Deep Learning for Geo-referenced Data
Case Study: Earth Observation

Nosheen Abid

Dept. of Computer Science, Electrical and Space Engineering
Luleå University of Technology
Luleå, Sweden

Supervisors:

Marcus Liwicki, Faisal Shafait, and György Kovács

To my Lord who created me...

Abstract

The thesis focuses on machine learning methods for Earth Observation (EO) data, more specifically, remote sensing data acquired by satellites and drones. EO plays a vital role in monitoring the Earth's surface and modelling climate change to take necessary precautionary measures. Initially, these efforts were dominated by methods relying on handcrafted features and expert knowledge. The recent advances of machine learning methods, however, have also led to successful applications in EO. This thesis explores supervised and unsupervised approaches of Deep Learning (DL) to monitor natural resources of water bodies and forests.

The first study of this thesis introduces an Unsupervised Curriculum Learning (UCL) method based on widely-used DL models to classify water resources from RGB remote sensing imagery. In traditional settings, human experts labeled images to train the deep models which is costly and time-consuming. UCL, instead, can learn the features progressively in an unsupervised fashion from the data, reducing the exhausting efforts of labeling. Three datasets of varying resolution are used to evaluate UCL and show its effectiveness: SAT-6, EuroSAT, and PakSAT. UCL outperforms the supervised methods in domain adaptation, which demonstrates the effectiveness of the proposed algorithm.

The subsequent study is an extension of UCL for the multispectral imagery of Australian wildfires. This study has used multispectral Sentinel-2 imagery to create the dataset for the forest fires ravaging Australia in late 2019 and early 2020. 12 out of the 13 spectral bands of Sentinel-2 are concatenated in a way to make them suitable as a three-channel input to the unsupervised architecture. The unsupervised model then classified the patches as either burnt or not burnt. This work attains 87% F1-Score mapping the burnt regions of Australia, demonstrating the effectiveness of the proposed method.

The main contributions of this work are (i) the creation of two datasets using Sentinel-2 Imagery, PakSAT dataset and Australian Forest Fire dataset; (ii) the introduction of UCL that learns the features progressively without the need of labelled data; and (iii) experimentation on relevant datasets for water body and forest fire classification. This work focuses on patch-level classification which could in future be expanded to pixel-based classification. Moreover, the methods proposed in this study can be extended to the multi-class classification of aerial imagery. Further possible future directions include the combination of geo-referenced meteorological and remotely sensed image data to explore proposed methods. Lastly, the proposed method can also be adapted to other domains involving multi-spectral and multi-modal input, such as, historical documents analysis, forgery detection in documents, and Natural Language Processing (NLP) classification tasks.

Contents

Acknowledgments

The research presented in this thesis is carried out at two institutes. Initial phases of the study are conducted in Deep Learning Lab (DLL) of School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Pakistan. The latter part of the research is done in the Machine Learning Group of EISLAB, Luleå University of Technology (LTU), Sweden.

First and foremost, my profound gratitude goes to my supervisors and mentors, Professor Faisal Shafait, Professor Marcus Liwiki, Professor Ulrich Schwanecke, Dr. Muhammad Shahzad, Dr. Muhammad Imran Malik, Dr. György Kovács, and Oluwatosin Adewumi for giving me the opportunities to conduct my research optimistically and guiding and educating me towards the right direction. Their incredible support has helped me make my way, define my directions, and bring the best of myself. I am very grateful for the care and meticulous supervision provided. The prompt support, guidance, and attention I have received from authorities like the Head of the Division, Head of the department, and administrative staff, especially Karin Rosengren and Monica Björnfot, have been significant particularly in the difficult times of COVID when I was hired. I am very thankful.

It is impossible to forget the support and efforts of my family, friends, colleagues, and professionals halfway through this journey from different parts of the world. I would like to sincerely thank all of them for their inputs in the form of moral or professional support, care, and guidance. The critique, encouragement, suggestions, advice, and laughter from my fellows, friends, and family are invaluable. I am grateful to those who made me feel at home in the new city with Saturday cooking and gatherings, those who taught me how to play pool, those who supported me at the beginning of my new journey in Luleå, those who made my rough days productive, those who added fun during my working hours and official trips, especially (in no specific order) Oluwatosin Adewumi, György Kovács, Lama Alkhaled, Hamam Mokayed, Sana Al-Azzawi, Maryam Pahalvan, András Bota, Pedro Alonso, Gustav Grund Pihlgren, Saleha Javed, Richa Upadhyay, Prakash Chandra Chhipa, Rajkumar Saini, Konstantina Nikolaidou, Foteini Liwicki, and all the members in the Machine Learning Group. It is a blessing to be a part of such a supportive and understanding group. Finally, I would like to credit all of them for contributing to my journey and making it fruitful.

Luleå, October 2021
Nosheen Abid

Part I

Chapter 1

Introduction

"I am just a guest born in this world,
to know the secrets that lie beyond it."

Jalāl ad-Dīn Muhammad Rūmī

Earth Observation (EO) aims at acquiring information about the surface of Earth (like physical, chemical, and biological details) via spaceborne, airborne, and ground-based technologies. spaceborne data covers the Remote Sensing (RS) imagery captured from satellites, airborne covers imagery captured from aircraft or Unmanned Aerial Vehicles (UAV), ground-based data covers CCTV footage, and geo-referenced meteorological data. EO also includes numerical data gathered from thermometers, wind gauges, ocean buoys, altimeters or seismometers, images from Radio Detection and Ranging (RADAR) and Light Detection and Ranging (LiDAR), information from ocean-based instruments decision-based tools based on processed information like maps, and many more. EO data plays a crucial role in monitoring natural resources, modeling climate change, and depicting land-cover and use changes. Land-cover represents the surface of Earth covered with natural resources, and land-use illustrates the built-ups and constructions [1]. The term EO is considered differently from place to place, creating confusion. In Europe, EO is mainly used in reference to satellite-based RS data [2]. The term EO has a broader coverage representing all kinds of geo-referenced data to monitor the Earth's surface.

EO is emerging with significant importance because of the dramatic impact of modern human civilization on the Earth's eco-system and global environment. EO data is extremely useful for estimating and evaluating the negative impacts on the Earth like deforestation, water scarcity, overpopulation, burning fossil fuels, and pollution. Such influences have precipitated climate change, poor air quality, soil erosion, and undrinkable water. EO data is used in many applications designed for the environment to mitigate adverse effects. Some of the specific EO applications include [3]:

- Weather forecasting.
- Measuring land use and land-cover change, like rural development, urban planning, and deforestation.

- Managing and mitigating natural disasters, like wildfires, floods, earthquakes, and tsunamis.
- Monitoring natural resources, like air, fresh water supply, oceans, soil, forestry, and agriculture.
- Addressing emerging diseases and other health risks.
- Predicting and responding to climate change

1.1 Motivation

This work mainly focuses on the issues of freshwater analysis and wildfire detection. Water and forests are the primary resources of the existence of life on this planet and have a significant ecological and economical importance [4, 5]. The United Nations 2020 report [6] stated that water consumption is increasing in line with population growth. According to the report increase in population, enhancement in agriculture, economic expansion are the primary causes of water scarcity and deforestation. Wildfire is another calamity that is destroying green forests to a great extent. Different continents have lost their forests because of gigantic wildfires. Australia is, more than any other, a fire continent [7] which has faced massive fires between late 2019 and early 2020, burning 5.2 million hectares of forests [1]. Similar challenges exist globally; for instance, at least 70 large wildfires have been recorded across the US in 2021 [2]. There has been about 270 massive fires fuelled by southern European heatwaves ravaged in 53 provinces across Turkey [3]. The 2021 wildfire in California has burnt 13,000 acres in 3 days [4]. In July 2021, hundreds of fires lighted up in Central Africa [5]. Recently in 2021, it has been a month now that the La Palma, the Canary Islands, is going through continuous volcanic eruption burning everything around including crucial natural resources' like forests [6].

The World Health Organization [7] and World Wildlife Fund for Nature (WWF) [8]

[1] "Fires in Victoria destroy estimated 300 homes, former police chief to lead Bushfire Recovery Victoria", *ABC News*, Jan. 2020, `https://www.abc.net.au/news/2020-01-06/bushfires-in-victoria-destroy-at-least-200-homes/11844292`

[2] "At least 70 large wildfires burning in US west as fears mount over conditions", *The Guardian, Guardian News and Media*, 2021, `https://www.theguardian.com/us-news/2021/jul/17/us-west-wildfires-bootleg-fire-oregon`

[3] "Information bulletin: Turkey wildfires – 10.08.2021 - Turkey", *ReliefWeb, Turkish Red Crescent*, 2021, `https://reliefweb.int/report/turkey/information-bulletin-turkey-wildfires-10082021`

[4] Tim Stelloh and Elisha Fieldstadt, "California wildfire balloons to 13,400 acres, jumps major highway", *NBCNews.com, NBCUniversal News Group*, 2021, `https://www.nbcnews.com/news/us-news/wildfire-closes-major-california-highway-prompts-evacuations-n1281298`

[5] "Fires in Central Africa", *Aeronautics and Space Administration (NASA)*, Aug. 2021, `https://modis.gsfc.nasa.gov/gallery/individual.php?db_date=2021-08-02`

[6] Philip Whiteside, "La Palma Volcano: Eruption on Canary Island shows no sign of slowing, officials say", *Sky News*, Oct. 2021, `https://news.sky.com/story/la-palma-volcano-eruption-on-canary-island-shows-no-sign-of-slowing-officials-say-12436618`

[7] "Wildfires", *World Health Organization*, 2020, `https://www.who.int/health-topics/wildfires#tab=tab_1`

report stated that wildfire and volcanic eruptions had affected 6.2 million people worldwide and almost 3000 deaths from suffocation, injuries, and burning in the last two decades. Even more alarmingly, the frequency of wildfires are increasing over time due to climate change. The weather is turning hotter and dryer with every passing year drying the ecosystem and leading to large wildfires. Wildfires are drastically deteriorating air quality by releasing massive quantities of carbon dioxide, carbon monooxide, and fine particulate matter polluting the air. The polluted air has significant effects on the respiratory system and creates cardiovascular problems. It also has a significant influence on mental health and psycho-social well-being. It also leads to the loss of wildlife, crops, properties, and resources.

Water scarcity, also known as the water crisis, implies the lack of freshwater resources in line with the standard demand. Many factors raise water stress like deforestation, increase in demand by humans, and climate change, including floods and droughts. The World Wildlife Fund for Nature (WWF) [9] reported that 70% of the Earth's surface is covered with water, out of which only 3% is freshwater. Two-third of the freshwater is frozen in glaciers or not in a usable state. This situation makes almost 1.1 billion people deprived of fresh water, and 2.7 billion people face water shortage at least for a month every year. Adding to it, 2.4 billion population suffer from inadequate sanitation resulting in exposure to diseases like cholera, typhoid fever, and many other water-borne diseases. About two million human beings, mainly kids, are dying every year because of water-borne illnesses. Natural sources like rivers and lakes are drying up with the change in weather conditions from climate change.

The highlighted issues and dependencies raise the need for efficient systems for monitoring and managing natural resources such as water and forests. Typically, the monitoring systems for EO analysis mainly rely on rich RS spatial and temporal data giving an aerial view of the Earth's surface. This thesis uses RS data to depict the burnt regions from wildfires and detect freshwater from aerial imagery using deep learning techniques.

1.2 Remote Sensing

RS data, often referred to as EO data, is a significant source of aerial imagery for observing the Earth's surface. Aerial (airborne) and satellite (spaceborne) imagery are the two main types of RS data. Unmanned Aerial Vehicles (UAV) are the primary capturing sources of airborne imagery. These crafts and drones capture the Earth's surface from low elevation and capture high spatial resolution aerial imagery when required. The spatial resolution may vary from a few meters to a sub-meter Ground Sample Distance (GSD) pixel resolution. The capturing process at low elevation and flexibility with capturing time helps mitigate the limitations of law regulation and weather conditions such as cloud cover [10, 11]. UAVs are mainly equipped with conventional cameras that provide only visual bands, i.e., red, green, blue, and sometimes Near Infrared (NIR). UAV-based airborne imagery is comparatively cheaper and faster to capture than spaceborne imagery. airborne imagery has actively been used for land-cover and land-use detection [12, 13], and natural disasters analysis like floods and earthquakes [14]. The high resolution of airborne imagery is quite helpful for detailed land feature analysis of any specified area.

Multiple commercial and open-source satellite missions have been launched since the last century for capturing spaceborne imagery. LandSat (launched in 1972 by the National Aeronautics and Space Administration - NASA) and Sentinel (launched in 2015 by European Space Agency - ESA) are the most successful open-source multi-spectral satellite projects. Landsat is an older project having multiple satellites launched over time from Landsat-1 to Landsat-9. Landsat-9 [15], a joint effort of NASA and the U. S. Geological Survey is recently released on 27th September 2021, and its data will be publicly available from early 2022 [16]. Landsat-9 shares the same orbit as Landsat-8. Landsat missions provide EO data of several decades with a trade-off of spatial resolution, i.e., 30 meters per pixel for visual, NIR, and Short Wavelength Infrared (SWIR) bands. In comparison, the Sentinel-2 mission provides better spatial resolution, i.e., 10 meters for visual and NIR bands, and 20 meters for SWIR bands (see Table 1.1). It can be seen in the table that Landsat-8/9 has lesser multi-spectral bands and spatial resolution than Sentinel-2. Moreover, Sentinel-2 provides better revisit time and wider swath [17]. See Table 1.2 for overall comparison of Landsat-8/9 and Sentinel-2A/2B.

Considering the advantages of Sentinel-2 over Landsat-8/9, this thesis has used Sentinel-2 imagery to create the dataset for Pakistani water bodies, PakSAT, and Australian wildfire dataset. Furthermore, this work used two publicly available datasets, EuroSAT [18] and SAT-6 [19] to keep the research open-source and reproducible.

1.3 Related Work

Numerous statistical thresholding methods are used in RS to classify various categories of land-use and land-cover from space data. Such as, Normalized Difference Vegetation Index (NDVI) [20] to mainly extract vegetation. Normalized Difference Water Index (NDWI)[21] and Modified Normalized Difference Water Index (MNDWI) [22] for water bodies segmentation. Normalized Burned Ratio (NBR) [23], the Mid-InfraRed Burn Index (MIRBI) [24], and the Modified Burned Area Index (BAIM) [25] for detecting burnt regions. These methods have been introduced for EO multispectral imagery but require handcrafted features and domain knowledge. Furhtermore, these approaches are sensitive to noise like cloud cover and require exhaustive prepossessing of massive corpus of data, making the task more challenging.

In the recent past, some researchers have also used machine learning techniques based on pattern recognition to detect burnt regions and water bodies on the surface of Earth from RS data. Example of ML algorithms are region-growing [26, 27], Support Vector Machine (SVM) [28, 29], Decision Tree (DT) [30, 31], Random Forest (RF) [32, 33], and Gradient Boosting [34] which have been experimented for detection tasks in RS. Despite improvements over time in the algorithms, there are still some limitations of the models that needs to be addressed and are beyond the capacity of the methods. More specifically, tools used to monitor land-use and land-cover would be better if they could handle larger spatial and temporal data, better estimation for uncertainty, and efficiently mapping the desired category on the earth surface [35]. It is raising the need of scalable method that

Landsat 8/9		Sentinel 2A/2B	
Bands	Resolution (meters)	Bands	Resolution (meters)
Band 1 - Coastal aerosol	30	**Band 1 – Coastal aerosol**	60
Band 2 - Blue	30	**Band 2 – Blue**	10
Band 3 - Green	30	**Band 3 – Green**	10
Band 4 - Red	30	**Band 4 – Red**	10
Band 5 - Near Infrared (NIR)	30	**Band 5 – Vegetation red edge**	20
Band 6 - SWIR 1	30	**Band 6 – Vegetation red edge**	20
Band 7 - SWIR 2	30	**Band 7 – Vegetation red edge**	20
Band 8 - Panchromatic	15	**Band 8 – NIR**	10
Band 10 - Thermal Infrared (TIRS) 1	100	**Band 9 – Water vapour**	60
Band 11 - Thermal Infrared (TIRS) 2	100	**Band 10 – SWIR – Cirrus**	60
		Band 11 – SWIR	20
		Band 12 – SWIR	20

Table 1.1: Landsat-8/9 and Sentinel-2A/2B bands names with their respective spatial resolution.

Parameters	Landsat-8/9	Sentinel-2A/2B
Spectral resolution (Count of Bands)	11	13
Spatial resolution (meters/pixel)	30-60	10-60
Revisit (Time Global Median Average in days)	8	3.7
Swath (km)	185	290

Table 1.2: Shows the comparison of Landsat-8/9 and Sentinel 2A/2B satellites.

is adaptable to inter and intra class variations. Deep Learning (DL) models have the capability to address the stated limitations [36].

DL is emerging as the state of the art solution for learning variation and complex features in the data and across various domains [37, 38]. Computer-vision problems like object detection, localization, and recognition are successfully being addressed with Convolutional Neural Networks (CNN) [39]. In RS, CNNs are widely being used in emerging applications of land use such as segmentation of buildings, roads [40] and small objects [41], land-cover classification [42], cloud-cover detection [43], and reconstruction of missing information in the data [44]. DL models are also used for effective utilization of Spatio-temporal data [45, 46]. Knopp et al. [47] used CNN-based U-Net to segment the burnt regions from mono-temporal Sentinel-2 data. Pinto et al. [48] combined U-Net and Long-Short Term Memorys (LSTM) neural networks to map and date the burnt regions using multispectral imagery. Fang et al. [49] used ResNet for identifying global water reservoirs.

