Tri-Band Assessment of Multi-Spectral Satellite Data for Flood Detection

Pallavi Jain, Bianca Schoen-Phelan, Robert Ross

Technological University Dublin, Dublin, Ireland
{pallavi.jain, bianca.schoenphelan, robert.ross}@tudublin.ie

Abstract. Multi-spectral satellite data provides vast resources for important tasks such as flood detection, but training and fine tuning models to perform optimally across multi-spectral data remains a significant research challenge. In light of this problem, we present a systematic examination of the role of tri-band deep convolutional neural networks in flood prediction. Using Sentinel-2 data we explore the suitability of different deep convolutional architectures in a flood detection task; in particular we examine the utility of VGG16, ResNet18, ResNet50 and EfficientNet. Importantly our analysis considers the questions of different band combinations and the issue of pre-trained versus non-pre-trained model application. Our experiment shows that a 0.96 F1 score is achievable for our task through appropriate combinations of spectral bands and convolutional neural networks. For flood detection, three-band combinations of RB8aB11 and RB11B outperformed 33 other combinations when trained with pre-trained ResNet18 and other models. Our analysis further demonstrates a strong performance by pre-trained models despite the fact that these pre-trained models were originally trained on different spectral bands.

Keywords: Remote Sensing · Deep Convolutional Neural Network · Multi-spectral · Flood Detection · Sentinel-2.

1 Introduction

Floods are natural hazards that occur throughout the year in many different parts of the world, and can occur for reasons including heavy rainfall, melting snow or tsunamis. Floods damages properties and agricultural areas, and are also highly hazardous to human life [10]. These factors together highlight the need for timely and accurate detection in order to make flood management operation more effective. To this end satellite data processing has become a vital diagnostic tool.

Satellite data provides much more than simple RGB (Red-Green-Blue) image data. Instead satellites are multi-spectral instruments (MSI) that generate multi-spectral and sometimes hyper-spectral data over very different wavebands to RGB. The importance of this is that different spectral bands are reflected or absorbed differently depending on geo-physical properties, e.g., short-wave infrared (SWIR) bands can discriminate between wet soil and dry soil whereas near infrared (NIR) wavelength are absorbed by water and reflects by vegetation. This property of MSI provides volumes of information in terms of reflectance at specific wavelengths across geographic areas.

Observing and mapping floods using satellite image data is an active area of research with many promising and effective techniques being developed [18, 15]. The main challenges using satellite data present in urban areas and in areas of high vegetation coverage. Challenges are due to water having variations in optical properties across different geographical features such as rivers, lakes, oceans, ponds and floods due to their different compositions, depths and local interactions. Local composition issues make shallow waters, for example rivers, difficult to detect.

Approaches to the multi-spectral imaging processing have evolved in recent years. Traditional approaches to multi-spectral data processing rely on handcrafted features of spectral reluctances. These indexing techniques are known to be sub-optimal for image processing. Recent trends in Deep Neural Networks provide an opportunity to learn optimal predictive functions that fully leverage the potential of available MSI satellite data. While deep networks and CNNs in particular are highly successful in image classification, they generally requires large amount of labelled data. For this reason pre-trained models are of great use, and have shown great performances in several domains.

One challenge with pre-trained models is that they have generally been trained only on 3 channel image data corresponding to Red, Green and Blue information. The specific types of features that they have been trained to identify are dependent on combinations of these spectral bands. It is not obvious whether the features that have been learned are easily applicable to other combinations of spectral bands such as what is required for flood detection.

Given the innate advantage of CNNs in a task such as flood detection, in this paper we examine the effectiveness of pre-trained vs non-pre-trained CNN models on multi-spectral satellite image data processing. We do this specifically for the task of flood detection, using Sentinel-2 data. Three concerns are particularly interesting here: how well do different pre-trained models trained on RGB compare to each other when used for to our multi-spectral flood detection task; is there a notable improvements over models trained from scratch on multi-spectral input data; and is there a specific three-band combination of spectral bands that outperforms others (pre-trained models have been trained on three bands which means that three bands need to be input when using such models). Before proceeding to introduce the specifics of our investigation, we first review some key related work.

