with the object of known dimensions (e.g., car, bus, and bicycle) to infer the water level. These examples highlight the promise of social media images for the development of CNNs.

2.1.2. Satellite and airborne imagery

Images retrieved from satellites are widely used for flood risk management. With recent algorithmic advances in image detection and segmentation, many recent studies have been conducted taking advantage of combining CNNs and satellite imagery for flood applications.

In a recent study, Sentinel-2 images acquired from the MediaEval 2019 competition were used to detect floods (Jain *et al.* 2020). The data consisted of 335 sets with 267 images used for training and 68 for testing. Each set contained varying 1- to 24-day time series of images prior and after flood events; i.e., 2,770 images in total. The images were of 512×512 pixels in size from all the 12 bands of sentinel-2 data. Indices such as the Normalized Difference Water Index (NDWI), the modified NDWI, and the Automated Water Extraction Index (AWEI) were computed and added to the initial Sentinel-2 images to train a CNN (VGG16) network with a fully connected layer. Another study proposed a novel approach named H2O-Net, integrating Sentinel-2 and height resolution commercial imagery from PlanetScope to delineate floodwater (Akiva *et al.* 2020). The Sentinel training dataset was comprised of 3,676 images of 512×512 pixels covering RGB and SWIR (shortwave infrared) data resampled to 10 m per grid cell. PlanetScope is a high-resolution optical satellite providing RGB and near infrared (NIR) bands at 3 m resolution.

Many other studies have harnessed Landsat and Sentinel-1 and -2 images, by deriving indices used to train CNN models to assess flood characteristics and susceptibility (e.g., Tien Bui *et al.* 2020; Wang *et al.* 2020a; Zhao *et al.* 2020; Panahi *et al.* 2021). For example, the Normalized Difference Vegetation Index (NDVI) derived from Landsat 8 images was used alongside maps of flood-generating factors to train a CNN for flash flood susceptibility (Panahi *et al.* 2021). Similarly, Zhao *et al.* (2020) integrated Landsat images and other factors such as elevation, slope, the NDVI, and rainfall to assess flood susceptibility within an urban catchment.

Other studies have integrated both satellite and camera-based images to strengthen the model training process. For example, crowdsourced images and satellite images were used in conjunction to detect the passability of roads during floods (Ahmad *et al.* 2019).

Synthetic aperture radar imagery (SAR) is another valuable data source that has been used in flood delineation and detection, as it has the capability to see through clouds and darkness. For instance, co-registered optical Sentinel-2 and SAR Sentinel-1 image time series were recently employed for the detection of floods using combined CNNs and recurrent neural networks (Rambour *et al.* 2020a).

Finally, the UAV technology has recently been recognized as an efficient and rapid way to quickly acquire high-resolution imagery for flood applications (Rivas Casado *et al.* 2018; Gebrehiwot *et al.* 2019b). Deep CNN approaches have been used to extract flooded areas from high-resolution UAV images (Gebrehiwot *et al.* 2019b). The authors used a manually labelled dataset alongside transfer learning, i.e., a pre-trained model.

In general, satellite-based imagery data have been used for flood detection, flood extent mapping, or flood damage assessment due to their wide spatial coverage but also the spectral information they contain beyond the visible bands. Most research examined here relayed on freely available mid-resolution data such as Sentinel-2 and-1 and Landsat products (Peng *et al.* 2019; Wieland & Martinis 2019; Rambour *et al.* 2020b). Ground images, on the other hand, are mainly harnessed for flood depth estimation as well as for detection and mapping of floods due to the resolution and the apparent flooded features that can be used as proxies for flood depth estimation (Chaudhary *et al.* 2020; Huang *et al.* 2020).

Both satellite and ground camera imagery have been used for damage detection and assessment (Rubio *et al.* 2019; Gupta & Shah 2020; Muhadi *et al.* 2020), but they both have associated challenges and limitations. While satellite imagery provides a valuable resource of images for CNN flood applications, the images are often provided in large image tiles and require preprocessing and transformation to smaller subsets of imagery (image chips) to be used as CNN inputs. Additionally, small spatial features of flooded areas and floods in small streams are difficult to map using low-resolution satellite imagery. The quality and availability of flood images, particularly optical data, are also hampered by sky conditions such as cloud coverage. For ground imagery data, finding a sufficient number of images relating to a certain flood event can be challenging (Barz *et al.* 2019).

One of the most challenging phases in developing a CNN flood application from both ground and satellite image data is the labelling of images for training. Different strategies have been developed to increase the number of annotated images.

2.2. Labelling strategies

Generally, the training datasets required for different types of deep CNN problems, such as image classification, semantic segmentation, and object detection, are composed of pairs of data (images) and labels, then the images are fed to the CNN model alongside their corresponding labels as outputs to train the model. Labels are generated from ground reference data and/or annotated imagery. Very few pre-labelled and benchmark datasets are currently available for visual flood data. Most (~57%) of the studies reviewed here used their own manually labelled images assisted by tools such as markGT, LabelMe, and DarkLabel (Figure 3). Some studies used a combination of manually annotated images and existing benchmark pre-labelled data (Supplementary Material, Table S1), such as MediaEval (Bischke *et al.* 2017, 2018, 2019) which is a benchmark dataset of satellite imagery and social media images and the European Flood 2013 datasets (Barz *et al.* 2019) which is an annotated dataset of images from the central European flood 2013 (Table 1).