Inspired by the excellent performance of supervised deep models, they still carry some limitations. i) In the absence of labeled data, they need a collection of data and expert knowledge to label the data, which is a time-consuming and tedious task. ii) Supervised models are domain-specific. Their performance dramatically decreases when applied to different domain data of the same problem. Recently, Generative Adversarial Neural network (GAN) are being used in a semi-supervised manner to address the problem of cross-domain adaptation in semantic segmentation ofRS imagery [50, 51]. Some researchers have explored the concept of Curriculum Learning (CL) to train supervised deep models [52] efficiently. The focus of our work is to overcome the stated issues of supervised approaches by introducing an unsupervised DL solution using the concept of CL for the classification of water bodies [53] and mapping burnt regions [54] using RS data.

1.4 Scope and Delimitation

The aim of this thesis is to investigate DL models in an unsupervised way to overcome the limitations of supervised architectures of data labeling and better domain adaptation. Unsupervised learning has a broad spectrum and an emerging field of Artificial Intelligence. Therefore, it is difficult for one thesis to cover this topic without delimitation. This thesis primarily considers the domain of RS for exploring unsupervised deep learning. It considers the geo-referenced RS imagery as a case study and mainly focuses on the patch-based classification of RS imagery. This work has used publicly available satellite data. Publicly available satellite data has comparatively lower resolution than commercial satellites leading to decreased intraclass variation. Currently, this work is designed for two-class classification, which will be extended to multi-class in future work. The research questions that arose and were addressed during this research are presented below.

Q1 How well do supervised methods work with RGB aerial imagery?

Q2 Is it possible to reach similar performance in RS tasks to supervised methods using unsupervised approaches?

Q3 Can unsupervised approach be extended to multispectral satellite imagery?

Chapter 3 explains in more detail the emergence and evolution of proposed questions throughout the research process and how they are addressed.

1.5 Outline

This thesis is divided into two main parts where Part I is composed of five chapters to provide a broader perspective and context of Part II. Paart II is composed of two different published research papers carrying more detailed in-site of technical work and results proposed in this thesis.

The remainder of this thesis is structured as follows. Chapter 2 describes the methodology that was followed to address the formulated research questions. It gives the technical details of the work. Chapter 3 describes how the methodology was evaluated, and the results of those evaluations. Chapter 4 presents the titles, abstracts, publication details, and contribution(s) of authors of two papers of this thesis. Chapter 5 concludes Part I of the thesis by giving a summary of the work and motivation for future directions.

Part II is composed of two published papers. Paper A, “UCL: Unsupervised Curriculum Learning for Water Body Classification from Remote Sensing Imagery”, is a journal publication representing the concept of unsupervised curriculum learning for classifying water bodies from airborne and spaceborne RGB imagery and its detailed evaluation on three different datasets from three different continents. Paper B, “Burnt Forest Estimation from Sentinel-2 Imagery of Australia using Unsupervised Deep Learning”, is a conference publication that used multi-spectral spaceborne imagery to do the unsupervised classification of burnt wildfire regions of Australia faced in late 2019 and early 2020.

Chapter 2

Methodology

"If we knew what it was we were doing, it would not be called research, would it?."

Albert Einstein

This chapter explains the research methodology used to address the designed research questions. It starts with explanation and formulation of research questions. It further explains the general research method used in the research when formulating the research questions and addressing them. The later section explains the detailed description of the technical approach used in this work which contains the datasets, unsupervised architecture, and implementation details.

2.1 Research Method

Deep Learning (DL) models are state-of-the-art methods capable of automatically learning the complex and prominent features from the data without any hand-crafting. However, they still have some limitations. Most of them are supervised models and need a huge labeled corpus that requires the efforts of data gathering and expert domain knowledge for labeling. These models are very domain-specific as trained to learn fine details from the data labeled by domain experts. The stated limitations could be addressed using clustering techniques to generate the pseudo-labels, which could remove the requirement of domain experts' knowledge and ease the tedious task of data labeling. Clustering techniques could be the right choice to generate pseudo-labels but have their trad-offs. The generated clusters are based on the prominent features present in the data, which may not address the outliers as no additional information from experts is used. The models using the pseudo-labels may be less domain-specific as they are not overfitted to the domain expert knowledge but only to the prominent features in the data.

This research keeps in mind the strengths of DL models and clustering techniques that could address some of the limitations of DL models. This research integrates both algorithms to introduce an unsupervised DL architecture. The proposed idea is inspired by the computer vision domain [55, 56, 57] which combines the benefits of transfer learning and latent space representation to enable cross-domain adaptation. Another source of inspiration is the concept of Curriculum Learning (CL) [58] where training samples are fed to the learning model in order of their difficulty. This difficulty is defined by the complexity of the samples' features. Where easy samples are fed in the initial iterations, and the complexity gradually increases in the subsequent iterations. It has been demonstrated that feeding the learner in this manner speeds up the learning as the DL model has to learn complex examples gradually. One of the significant challenges of CL is how to define the difficulty level in the samples. Different heuristics have been used in different domains for successfully defining the complexity of the samples (e.g., [52]). In this research, proximity is used from cluster centroids as a selection criterion for representing "easy" samples, called "reliable samples". The proposed research methodology comprises the DL model, clustering mechanism, and selection operation inspired by Curriculum Learning.

2.1.1 Research Questions

This research is composed of three primary research questions. Before jumping into the unsupervised domain for RS data, it was essential to have some technique as a benchmark for the comparison. As DL models are state-of-the-art for learning the complex features from the data, this inspired to the first research question:

Q1 How well do supervised methods work with RGB aerial imagery?

CNN-based DL models are considered state-of-the-art in learning complex features from images. This encouraged the use of these models as a benchmark to evaluate the RS datasets, see Paper A. The state-of-the-art pre-trained CNN-based DL models are designed from visual bands, i.e., red, green, and blue. So, to address this question, initial experiments are designed to consider only red, green, and blue bands of RS Imagery. As supervised models used labeled datasets, it is shown that they outperform the other techniques in terms of accuracy, which do not use domain knowledge in the training process. However, accuracy on the data set used for training purposes is not the only metric to evaluate models. There are other factors, too, for example, domain adaptation. This gave rise to the second research question of this work:

Q2 Is it possible to reach similar performance as supervised methods using unsupervised approaches?

The proposed unsupervised curriculum learning based architecture is evaluated on the datasets and are compared with the supervised methods used as a benchmark in this study. It is shown in Paper A that the proposed unsupervised model was able to learn the RS features to classify the patches. In contrast, the supervised methods achieved higher accuracy. However, the performance of supervised models decreased dramatically when applied to other datasets of the same problem statement. In comparison, the proposed unsupervised architecture was able to perform better than supervised models in cross-domain adaptation.

One thing should not be ignored, i.e., the more the data, the better the performance of DL models. Here, the RS data from the satellite has multispectral bands. This brought inline the third research question of this thesis:

Q3 Can the unsupervised approach be extended to multispectral satellite imagery?

The pre-trained state-of-the-art CNNs are based on three-channel input. However, satellite imagery is composed of multispectral bands. This inspired researchers to to use the other bands of the satellite imagery along with visual bands in the proposed unsupervised architecture. In Paper B, 12 out of 13 bands of the Sentinel-2 satellite are integrated to match the requirement of pre-trained models of three-channel input. The proposed unsupervised technique is fine-tuned for multispectral imagery to classify burnt regions from Sentinel-2 data. The 12 bands are selected based on their relevance to the problem being addressed, i.e., classifying burnt regions from Australian Sentinel-2 imagery.

Both research papers have explored and addressed the raised research questions in a limited setup, proving the hypothesis with simple binary classification rather than exploring other aspects, like multi-class classification and other domains having multi-model input, which will be considered future directions of this work.

2.2 Technical Approach

This section covers some of the technical details of the research carried out for this thesis and explains the choices made in this work. It describes the details of datasets created and used in this study and a bit of insight of the introduced unsupervised curriculum learning based approach and its implementation details.

2.2.1 Datasets

The two papers of this work have used four datasets. Paper A has used three datasets for water bodies classification from RS data. Two of the datasets, EuroSAT [18] and SAT-6 [19] are open-source. EuroSAT is composed of Sentinel-2 Imagery with a patch size of 64 × 64. It has been captured from different cities of 34 European countries. It comprises ten different classes, out of which two were of water categories, “sea and lake” and “rivers”. The SAT-6 dataset is composed of aircraft image patches of California. It has a patch size of 28 × 28 with a high spatial resolution (1-meter per pixel). It has six different categories, and only one category is of water bodies. Both of the datasets were pre-processed into two classes of water and not water categories to carry equal ratios of samples. The third dataset used in Paper A, PakSAT, is a newly introduced dataset designed for water body segmentation. It is composed of Sentinel-2 image patches of Pakistan having size 64 × 64. Hence, Paper A has used three different datasets having RS imagery. Paper B has used the Australian Wildfire dataset to classify burnt regions of Australia from late 2019 to early 2020. This dataset is newly introduced in this work using Sentinel-2 Imagery having patch size 64 × 64. Further details of the datasets used and produced are provided in Paper A and B.

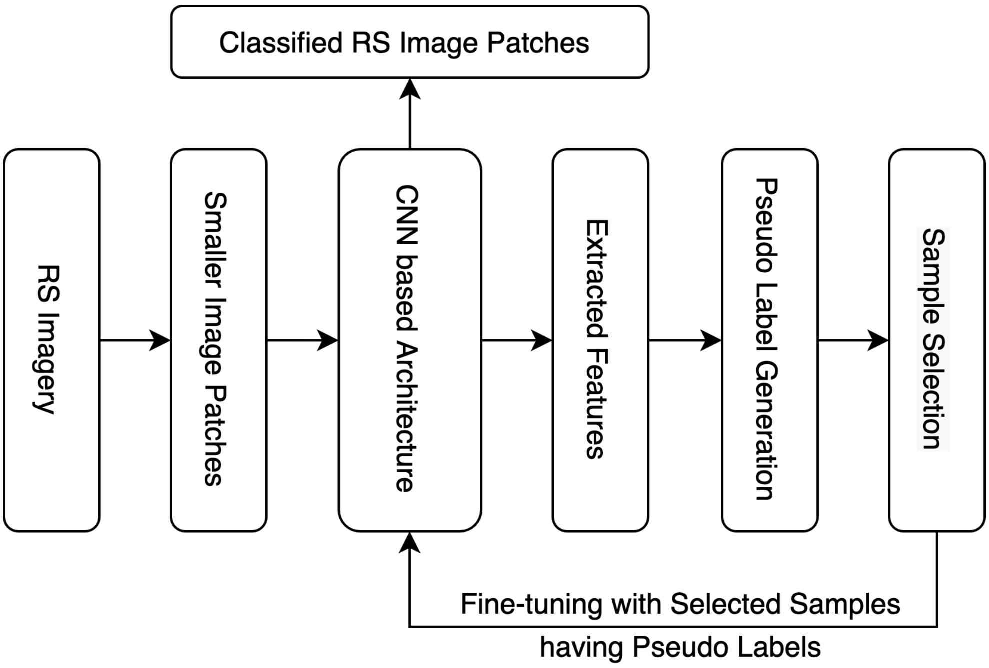


Figure 2.1: Proposed unsupervised deep learning based approach.

2.2.2 Unsupervised Curriculum Learning

Unsupervised Curriculum Learning (UCL) is composed of three different techniques inspired from the literature. The widely used CNN model, VGG-16, is used for feature extraction from the RS data. The clustering mechanism used to generate pseudo-labels plays a vital role in making the proposed approach unsupervised. A sample selection criterion is applied to extract reliable samples from the generated clusters is inspired by the concept Curriculum Learning.

The proposed pipeline of UCL is shown in Figure 2.1. In preprocessing, the huge tiles of RS imagery are broken down into smaller patches suitable for the DL model. The proposed unsupervised curriculum learning based pipeline can be described in three phases; 1) the dDL model for extracting features from the corpus, 2) clustering technique to generate clusters from extracted features to assign pseudo-labels, and 3) selection operation to extract reliable samples from clusters to fine-tune the deep model. A CNN based DL model extracts features from RS image patches. The considerDL model is pre-trained on ImageNet weights [59] (an irrelevant domain) that does not contain the RS imagery. As a result, the extracted features from the pre-trained model may not be a good representation of the RS data. Thus, the clusters they form may be loosely packed. A selection operation is applied that extracts the samples present near the centroids as reliable samples. These reliable samples are used to fine-tune the DL model. This process of extracting features and clustering them, extracting reliable samples, and fine-tuning the deep model with reliable samples is done iteratively till the DL model has converged. This iterative process of learning features with pseudo-label makes the proposed model unsupervised. For a better understanding of UCL, please see the algorithm 1. The UCL model is explained in more detail in Paper A.

Algorithm 1 UCL Alogrithm

```
Input: Unlabeled data $\{x_i\}_{i=1}^{N}$, Reliability Threshold $\lambda$, pre-trained DL model $\phi(., \theta_0)$
Output: Trained model $\phi(., \theta_t)$
Initialization: $\theta_t \leftarrow \theta_0$
while model not converged do
    $\{f_{Map_i}\}_{i=1}^{N} \leftarrow \phi(\{x_i\}_{i=1}^{N}, \theta_t)$    ▷ Feature extraction from deep model $\theta_t$ at time $t$
    $\{f_i\}_{i=1}^{N} \leftarrow$ Flatten($\{f_{Map_i}\}_{i=1}^{N}$)    ▷ Vector representation of extracted features
    $\{c_k\}_{k=1}^{2} \leftarrow$ Clustering($\{f_i\}_{i=1}^{N}$)    ▷ Clustering the features in 2 clusters
    $\{\hat{c}_k\}_{k=1}^{2} \leftarrow$CentroidFeature($\{f_i\}_{i=1}^{N}, \{c_i\}_{i=1}^{2}$)    ▷ Centroids of 2 clusters
    for $k := 1$ to 2 do
        for $i := 1$ to $N$ do
            $\gamma_i \leftarrow f_i \cdot \hat{c}_k$
            if $\gamma_i > \lambda$ then
                $x_i^{'} \leftarrow x_i$    ▷ Selection of reliable samples
            end if
        end for
    end for
    $\theta_{t+1} \leftarrow$ Finetune($x^{'}, \theta_t$)    ▷ Fine-tune deep model $\theta_t$ with reliable samples
    $\theta_t \leftarrow \theta_{t+1}$    ▷ Update the deep model $\theta_t$ with fine-tuned one $\theta_{t+1}$
end while
```

Some empirical experiments are performed before finalizing the proposed unsupervised architecture. Multiple widely used CNN models, namely VGG-16[60], ResNet-50 [61], Inception Network [62], Xception network [63], and DensNet [64], have been explored in this study of unsupervised learning. Out of all these architectures, ResNet and VGG-16 performed better than other architectures. The performance scores attained by VGG-16 and ResNet-50 are not significantly different. Whereas ResNet was taking longer to train than VGG-16. Considering the performance efficiency and minimum time consumption for training, VGG-16 is used in the proposed UCL architecture. Similarly, three different clustering techniques are explored; namely, K-Means [65], Fuzzy C-Means (FCM) [66] and Hierarchical Clustering [67]. K-Means and FCM performed better than the Hierarchical clustering technique, but FCM took longer to generate clusters than K-Means. As a result, the K-Means clustering algorithm is used in UCL. The final architecture of UCL is composed of VGG-16, K-Means clustering algorithm, and a selection criterion for reliable samples.

2.2.3 Implementation Details

UCL is implemented in Python using TensorFlow [68] and Keras [69] libraries. JupyterLab is used as a development framework. In all cases of the downstream task, the dataset is shuffled and split into training and test corpus in the ratio 80:20. Out of this 80% of the data, 20% is used for validation, and 80% is used for training. In every iteration of UCL fine-tuning, the training and validation corpus are shuffled and recreated. So for each fine-tuning iteration of UCL, there is a different validation set. Stochastic Gradient Descent (SGD) [70] optimizer, Softmax activation function [71] and Cross-entropy loss function are used in the training process of deep model. The early-stopping criterion is used to stop the training. The criterion says if the validation is not reduced in the next five epochs, stop the training and store the model of the epoch reporting minimum validation loss. All the experiments reported in this thesis are conducted on a GPU machine having an NVIDIA Titan-X GPU for training and fine-tuning with 32 GB RAM and Linux operating system.

Chapter 3

Experiments and Results

"Look for the answer inside your question."

Jalāl ad-Dīn Muhammad Rūmī

This chapter summarizes the common areas of the experimental details of both papers. In the beginning of the chapter, the proposed experiments for Paper A and B are described. It also explains the evaluation metrics used to analyze the results. Lastly, the results attained in two papers are summarized.