2 Related Work

It is long known that different wavelengths vary considerably in their reflectances for different geo-physical properties such as presence of water bodies or vegetation, and indeed the presence of specific chemicals. Given these reflectance properties, many thresholding techniques [2, 12] and water indexing techniques [13, 19, 6] have been proposed in the past to identify pixels corresponding to water in multi-spectral images.

Since these handcrafted indexing techniques have been developed to detect water bodies in general, they are not useful for distinguishing flood water from permanent water bodies. One of the reasons for their inefficiency is that flood water is usually shallow and possesses a similar reflectance value as built-up areas or cloud shadows, and that their images contain mixed pixel values due to sediments and other objects [3]. Ideally, in order to detect a flood these techniques require pre-flood and post-flood images, which may not be always available. Since water indices are based on the bands capabilities of highlighting certain geological properties, it is reasonable to analyse different combinations of bands. Additionally, this provides a wider and more robust detection criteria, using automatic detection without any handcrafted features or static thresholding, as these are highly sensitive to several factors such as region, and similar spectral signature.

Rather than having to rely on handcrafted features, in recent years it has been shown that deep learning based CNNs provide great performance in several remote sensing tasks, for example scene classification, building mapping or detection of passable roads [8, 1]. There have been many variants on CNN architectures over the last few years with models such as VGG [16], ResNet [7], EfficientNet [17] on the forefront. Unfortunately, there is no universal CNN architecture that could be applied to every task, and a model that is state-of-the-art in one domain, is not necessarily guaranteed to be effective in another domain. Since CNNs have the ability to learn spatial as well as spectral features of the images [20], this is valuable in multi-spectral satellite imaging where spatial variance is also important. For example, EuroSat data contains a large amount of high resolution images for land cover classification. ResNet50 is the most successful model on this data, classifying with 98.57% accuracy [8].

Training a residual network requires a large amount of data, which may not be available. Transfer learning, which means pre-training the network on huge amount of available datasets such as ImageNet on the ILSVRC challenge [5], and then adjusting the model by inputting domain specific data, may be the solution in those cases and have demonstrated very good performance in the past [9, 11]. The drawback of transfer learning is that all available pre-trained models use three channels, i.e. RGB. As multi-spectral satellite data typically offers around 10-13 bands, specific three band combinations need to be found that are suitable for flood detection from satellite images, which is where this work seeks to make a contribution. It is currently unclear whether wavelength dependent deep learning models generalise well to other wavelength combinations.

3 Study Design

Below we set out the detail of our study including the selection of data, the selection of candidate CNN models, the training processing, and evaluation methods.

3.1 Data

SENTINEL-2 provides high-resolution, multi-spectral images, and monitors land, vegetation, soil and water cover, as well as observation of inland waterways and coastal areas. It has multi-spectral instrument (MSI) samples with 13 spectral bands. Among the 13 spectral bands, four bands are at 10 metres, six bands at 20 metres and three bands at 60 metres spatial resolution.

We leverage the annotated dataset provided by the MediaEval[1] 2019 competition [4]. The data consists of 335 image sets with 267 identified as development sets and 68 as test sets. Each set consists of between 1 to 24 day time series images of before and after flood events; this provides a total of 2,770 images. We filtered images which did not had full coverage or had full cloud coverage, which made our dataset total with 2180 images. The data has 12 bands instead of the 13 bands available on SENTINEL-2, which comes in three different sets of resolutions: 10 metres, 20 metres and 60 metres. Each 10 metre resolution images 512 X 512 pixels in size, 20 metre resolution images are of 256 X 256 pixels, and 60 metre images are of 128 X 128 pixels in size. Among 12 bands, we excluded 60m band images, which left us with 10 bands in total.

The development dataset is split into three parts: training, validation, and test, in the ratio of 80:10:10. The split was done based on location that is 267 set of images, to remove any learning bias in our test data. Although the data provides us with time-series data but we haven't utilise the temporal information in our current work. Additionally, we performed image augmentation by shifting, rotating, and flipping the images with batch sizes of 16 and 8 in order to increase our training dataset and remove any biases.

3.2 Image Pre-Processing and Processing

As the reflectance value range varies considerably for different bands, we normalised image's each band's pixel or reflectance value to a standard range from 0-255. Furthermore, we up-scaled the 20m resolution band images to 10m resolution images due to the fact that images were provided with two different spatial resolutions. After normalising and up-scaling we stacked three different bands together to form various three channel combinations. The process of selecting the three band combinations was based on the selection of bands rather than sequence they are stacked.