Some studies used automated techniques to minimize the effect of a limited annotated training dataset. For example, to examine the role of SAR intensity and interferometric coherence combined with CNN in urban flood detection, Li *et al.* (2019) introduced an active self-learning CNN framework that selects informative unlabelled samples based on a temporal-ensembling CNN model. Subsequently, these samples are pseudo-labelled by a multi-scale spatial filter. Chaudhary *et al.* (2020) proposed a multi-task deep CNN approach for water depth estimation based on social media images for flood mapping. They trained the model for a small set of annotated water levels (regression task) and a larger set of weak annotated dataset (ranking task) to save annotation effort. The authors suggest that weak supervision using pairwise ranking can be a promising alternative to costly and time-consuming, fine-grained labelling.

2.3. CNN algorithms

CNNs are one of the most successful algorithms for tackling computer vision problems and have achieved high performance for image-related problems such as the detection of features on images. Although modern CNNs can be traced back to the 1990 s following the work of LeCun *et al.* (1989) and Lecun *et al.* (1998), their wide expansion came after 2012 as the data availability and computation capacities increased, but also as deeper architectures improved their representational capacity. This is marked by the work of Krizhevsky *et al.* (2012), who developed the AlexNet architecture that achieved notable performance labelling pictures in the ImageNet challenge (Russakovsky *et al.* 2015).

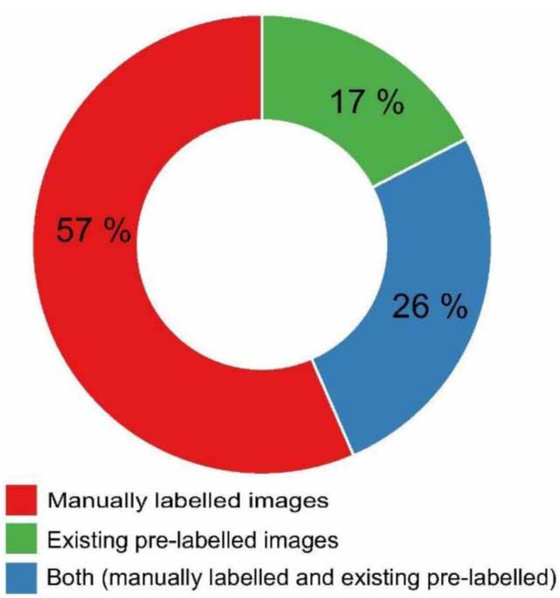

Figure 3 | Training data-labelling approaches employed in the reviewed literature (70 papers).

Table 1 | Summary of publicly-accessible benchmark datasets employed in the literature for flood applications

Data name	Types	References	
MediaEval Multimedia Satellite task 2017	Satellite Imagery and Social Media	Bischke *et al.* (2017)	Combination of satellite and social media images and segmentation masks.
MediaEval Multimedia Satellite task 2018	Satellite Imagery and Social Media	Bischke *et al.* (2018)	Combination of satellite and social media images and segmentation masks.
MediaEval Multimedia Satellite task 2019	Satellite Imagery and Social Media	Bischke *et al.* (2019)	Combination of satellite and social media images and segmentation masks.
European Flood 2013	RGB ground images	Barz *et al.* (2019)	Annotated dataset of images from the central European flood 2013.
SEN12-FLOOD	SAR and Multispectral images	Rambour *et al.* (2020b)	Optical Sentinel-2 and SAR Sentinel-1 image sequences for the detection of flooded areas.
Ms COCO	RGB ground images	Lin *et al.* (2014)	General object recognition datasets.
Cityscapes	RGB ground images	Cordts *et al.* (2016)	Semantic understanding of urban street scenes.

Ms COCO and Cityscapes are general purpose training images used to support model training.

An example of a simple CNN architecture includes one input layer, one convolutional layer, one pooling layer, and one fully connected neural network layer. There are many variants of CNNs as their architectures vary with different aspects such as processing units, parameter and hyper-parameter optimization strategies, design patterns, and connectivity of layers. For an in-depth comparison of various CNN architectures and their evolution, see the review survey conducted by Khan *et al.* (2020). The vast majority of the reviewed work here (~80%) used existing CNN pre-trained architectures such as Alexnet (Krizhevsky *et al.* 2012), VGG (Simonyan & Zisserman 2015), ResNet (He *et al.* 2016), DenseNet (Huang *et al.* 2017), GoogleNet (Szegedy *et al.* 2015), InceptionV3 (Szegedy *et al.* 2016), and Xception (Chollet 2017) (Supplementary Material, Table S1). Some studies developed problem-specific CNN models (e.g., H2O-Network), with the aim of achieving better predictive performance (Li *et al.* 2019; Akiva *et al.* 2020). One of the most used pre-trained networks is the VGG-16 network (Simonyan & Zisserman 2015) that is trained on the ImageNet dataset which contains over a million images and 1,000 labels (Russakovsky *et al.* 2015). VGG16 is composed of 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers. Therefore, the number of layers having tunable parameters is 16 with 13 convolutional layers and 3 fully connected layers. A pre-trained network is used on new data by using layers as feature extractors where each data point can be mapped into a new representation. Lower layers can be considered as low-level feature extractors and are applicable across domains. Whereas higher layers are more domain-specific (Figure 4). Another approach is to use existing weights of the pre-trained network as initialization weights for the new dataset. This approach could prevent overfitting but it is computationally demanding.