Experiments in Paper A are designed to evaluate the Unsupervised Curriculum Learning (UCL) for water body classification using three different datasets. As each dataset is composed of RS imagery of places from different continents, this addresses regional independence. Further, these datasets are composed of images captured from different sources adding variability in spatial resolution. These variations in the datasets make the results more robust. These experiments are designed for visual bands of RS imagery catering to the air-borne data from aircraft and UAVs having primarily visual bands only. Later, this technique is extended to multi-spectral imagery of space-borne data in Paper B, where the task is to classify burnt regions using multi-spectral imagery.

3.1 Evaluation Metrics

The results produced in the two papers included in this thesis are evaluated using various performance metrics. The compactness of the generated clusters is evaluated by using purity and the Silhouette Score. The Sum of Squared Error (SSE) is calculated for added cluster analysis, an objective function of K-Means clustering. Along with these metrics, the reliable samples extracted at each iteration of fine-tuning are also analyzed, such as the growing count in the reliable set with every iteration of fine-tuning and how consistent the pseudo-labels are in progressive fine-tuning iterations. The training curves of the deep model with clustering pseudo-labels are also monitored. After the fine-tuning of UCL with pseudo-labels, the model is evaluated on a test dataset where precision, recall, and F1-Score performance are used.

3.1.1 Purity

Purity is an external evaluation criterion for analyzing the quality of clusters. It uses the true labels and pseudo-labels the calculate the purity score. Its value ranges between 0 and 1. The higher the value of purity, the better the clusters in quality. In other words, it calculates the ratio of the total number of samples in the clusters that are correctly classified in the unit range [0..1]. The following Equation 3.1 is used to calculate the purity score.

$$Purity = \frac{1}{N}\sum_{i=1}^{k} \max_{j} |c_i \cap t_j| \tag{3.1}$$

Where N is the total count of samples in the corpus, k is the total count of clusters. As this thesis addresses binary classification, $k = 2$, i.e., in Paper A, either the generated cluster is of water bodies or not water bodies, and in Paper B, either the generated cluster is of burnt regions or not burnt regions. c_i is the set of clusters, and t_j is the set of classes with a maximum count for cluster c_i. Equation 3.1 is used to calculate the purity of the clusters.

3.1.2 Silhouette Coefficient

Silhouette Coefficient or Silhouette Score is a metric used to evaluate the saturation in the clusters. It calculates the ratio based on the distance between each sample within the cluster (Equation 3.2) and the neighboring clusters (Equation 3.3). Its value ranges from -1 to 1. Where 1 means the clusters are well separated from each other and can clearly be distinguished, and 0 means that the cluster is indifferent or the distance between the clusters is insignificant. The Silhouette Score for each sample in the dataset is calculated by using the following set of Equation 3.2 to 3.5.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j) \tag{3.2}$$

Where $a(i)$ represents the intra-cluster distance of sample i. In other words, the distance of sample i from other samples present in the same cluster of i, where C_i are the sample points in cluster i. $d(i, j)$ is the distance between each sample i and j present in cluster i. $a(i)$ measures how well a sample i belongs to the cluster. The smaller the value is, the better the association.

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \tag{3.3}$$

Where $b(i)$ represent the inter-cluster distance. In other words, the mean distance of sample i from other samples present in any other cluster C_k. It calculates the dissimilarity of a sample i with other clusters. The cluster with the smallest dissimilarity could be considered as the closest neighbors cluster of i.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}\,, \quad \text{if } |C_i| > 1 \tag{3.4}$$

$$s(i) = 0\,, \quad \text{if } |C_i| = 1 \tag{3.5}$$

The calculated intra and inter cluster distances ($a(i)$ and $b(i)$) are used to calculate the Silhouette Score $s(i)$ of sample i using Equation 3.4. It can clearly be seen in Equation 3.2 that intra cluster distance $a(i)$ can not be calculated when cluster C_i has only one sample. To avoid this situation, $s(i)$ is set to 0, represented in Equation 3.5.

3.1.3 Confusion Matrix

Each dataset used in this study has a text set. Once the UCL is fine-tuned on the datasets, it is evaluated on each dataset using their respective test corpus. Confusion matrix is a key measure for other evaluation metrics like Precision, Recall and F1=Score. In both of the papers, following format in Table 3.1 is used for confusion Matrix.

		Predicted	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Table 3.1: Confusion matrix used to calculate Precision, Recall and F1-Score.

3.1.4 Precision

Precision is one of the evaluation metric that is based on the confusion matrix and is used in this work. It calculates the ratio of correctly classified among the samples that were classified as true.

$$Precision = \frac{TP}{TP + FP} = \frac{\text{Correctly True Classified}}{\text{Total True Classified}} \tag{3.6}$$

3.1.5 Recall

Recall is an evaluation metric that calculates the ratio of correctly classified among the actual positive samples.

$$Recall = \frac{TP}{TP + FN} = \frac{\text{Correctly True Classified}}{\text{Total True Classified}} \tag{3.7}$$

3.1.6 F1-Score

In simple terms, the F1-Score is the harmonic mean of Precision and Recall. It is one of the most suitable measuring scales as it deals with the non-uniform distribution of data among the classes. It is quite useful to seek a balance between Precision and Recall. This work also uses the macro and micro F1-Scores. The F1-Score is calculated using following Equation 3.8

$$Precision = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.8}$$

Lastly, in Paper B, the generated results from different fine-tuned iterations of UCL are visualized on the map by georeferencing. Georeferencing is mapping the digital map (vector data), RS imagery, or processed aerial imagery (raster data) using an internal coordinate system. The subject imagery to the map can be related to the ground system of geographic coordinates.

3.2 Experiments and Results

This section summarizes the results of both papers. The following subsections present results from each paper in a very brief format.

3.2.1 Paper A: UCL: Unsupervised Curriculum Learning for Water Body Classification from Remote Sensing Imagery

Before moving towards the evaluation of UCL, some initial experiments have been conducted. In these experiments, direct testing of VGG-16 architecture with ImageNet weights is done using RGB RS imagery of the EuroSAT, PakSAT, and SAT-6 datasets. Table 3.2 describes the results of these experiments. It is assumed in this experiment that there is no labeled data provided. As a result, the classification using VGG-16 can either be done using a random classification layer or some clustering technique. This

Initial Weights	Test Set	F1-Score (%)	
		Random FC	K-Means
ImageNet	EuroSAT	44.46	54.45
ImageNet	PakSAT	51.98	57.13
ImageNet	SAT-6	58.20	65.50

Table 3.2: Describes the direct inference of EuroSAT, PakSAT, and SAT-6 datasets on VGG-16 considering three different techniques. Random FC indicates the results with random initialization of the FC layer. K-Means shows the results for clustering the features extracted from VGG-16 and classified by K-Means clustering.

experiment has explored both techniques for the classification of RS image patches. It has been shown in Table 3.2 that that ImageNet weights with K-Means clustering gave better results than ImageNet weights with randomly initialized classification layer.

Fine-Tuned & Tested	F1-Score (%)	
	Supervised	UCL
EuroSAT	99.49	84.05
SAT-6	99.53	90.89
PakSAT	96.07	87.66

Table 3.3: F1-Scores of VGG-16 supervised fine-tuned and tested on each dataset; EuroSAT, PakSAT, and SAT-6. The last column reports the results of UCL fine-tuned and tested on each considered dataset.

In the next phase of experiments the performance of VGG-16 is evaluated in a way that for each data sets, the model is fine-tuned in a supervised or unsupervised manner. The resulting F1-Scores are reported in Table 3.3. It can be seen in the table that supervised models reported F1-Score above 99% for EuroSAT and SAT-6, and 96% for the PakSAT dataset showing that the model has learned the classes quite well. Although, the UCL results are comparatively compromised. They still reported F1-Scores between 84% and 91% for the three datasets examined, showing that the model can learn the classes from the data without the need for labeled corpora for training purposes, which is an achievement in itself.

Fine-Tuned	Tested	F1-Score (%)	
		Supervised	UCL
EuroSAT	PakSAT	70.68	78.00
	SAT-6	32.84	75.76
SAT-6	PakSAT	63.16	68.97
	EuroSAT	42.83	72.43
PakSAT	EuroSAT	62.70	79.00
	SAT-6	59.92	71.72

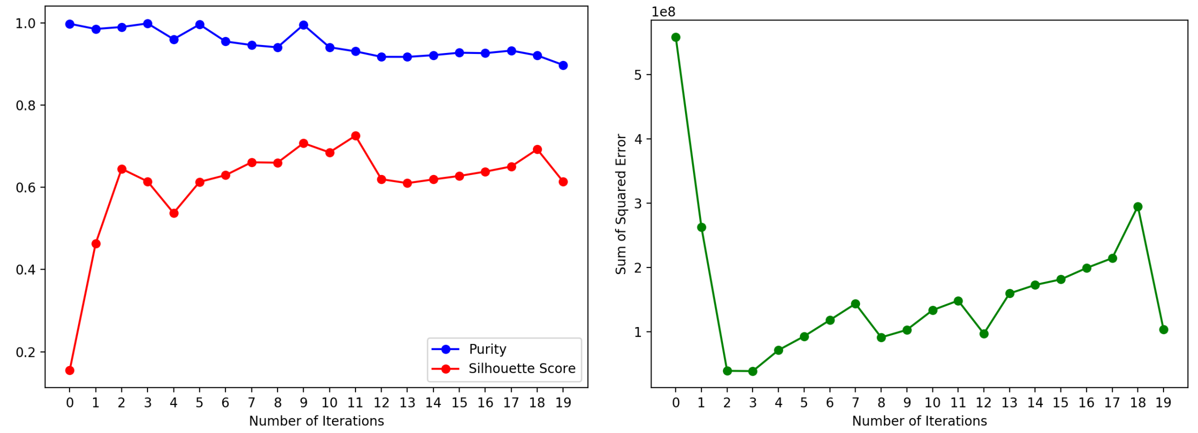
Table 3.4: F1-Scores of VGG-16 fine-tuned in a supervised manner on each dataset and tested on the other two datasets. The last column reports the results of UCL fine-tuned on each dataset and tested on the other two datasets. The considered datasets are EuroSAT, PakSAT, and SAT-6.

In the last phase, the conducted experiments evaluate the cross-domain adaptation for both approaches; the supervised fine-tuned VGG-16 and UCL. The results show that supervised fine-tuned VGG-16 on one dataset when tested on the other two considered datasets, its performance decreased dramatically. In comparison, UCL performed better in domain-adaption than the supervised models. The results of these experiments are shown in Table 3.4.

	Predicted →		
Actual ↓	Water	Non-water	Total
Water	1,460	0	1,460
Non-water	20	1,440	1,460
Total	1,480	1,440	2,920

Table 3.5: Confusion Matrix for SAT-6 test corpus on best iteration of fine-tuning of VGG-16.

Table 3.5 shows the confusion matrix for UCL fine-tuned model on SAT-6 dataset. As we can see in the matrix, only 20 samples out of 2,920 from the test corpus were misclassified, showing that UCL efficiently learned the water and not water categories from the SAT-6 dataset.



(a) (b)

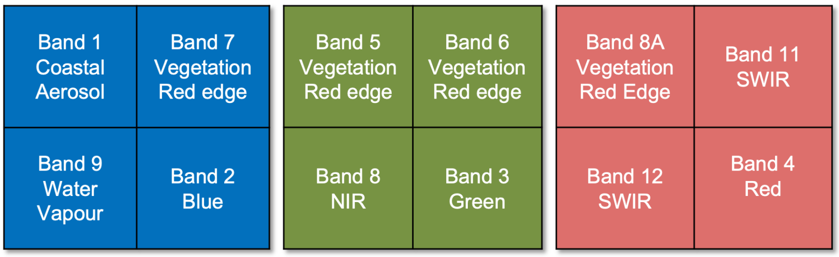
Figure 3.1: (a) Purity and Silhouette Score and (b) Sum of Squared Error of generated clusters for water bodies and other regions of SAT-6 over every iteration of fine-tuning of VGG16.

In Figure 3.1, two graphs are presented; (a) showing the Purity and Silhouette Score and (b) showing Sum of Squared Error calculated for the clusters generated during the fine-tuning interactions of UCL model for SAT-6 dataset. It can be seen in the graph (a) that purity remains quite consistent whereas the Silhouette Score, representing the saturation of the clusters, is quite low at the beginning. However, with fine-tuning iterations, the score increased indicating that the clusters are getting saturated with iterations. A small change in purity can be because of the very few reliable samples in the initial fine-tuning iterations which later grew to the count of training corpus.

This is the gist of the experiments explained above. All these experiments are performed on RGB RS imagery for water body classification. The detailed experiments with in-depth analysis are provided in Paper A.

3.2.2 Paper B: Burnt Forest Estimation from Sentinel-2 Imagery of Australia using Unsupervised Deep Learning

In this paper, a new dataset based on multi-spectral imagery of Sentinel-2 is created for Australian Wildfires classification. UCL model is designed in a way that it takes three-channel input. Whereas, the Sentinel-2 imagery is composed of 13 bands. 12 out of 13 bands are selected on the basis of their co-relation with the fire detection. These 13 bands are concatenated in a way that each input channel carries 4 bands of Sentinel-2 Imagery. The concatenation on three-channels can be seen in Figure 3.2.



Band 1 Coastal Aerosol	Band 7 Vegetation Red edge	Band 5 Vegetation Red edge	Band 6 Vegetation Red edge	Band 8A Vegetation Red Edge	Band 11 SWIR
Band 9 Water Vapour	Band 2 Blue	Band 8 NIR	Band 3 Green	Band 12 SWIR	Band 4 Red

Figure 3.2: The considered 12 bands of multispectral satellite imagery of Sentinel-2 into three-channel input. Each channel contains four bands, 1 in each quarter. Each of the red, green, and blue bands is kept in each channel, considering the input configuration of the Convolutional Neural Networks (CNN) model. All the patches are preprocessed in this way to make them suitable for the input of the DL model.

This input is fed to UCL for fine-tuning. Similar to Paper A, cluster generation saturation and quality is monitored using purity and Silhouette Score. The analysis is shown in Figure 3.3. It can be seen that the clusters are pretty saturated in the beginning, and with fine-tuning iterations, the saturation score, i.e., Silhouette Score, is getting improved.

Table 3.6 summarizes the performance of fine-tuned UCL by reporting Precision, Recall, F1-Score, and Macro and Weighted Avg F1-Score. UCL has learned the classes by reporting an F1-Score 85%.

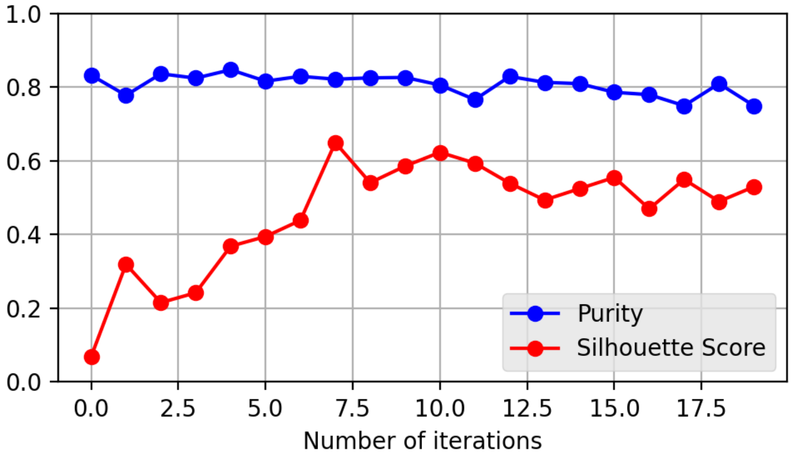


Figure 3.3: Graph showing the purity and Silhouette Score of generated clusters for burnt forest and other regions over every iteration of fine-tuning of VGG16.

	Precision	**Recall**	**F1-Score**
Class 0	0.86	0.83	0.84
Class 1	0.83	0.86	0.85
Accuracy	-	-	0.85
Macro Avg	0.85	0.85	0.85
Weighted Avg	0.85	0.85	0.85

Table 3.6: The table shows the average precision, recall, F1-Score and accuracy for best 5 iterations on the test corpus.

Figure 3.4 shows the georeferenced results of UCL of best 5 fine-tuned iterations. The area considered for this visualization is the Australian Capital Territory. Among the considered fine-tuned iterations, sub-figure (b) in Figure 3.4 reports the best performance in deployment. The detailed results of this study are explained in Paper B.

3.3 Critical Analysis

UCL has performed exceptionally well on the considered task, but it carries some limitations too. The proposed solution is fed with unbiased data, which means that both classes carry an almost equal count of samples. This data distribution may enforce the clustering technique to generate the clusters of the type of classes present in the data. Also, both classes have a clear distinction of either to be of one class or not. Currently, this distribution of samples among the classes is done manually. This distinction could also be made using indices to get the equal distribution of samples in both classes. It will be worthy of exploring how UCL with class distribution in the data using indices.