Sentinel-2 bands	Wavelength (nm)	Resolution(m)
B – Blue	492	10
G – Green	559	10
R – Red	664	10
B5 – Vegetation red edge	704	20
B6 – Vegetation red edge	740	20
B7 – Vegetation red edge	782	20
B8 – NIR	832	10
B8a – Narrow NIR	864	20
B11 – SWIR	1613	20
B12 – SWIR	2202	20

Table 1: Band Notations Used and their Wavelength and Spatial Resolution

[1] http://www.multimediaeval.org/mediaeval2019/

This made 120 combinations from 10 base bands, out of which we showed only 33 combinations in this work, whose overall F1 score was greater than threshold value of 75.

Available spectral bands, their names, and shorthand notations for each band are provided in Table 1; here we used R, G, B as is, while other bands notations were used as per their wavelength positions. Given this notation, and the 10 bands, we analysed 33 band combinations as follows: RB8aB11, RB11B, B7B11B, B7GB11, B8B11R, B6GB, B8aGB, RB8aB, B7GB, B7GB8, RGB11, B7B8aB11, B7B8aB, B8GB, RB8B, RB8aB7, RGB8, B7B8B, B11GB, B7B8B11, RB11B7, RB7B, B8aB11B, B8aGB11, B5GB, B7GB8a, RGB7, RGB8a, RB12B, B8GB11, RB8B7, RGB and B8B11B.

3.3 CNN based Training Models

The basic building block of a CNN is a multi-tier network which includes convolution layer, pooling layer, and fully connected layer, which extracts the features such as lines and edges. Over time CNNs have evolved and are now able to learn more features by increasing the depth and width of the network, which we call a deep CNN. There are many variants of CNN architectures available. We examine three important variants: VGG [16], Deep Residual Network ResNet [7], and EfficientNet [17]. These networks are popular due to their performance in image classification tasks. All considered networks are very deep and released in several depth variants. For example, VGG has an architecture with 16 and an architecture with 19 layers.

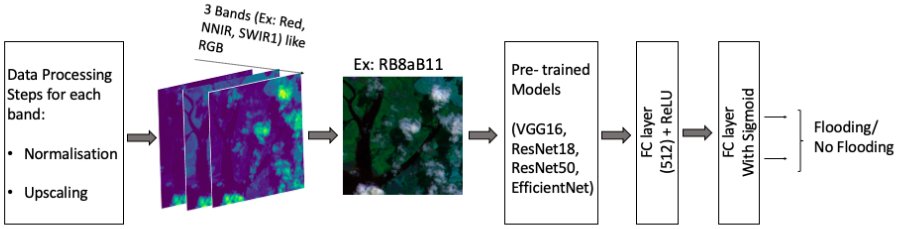


Fig. 1: Experiment Design

This study uses ImageNet pre-trained variants of VGG16, ResNet18, ResNet50 and EfficientNetB0. Global average pooling was used at the output of each pre-trained model, which then fed to a fully connected layer of 512 units with a ReLU activation function. In order to avoid over-fitting during training, a dropout of 0.5 was used. The final output layer used a sigmoid function for binary classification.

Pre-trained models are compared with their respective versions trained from scratch on our available dataset, described above. A VGG16, ResNet18 and

ResNet50 models was trained. For both pre-trained and 'from scratch' models, rectified linear unit (ReLU) has been used as activation function and the Adam optimiser was applied to guide the training process. Flood detection is a binary classification problem, i.e. flood present and flood not present. Therefore, a binary cross entropy loss function was used in order to calculate the loss. For training of the model an initial learning rate of 5e-6 was used. All pre-trained models were trained for 100 epochs with a batch size of 16. For VGG16, ResNet18 and ResNet50 models from scratch, a batch size of 8 was used and training occurred for a variable number of epochs, as per their best performance. For the training phase an NVIDIA Tesla K40m GPU was used.

3.4 Evaluation Method

For evaluation of the binary classification i.e. flooded and non-flooded region, F1 score and Kappa coefficient are used. The F1 score is a harmonic mean of the precision and recall performance metrics. Cohen's Kappa meanwhile compares observed accuracy with expected accuracy. Kappa provides the fair comparison when the classes are imbalanced as in the case for our data.