For example, Gebrehiwot *et al.* (2019b) used a pre-trained VGG-based fully CNN model to map flood extent using UAV images. The classification results showed classification accuracy with more than 90% compared to 89% using the SVM classification approach.

Some studies used an ensemble of networks to improve performances, for example, Lopez-Fuentes *et al.* (2020) used ensemble of InceptionV3, Xception, VGG16, VGG19, InceptionResNetV2, MobileNet, DenseNet121, DenseNet201, and NaSNetLarge to improve predictive performances. U-net (Ronneberger *et al.* 2015) is also one of the popular architectures used for image segmentation to map flood extent (e.g., Moy De Vitry *et al.* 2019; Wieland & Martinis 2019).

2.4. Types of flood problems

Deep CNNs have increasingly been used in many areas of flood management research in the last years. We identified the following three main flood-related problems where CNNs have been actively employed: (1) the detection of flood presence and extent; (2) flood depth or water level estimation; (3) flood susceptibility mapping; and (4) damage assessment of flood protection structures (Supplementary Material, Table S1).

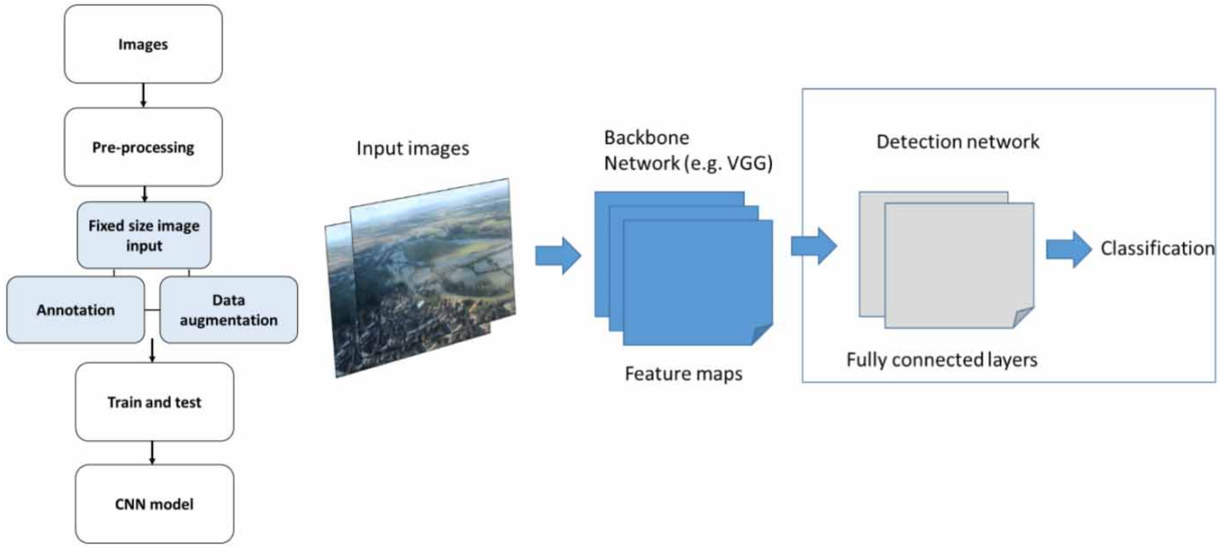

Figure 4 | Typical workflow and CNN architecture.