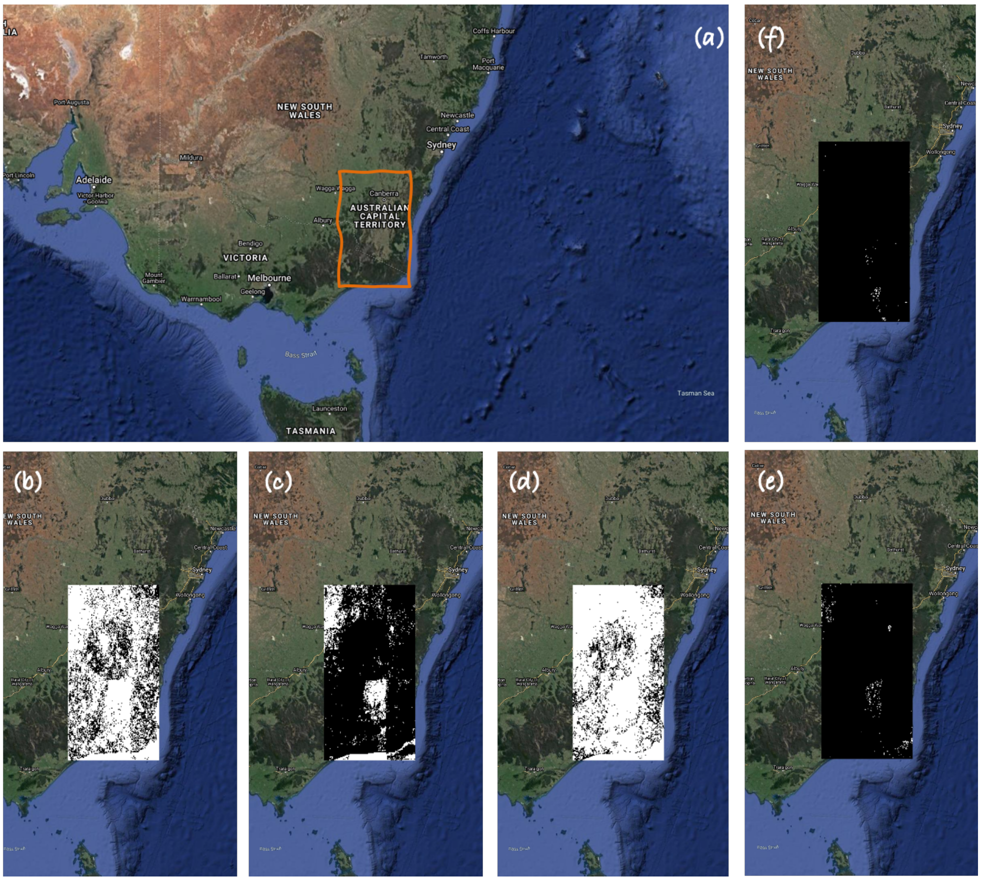


Figure 3.4: **(a)** The region considered, covering the Australian Capital Territory and South of it in an orange polygon, for testing the fine-tuned models on Sentinel-2 median image of three months (Feb 2020 - Apr 2020). **(b)** The prediction results for the fine-tune iteration of the deep learning model, reporting the highest accuracy on the test set. **(c)** The prediction results for iteration reporting the 2nd highest accuracy on the test set. **(d)** The prediction results for iteration reporting the 3rd highest accuracy on the test set. **(e)** The prediction results for iteration reporting the 4th highest accuracy on test corpus. **(f)** The prediction results for iteration reporting the 5th highest accuracy on test corpus.

Chapter 4

Contributions

"Shine like the whole universe is yours."

Jalāl ad-Dīn Muhammad Rūmī

4.1 Paper A: UCL: Unsupervised Curriculum Learning for Water Body Classification from Remote Sensing Imagery

Title: UCL: Unsupervised Curriculum Learning for Water Body Classification from Remote Sensing Imagery
Authors: Nosheen Abid, Muhammad Shahzad, Muhammad Imran Malik, Ulrich Schwanecke, Adrian Ulges, György Kovács and Faisal Shafait
Submitted in: International Journal of Applied Earth Observation and Geoinformation
Abstract: This paper presents a Convolutional Neural Networks (CNN) based Unsupervised Curriculum Learning approach for the recognition of water bodies to overcome the stated challenges for remote sensing based RGB imagery. The unsupervised nature of the presented algorithm eliminates the need for labelled training data. The problem is cast as a two class clustering problem (water and non-water), while clustering is done on deep features obtained by a pre-trained CNN. After initial clusters have been identified, representative samples from each cluster are chosen by the unsupervised curriculum learning algorithm for fine-tuning the feature extractor. The stated process is repeated iteratively until convergence. Three datasets have been used to evaluate the approach and show its effectiveness on varying scales: (i) SAT-6 dataset comprising high resolution aircraft images,(ii) Sentinel-2 of EuroSAT, comprising remote sensing images with low resolution, and (iii) PakSAT, a new dataset we created for this study. PakSAT is the first Pakistani Sentinel-2 dataset designed to classify water bodies of Pakistan. Extensive experiments on these datasets demonstrate the progressive learning behaviour of UCL and reported promising results of water classification on all three datasets. The obtained accuracies outperform the supervised methods in domain adaptation, demonstrating the

effectiveness of the proposed algorithm.

Personal Contribution: Conceptualization, methodology, and experimentation by Nosheen Abid. Refining scope, contribution, and methodology by Muhammad Shahzad and Faisal Shafait. Original draft written by Nosheen Abid. Review and supervision by Muhammad Imran Malik, Ulrich Schwanecke, Adrian Ulges, György Kovács and Faisal Shafait.

4.2 Paper B: Burnt Forest Estimation from Sentinel-2 Imagery of Australia using Unsupervised Deep Learning

Title: Burnt Forest Estimation from Sentinel-2 Imagery of Australia using Unsupervised Deep Learning

Authors: Nosheen Abid, Muhammad Imran Malik, Muhammad Shahzad, Faisal Shafait, Haider Ali, Muhammad Mohsin Ghaffar, Christian Weis, Norbert Wehn and Marcus Liwicki

Published in: International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2021.

Abstract: Massive wildfires not only in Australia but also across the globe burn millions of hectares of forests and green land affecting the social, ecological and economical situation worldwide. Widely used indices-based threshold methods like Normalized Burned Ratio (NBR) require a huge amount of data pre-processing and are specific to the data capturing source. The state-of-the-art Deep Learning models are supervised and require domain experts knowledge for labeling the data in huge quantity. These limitations make the existing models difficult to be adaptable to the new variations in the data and capturing sources. In this work, we have proposed an unsupervised Deep Learning based architecture to map the burnt regions of forests. The model considers small patches of satellite imagery and classifies them into burnt and not burnt. These small patches are concatenated into binary masks to segment out the burnt region of the forests. The proposed system is composed of two modules: 1) a state-of-the-art Deep Learning architecture for feature extraction and 2) a clustering algorithm for the generation of pseudo labels to train the Deep Learning architecture. The proposed method is capable of learning the features progressively in an unsupervised fashion from the data with pseudo labels reducing the exhausting efforts of data labeling that requires expert knowledge. We have used the real-time data of Sentinel-2 for training the model and mapping the burnt regions. The obtained F1-Score of 0.87 demonstrates the effectiveness of the proposed model.

Personal Contribution: Conceptualization, methodology, and experimentation by Nosheen Abid. Refining methodology by Muhammad Imran Malik, Muhammad Shahzad, Faisal Shafait, and Haider Ali. Original draft written by Nosheen Abid. Review and Supervision by Faisal Shafait, Muhammad Mohsin Ghaffar, Christian Weis, Norbert Wehn and Marcus Liwicki.

Chapter 5

Conclusion and Future Work

"Your heart knows the way, run in that direction."

Jalāl ad-Dīn Muhammad Rūmī

Remote Sensing and Earth Observation have been evolving from providing efficient sensors for capturing detailed information about the planet to designing EO algorithms. Many algorithms are designed and developed for monitoring the surface of the Earth, considering the specifications of the sensors, making them very specific to capturing the source. One of the famous and widely used solutions for RS data analysis are thresholding methods based on indices. In literature, there are several designed indices for a different types of classification tasks from RS data. Namely, Automated Water Extraction Index (AWEI) [72] for detecting water, and Modified Burned Area Index (BAIM) [25] for detecting burnt areas. These, methods, however, largely rely on multspectral imagery, which is not always available (particularly, when working with aerial imagery that provides high spatial resolution). Moreover, thresholding method are designed by domain experts, and rely on their expert knowledge, and hand-crafted features. Many classical machine learning algorithms have already been used in working with RS data, including Decision Trees [30], Random Forests [73], Support Vector Machines [29], and different types of Neural Networks. These methods, however, still largely rely on the use of hand-crafted features. It was the emergence of Deep Learning (DL) techniques that removed the need of these features. These methods are considered state-of-the-art in analyzing and classifying the data in many fields, including RS.

This thesis explores DL methods for RS data to monitor natural resources, namely, water bodies and forests. DL methods are efficient in learning features from the data but require a massive corpora of labeled data for training the model. In such a scenario, unsupervised approaches are pretty suitable. This thesis is composed of two published papers exploring the combination of DL models with unsupervised algorithms to take advantage of both methods. The research questions addressed in this thesis are:

Q1 How well do supervised methods work with RGB aerial imagery?

Q2 Is it possible to reach similar performance to supervised methods using unsupervised approaches?

Q3 Can the unsupervised approach be extended to multispectral satellite imagery?

Paper A addresses the first two research questions by introducing an Unsupervised Curriculum Learning (UCL) method for classifying water bodies from three different continents to the classification of space-borne and air-borne RGB imagery. It is composed of the widely used VGG-16 architecture with clustering technique, K-Means. UCL progressively learns the features from the data in an unsupervised manner with pseudo-labels generated by K-Means clustering. The unsupervised nature of the UCL removes the requirement of data labeling that demands domain experts' knowledge. UCL is evaluated using three datasets; SAT-6, EuroSAT, and PakSAT. Passat is a novel contribution of this work designed for Pakistani water bodies. The evaluated results on these datasets showed that UCL outperformed supervised models in domain adaptation.

Paper B builds on the previous study by extending UCL to multispectral imagery to classify Australian wildfires of late 2019 and early 2020. This study has used multispectral imagery of Sentinel-2 satellite to create the patch-based dataset forest fire damaging Australia. It concatenates 12 out of the 13 bands of Sentinel-2 imagery in such a way that they are suitable as a three-channel input of UCL. It classifies the patches into burnt and not burnt categories. This work achieved an F1-Score 87% mapping the burnt and not burnt regions of Australia, demonstrating the effectiveness of the proposed model.

5.1 Future directions

Unsupervised learning for classification covers a wide range. This thesis touches on one of its aspects and introduces the concept of Unsupervised Curriculum Learning (UCL) for classification of image patches. The work presented in this thesis can be extended in many ways in the domain of Earth Observation (EO) (e.g. by increasing the number of classes, modifying the method developed for multi-class problems, or by moving from patch-based classification to pixel-based classification), and I describe these directions in more detail in Section 5.1.1. The present work can also be extended beyond Earth Observation (EO), to other domains, like document analysis and NLP, as it is discussed in more detail in Section 5.1.2.

5.1.1 Expansion of UCL

To demonstrate the viability of the UCL model on a relatively simple task first, the research work so far was mainly focusing on the binary classification of patches. The proposed solution, however, can be extended to the multi-class classification of patches. One way to achieve this would be by adding for each class a representative sample (that is a sample one can think of as the prototypical representation of its class) to the training set. In the first iteration of fine-tuning, clusters would be created considering these samples as the centroids so that the model could distinguish the multiple classes. This makes the model semi or weakly supervised.

Another possible direction can be extending the patch-based classification to pixel-based one. Initial experiments are in process to analyze the performance of widely used supervised deep learning-based segmentation algorithms, namely, U-Net, SegNet, Masked R-CNN [74] and RetinaNet [75]. Further direction is to explore semi or unsupervised domains to develop a better-unsupervised approach for pixel-based classification. Moreover, as EO can also include other types of data than just images (e.g. meteorological data), another possible future direction for this work is to include other modalities in the proposed framework.

5.1.2 Beyond Earth Observation

The proposed method can also be applied to other domains involving multispectral and multi-model data. It can be adapted to the domain of historical document analysis, forgery detection from documents, and Natural Language Processing (NLP) classification.

UCL is a patch-based binary classifier. Some areas of historical document analysis like font classification could be solved with the proposed unsupervised approach where data is processed into small patches carrying important information of the font text. Later, these patches will be fed to UCL to learn the different fonts from the data.

The same iterative approach can be applied even outside the domain of image processing too. Here, one would only need to replace the current model (e.g., VGG16) with one that would learn a latent representation for other modalities. One possible candidate for this is Natural Language Processing (NLP) and, in particular, text classification. Here, instead of the Convolutional Neural Network for image processing, one can use text representation models, like the State-of-the-art transformers (e.g., BERT [76], or T5 [77]). Then, the same clustering as before can be applied to the latent representations generated by these models for text data for the classification of text. This integration of NLP deep models will make UCL framework suitable for NLP binary problems like hate speech classification. The initial experiments are in process for binary hate speech classification with the initial UCL framework. Hate speech classification could prove to be a potential problem to explore the effectiveness UCL in NLP domain.

References

[1] N. NOAA, "What is the difference between land cover and land use," 2015.

[2] F. team, "Earth observation," Jun 2021. [Online]. Available: https://ec.europa.eu/jrc/en/research-topic/earth-observation

[3] "Geo - Group on earth observations: FAQ." [Online]. Available: https://www.earthobservations.org/g_faq.html

[4] N. B. Mishra, "Wetlands: remote sensing," in *Wetlands and Habitats*. CRC Press, 2020, pp. 201–212.

[5] D. W. Pearce, "The economic value of forest ecosystems," *Ecosystem health*, vol. 7, no. 4, pp. 284–296, 2001.

[6] *United Nations World Water Development Report 2020: Water and Climate Change*, 2019.

[7] S. J. Pyne, *World fire: the culture of fire on earth.* University of Washington press, 1997.

[8] "Fires, forests and the future: A crisis raging out of control?" [Online]. Available: https://wwfeu.awsassets.panda.org/downloads/wwf_fires_forests_and_the_future_report.pdf

[9] "Water Scarcity." [Online]. Available: https://www.worldwildlife.org/threats/water-scarcity

[10] K. Iizuka, M. Itoh, S. Shiodera, T. Matsubara, M. Dohar, and K. Watanabe, "Advantages of unmanned aerial vehicle (UAV) photogrammetry for landscape analysis compared with satellite data: A case study of postmining sites in Indonesia," *Cogent Geoscience*, vol. 4, 2018.

[11] M. Pásler, J. Komárková, and P. Sedlák, "Comparison of possibilities of UAV and Landsat in observation of small inland water bodies," in *International Conference on Information Society, i-Society 2015*, 2015.

[12] E. Salamí, C. Barrado, and E. Pastor, "UAV flight experiments applied to the remote sensing of vegetated areas," *Remote Sensing*, vol. 6, 2014.

[13] Q. Feng, J. Liu, and J. Gong, "UAV Remote sensing for urban vegetation mapping using random forest and texture analysis," *Remote Sensing*, vol. 7, 2015.

[14] M. R. Casado, T. Irvine, S. Johnson, M. Palma, and P. Leinster, "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*, vol. 10, 2018.

[15] J. G. Masek, M. A. Wulder, B. Markham, J. McCorkel, C. J. Crawford, J. Storey, and D. T. Jenstrom, "Landsat 9: Empowering open science and applications through continuity," *Remote Sensing of Environment*, vol. 248, p. 111968, 2020.

[16] "Landsat 9." [Online]. Available: https://landsat.gsfc.nasa.gov/landsat-9

[17] J. Li and B. Chen, "Global revisit interval analysis of landsat-8-9 and sentinel-2a-2b data for terrestrial monitoring," *Sensors*, vol. 20, no. 22, p. 6631, 2020.

[18] P. Helber, B. Bischke, A. Dengel, and D. Borth, "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*, vol. 12, 2019.

[19] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "Deepsat: a learning framework for satellite imagery," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2015, pp. 1–10.

[20] C. Domenikiotis, A. Loukas, and N. R. Dalezios, "The use of NOAA/AVHRR satellite data for monitoring and assessment of forest fires and floods," *Natural Hazards and Earth System Science*, vol. 3, 2003.

[21] S. K. McFeeters, "The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features," *International Journal of Remote Sensing*, vol. 17, 1996.

[22] H. Xu, "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," *International Journal of Remote Sensing*, vol. 27, 2006.

[23] C. H. Key and N. C. Benson, "Measuring and remote sensing of burn severity," in *Proceedings joint fire science conference and workshop*, vol. 2. University of Idaho and International Association of Wildland Fire Moscow, ID, 1999, p. 284.

[24] S. Trigg and S. Flasse, "An evaluation of different bi-spectral spaces for discriminating burned shrub-savannah," *International Journal of Remote Sensing*, vol. 22, no. 13, pp. 2641–2647, 2001.

[25] M. Martín, I. Gómez, and E. Chuvieco, "Performance of a burned-area index (BAIM) for mapping Mediterranean burned scars from MODIS data," in *Proceedings of the 5th International Workshop on Remote Sensing and GIS Applications to forest fire management: fire effects assessment.* Universidad de Zaragoza, GOFC GOLD, EARSeL, 2005.

[26] I. Alonso-Canas and E. Chuvieco, "Global burned area mapping from ENVISAT-MERIS and MODIS active fire data," *Remote Sensing of Environment*, vol. 163, pp. 140–152, 2015.

[27] D. Stroppiana, G. Bordogna, P. Carrara, M. Boschetti, L. Boschetti, and P. Brivio, "A method for extracting burned areas from Landsat TM/ETM+ images by soft aggregation of multiple Spectral Indices and a region growing algorithm," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 69, pp. 88–102, 2012.