4 Results

Results focus on the top 10 waveband combinations, an analysis of the top 2 performing results across the pre-trained CNN variants, and finally, an analysis of pre-trained models versus those trained from scratch.

4.1 Three-Band Combinations

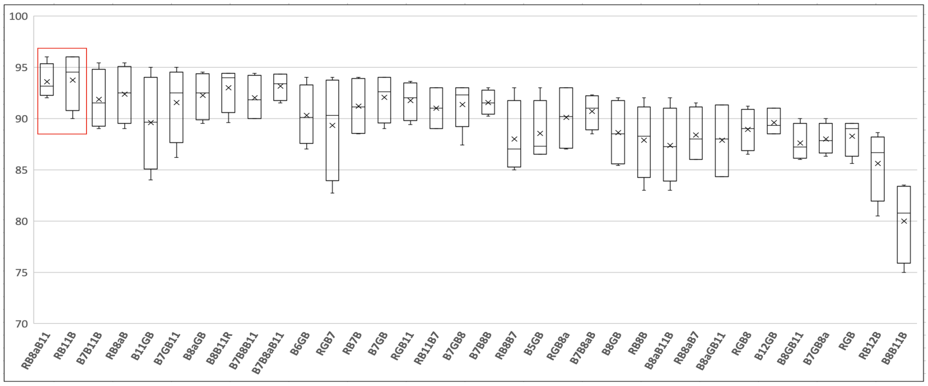


Fig. 2: Comparative Results of 33 Combinations from VGG16, ResNet18, ResNet50, and EfficientNetB0

In order to have the best performing combinations for detecting flood events, 33 different waveband combinations were run. Although each pre-trained model has its own best combination, some combinations performed better overall on all the models. Figure 2 shows the boxplot result for each combination's F1 score spread across the four pre-trained models. Along with boxplot, mean lines are included to clearly show the average for each combination. Overall competitive results are produced by RB11B, RB8aB11, and B7B11B. While RGB, RB12B, and B8B11B under-perform across all models. This is consistent with previous research demonstrating that NIR and SWIR bands both are good at highlighting water pixels [13, 14, 19]. The performance of RB8aB11 shows that the NNIR band is also capable of providing good results when combined with Red and SWIR bands. Furthermore, with respect to the performance of RB11B, the Blue band is useful for mapping depths and shapes of underwater terrain, and distinguishes soil from vegetation. This highlights water pixels in vegetated and urban areas specifically when combined with the other successful bands.

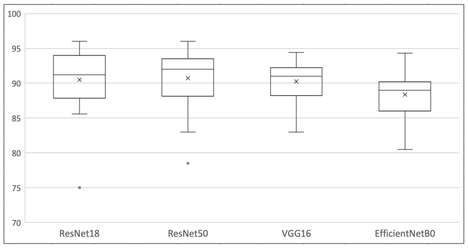


Fig. 3: Comparison of VGG16, ResNet18, ResNet50, and EfficientNetB0 model's F1 Score Across All the Combinations

Combination	Model	F1	Kappa	TN	TP
RB8aB11	**ResNet18**	**96**	**91.3**	**95**	**97**
RB11B	**ResNet18**	**96**	**91.2**	**97**	**94**
B7B11B	**ResNet18**	**95.4**	**90.5**	**95**	**95**
RB8aB	ResNet50	95.4	90	96	94
B11GB	ResNet50	95	89	95	94
B7GB11	ResNet18	95	89	96	93
B8aGB	ResNet50	94.5	88.4	95	94
RB8B11	ResNet18	94.4	88.2	92	98
B7B8B11	VGG16	94.4	88	97	90
B7B8aB11	EfficientNetB0	94.3	88	94	95

Table 2: Top-10 Best Performing Combinations in Terms of F1 and Kappa

4.2 Pre-Trained Model Performance

Four pre-trained models are run across 33 different three-band combinations: VGG16, ResNet18, ResNet50 and EfficientNetB0. According to figure 3, VGG16 and EfficientNetB0 performed relatively poorly in comparison to ResNet18 and ResNet50. ResNet50 achieved slightly higher median and average compared to ResNet18 but ResNet18 has the more compact spread compared to ResNet50. Increasing the layers improves the performance. However, ResNet18 and ResNet50 perform among the top 3 in terms of accuracy, as illustrated in table 2. This indicates that the water identification process may not necessarily require much deeper models than 18, which has the added benefit of being detected at earlier stages as the model finishes faster.