2.4.1. Detection of flood presence and extent

Distinguishing flooded and non-flooded areas and their extent in image data is one of the widely used applications of CNNs in the reviewed papers. Witherow *et al.* (2019) proposed an image processing pipeline for detecting inundated roadways using crowdsourced pairs of dry and flooded images. The pre-processing approach includes training a region-based R-CNN for vehicle detection on images before applying the water edge detection, image inpainting, and contrast correction algorithms to segment inundated areas. The authors obtained satisfactory performances but they highlighted challenges in recreating the dry images from the same image angle in addition to the drawback of several manually selected model parameters in their proposed approach. Wieland & Martinis (2019) introduced a processing workflow for automated flood monitoring from multispectral satellite data selected from various parts of the world. They used CNN U-Net architecture to discriminate water from shadow, snow, land, and cloud classes and then evaluated the performance of the model with other benchmark approaches such as random forests and a simple NDWI thresholding method. The water class is then further analyzed to separate floodwater from permanent water. Multispectral imagery is also used by Peng *et al.* (2019) to develop a CNN-based data fusion framework for mapping urban flood extent with pre- and post-flooding surface reflectance imagery. The authors proposed the patch similarity CNN with two variants to estimate the similarity between pre- and post-flooding satellite multispectral surface reflectance image patches, and then to determine whether the post-flooding patch under test is flooded. The results indicated good performances and showed that surface reflectance-based data are better than digital numbers-based data as surface reflectance is more stable under varied inconsistent illumination conditions. Similarly, a CNN classification model to map the extent of flooded areas from Landsat satellite images is proposed by Sarker *et al.* (2019). The authors utilized the spatial information from the neighbouring area of target pixel in classification. The proposed classification model was compared to a conventional SVM classification model with the CNN model outperforming the Support Vector Machines (SVM) classification model. Extracting flooded areas from high-resolution UAV imagery is gaining popularity in recent years, Gebrehiwot *et al.* (2019a) fine-tuned a pre-trained CNN model (FCN-16 s) using a small UAV dataset to distinguish floodwater from other classes. They compared CNN performance with more classic machine learning models, such as SVM, and found CNN-based classifier more suitable in flood imagery segmentation with an overall accuracy of 95%.

2.4.2. Flood depth or water level estimation

Several approaches have been developed to estimate food depth from visual images with CNNs. Chaudhary *et al.* (2019) proposed a method to estimate the flood-water level from images gathered from social media by assessing features in the images that are submerged in water. Their workflow includes a first Mask R-CNN to define a set of regions of interest within the image then, for the objects belonging to specific categories (objects with known dimensions) estimate how much objects are submerged in water. From that, a dataset was built with annotating images with flood-water level information that was used to train another CNN model for the instance segmentation and flood level in the image.

This work was extended by Chaudhary *et al.* (2020) by training the model for a small set of annotated water levels (regression task) and a larger set of weak annotated dataset (ranking task) generating an annotated dataset of 8,145 images. The water level was estimated using a multi-task approach with the VGG16 CNN.

CCTV systems and CNNs have been also used to provide qualitative flood level trend information. For example, Moy De Vitry *et al.* (2019) proposed an approach to detect floodwater in CCTV footage and devised a qualitative flood index as a proxy for water level fluctuations visible from a surveillance camera's view point. They first trained a CNN model on 1,218 flooding images collected from the Internet to detect flooded extent area and applied the model to six CCTV videos representing different flooding and lighting conditions. The authors then inferred the water level fluctuations using an index named *static observer flooding index* which is a dimensionless proxy that can be extracted from segmented images of stationary surveillance cameras. The index was computed as the ratio of flooded pixels to the total pixels in the segmented stationary scene. The index value can vary between 0% with no flooding visible and 100% with only flooding visible. When this index is evaluated at multiple consecutive moments in time, the variation of its value provides information about fluctuations of the actual water level.

2.4.3. Flood susceptibility mapping

The CNN has also been employed for flood susceptibility mapping, for instance, Wang *et al.* (2020b) used a spatial database of 13 flood influencing factors such as land use, rainfall, lithology, slop, and 108 historical flood events to develop CNN models to classify and map flood susceptibility in Shangyou, China. The influencing factors layers were converted into a stack of raster of 30×30 m grid size then to small 2D and 3D matrices. Similar susceptibility mapping approach is used by Zhao *et al.* (2020) to assess flood susceptibility for an urban catchment and by Tien Bui *et al.* (2020) for flash floods susceptibility.

2.4.4. Damage assessment of flood protection structures

Damage assessment of flooding structures is one the growing areas in CNN flood applications. Rubio *et al.* (2019) used CNNs (VGG16) to automatically assess damage on concrete bridges. The authors used a database of 734 images bridge damage to detect 'delamination' and 'rebar exposure'. Their model showed mean accuracy of 89.7% for delamination and 78.4% for rebar exposure. The results suggest also that a VGG-based network pre-trained on ImageNet is well suited for texture recognition in the scope of civil infrastructure surface damage semantic segmentation.

More recently, Mishra *et al.* (2020) trained a CNN model (VGG16) to detect and classify drainage blockages based on images of drains and gullies from Google Images and YouTube. The images were classified according to different blockage severity classes. The authors suggested that the trained model could then be used with an IoT-enabled camera to monitor gullies and drainage.

3. BENEFITS OF OPEN ANNOTATED DATA FOR CNNS FLOOD APPLICATIONS

This mini-review demonstrates the growing importance of deep CNNs applied to flood problems. A steady increase of deep CNNs use cases in the literature covering different problems ranging from surface water detection, flood extent delineation, depth or water level estimation, flood prediction, and damage assessment of flood protection structures. The review reveals also various challenges of existing approaches and highlights a need for community collaborations to provide dataset benchmarks that can be used to train and validate models, and improve the model performance and generalization.