[28] A. A. Pereira, J. Pereira, R. Libonati, D. Oom, A. W. Setzer, F. Morelli, F. Machado-Silva, and L. M. T. De Carvalho, "Burned area mapping in the Brazilian Savanna using a one-class support vector machine trained by active fires," *Remote Sensing*, vol. 9, no. 11, p. 1161, 2017.

[29] Z. Wang, J. Liu, J. Li, and D. D. Zhang, "Multi-SpectralWater Index (MuWI): A Native 10-m Multi-SpectralWater Index for accuratewater mapping on sentinel-2," *Remote Sensing*, vol. 10, 2018.

[30] T. D. Acharya, D. H. Lee, I. T. Yang, and J. K. Lee, "Identification of water bodies in a landsat 8 oli image using a j48 decision tree," *Sensors (Switzerland)*, vol. 16, 2016.

[31] G. Lefebvre, A. Davranche, L. Willm, J. Campagna, L. Redmond, C. Merle, A. Guelmami, and B. Poulin, "Introducing WIW for detecting the presence of water in wetlands with Landsat and Sentinel satellites," *Remote Sensing*, vol. 11, 2019.

[32] B. C. Ko, H. H. Kim, and J. Y. Nam, "Classification of potential water bodies using landsat 8 OLI and a combination of two boosted random forest classifiers," *Sensors (Switzerland)*, vol. 15, 2015.

[33] B. D. Vries, C. Huang, M. W. Lang, J. W. Jones, W. Huang, I. F. Creed, and M. L. Carroll, "Automated quantification of surface water inundation in wetlands using optical satellite imagery," *Remote Sensing*, vol. 9, 2017.

[34] S. Mahdavi, B. Salehi, J. Granger, M. Amani, B. Brisco, and W. Huang, "Remote sensing for wetland classification: a comprehensive review," *GIScience and Remote Sensing*, vol. 55, 2018.

[35] F. Mouillot, M. G. Schultz, C. Yue, P. Cadule, K. Tansey, P. Ciais, and E. Chuvieco, "Ten years of global burned area products from spaceborne remote sensing—a review: Analysis of user needs and recommendations for future developments," *International Journal of Applied Earth Observation and Geoinformation*, vol. 26, 2014.

[36] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *International conference on artificial neural networks.* Springer, 2018, pp. 270–279.

[37] G. Hinton, Y. LeCun, and Y. Bengio, "Deep learning," *Nature*, 2015.

[38] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "DeepSat - A learning framework for satellite imagery," in *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 2015.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[40] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. Dalla Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 139–149, 2017.

[41] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 1–9.

[42] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 96–107, 2018.

[43] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote sensing of environment*, vol. 229, pp. 247–259, 2019.

[44] F. Zhang, J. Li, B. Zhang, Q. Shen, H. Ye, S. Wang, and Z. Lu, "A simple automated dynamic threshold extraction method for the classification of large water bodies from landsat-8 OLI water index images," *International Journal of Remote Sensing*, vol. 39, 2018.

[45] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais *et al.*, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, pp. 195–204, 2019.

[46] C. Pelletier, G. I. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote Sensing*, vol. 11, no. 5, p. 523, 2019.

[47] L. Knopp, M. Wieland, M. Rättich, and S. Martinis, “A Deep Learning Approach for Burned Area Segmentation with Sentinel-2 Data,” *Remote Sensing*, vol. 12, no. 15, p. 2422, 2020.

[48] M. M. Pinto, R. Libonati, R. M. Trigo, I. F. Trigo, and C. C. DaCamara, “A deep learning approach for mapping and dating burned areas using temporal sequences of satellite images,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 160, pp. 260–274, 2020.

[49] W. Fang, C. Wang, X. Chen, W. Wan, H. Li, S. Zhu, Y. Fang, B. Liu, and Y. Hong, “Recognizing global reservoirs from Landsat 8 images: A deep learning approach,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, 2019.

[50] R. Zhu, L. Yan, N. Mo, and Y. Liu, “Semi-supervised center-based discriminative adversarial learning for cross-domain scene-level land-cover classification of aerial images,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 155, 2019.

[51] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, “Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, 2021.

[52] A. Ul-Hasan, F. Shafaity, and M. Liwicki, “Curriculum learning for printed text line recognition of ligature-based scripts,” in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2015.

[53] N. Abid, M. Shahzad, M. I. Malik, U. Schwanecke, A. Ulges, G. Kovács, and F. Shafait, “Ucl: Unsupervised curriculum learning for water body classification from remote sensing imagery,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 105, p. 102568, 2021.

[54] N. Abid, M. I. Malik, M. Shahzad, F. Shafait, H. Ali, M. M. Ghaffar, C. Weis, N. Wehn, and M. Liwicki, “Burnt forest estimation from sentinel-2 imagery of australia using unsupervised deep learning,” in *2021 Digital Image Computing: Techniques and Applications (DICTA).* IEEE, 2021, pp. 01–08.

[55] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in *The European Conference on Computer Vision (ECCV)*, September 2018.

[56] H. Fan, L. Zheng, C. Yan, and Y. Yang, “Unsupervised person re-identification: Clustering and fine-tuning,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 4, 2018. [Online]. Available: https://doi.org/10.1145/3243316

[57] R. M. S. Bashir, M. Shahzad, and M. M. Fraz, “VR-PROUD: Vehicle Re-identification using PROgressive Unsupervised Deep architecture,” *Pattern Recognition*, vol. 90, pp. 52–65, 2019.

[58] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ACM International Conference Proceeding Series*, vol. 382, 2009.

[59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[62] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[63] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[64] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[65] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[66] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[67] C. C. Bridges Jr, "Hierarchical cluster analysis," *Psychological reports*, vol. 18, no. 3, pp. 851–854, 1966.

[68] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[69] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[70] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[71] I. Goodfellow, Y. Bengio, and A. Courville, "6.2. 2.3 softmax units for multinoulli output distributions," *Deep learning*, no. 1, p. 180, 2016.

[72] G. L. Feyisa, H. Meilby, R. Fensholt, and S. R. Proud, "Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery," *Remote Sensing of Environment*, vol. 140, 2014.

[73] C. Huang, L. S. Davis, and J. R. Townshend, "An assessment of support vector machines for land cover classification," *International Journal of Remote Sensing*, vol. 23, 2002.

[74] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[75] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

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

[77] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

Part II

Paper A

UCL: Unsupervised Curriculum Learning for Water Body Classification from Remote Sensing Imagery

Authors:
Nosheen Abid, Muhammad Shahzad, Muhammad Imran Malik, Ulrich Schwanecke, Adrian Ulges, György Kovács, and Faisal Shafait

Reformatted version of paper originally submitted in:
International Journal of Applied Earth Observation and Geoinformation

UCL: Unsupervised Curriculum Learning for Water Body Classification from Remote Sensing Imagery

Nosheen Abid, Muhammad Shahzad, Muhammad Imran Malik, Ulrich Schwanecke, Adrian Ulges, György Kovács and Faisal Shafait

Abstract

This paper presents a Convolutional Neural Networks (CNN) based Unsupervised Curriculum Learning approach for the recognition of water bodies to overcome the stated challenges for remote sensing based RGB imagery. The unsupervised nature of the presented algorithm eliminates the need for labelled training data. The problem is cast as a two class clustering problem (water and non-water), while clustering is done on deep features obtained by a pre-trained CNN. After initial clusters have been identified, representative samples from each cluster are chosen by the unsupervised curriculum learning algorithm for fine-tuning the feature extractor. The stated process is repeated iteratively until convergence. Three datasets have been used to evaluate the approach and show its effectiveness on varying scales: (i) SAT-6 dataset comprising high resolution aircraft images,(ii) Sentinel-2 of EuroSAT, comprising remote sensing images with low resolution, and (iii) PakSAT, a new dataset we created for this study. PakSAT is the first Pakistani Sentinel-2 dataset designed to classify water bodies of Pakistan. Extensive experiments on these datasets demonstrate the progressive learning behaviour of UCL and reported promising results of water classification on all three datasets. The obtained accuracies outperform the supervised methods in domain adaptation, demonstrating the effectiveness of the proposed algorithm.

1 Introduction

Waterbody detection from the water surface is a fundamental module in many remote sensing studies such as land cover [1] and land use [2], estimating water scarcity [3], controlling flood hazard [4], predicting aquatic widespread disease, and measuring water quality [5]. In this paper, we cast water detection as a two-class problem where input images are divided into smaller sized image patches, and each image patch is classified as water or a non-water patch. To solve this classification problem, we present an efficient and robust deep Unsupervised Curriculum Learning (UCL) based algorithm. Specifically, in this work, we propose unsupervised curriculum learning such that representative sample selection for water and non-water category is done from unlabelled data.

Extraction of representative samples for each cluster to fine-tune the deep network is thus a key step in our approach. For this, we take inspiration from curriculum learning [6] in which the learning algorithm is fed by training samples in the order of their difficulty.

Difficulty depends upon the complexity of the samples' features. Easy samples are fed in the starting iterations and the complexity is gradually increased in the subsequent iterations. It has been shown that feeding the learner in this manner improves the learning speed as deep learning algorithm has to gradually learn complex examples. The major challenge in curriculum learning is how to define the difficulty level of a sample. Different heuristics have been used successfully in different domains to establish the complexity of a data sample. Here, we use proximity from cluster centroids as the criteria for the selection of representative "easy" samples. In the beginning, the deep learning model used for feature extraction is pre-trained on the ImageNet dataset, which represents a different domain. Hence, the resulting model may not extract good features of the water bodies and non-water bodies from remote sensing imagery, leading to loose clusters. When the clusters are loosely packed, only a few easy samples are selected. The deep learning model is fine-tuned on these samples using their pseudo labels. With every iteration, the fine-tuned deep learning model extracts better features resulting in improved clusters. Progressively, we increase the complexity of the chosen samples via curriculum learning by allowing more samples to be selected for fine-tuning deep model on water bodies using pseudo labels. This progressive-leaning behaviour can be called Unsupervised Curriculum Learning. This idea has been exploited in natural images [7, 8] and the proposed framework adopted this idea for the classification of remote sensing imagery.

In this context, the contributions proposed in this paper are two-fold. (1) An easy-to-implement unsupervised progressive deep learning model for water body classification from RGB remote sensing imagery. The integration of clustering with curriculum learning leads to unsupervised learning of the deep model by using pseudo labels.(2) The evaluation of the proposed strategy is performed using three datasets, out of which two are benchmark datasets, space-borne EuroSAT [9] and air-borne Sat-6 [10], and our newly introduced custom space-borne dataset, PakSAT, which shall be made open for the public. The statistics of the PakSAT dataset are detailed in Section 4.3.

The rest of the paper is structured as follows: In Section 2 we present a brief review of the techniques related to our approach and waterbodies detection using remote sensing data. In Section 3 we describe the designed unsupervised methodology. Then, in Section 4 we discuss the datasets used "as-is" as well as the dataset generated for this paper. Following this, in Section 5 we present the experiments conducted and give a detailed analysis of the obtained results. Lastly, in Section 6 we provide our conclusions and give an outlook to future work.

2 Related Work

Numerous techniques have been introduced to detect water utilizing radar and optical imagery [11]. Radar data have the advantage of capturing the information in almost every weather and day-night condition. However, the prominent features of vegetation [12], waves [13, 14], sand [15] and radar shadows produced by landscape features [16] deterrent the effective separation of water from the land surface. Therefore, the extraction of water bodies from remotely sensed data is more effective from optical imagery than from radar data [17].

2.1 Sensing Modalities

Satellite Imagery (Space-borne)

The satellite imagery (spaceborne) of the Landsat Program (launched in 1972) gradually became the most popular source of optical imagery in remote sensing. The optical sensor of the Landsat satellites has a typical spatial resolution of 30 m which was better than other freely available coarse sensors like MERIS with spatial resolution 300 m, NOAA/AVHRR with spatial resolution 1100 m, and MODIS with spatial resolution $250 - 1000$ m. Therefore, most of the global digital maps for water are designed using Landsat imagery, some of the datasets are Global Inland Water Body (GIW) [18], Global Surface Water [19], and Global Water Bodies Database (GLOWABO) [20]. Similarly, most of the water detection based applications are designed using Landsat Imagery [21, 22, 23]. The Sentinel-2 which was launched by the European Space Agency (ESA) in 2015 became an alternative to Landsat. Compared to the Landsat imagery, Sentinel-2 provides 10 meter spatial resolution imagery of RGB bands with less revisit time and wider swath.

The data captured from satellites has a low ground resolved distance. This leads to a blurring of the images and making the detailed land features like the boundaries of water bodies difficult or almost impossible to identify. The lack of detailed land feature analysis impacts the results of environmental assessment [24]. Despite the availability of high-resolution sensors like GeoEye ($0.46 - 1.84$ m), WorldView ($0.31 - 2.40$ m), and IKONOS ($1 - 4$ m) [25] that are used in different applications, the difficulty of accessing the data creates hindrance in the development of pervasive applications. Alternative solutions are required to overcome the stated limitation to capture the land information in detail for analyzing inter and intra class variation.

Aerial Imagery (Air-borne)

The multispectral satellite imagery allows reliable extraction of water using various water indices [26, 27, 28] and specific bands based threshold methods. However, usage of optical imagery in the presence of clouds prevents the observation of the earth's surface [29]. For this, some approaches have used radar data during the period of intense cloud cover to overcome the limitations of optical imagery [19, 21]. On the contrary, most air-crafts and Unmanned Aerial Vehicles capture the data from very low elevation and provide high-resolution air-borne imagery when required considering the limitations of law regulation and weather condition [30, 31]. Air-borne imagery is actively used to detect land cover and land use changes [32, 33], and natural disasters like floods and earthquakes [34, 35]. Baker et al. [36] presented autonomous shoreline navigation using UAVs. The use of UAVs to capture air-borne imagery is comparatively cheap and fast for detailed land feature analysis of a specific area.

2.2 Relevant Approaches

Threshold Methods

In remote sensing studies, most of the water detection algorithms are based on water indices. In 1996, McFeeters [26] firstly designed a popular water index, the Normalized Difference Water Index (NDWI) for water mapping from satellite imagery. He has used the near-infrared (NIR) and the green bands of the Landsat Thematic Mapper (TM) for depicting water features. Xu [37] modified the NDWI by replacing the NIR band with shortwave-infrared (SWIR) and named it Modified Normalized Difference Water Index (MNDWI). MNDWI partly reduced the error rate generated by soil, vegetation, and urbanized areas. Feyisa [38] introduced the Automated Water Extraction Index (AWEI) to cater for the misclassification of shadow as water by using multispectral bands. A new water index was created with linear discriminant analysis which Fisher [39] revised by using five surface reflectance (SR) bands of Landsat. She also provided a thorough comparison of water indices for Landsat imagery.

Water detection requires rich spatial information to design threshold methods like Near Infra-Red (NIR) band that separates the water from the land. Most of the low-cost off-the-shelf UAVs are only equipped with conventional cameras that provide only the spectral bands red, green, and blue. RGB based threshold methods are only used for vegetation detection and observing its growth, e.g., Colour Index of Vegetation Extraction (CIVE) [40], Excess Green (ExG) [41], Excess Red (ExR) [42], Green Leaf Index (GLI) [43], Normalized Green-Red Difference Index (NGRDI) [44], Red-Green-Blue Vegetation Index (RGBVI) [45], and many others. The lack of rich spectral information in imagery provided by most UAVs limits the use of indices for water body classification [46]. In such scenarios, an effective machine learning algorithm is needed based on water classification for remote sensing imagery having limited spectral information but substantial scale variations.

Machine Learning Methods

Machine learning algorithms for water estimation can be categorized into supervised and unsupervised methods. Many waterbody classification algorithms have been designed using supervised methods, such as Support Vector Machines (SVM) [47], Decision Trees [22, 48, 49], Random Forests [50, 51, 52], Gradient Boosting [53], and Deep Neural Networks [54]. In the past few years, bag-of-visual-words based methods employing K-Means and SVM have been used in several classification and target detection techniques leading to better accuracy [55, 56]. It is noteworthy that the essential semantic information is stored within the spatial relationship of pixels instead of individual pixel intensity values. Many methods have been introduced involving image context to make the class information more explicit [33]. Luo et al. [57] proposed a hierarchical generative model, the Author-Genre-Topic Model (AGTM), to introduce context information. It was designed to perform annotation of satellite images. Recently, Generative Adversarial Neural networks (GANs) are being used in a semi-supervised manner to address the problem of cross-domain adaptation in semantic segmentation of remote sensing imagery [58, 59].

Deep learning has become a state-of-the-art method to extract more abstract features from lower layers to higher layers of the model. Comparing deep learning methods with shallow classification methods like SVM, deep learning solutions result in better learning models [60]. Cheng et al. [61] replaced hand-crafted features with CNN for water bodies segmentation. Lin et al. [62] used Fully Convolutional Network (FCN) to add multi-scale information. Noh et al. [63] designed a multi-layer deconvolutional network to address the scale challenge. Wei et al. [64] and Miao et al. [65] used auto-encoders to extract high-level feature maps from high-resolution images. Fang et al. [66] used the ResNet model to identify global water reservoirs. Yagmur et al. [67] combined residual blocks in the inception network to detect shallow water areas. In spite of their excellent performance, supervised methods have some limitations. 1) If labelled data is not available, supervised methods require the collection of data and expert knowledge for data labelling which is a time-consuming and tedious task. 2) Supervised methods are domain specific. Their accuracy often decreases drastically when applied to different domain data about the same problem. Some researchers have also explored and used the concept of Curriculum Learning to efficiently train supervised deep neural networks [68]. The focus of our work is to overcome the stated issues by introducing an unsupervised deep learning solution using the concept of curriculum learning for water body classification for RGB data.