4.3 Pre-Trained vs Models from Scratch

Given that RB8aB11 showed overall best results, VGG16, ResNet18 and ResNet50 models were trained from scratch for this specific target combination. The number of epochs varied depending on performance of the model. Result are illustrated in table 3. Among the three models trained from scratch, ResNet50 showed the overall best results but it could not compete with pre-trained models. Moreover, ResNet50's accuracy for training and validation flat-lined after 500 epochs.

Model	F1	Kappa	Epoch
ResNet18	89	76.5	500
ResNet50	**93**	**84**	**500**
VGG16	89	76	300
VGG16	86.4	71.6	400

Table 3: Results for Models Trained from Scratch on RB8aB11

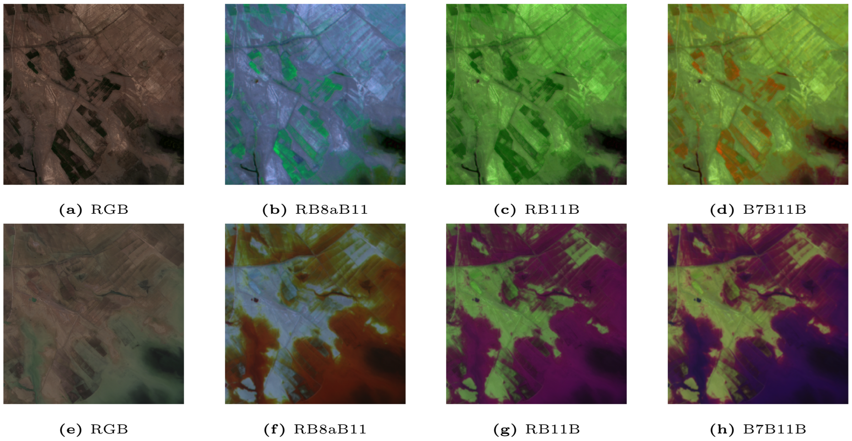
As seen earlier in table 2, the top-3 accuracy was achieved by the ResNet18 pre-trained model, while ResNet50 shows the best result among the 'from scratch' models. It could be argued that pre-trained models require less depth, while training from scratch requires deeper models in order to achieve the competitive results. This also highlights that it could be possible to improve the accuracy by training networks deeper than Resnet50 from scratch in order to compete with pre-trained model performance. This affects computational cost. Among the four models VGG16 takes the longest time to train while ResNet18 and EfficientNetB0 were the fastest.

5 Discussion

SWIR is known for its ability to penetrate thin clouds, smoke and haze better than visible bands. Considering the low reflectance property of water in SWIR bands, it is perhaps not a surprise that SWIR1 is the most common band seen in the Top-10 accuracy results in table 2. Consequently, SWIR1 (B11) is a good separator of water when bands with different or similar reflectance properties are combined. It is also noticeable, that not all combinations with SWIR1 (B11) perform well, as B8B11B achieved the poorest result. This could be due to the fact that none of the three bands distinguish vegetation. Instead, they provide predominantly water and cloud identifiers. This also indicates the need for balance among bands in order to highlight a particular feature. Therefore, it is crucial for further analysis to choose the right combination of bands.

6 Conclusion

This study illustrates that some particular spectral bands combinations excel with almost all deep CNN models in the flood classification task. The best three combinations observed were RB8aB11, RB11B and B7B11B. We suggest that this is due to their spectral sensitivity to distinguish water and soil/vegetation clearly. Significantly, our analysis also showed that pre-trained models easily outperform models trained from scratch. Amongst the examined architectures ResNet18 performed best in the case of pre-trained data, while ResNet50 performed best in the case of training models from scratch. On hand hand this is unsurprising since pre-trained models trained on large volumes of data are



(a) RGB **(b)** RB8aB11 **(c)** RB11B **(d)** B7B11B

(e) RGB **(f)** RB8aB11 **(g)** RB11B **(h)** B7B11B

Fig. 4: Pre and Post Flood Images for RGB and Top-3 Band Combination , where (a),(b),(c),(d) are Pre-Flood images and (e),(f),(g),(h) are Post-Flood images

known to give benefit to domains with low volumes of data. On the other hand it is somewhat surprising that the differences in wavebands and not only training task does not provide a greater limitation when the pre-trained models are applied to new domains. We suggest that this will be due to the relative dominance of spatial features rather than specific spectral combinations in the pre-trained networks.