3.1. Lack of visual open datasets

Availability of benchmark image datasets is found lacking across the literature. The majority of papers reviewed employ relatively small sets (<1,000 images) of data often manually annotated for training. Though deficiency in benchmark images is a wide limitation in machine learning and CNN applications across fields, it is particularly obvious in flooding applications, and there few libraries of training images for community collaboration. This lack of open datasets leads to reproducibility and replicability issues: currently it is difficult to reproduce CNN-flood research given the inaccessibility of much of the data. Additionally, it is difficult to benchmark the performances of developed models.

3.2. Benefits of an open library

CNNs and machine learning research in general depends on both the availability of training datasets and of objectively interpretable, comparable, and reproducible algorithm benchmarks. Availability of training and benchmark datasets are critical for generating long-term advances in deep learning methods, and their application to flood risk management challenges. The creation of an open flood library for model training is likely to facilitate the creation of datasets, increase collaboration, and improve CNN applications to flooding challenges. Such a library could be used to standardize the setup of data and reporting of benchmarks. Datasets could be uniformly formatted in standardized formats and could be easily downloaded programmatically through Application Programming Interface (APIs) and client libraries. The library could provide standardized train-test to ensure that results can be objectively compared. Testing different models on the same datasets is crucial, users could have possibility to upload their own datasets to the database. Such an approach would reduce the time and efforts that researchers spend implementing existing baselines to conduct comparative experiments.

In terms of the economic benefits of open data, many countries are assessing the benefits of open data in general and recent research showing that higher economic benefits are expected from open data. For example, the European data market was estimated at 54,351 and 59,539 million Euros in 2015 and 2016, respectively (Berends *et al.* 2020). The same study estimated a Gross Domestic Product (GDP) growth of 36.9% in 2016–2020, which saves 2,549 man-hours and reduces energy use by 16%. Furthermore, open data jobs in Europe are estimated to increase from 75,000 to 100,000 jobs between 2016 and 2020 (Berends *et al.* 2020). Open data do not just drive the economic growth but also enables for transparency, accountability, innovation, and knowledge. In the UK, the generated economic output of the data economy was £73.3 billion which is set to grow to £94.6 billion in 2025 (Waterman *et al.* 2021). Public data such as open flood training data can have a high potential for re-use in new services and applications, addressing societal challenges, increasing transparency and accountability, and achieving efficiency gains by sharing data.

There are many flood management applications that would benefit from the availability of open and benchmarked datasets for instance, the detection of blockage of drainage in urban areas. A rich and easily accessed training and benchmark dataset would improve the model performance and enable the automatic and real-time detection of blockages. Another example of a use case where a collaborative and an open library might be beneficial is in the monitoring of the performance of flood prevention structures such as dams. Enhanced availability of such datasets could also improve the performance of the use of UAV technologies for flood detection; for the identification of humans during the search and rescue; or the detection of damage for insurance purposes.

4. CONCLUSIONS

In this review, we provide an overview of the use cases of the CNN for flood management with a focus on the evaluation of the availability visual (images) annotated data. We reviewed the types of visual datasets used and common model architectures applied. We highlighted the need of an open library of training datasets for flood applications and discussed the benefits of such open library. The main highlights from this review are as follows:

- Most CNN flood applications that are widely used across the literature are the detection of flood presence and extent, flood depth or water level estimation, and damage assessment of flood protection structures.
- As training datasets, about 43% of the 70 reviewed papers used ground camera images, while ~37% of papers used satellite images such as Landsat and Sentinel imagery. Other papers used a combination of both satellite and camera images. Only few (7%) of reviewed papers used UAV imagery.
- The majority of papers reviewed employ relatively small sets of data often manually annotated for training. This lack of open datasets leads to reproducibility and replicability issues: currently it is difficult to reproduce CNN-flood research given the inaccessibility of much of the data. Additionally, it is difficult to benchmark the performances of developed models.
- Very few pre-labelled and benchmark datasets are currently available for flood analyses. Most of the studies reviewed used their own manual labelling. Some studies used a combination of existing benchmark pre-labelled data such as MediaEval and the European Flood 2013 datasets. Few studies employed techniques to help automate image annotation.
- Most reviewed studies used existing pre-trained CNN models to extract main image features with few studies employed problem-specific CNN architectures.

Finally, this review highlights the need for more training datasets to accelerate the progress of CNNs applications for flood management. This can be done through the creation of collaborative open training datasets for flood management applications. Building an open dataset is a community effort and requires close collaboration with researchers and organizations generating the data. Furthermore, such an open library would improve access, sharing of datasets, and contribute to overcoming some of the challenges in a collaborative manner, thereby accelerating future advances of CNNs for flood application.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Ahmad, K., Pogorelov, K., Riegler, M., Ostroukhova, O., Halvorsen, P., Conci, N. & Dahyot, R. 2019 Automatic detection of passable roads after floods in remote sensed and social media data. *Signal Processing: Image Communication* **74**, 110–118. https://doi.org/10.1016/j.image.2019.02.002.

Akiva, P., Purri, M., Dana, K., Tellman, B. & Anderson, T. 2020 H2O-Net: Self-supervised flood segmentation via adversarial domain adaptation and label refinement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 111–122.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M. & Farhan, L. 2021 Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data* **8** (1). https://doi.org/10.1186/s40537-021-00444-8.