3 Methodology

The proposed deep learning based UCL model learns the features of water bodies from remote sensing imagery in an unsupervised manner using the pseudo-labels generated by the clustering technique. The outline of the proposed method is shown in Figure 1. UCL is composed of two main modules (i) A pre-trained deep learning architecture (CNN) to extract and learn features from the remote sensing data, (ii) an unsupervised clustering technique to cluster the extracted features. A UCL based selection operation is added between clustering and fine-tuning to extract the samples present near the clusters' centroids, called "reliable samples". The deep learning model is fine-tuned on the extracted reliable samples. UCL is composed of the following steps:

Step 1: Extract features of remote sensing imagery of water bodies and non-water bodies using a pre-trained deep learning architecture.

Step 2: Create two clusters on the extracted features of remote sensing imagery, assigning them pseudo-labels of water-bodies and non-water-bodies clusters.

Step 3: From each cluster, select the reliable images using the UCL based selection operation.

Step 4: Fine-tune the deep learning module on reliable samples using pseudo labels given by the clusters.

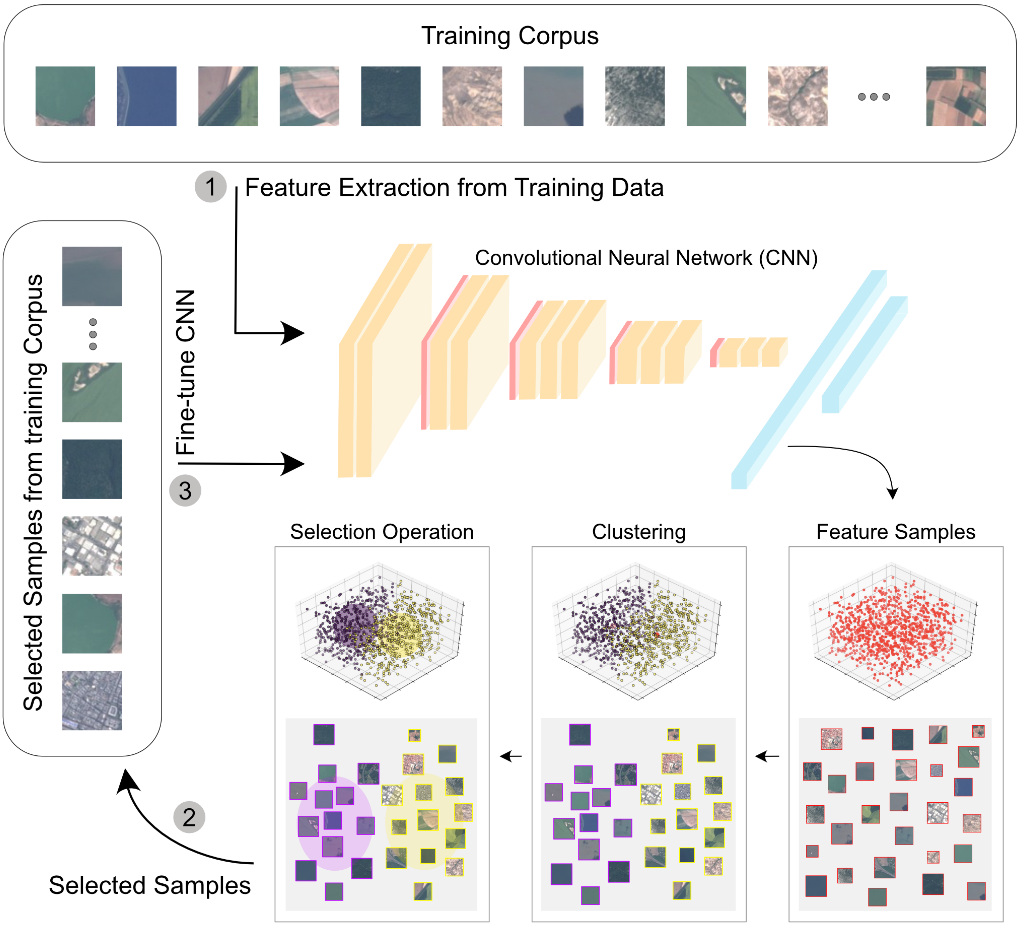

Training Corpus
1 Feature Extraction from Training Data
Convolutional Neural Network (CNN)
Selected Samples from training Corpus
Fine-tune CNN
3
Selection Operation
Clustering
Feature Samples
2
Selected Samples

Figure 1: UCL: Deep learning based Unsupervised Curriculum Learning for water classification. UCL (1) extracts the features of a training corpus using Convolutional Neural Network (CNN). It clusters the features into two classes. (2) It applies a selection operation to remove the noisy samples from the clusters. (3) The selected samples from the training corpus are used to fine-tune the CNN model. Once, the CNN model is fine-tuned, the steps are repeated until the CNN model has learned the patterns in the training corpus.

Step 5: Extract features of the whole training corpus of remote sensing imagery using the pre-trained model of the previous step.

Step 6: Repeat steps 2 to 5 until the deep learning model is converged.

In the beginning, the CNN, pre-trained on a different domain (ImageNet) is used to extract features from remote sensing imagery. These features are clustered into two, assuming them to be of water and non-water category. As the clusters are created from the features generated from a deep learning model trained on a different domain, we may obtain noisy clusters (for water classification of the remote sensing imagery). To filter out the reliable samples from the clusters, a UCL based selection operation is used to extract a small number of reliable samples. These are the samples present near the centroid of the clusters, containing the prominent features. The CNN model is fine-tuned on these reliable samples with pseudo labels assigned by clustering. The reliable samples restrict the CNN to learn only the prominent features of the clusters by avoiding unnecessary noise. The updated CNN is used for feature extraction in the preceding iteration. With every iteration, the model learns the features of remote sensing imagery with pseudo labels of clusters resulting in comparatively better clusters than the previous iteration. The process iterates until the CNN model has converged. This process is called unsupervised because it only needs the pseudo labels of the clusters to fine-tune the CNN model.

3.1 Deep Learning Model

In UCL, a deep learning module is used for feature extraction from remote sensing imagery. Later, this module is fine-tuned on selected reliable samples, say, of water bodies and non-water bodies. Several deep learning models like VGG-16, ResNet-50, DenseNet, Inception Net, and Xception Net have been explored in this work to demonstrate the generalization of the introduced framework. VGG-16 and ResNet-50 outperformed the other networks for water bodies classification from remote sensing imagery. Hence, we have used VGG-16 for our final UCL algorithm due to its lower computational complexity compared to ResNet-50. The described training process is general and independent of the CNN used. It will work with VGG as well as with Inception Net, and Xception Net. The training process of VGG-16 in UCL can be decomposed into two parts:

1. Feature Extraction: VGG-16, pre-trained on the ImageNet dataset is used for feature extraction of remote sensing imagery in the first iteration. The output of the last convolutional layer is extracted to get feature maps of each sample in the dataset. The extracted feature maps are flattened to get the feature vectors. These feature vectors are clustered into two, assuming them to be of water and non-water bodies. From these clusters, reliable feature samples are selected using the UCL based selection operation for fine-tuning the deep model.

2. Fine-tuning of VGG-16 with reliable images: The model is fine-tuned on the training set of reliable samples considering their cluster as their pseudo label. We have modified the input layer of VGG-16 according to our aerial image patch size and the output layer to the number of clusters we generate. In the current scenario, the considered two classes for training are; water-body and non-water-body.

3.2 Clustering

The features extracted from the deep learning module are clustered using an unsupervised clustering technique. We have explored three different types of clustering techniques, K-Means, hierarchical clustering, and Fuzzy C-Means (FCM). These techniques are suitable for the considered problem of water and non-water classification of remote sensing imagery patches as they have a fixed number of classes. FCM and K-Means clustering generated better clusters than hierarchical clustering (see Section 5). We have deployed K-Means clustering as it has lower time complexity than FCM.

Suppose the features extracted from remote sensing imagery patches $\{x_i\}_{i=1}^{N}$ using the deep learning model $\phi(., \theta_i)$ are represented by $\{f_i\}_{i=1}^{N}$.

$$\{f_i\}_{i=1}^{N} \leftarrow \phi(\{x_i\}_{i=1}^{N}, \theta) \tag{1}$$

These features are clustered such that each feature vector is assigned a cluster label $\{y_i\}_{i=1}^{N}$ where $y_i \in \{1, \ldots, k\}$ on the basis of a minimum distance from the centroid c_k, where c is the centroid of k^{th} cluster. In the current scenario, $k = 2$ to generate two clusters, assuming them to be of water bodies and non-water bodies.

$$\{y_i\}_{i=1}^{N} \leftarrow \min \sum_{i=1}^{N} \sum_{k=1}^{2} |f_i - c_k| \tag{2}$$

These generated pseudo-labels, $\{y_i\}_{i=1}^{N}$, are later used for fine-tuning the CNN model.

3.3 UCL based Selection Operation

Using the concept of CL we want to extract reliable samples for fine-tuning the CNN. We achieve this in an unsupervised manner, by selecting features near the centroid of a cluster. More specifically, we select all features that are at a distance λ from the centroid of the cluster (see Selection Operation in Figure 1). The parameter λ is a constant value that can be adjusted according to the requirement. We have used $\lambda = 0.85$ after empirical evaluations.

The closest feature vector to the centroid is considered as a centroid feature vector, $\{f_k\}_{k=1}^{2}$ where k represents the cluster.

$$\{f_k\}_{k=1}^{2} \leftarrow \min\{|f_{ik} - c_k|\}_{i=1}^{N} \tag{3}$$

We calculate the similarity between a specific feature vector f_i belonging to a cluster k and the centroid feature vector f_k using the inner product, i.e.

$$\{\gamma_i\}_{i=1}^{N} \leftarrow \{f_{ik} \cdot f_k\}_{i=1}^{N} \tag{4}$$

If the calculated similarity is greater than the λ, the sample of the considered feature vector is declared as a reliable sample $x_i^{'}$ and the cluster label is considered as the pseudo sample label for the next training cycle.

$$\{x_i^{'}\}_{i=1}^{M} \leftarrow \{\gamma_i\}_{i=1}^{N} < \lambda \tag{5}$$

The number of extracted reliable samples vary at every iteration of fine-tuning.

3.4 Progressive Learning

Initially, when the model is not trained on remote sensing imagery, the features extracted from this model may result in loose clusters. These clusters are less dense and result in a set of only a few reliable images. In the beginning, the network is fine-tuned on this set of a few reliable images considering their cluster as their pseudo label, either water or non-water. Then the fine-tuned network is used to extract features of the whole training corpus and new clusters are generated. The selection operation is performed on these clusters to extract reliable samples. This time the clusters might be comparatively dense than they were in the previous iteration, and we get more reliable images. Progressively, the model gets stronger by learning the patterns from the data and the set of reliable images iteratively grows leading to self-paced learning.

The proposed UCL technique is an unsupervised binary classifier where the model has learnt the features and is able to create clusters of water and non-water. Now the question is how to know which cluster indicates which class? The automation of mapping the pseudo-labels to true labels is beyond the scope of this work and can be considered as one of the future directions.

3.5 Implementation Details

The experiments were conducted on a GPU machine having an NVIDIA Titan-X GPU for training and fine-tuning with 32 GB RAM and Linux operating system. It took about 4.5 h for training the model on the considered datasets. We used a Stochastic Gradient Descent (SGD) optimizer and categorical cross-entropy loss. Learning rate was set to 0.0001 , momentum to 0.9 and batch size to 16 images. The training dataset with pseudo labels is split into 80% for training and 20% for validation for progressive learning of UCL. The input layer of the deep learning model is set to 64x64x3. As the considered three datasets have a difference in their image patch sizes (28x28, 61x61, and 64x64), the patches of all the datasets were interpolated to patch size 64x64 to make them suitable as an in input to the deep learning model.

4 Datasets

UCL takes the input in the form of an image patch and classifies it either to be of water or non-water category. Therefore, huge tiles of aerial images were broken down into smaller patches on the basis of the requirement for better classification. Only RGB bands of the images were considered to model the situation where only these bands are available, like high-resolution UAV data which usually does not contain multispectral bands. We have used two publicly available datasets, namely, EuroSAT [9] and SAT-6 [10], and our newly created PakSat dataset to demonstrate the effectiveness of the proposed architecture.

This study has considered the imagery from different parts of the world to address variations in the spectral responses of water depending on the region of the globe. EuroSAT carries Sentinel-2 imagery patches of 34 European countries, SAT-6 covers the area of California, and PakSAT is composed of Sentinel-2 imagery patches of Pakistan.

4.1 SAT-6 Dataset

SAT-6 [10] is a high-resolution dataset captured from an aircraft providing 1 meter GSD pixel resolution covering different parts of California. It is composed of small patches of size 28 × 28 divided into six classes. Initially, it has four bands; red, green, blue, and infrared. The dataset was preprocessed to remove the infrared band from the corpus to make it suitable for RGB input. The patches were divided into two categories i.e. water and non-water class. Both classes contain an almost equal number of patches, having a corpus of 7500 image patches for fine-tuning and 3000 for testing. This high-resolution dataset was used to evaluate the robustness of UCL for scale variation and to prove the hypothesis of the progressive learning behaviour of the model.

4.2 EuroSAT Dataset

EuroSAT [9] is composed of image patches from different regions of 34 European countries. The image parches are of sentinel-2 satellite having a resolution of 10 m per pixel for red, green, and blue bands. It consists of 10 classes of land cover and land use. We have only used the red, green, and blue bands of the data and divided the dataset into two classes by considering the “river” and “sea & lake” as water class and the rest of the classes as the non-water class in such a way that there are almost 50 % samples of water and 50 % of non-water category. The purpose behind balancing the two classes is the unbiased training of the deep learning model. The size of each patch is in EuroSAT dataset is 64 × 64. We have considered 8000 image patches for fine-tuning and 3000 for testing.

4.3 PakSAT Dataset

With the PakSat dataset, we have developed the dataset for water bodies segmentation from Sentinel-2 satellite imagery. This dataset is composed of water bodies of Pakistan. We have applied the threshold methods to get a roughly estimated mask of water pixels in the image. Later, the correction of the generated mask is done manually. The PakSat dataset is composed of 61 × 61 sized patches of water, non-water, and mixed classes. Each patch has 14 bands, the first 13 bands are of Sentinel-2 and the 14 th is the generated mask. In this study, we have only considered the red, green, and blue bands of the data having a pixel resolution of 10 m. The dataset consists of three classes; namely water, non-water, and mixed. The patches containing more than 75 % of water pixels are declared as water patches. Patches with less than 75 % of water are placed in the mixed category. The patches with no water pixel were classified as non-water patches. Lets have a look into the process of PakSAT dataset creation.

Process for creating the PakSAT dataset

The creation of PakSAT dataset was started with the downloading Sentinel-2 tiles of water reservoirs and part of the Indus river of Pakistan from June 2015 till October 2019. After downloading the required tiles of Sentinel-2 imagery, the following steps were taken one by one to create the ground truth of the PakSAT dataset.

Resampling

The Sentinel-2 bands come in varying spatial resolution of 10m, 20m, and 60m. For overlying data, the cell resolution should be the same. For this, each band was upsampled to achieve 10 m resolution using the Bilinear Resampling technique. Bilinear is an interpolation technique that considers the values of the four nearest pixels to calculate the value for the current pixel on the output image. The calculated new values on the output raster are the weighted average of the considered four nearest values. The four values are considered on the basis of their distance from the center of the output pixel. Resampling is processed by Data Management Tools in ArcGIS. All bands of the spatial resolution of 20m e.g. Band5, Band6, Band7, Band8A, Band11, and Band12 of Sentinel-2 and bands of 60m resolution e.g. Band1, Band9, and Band10 of Sentinel-2 resolution was resampled to 10m by using the Bilinear Resampling technique. We have performed this step in the creation of the PakSAT data to make it more practical for more interested researchers who intended to use the dataset with multispectral bands.

Labelling Water Features

Clear and greenish reflection water bodies show high reflectance values for the green band than red and blue. Whereas muddy water bodies show high reflectance on the red band than green and blue. Hence, the Normalized Difference Water Index (NDWI), uses the green band along with the near-infrared (NIR) band, whereas the Normalized Difference Vegetation Index (NDVI), uses the red band with near-infrared (NIR). Both of the approaches map the water bodies differently according to the difference in their color. We have used both of the indices to estimate the water bodies in Sentinel-2 imagery.