One limitation of this study is that only three-band combinations were considered. This is due to our goal of examining pre-trained networks that to this point are by definition trained using three channels and consequently expect three channels as input data. Examining larger combinations of spectral bands is naturally important. A natural extension of this work is to investigate the use of actual MSI data in the construction of suitable pre-trained networks that can be applied to multi-spectral tasks. While we leave some of these issues for future work, we see the current work as a useful contribution in confirming the overall applicability of pre-trained networks in satellite based imaging tasks such as flood detection.

References

1. Ahmad, K., Pogorelov, K., Riegler, M., Ostroukhova, O., Halvorsen, P., Conci, N., Dahyot, R.: Automatic detection of passable roads after floods in remote sensed and social media data. Signal Processing: Image Communication **74**, 110–118 (2019)
2. Ban, H.J., Kwon, Y.J., Shin, H., Ryu, H.S., Hong, S.: Flood monitoring using satellite-based rgb composite imagery and refractive index retrieval in visible and near-infrared bands. Remote Sensing **9**(4), 313 (2017)
3. Bangira, T., Alfieri, S.M., Menenti, M., van Niekerk, A.: Comparing thresholding with machine learning classifiers for mapping complex water. Remote Sensing **11**(11), 1351 (2019)

4. Benjamin Bischke, Patrick Helber, Erkan Basar, Simon Brugman, Zhengyu Zhao and Konstantin Pogorelov: The Multimedia Satellite Task at MediaEval 2019: Flood Severity Estimation. In: Proc. of the MediaEval 2019 Workshop. Sophia-Antipolis, France (2019)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Feyisa, G.L., Meilby, H., Fensholt, R., Proud, S.R.: Automated water extraction index: A new technique for surface water mapping using landsat imagery. Remote Sensing of Environment **140**, 23–35 (2014)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019)
9. Hu, F., Xia, G.S., Hu, J., Zhang, L.: Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sensing **7**(11), 14680–14707 (2015)
10. Jonkman, S.N., Kelman, I.: An analysis of the causes and circumstances of flood disaster deaths. Disasters **29**(1), 75–97 (2005)
11. Liu, X., Chi, M., Zhang, Y., Qin, Y.: Classifying high resolution remote sensing images by fine-tuned vgg deep networks. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. pp. 7137–7140. IEEE (2018)
12. Matgen, P., Hostache, R., Schumann, G., Pfister, L., Hoffmann, L., Savenije, H.: Towards an automated sar-based flood monitoring system: Lessons learned from two case studies. Physics and Chemistry of the Earth, Parts A/B/C **36**(7-8), 241–252 (2011)
13. McFeeters, S.K.: The use of the normalized difference water index (ndwi) in the delineation of open water features. International journal of remote sensing **17**(7), 1425–1432 (1996)
14. Mishra, K., Prasad, P.: Automatic extraction of water bodies from landsat imagery using perceptron model. Journal of Computational Environmental Sciences **2015** (2015)
15. Notti, D., Giordan, D., Caló, F., Pepe, A., Zucca, F., Galve, J.P.: Potential and limitations of open satellite data for flood mapping. Remote sensing **10**(11), 1673 (2018)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114 (2019)
18. Wieland, M., Martinis, S.: A modular processing chain for automated flood monitoring from multi-spectral satellite data. Remote Sensing **11**(19), 2330 (2019)
19. Xu, H.: Modification of normalised difference water index (ndwi) to enhance open water features in remotely sensed imagery. International journal of remote sensing **27**(14), 3025–3033 (2006)
20. Yue, J., Zhao, W., Mao, S., Liu, H.: Spectral–spatial classification of hyperspectral images using deep convolutional neural networks. Remote Sensing Letters **6**(6), 468–477 (2015)