Barz, B., Schröter, K., Münch, M., Yang, B., Unger, A., Dransch, D. & Denzler, J. 2019 Enhancing flood impact analysis using interactive retrieval of social media images. *ArXiv*. https://doi.org/10.5445/KSP/1000087327/06.

Berends, J., Wendy, C. & Cosmina, R. 2020 *Analytical Report 9: The Economic Benefits of Open Data*. https://doi.org/10.2830/081754.

Bischke, B., Helber, P., Schulze, C., Srinivasan, V., Dengel, A. & Borth, D. 2017 The multimedia satellite task at mediaeval 2017: emergency response for flooding events. *CEUR Workshop Proceedings* **1984**, 13–15.

Bischke, B., Helber, P., Zhao, Z., De Bruijn, J. & Borth, D. 2018 The multimedia satellite task at MediaEval 2018 emergency response for flooding events. *CEUR Workshop Proceedings* **2283**, 4–6.

Bischke, B., Helber, P., Brugman, S., Basar, E., Zhao, Z., Larson, M. & Pogorelov, K. 2019 The multimedia satellite task at mediaeval 2019 estimation of flood severity. *CEUR Workshop Proceedings* **2670**, 4–6.

Brown, A. G., Gregory, K. J. & Milton, E. J. 1987 The use of Landsat multispectral scanner data for the analysis and management of flooding on the river Severn, England. *Environmental Management* **11** (5). https://doi.org/10.1007/BF01880168.

Chaudhary, P., D'Aronco, S., Moy de Vitry, M., Leitão, J. P. & Wegner, J. D. 2019 Flood-water level estimation from social media images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **IV-2/W5** (2/W5), 5–12. https://doi.org/10.5194/isprs-annals-IV-2-W5-5-2019.

Chaudhary, P., D'Aronco, S., Leitão, J. P., Schindler, K. & Wegner, J. D. 2020 Water level prediction from social media images with a multi-task ranking approach. *ISPRS Journal of Photogrammetry and Remote Sensing* **167**, 252–262. https://doi.org/10.1016/j.isprsjprs.2020.07.003.

Chollet, F. 2017 Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1251–1258.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. & Schiele, B. 2016 The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem*, pp. 3213–3223. https://doi.org/10.1109/CVPR.2016.350.

Deutsch, M. & Ruggles, F. 1974 Optical data processing and projected applications of the erts-1 imagery covering the 1973 Mississippi river valley floods. *JAWRA Journal of the American Water Resources Association* **10** (5), 1023–1039. https://doi.org/https://doi.org/10.1111/j.1752-1688.1974.tb00622.x.

Fohringer, J., Dransch, D., Kreibich, H. & Schröter, K. 2015 Social media as an information source for rapid flood inundation mapping. *Natural Hazards and Earth System Sciences* **15** (12), 2725–2738. https://doi.org/10.5194/nhess-15-2725-2015.

Gebrehiwot, A., Hashemi-Beni, L., Thompson, G., Kordjamshidi, P. & Langan, T. E. 2019a Deep convolutional neural network for flood extent mapping using unmanned aerial vehicles data. Sensors *(*Switzerland*)* **19** (7). https://doi.org/10.3390/s19071486.

Gebrehiwot, A., Hashemi-Beni, L., Thompson, G., Kordjamshidi, P. & Langan, T. E. 2019b Deep convolutional neural network for flood extent mapping using unmanned aerial vehicles data. Sensors *(*Switzerland*)*. https://doi.org/10.3390/s19071486.

Gupta, R. & Shah, M. 2020 Rescuenet: Joint building segmentation and damage assessment from satellite imagery. In *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, pp. 4405–4411.

Hashemi-Beni, L. & Gebrehiwot, A. 2021 Flood extent mapping: an integrated method using deep learning and region growing using UAV optical data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **14**, 2127–2135. https://doi.org/10.1109/JSTARS.2021.3051873.

He, K., Zhang, X., Ren, S. & Sun, J. 2016 Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778.

Hoeser, T. & Kuenzer, C. 2020 Object detection and image segmentation with deep learning on earth observation data: a review-part I: evolution and recent trends. *Remote Sensing* **12** (10), 1667. https://doi.org/10.3390/rs12101667.

Horritt, M. S., Mason, D. C. & Luckman, A. J. 2001 Flood boundary delineation from synthetic aperture radar imagery using a statistical active contour model. *International Journal of Remote Sensing* **22** (13). https://doi.org/10.1080/01431160116902.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. 2017 Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4700–4708.

Huang, J., Kang, J., Wang, H., Wang, Z. & Qiu, T. 2020 A novel approach to measuring urban waterlogging depth from images based on mask region-based convolutional neural network. *Sustainability* **12** (5), 2149. https://doi.org/10.3390/su12052149.

Inman, V. L. & Lyons, M. B. 2020 Automated inundation mapping over large areas using landsat data and google earth engine. *Remote Sensing* **12** (8). https://doi.org/10.3390/RS12081348.