Normalized Water Index (NDWI) is another threshold method to detect water from remote sensing data. NDWI absorbs the NIR reflectance and emphasizes the green band reflectance to detect water bodies. Thus, water pixels become prominent having positive values, and other categories like soil and vegetation carry zero or negative values. A threshold on these values gives an estimated binary mask of water bodies.

$$NDWI = \frac{Green - NIR}{Green + NIR} \tag{6}$$

The values of NDWI range between -1 and 1. In most cases, NDWI was used to detect water features but in some tiles, where there are some problems in detecting features due to shadows, clouds, or muddy water, NDVI was deployed. The values of NDVI as well range between -1 and 1 as both are normalized indices. For NDVI, the negative values close to -1 represent water/ The values close to zero usually correspond to barren areas like rocky or sandy land or snow category. Low positive values roughly up to 0.4 classify green lands like grass and shrubs. Whereas positive values approaching 1 represent tropical and temperate rainforests.

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{7}$$

The pixels of water were labelled by reclassifying the pixels. The pixel values representing water were labelled as 1 and 0 otherwise. This resulted in a binary mask for water bodies. These auto-generated binary masks contained some errors like cloud cover, shadows, and boundary pixels of the water bodies.

Manual Correction

Manual Correction included the step of locating the errors by visualizing the satellite image and the corresponding generated binary mask for a specific area of interest containing water body (see Figure 2). Wherever the pixel(s) was misclassified, it was manually corrected by changing the binary mask value from 0 to 1 and vice versa.

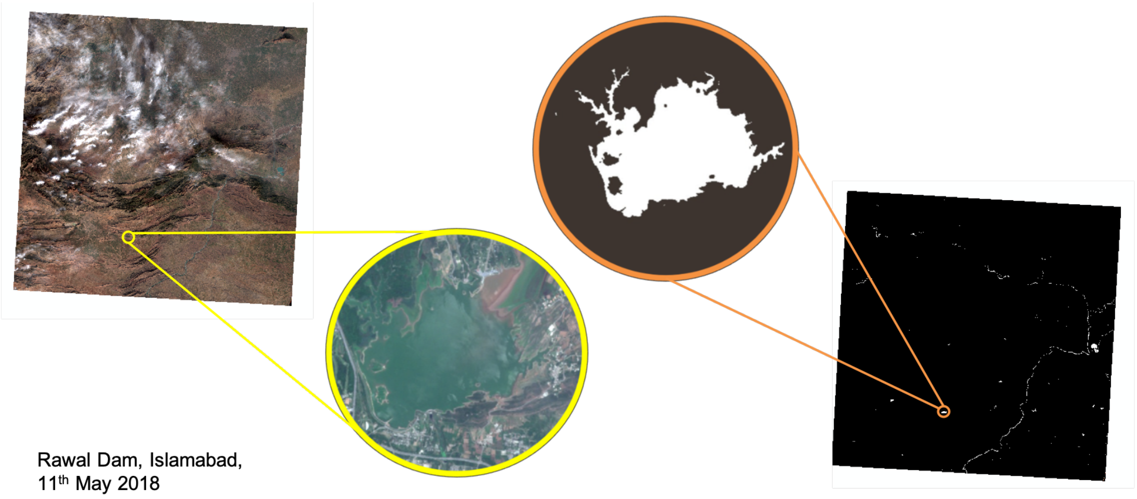


Figure 2: Sentinel-2 tile captured on 11th May 2018 containing Rawal Dam along with its approximated binary mask generated by using NDVI and NDWI.

Bands Composition

Band composition is similar to layer stacking in which after resampling into the same resolution of 10m, all the 13 bands are stacked together in the standard order of Sentinel-2. Lastly, the generated ground-truth mask is added to it as a 14th band and stored as a single raster file.

Water Bodies	Capturing Date	Water	Land	Mixed	Total
Chashma Reservior	2016-09-05	877	160	-	1037
	2017-11-06	315	296	-	611
Darawat Dam	2019-10-19	118	510	131	759
Ghazi Barotha Reservior	2019-04-06	9	778	137	924
Gomal zam Dam	2016-08-15	372	-	-	372
	2019-06-08	58	189	76	323
Indus River	2016-07-22	287	-	-	287
Manchar Lake	2017-06-08	650	215	-	865
	2017-08-27	570	1651	173	2394
	2018-08-01	554	-	-	554
	2019-10-16	719	1694	242	2655
Mangla Dam	2016-02-06	455	3186	228	3869
	2016-10-20	673	-	-	673
	2017-12-07	206	2299	402	2907
	2018-08-19	492	3336	660	4488
	2018-12-11	452	-	-	452
Rawal Dam	2016-11-12	75	-	-	75
	2018-10-07	73	-	-	73
	2019-06-10	48	951	171	1170
Tarbela Dam	2016-07-04	347	-	-	347
	2017-07-07	235	2102	212	2549
	2019-04-06	389	1954	335	2678
Total Count		7974	19321	2767	29389

Table 1: The count of patches in each category generated from Sentinel-2 tile captured on a specified date containing water body.

Splitting into Patches

As each raster file was huge in size, they should be divided into smaller patches to make it suitable for UCL. Each raster file composed of 14 channels (13 Sentinel-2 bands and 14th ground-truth binary mask) is split into smaller non-overlapping patches of size $61 \times 61 \times 14$. The size of the patch is an important parameter as UCL takes the patch and does classification at the patch level.

Categorization of Patches

Some of the generated patches contain only water, some do not contain water or contain the coastline of the water having a portion of water and non-water part. We segmented the patches into three different categories on the basis of the water pixels they contain. If a patch has no water pixel, it belongs to the ”Land” category. If a patch has 75% or more water pixels then it belongs to the ”Water” category. If a patch has water less than 75%, it is categorized as ”Mixed” containing both water and non-water regions.

In total 29 389 patches were created. Out of which 7974 patches belong to the Water category, 19 321 belong to the Land category, and 2767 belong to the Mixed category. The details of the water patches can be seen in Table 1.

In this work, we have used the water and non-water patches with RGB bands for unsupervised progressive learning of water classification from satellite imagery. Figure 3 provides a visualization of the patches of the considered three datasets.

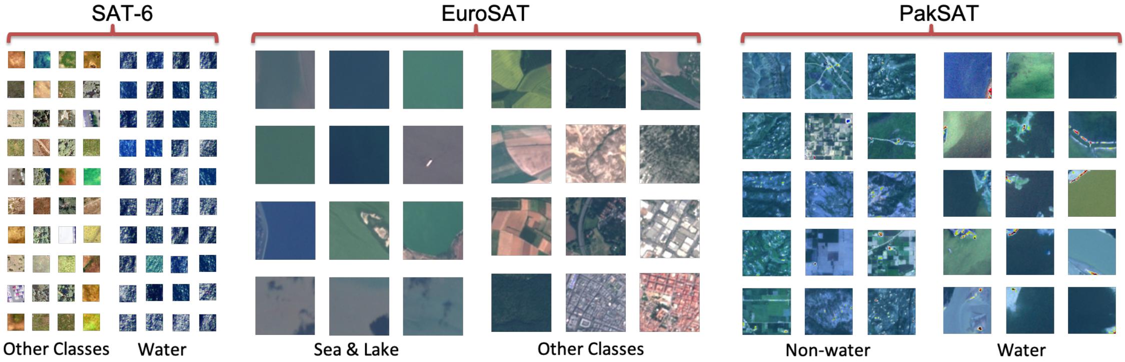


Figure 3: Some patches of Sat-6, EuroSAT and PakSAT Datasets.

The size of the input patch has great significance. If the patch size is large, there are chances of mixing the boundary pixels of multiple land covers. The smaller the patch size the better the results. Our main focus is to develop an unsupervised deep learning based approach considering the prominent features of each class. This could be done by avoiding the patches containing the features of both classes significantly like 50% of each class in the patch. It may confuse the network at the time of training for binary classification. If the network is designed for multiple classes and a new class can be added called ”mix class”, the network will be able to learn such patches containing features of multiple classes which can be considered as a future direction of the work. Our major focus is on learning the prominent features for binary classification.

5 Experimental Validation

The progressive learning behaviour of UCL was analyzed by conducting multiple experiments on the considered datasets, namely Sat-6, EuroSAT, and PakSAT. The experiments are divided into 4 subsections; i) direct testing of the considered dataset on ImageNet weights, ii) fine-tuning of VGG-16 and UCL supervised, iii) their cross domain adaptation, iv) clustering analysis and v) error analysis.

5.1 Direct Testing on ImageNet Weights

Before conducting the experiments of UCL progressive learning and supervised comparison, direct testing on VGG-16 with ImageNet weights of EuroSAT, SAT-6 and PakSAT is done (see Table 2).

Initial Weights	Test Set	F1-Score (%)	
		Random FC	K-Means
ImageNet	EuroSAT	44.46	54.45
ImageNet	PakSAT	51.98	57.13
ImageNet	Sat-6	58.20	65.50

Table 2: Describes the direct inference of EuroSAT, PakSAT, and SAT-6 datasets on VGG-16 considering three different techniques. Random FC indicates the results with random initialization of the FC layer. K-Means shows the results for clustering the features extracted from VGG-16 and classified by K-Means clustering.

VGG-16 with ImageNet weights is directly tested on EuroSAT, PakSAT, and SAT-6 datasets. Assuming the dataset to be unlabelled, two strategies were followed; 1) the classification layer of VGG-16 is randomly initiated with 0 mean and 0.001 standard deviations, and 2) VGG-16 with ImageNet weights are used as a feature extractor for remote sensing data. In Table 2, Random FC represents the results of randomly initializing the fully connected binary classification layer. As the random initialization of the classification layer is not aware of the considered datasets, it makes the classification quite challenging leading to unsatisfactory results. Whereas, in the second approach, K-Means has been used for the classification of deep extracted features which lead to comparatively better performance for all three datasets. The EuroSAT gives the lowest performance, it is because we have combined the River class with Sea & Lake. Sea & Lake class carries the Sentinel-2 patches of only water. Whereas, in River class patches we can observe a great part of the land with the river stream. This intermixes the river patches with non-water class leading to poor intra class variation for clustering and compromised F1-Score. The PakSAT and SAT-6 datasets carry prominent patches of water and non-water categories.

5.2 Fine-tuning of VGG-16 and UCL

In general, supervised deep learning models have better performance than unsupervised models as they are trained using ground-truth labels. UCL's performance has been analyzed considering the supervised model's performance as the benchmark. VGG-16 with ImageNet weights is fine-tuned in a supervised manner on EuroSAT, SAT-6 and PakSAT datasets that reported the F1-Score of 99.49 %, 99.53 % and 96.07 %, respectively, see Table 3. All the fine-tuned models have learned the remote sensing features to classify the water patches for the respective datasets.

Fine-Tuned & Tested	F1-Score (%)	
	Supervised	UCL
EuroSAT	99.49	84.05
SAT-6	99.53	90.89
PakSAT	96.07	87.66

Table 3: F1-Scores of VGG-16 supervised fine-tuned and tested on each dataset; EuroSAT, PakSAT, and SAT-6. The last column reports the results of UCL fine-tuned and tested on each considered dataset.

We have analyzed the progressive learning behaviour of UCL that is capable of learning the variations in the new dataset, progressively. For UCL training, we assume that there are no labels available for the training process. UCL uses the clustering technique to generate the pseudo labels to train the deep learning model. The CNN based model of UCL with ImageNet weights has been fine-tuned with EuroSAT, PakSAT, and SAT-6 datasets in unsupervised progressive learning behaviour. The UCL reported promising results on all the three datasets by giving F1-Score above 80%, see Table 3. The supervised model performed better than UCL but there is a huge trade-off of data labelling which is quite an exhaustive task.

Considering the dataset with no labels, the direct testing on ImageNet weights with K-Mean clustering performed comparatively better than the random classification layer (see Table 2). It reported the F1-Score of 54.45 % for EuroSAT, 57.13 % for PakSAT, and 65.5 % for SAT-6 dataset. Later, deploying the UCL for progressive unsupervised learning from the data, it is able to learn the features from the unlabelled data by reporting the considerable improvement in the F1-Score. The evaluated F1-Score on UCL is 84.05 % for EuroSAT, 87.66 % for PakSAT, and 90.89 % for SAT-6 dataset, see Table 3. The F1-Score is improved by around 20 % for EuroSAT and PakSAT datasets, and 25 % for SAT-6 dataset.

5.3 Cross-domain Adaptation

To analyze the domain adaptation, we have tested the model trained on one dataset on the other two datasets (see Table 4). It has been observed that the test accuracy for supervised fine-tuned VGG-16 is quite low on the other datasets, indicating that the supervised models are data-specific and lack adaptability in the cross-domain adaptation scenario. Here, the variation among the datasets is due to different data acquisition platforms (space-borne/air-borne), resulting in different image properties with varying image resolution.

To further analyze the progressive learning behaviour of the proposed model and

Fine-Tuned	Tested	F1-Score (%)	
		Supervised	UCL
EuroSAT	PakSAT	70.68	78.00
	SAT-6	32.84	75.76
SAT-6	PakSAT	63.16	68.97
	EuroSAT	42.83	72.43
PakSAT	EuroSAT	62.70	79.00
	SAT-6	59.92	71.72

Table 4: F1-Scores of VGG-16 fine-tuned in a supervised manner on each dataset and tested on the other two datasets. The last column reports the results of UCL fine-tuned on each dataset and tested on the other two datasets. The considered datasets are EuroSAT, PakSAT, and SAT-6.

its adaptation to the new dataset, each UCL model trained on one dataset was tested on the other two considered datasets. It can be seen in Table 4, the supervised model trained on EuroSAT does not perform that well on PakSAT and SAT-6 datasets with an F1-Score of 70.68 % and 32.84 %, respectively. Whereas, UCL fine-tuned with EuroSAT achieved comparatively better F1-Scores on the other datasets i.e., 78.00 % for PakSAT and 75.76 % for SAT-6 datasets. Similarly, the supervised trained models and the UCL trained models of SAT-6 and PakSAT were tested on the other two datasets and a similar trend was observed (as reported in Table 4).

5.4 Clustering Analysis

The UCL results of the SAT-6 dataset are further analyzed by observing the clusters generated using K-Means during training. We have observed the training procedure of UCL for 20 iterations of fine-tuning. Initially, the clusters are created using the features extracted from a model trained on a different domain, ImageNet, which does not contain any remote sensing data. Consequently, the generated clusters are not compact and loosely packed for remote sensing imagery. To evaluate the compactness of the clusters, purity and Silhouette Score are computed, see Figure 4. The purity is a supervised measure that calculates the correctly classified samples over the total number of samples in the cluster. Whereas, Silhouette Score is an unsupervised measure that calculates the compactness of the clusters on the basis of the distance between each sample within the cluster and the neighboring clusters.

It can be seen in Figure 4 that the Silhouette Score is as low as 0.15 at the start, showing the lack of compactness in the clusters. Later, with iterations of fine-tuning, the value grows indicating the saturation in the compactness. Whereas, the purity remains at a good score in the range of 0.90 to 1.00 over iterations.

The Sum of Squared Error (SSE) is evaluated for the clusters generated from SAT-6 dataset which is also an objective function of K-Means clustering. In Figure 5, the

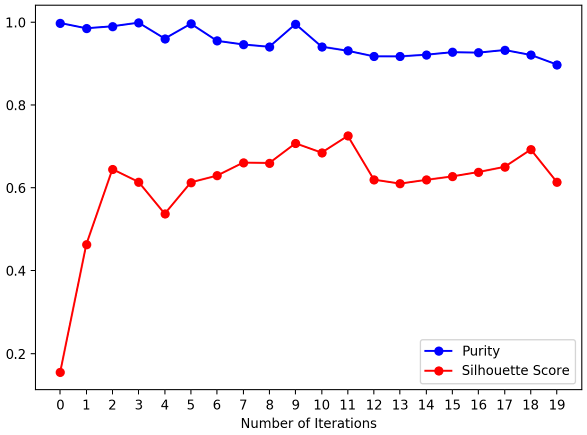


Figure 4: Graph showing the purity and Silhouette Score of generated clusters for water bodies and other regions of SAT-6 over every iteration of fine-tuning of VGG16.

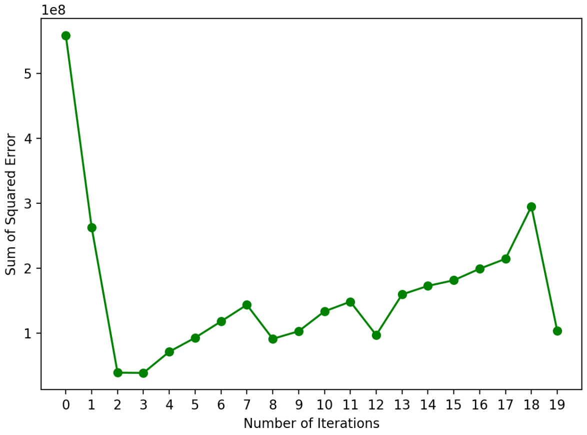


Figure 5: Graph showing Sum of Squared Error of generated clusters for water bodies and other regions of SAT-6 over every iteration of fine-tuning of VGG16.

graph represents the values of SSE for generated clusters at every fine-tuning iteration. The vertical axis represents the values of SSE and the horizontal axis are the fine-tuning iterations of the model. The graph shows a quite high value of SEE at the start when clustering is done with features extracted from a pre-trained model of ImageNet. As soon as the model is fine-tuned on SAT-6 patches, the SEE score is exponentially decreased. We can see a small peak at the end of the curve showing that the progressive fine-tuning of the model has led to somewhat over-fitting and fine-tuned models of these iterations can be ignored. This graph helps to choose the candidate fine-tuned models over iterations. The fine-tuning iterations having minimum SSE can be the potential fine-tuned model for the deployment.