Iqbal, U., Perez, P., Li, W. & Barthelemy, J. 2021 How computer vision can facilitate flood management: a systematic review. *International Journal of Disaster Risk Reduction* **53**, 102030. https://doi.org/10.1016/j.ijdrr.2020.102030.

Jain, S. K., Saraf, A. K., Goswami, A. & Ahmad, T. 2006 Flood inundation mapping using NOAA AVHRR data. *Water Resources Management* **20** (6). https://doi.org/10.1007/s11269-006-9016-4.

Jain, P., Schoen-Phelan, B. & Ross, R. 2020 Automatic flood detection in SentineI-2 images using deep convolutional neural networks. In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pp. 617–623. https://doi.org/10.1145/3341105.3374023.

Khan, A., Sohail, A., Zahoora, U. & Qureshi, A. S. 2020 A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review* **53** (8), 5455–5516. https://doi.org/10.1007/s10462-020-09825-6.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. 2012 Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* **25**, 1097–1105.

Le Coz, J., Hauet, A., Pierrefeu, G., Dramais, G. & Camenen, B. 2010 Performance of image-based velocimetry (LSPIV) applied to flash-flood discharge measurements in Mediterranean rivers. *Journal of Hydrology* **394** (1–2). https://doi.org/10.1016/j.jhydrol.2010.05.049.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W. & Jackel, L. 1989 Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems* **2**.

Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. 1998 Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86** (11), 2278–2324. https://doi.org/10.1109/5.726791.

Li, Y., Martinis, S. & Wieland, M. 2019 Urban flood mapping with an active self-learning convolutional neural network based on TerraSAR-X intensity and interferometric coherence. *ISPRS Journal of Photogrammetry and Remote Sensing* **152**, 178–191. https://doi.org/10.1016/j.isprsjprs.2019.04.014.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. 2014 Microsoft COCO: common objects in context. Lecture Notes in Computer Science *(*Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*)* **8693** (PART 5), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

Lopez-Fuentes, L., Farasin, A., Zaffaroni, M., Skinnemoen, H. & Garza, P. 2020 Deep learning models for road passability detection during flood events using social media data. *Applied Sciences* **10** (24), 8783. https://doi.org/10.3390/app10248783.

Mishra, B. K., Thakker, D., Mazumdar, S., Neagu, D., Gheorghe, M. & Simpson, S. 2020 A novel application of deep learning with image cropping: a smart city use case for flood monitoring. *Journal of Reliable Intelligent Environments*. https://doi.org/10.1007/s40860-020-00099-x.

Moy De Vitry, M., Kramer, S., Dirk Wegner, J. & Leitao, J. P. 2019 Scalable flood level trend monitoring with surveillance cameras using a deep convolutional neural network. *Hydrology and Earth System Sciences* **23** (11), 4621–4634. https://doi.org/10.5194/hess-23-4621-2019.

Muhadi, N. A., Abdullah, A. F., Bejo, S. K., Mahadi, M. R. & Mijic, A. 2020 Image segmentation methods for flood monitoring system. *Water (Switzerland)* **12** (6), 1–10. https://doi.org/10.3390/w12061825.

Panahi, M., Jaafari, A., Shirzadi, A., Shahabi, H., Rahmati, O., Omidvar, E., Lee, S. & Bui, D. T. 2021 Deep learning neural networks for spatially explicit prediction of flash flood probability. *Geoscience Frontiers* **12** (3), 1076. https://doi.org/10.1016/j.gsf.2020.09.007.

Park, S., Lee, N., Han, Y. & Hahn, H. 2009 The water level detection algorithm using the accumulated histogram with band pass filter. *World Academy of Science, Engineering and Technology* **56**. https://doi.org/10.5281/zenodo.1072871.

Peng, B., Meng, Z., Huang, Q. & Wang, C. 2019 Patch similarity convolutional neural network for urban flood extent mapping using bi-temporal satellite multispectral imagery. *Remote Sensing* **11** (21). https://doi.org/10.3390/rs11212492.

Profeti, G. & Macintosh, H. 1997 Flood management through landsat TM and ERS SAR data: a case study. *Hydrological Processes* **11** (10). https://doi.org/10.1002/(sici)1099-1085(199708)11:10<1397::aid-hyp530>3.0.co;2-y.

Qi, S., Brown, D. G., Tian, Q., Jiang, L., Zhao, T. & Bergen, K. M. 2009 Inundation extent and flood frequency mapping using LANDSAT imagery and digital elevation models. *GIScience and Remote Sensing* **46** (1). https://doi.org/10.2747/1548-1603.46.1.101.

Rambour, C., Audebert, N., Koeniguer, E., Le Saux, B., Crucianu, M. & Datcu, M. 2020a Flood detection in time series of optical and sar images. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences – ISPRS Archives* **43** (B2), 1343–1346. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1343-2020.

Rambour, C., Audebert, N., Koeniguer, E., Le Saux, B., Crucianu, M. & Datcu, M. 2020b *SEN12-FLOOD: A SAR and Multispectral Dataset for Flood Detection*. IEEE Dataport. https://doi.org/https://dx.doi.org/10.21227/w6xz-s898.

Rango, A. & Anderson, A. T. 1974 Flood hazard studies in the Mississippi river basin using remote sensing. *JAWRA Journal of the American Water Resources Association* **10** (5). https://doi.org/10.1111/j.1752-1688.1974.tb00625.x.

Rivas Casado, M., Irvine, T., Johnson, S., Palma, M. & Leinster, P. 2018 The use of unmanned aerial vehicles to estimate direct tangible losses to residential properties from flood events: a case study of Cockermouth following the Desmond storm. *Remote Sensing* **10** (10), 1548. https://doi.org/10.3390/rs10101548.

Ronneberger, O., Fischer, P. & Brox, T. 2015 U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, pp. 234–241.

Rubio, J. J., Kashiwa, T., Laiteerapong, T., Deng, W., Nagai, K., Escalera, S., Nakayama, K., Matsuo, Y. & Prendinger, H. 2019 Multi-class structural damage segmentation using fully convolutional networks. *Computers in Industry* **112**. https://doi.org/10.1016/j.compind.2019.08.002.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. & Fei-Fei, L. 2015 Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* **115** (3). https://doi.org/10.1007/s11263-015-0816-y.

Sarker, C., Mejias, L., Maire, F. & Woodley, A. 2019 Flood mapping with convolutional neural networks using spatio-contextual pixel information. *Remote Sensing* **11** (19), 1–25. https://doi.org/10.3390/rs11192331.

Sazara, C., Cetin, M. & Iftekharuddin, K. M. 2019 Detecting floodwater on roadways from image data with handcrafted features and deep transfer learning. In: *2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019*, pp. 804–809. https://doi.org/10.1109/ITSC.2019.8917368.

Simonyan, K. & Zisserman, A. 2015 Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 – Conference Track Proceedings*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. 2015 Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. 2016 Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2818–2826.

Tien Bui, D., Hoang, N. D., Martínez-Álvarez, F., Ngo, P. T. T., Hoa, P. V., Pham, T. D., Samui, P. & Costache, R. 2020 A novel deep learning neural network approach for predicting flash flood susceptibility: a case study at a high frequency tropical storm area. *Science of the Total Environment* **701**, 134413. https://doi.org/10.1016/j.scitotenv.2019.134413.

Tkachenko, N., Jarvis, S. & Procter, R. 2017 Predicting floods with Flickr tags. *PLOS ONE* **12** (2), e0172870. https://doi.org/10.1371/journal.pone.0172870.

Wang, R.-Q., Mao, H., Wang, Y., Rae, C. & Shaw, W. 2018 Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Computers & Geosciences* **111**, 139–147. https://doi.org/10.1016/j.cageo.2017.11.008.

Wang, Y., Fang, Z., Hong, H. & Peng, L. 2020a Flood susceptibility mapping using convolutional neural network frameworks. *Journal of Hydrology* **582**, 124482. https://doi.org/10.1016/j.jhydrol.2019.124482.

Wang, Y., Fang, Z., Hong, H. & Peng, L. 2020b Flood susceptibility mapping using convolutional neural network frameworks. *Journal of Hydrology* **582**, 124482. https://doi.org/10.1016/j.jhydrol.2019.124482.

Waterman, L., Casado, R., Bergin, E. & Mcinally, G. 2021 *A Mixed-Methods Investigation into Barriers for Sharing Geospatial and Resilience Flood Data in the UK*, pp. 1–16.

Wieland, M. & Martinis, S. 2019 A modular processing chain for automated flood monitoring from multi-spectral satellite data. *Remote Sensing* **11** (19). https://doi.org/10.3390/rs11192330.

Witherow, M. A., Sazara, C., Winter-Arboleda, I. M., Elbakary, M. I., Cetin, M. & Iftekharuddin, K. M. 2019 Floodwater detection on roadways from crowdsourced images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization* **7** (5–6), 529–540. https://doi.org/10.1080/21681163.2018.1488223.

Yang, H. C., Wang, C. Y. & Yang, J. X. 2014 Applying image recording and identification for measuring water stages to prevent flood hazards. *Natural Hazards* **74** (2). https://doi.org/10.1007/s11069-014-1208-2.

Zhao, G., Pang, B., Xu, Z., Peng, D. & Zuo, D. 2020 Urban flood susceptibility assessment based on convolutional neural networks. *Journal of Hydrology* **590**, 125235. https://doi.org/10.1016/j.jhydrol.2020.125235.

First received 1 September 2021; accepted in revised form 26 April 2022. Available online 17 May 2022

The need for training and benchmark datasets for convolutional neural networks in flood applications

Khouakhi, Abdou

2022-05-17

Khouakhi A, Zawadzka J. Truckell I. (2022) The need for training and benchmark datasets for convolutional neural networks in flood applications, Hydrology Research, Volume 53, Issue 6, pp. 795-806

https://doi.org/10.2166/nh.2022.093