As the clusters are loosely packed at the start, we extract the reliable samples present near the centroid of the clusters using the selection operation of UCL. The deep learning model is fine-tuned on these samples. This step avoided the noisy samples of the clusters and restricted the model to learn random features.

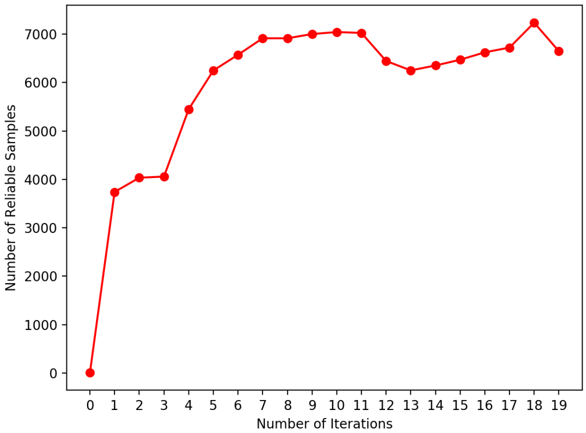


Figure 6: Graph showing the count of reliable samples of SAT-6 over every iteration of fine-tuning of VGG16.

It can be seen in Figure 6 that only a few reliable samples (exactly 4) are extracted at the start using the pre-trained model further indicating that the samples are loosely packed in the clusters. With iterations of fine-tuning, the count of reliable samples is significantly growing. After a few iterations, the growth in the count of the reliable samples seemed saturated. After 4 iterations, the count of reliable samples remains above 6000 indicating that the reliable set contained the majority of the samples from the corpus of size 7000.

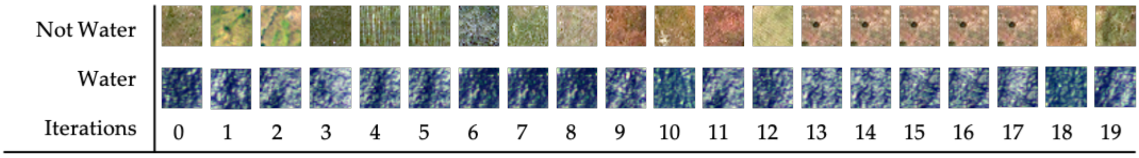


Figure 7: The centroids of the clusters of SAT-6 at each iteration of fine-tuning of VGG-16.

In UCL, VGG-16 is trained with pseudo-labels generated by the clustering technique, K-Means. The model has reported almost 0% error over cross-validation on every iteration of fine-tuning. The cross-validation corpus is a fraction of the training corpus where the generated pseudo-labels are used for cross-validation purpose as well. The model is fine-tuned end-to-end to the classification layer. Figure 7 shows a visualization of centroid patches of the clusters at every iteration of fine-tuning for the SAT-6 dataset. It can be seen that the centroids of both clusters at every iteration are belonging to the water and non-water class indicating that the model is able to distinguish both classes.

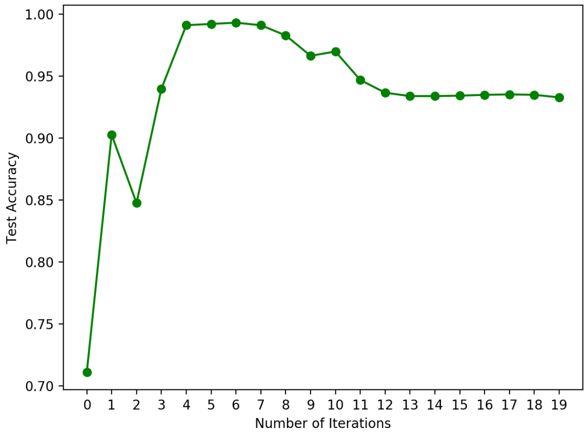


Figure 8: Graph showing the test accuracy of SAT-6 over every iteration of fine-tuning of VGG16.

Predicted →

Actual ↓	Water	Non-water	Total
Water	1,460	0	1,460
Non-water	20	1,440	1,460
Total	1,480	1,440	2,920

Table 5: Confusion Matrix for SAT-6 test corpus on best iteration of fine-tuning of VGG-16.

The fine-tuned models of 20 iterations are evaluated on the test corpus, see Figure 8. It can be seen that the model has converged well on the 4th iteration, reporting the F1-Score of 0.99. The accuracy remained consistent for the next two iterations. After that, the accuracy tends to decrease indicating a sign of overfitting. It may be because of the multiple interactions of fine-tuning the model over the same dataset. The model reported the highest accuracy 99.31 % at the 6th iteration. Table 5 shows the confusion matrix of the 6th iteration of fine-tuning. It can be seen that most of the patches are correctly classified with the exception of 20 false-negative patches that are belonging to other regions and are declared as water by the model.

To analyze the change of pseudo-labels among the patches, we have evaluated the count of the patches whose pseudo-labels are changed in the next iteration, see Figure 9. It can be seen that at the 0th iteration, all the patches are predicted with a specific class. Later, with every iteration, the count of change in the labels of patches tends to decrease and gets converged after a few iterations.

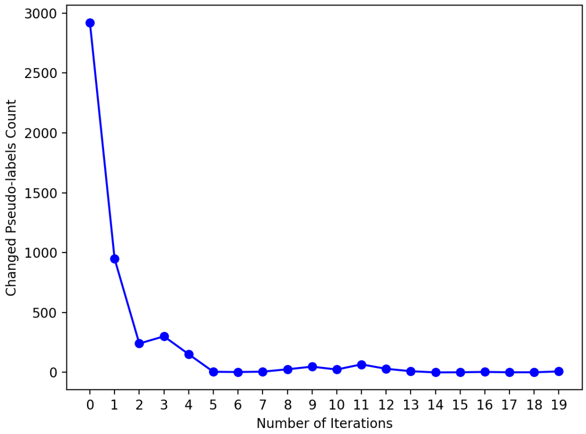


Figure 9: Graph showing the count of the patches whose pseudo-labels are changed in the next iteration of fine-tuning of VGG16.

5.5 Error Analysis

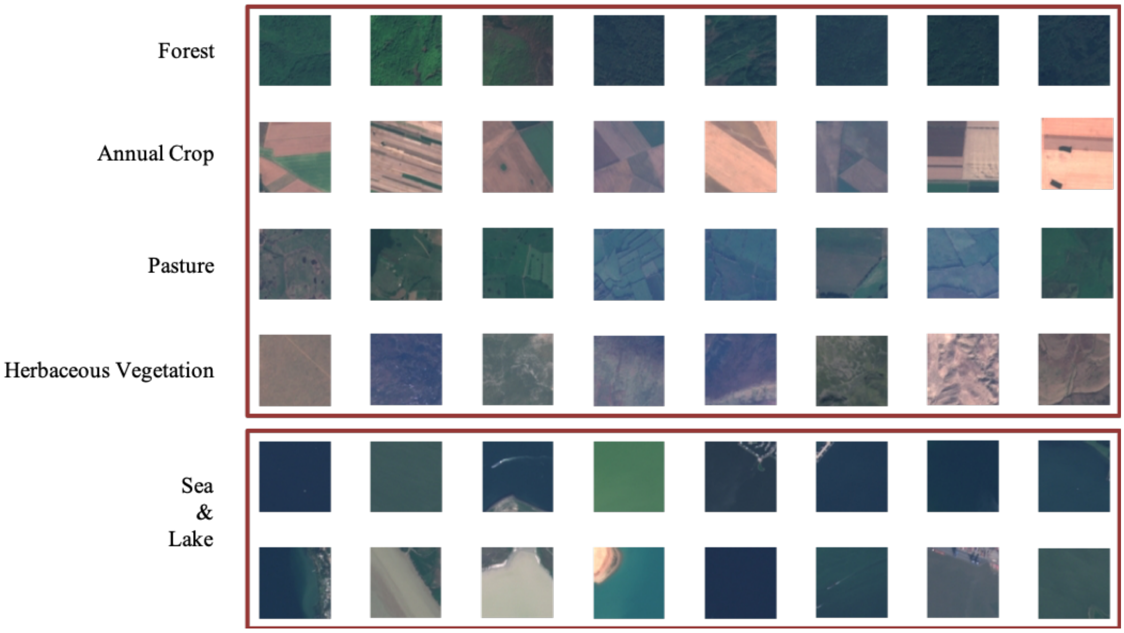


Figure 10: These are some failure cases that where intermixed with water class. The generated clusters were intermixing forest, annual crop, pasture and herbaceous vegetation with water class because of resemblance in the appearance.

The extracted reliable samples of the best fine-tuning iteration of EuroSAT demonstrate that the visual features of residential, industry, highways, annual and permanent crops were properly classified as non-water categories. Whereas, the forest patches show a similar colour as the reflection of some lakes and seawater which resulted in intermixing of some forest patches with water and vice versa. Herbaceous vegetation patches also showed a resemblance to water. It could be because of the low resolution of Sentinel-2 to appropriately distinguish the intra-class variation in different lands. Some of the failure cases have been visualized in Figure 10.

6 Conclusion and Outlook

In this paper, we have introduced an unsupervised method UCL to categorize water bodies from remote sensing imagery and showed that the unsupervised deep learning approach can learn the desired features and have the tendency to outperform the supervised model with respect to domain adaptation. The supervised models of deep learning need a massive dataset of labelled images to train the architecture, which is a tedious, exhausting, and time-consuming task. The unsupervised algorithm removes the requirement of a labelled dataset for training the architecture and perform classification. The datasets used to prove the hypothesis are, EuroSAT (covering 34 European countries), PakSAT (covering Pakistan), and SAT-6 (covering California). This paper has shown the efficiency of unsupervised architecture by reporting the F1-Score around 85% to 91% for considered datasets (see Table 3). We have also analyzed the domain adaptation of UCL to check its robustness using EuroSAT, PakSAT, and SAT-6 datasets. We have trained the UCL on one dataset and tested its performance on the other datasets. UCL was able to give better domain adaptation performance on other datasets than supervised models with a considerable difference in the F1-Score from 8% to 42% (see Table 4). However, a big room of work is still to be done in the field. We worked on the patch-wise classification of the images considering red, green, and blue bands only which can be extended to pixel-based segmentation. This work also focused only on binary classification and could be further extended to the multi-class classification of aerial photographs.

References

[1] F. Zhang, J. Li, B. Zhang, Q. Shen, H. Ye, S. Wang, and Z. Lu, "A simple automated dynamic threshold extraction method for the classification of large water bodies from landsat-8 OLI water index images," *International Journal of Remote Sensing*, vol. 39, 2018.

[2] W. G. Buma, S. I. Lee, and J. Y. Seo, "Recent surface water extent of lake Chad from multispectral sensors and GRACE," *Sensors (Switzerland)*, vol. 18, 2018.

[3] Q. Guo, R. Pu, J. Li, and J. Cheng, “A weighted normalized difference water index for water extraction using landsat imagery,” *International Journal of Remote Sensing*, vol. 38, 2017.

[4] A. Kyriou and K. Nikolakopoulos, “Flood mapping from Sentinel-1 and Landsat-8 data: a case study from river Evros, Greece,” in *Earth Resources and Environmental Remote Sensing/GIS Applications VI*, vol. 9644, 2015.

[5] M. Sadeghi, E. Babaeian, M. Tuller, and S. B. Jones, “The optical trapezoid model: A novel approach to remote sensing of soil moisture applied to Sentinel-2 and Landsat-8 observations,” *Remote Sensing of Environment*, vol. 198, 2017.

[6] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in *ACM International Conference Proceeding Series*, vol. 382, 2009.

[7] H. Fan, L. Zheng, and Y. Yang, “Unsupervised person re-identification: Clustering and fine-tuning,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 4, 2018.

[8] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11218, 2018.

[9] P. Helber, B. Bischke, A. Dengel, and D. Borth, “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*, vol. 12, 2019.

[10] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, “DeepSat - A learning framework for satellite imagery,” in *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 2015.

[11] M. Guo, J. Li, C. Sheng, J. Xu, and L. Wu, “A review of wetland remote sensing,” *Sensors*, vol. 17, no. 4, p. 777, 2017.

[12] W. Huang, B. DeVries, C. Huang, M. W. Lang, J. W. Jones, I. F. Creed, and M. L. Carroll, “Automated extraction of surface water extent from Sentinel-1 data,” *Remote Sensing*, vol. 10, 2018.

[13] J. Töyrä and A. Pietroniro, “Towards operational monitoring of a northern wetland using geomatics-based techniques,” *Remote Sensing of Environment*, vol. 97, 2005.

[14] B. Marti-Cardona, J. Dolz-Ripolles, and C. Lopez-Martinez, “Wetland inundation monitoring by the synergistic use of ENVISAT/ASAR imagery and ancilliary spatial data,” *Remote Sensing of Environment*, vol. 139, 2013.

[15] S. Martinis, S. Plank, and K. Ćwik, "The use of Sentinel-1 time-series data to improve flood monitoring in arid areas," *Remote Sensing*, vol. 10, 2018.

[16] L. Giustarini, R. Hostache, P. Matgen, G. J. Schumann, P. D. Bates, and D. C. Mason, "A change detection approach to flood mapping in Urban areas using TerraSAR-X," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, 2013.

[17] G. Schumann, G. D. Baldassarre, and P. D. Bates, "The utility of spaceborne radar to render flood inundation maps based on multialgorithm ensembles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, 2009.

[18] A. Veloso, S. Mermoz, A. Bouvet, T. L. Toan, M. Planells, J. F. Dejoux, and E. Ceschia, "Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications," *Remote Sensing of Environment*, vol. 199, 2017.

[19] K. N. Markert, F. Chishtie, E. R. Anderson, D. Saah, and R. E. Griffin, "On the merging of optical and SAR satellite imagery for surface water mapping applications," *Results in Physics*, vol. 9, 2018.

[20] N. Pahlevan, S. Sarkar, B. A. Franz, S. V. Balasubramanian, and J. He, "Sentinel-2 MultiSpectral Instrument (MSI) data processing for aquatic science applications: Demonstrations and validations," *Remote Sensing of Environment*, vol. 201, 2017.

[21] I. Manakos, G. A. Kordelas, and K. Marini, "Fusion of Sentinel-1 data with Sentinel-2 products to overcome non-favourable atmospheric conditions for the delineation of inundation maps," *European Journal of Remote Sensing*, vol. 53, 2020.

[22] A. Davranche, B. Poulin, and G. Lefebvre, "Mapping flooding regimes in Camargue wetlands using seasonal multispectral data," *Remote Sensing of Environment*, vol. 138, 2013.

[23] S. Puliti, S. Saarela, T. Gobakken, G. Ståhl, and E. Næsset, "Combining UAV and Sentinel-2 auxiliary data for forest growing stock volume estimation through hierarchical model-based inference," *Remote Sensing of Environment*, vol. 204, 2018.

[24] K. Sanga-Ngoie, K. Iizuka, and S. Kobayashi, "Estimating CO2 sequestration by forests in oita prefecture, Japan, by combining LANDSAT ETM+ and ALOS satellite remote sensing data," *Remote Sensing*, vol. 4, 2012.

[25] T. M. Tu, P. S. Huang, C. L. Hung, and C. P. Chang, "A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 1, 2004.

[26] S. K. McFeeters, "The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features," *International Journal of Remote Sensing*, vol. 17, 1996.

[27] C. Domenikiotis, A. Loukas, and N. R. Dalezios, "The use of NOAA/AVHRR satellite data for monitoring and assessment of forest fires and floods," *Natural Hazards and Earth System Science*, vol. 3, 2003.

[28] G. A. Kordelas, I. Manakos, G. Lefebvre, and B. Poulin, "Automatic inundation mapping using sentinel-2 data applicable to both camargue and do'ana biosphere reserves," *Remote Sensing*, vol. 11, 2019.

[29] X. Shen, D. Wang, K. Mao, E. Anagnostou, and Y. Hong, "Inundation extent mapping by synthetic aperture radar: A review," *Remote Sensing*, vol. 11, 2019.

[30] K. Iizuka, M. Itoh, S. Shiodera, T. Matsubara, M. Dohar, and K. Watanabe, "Advantages of unmanned aerial vehicle (UAV) photogrammetry for landscape analysis compared with satellite data: A case study of postmining sites in Indonesia," *Cogent Geoscience*, vol. 4, 2018.

[31] M. Pásler, J. Komárková, and P. Sedlák, "Comparison of possibilities of UAV and Landsat in observation of small inland water bodies," in *International Conference on Information Society, i-Society 2015*, 2015.

[32] E. Salamí, C. Barrado, and E. Pastor, "UAV flight experiments applied to the remote sensing of vegetated areas," *Remote Sensing*, vol. 6, 2014.

[33] Q. Feng, J. Liu, and J. Gong, "UAV Remote sensing for urban vegetation mapping using random forest and texture analysis," *Remote Sensing*, vol. 7, 2015.

[34] M. Abdelkader, M. Shaqura, C. G. Claudel, and W. Gueaieb, "A UAV based system for real time flash flood monitoring in desert environments using Lagrangian microsensors," in *2013 International Conference on Unmanned Aircraft Systems, ICUAS 2013 - Conference Proceedings*, 2013.

[35] M. R. Casado, T. Irvine, S. Johnson, M. Palma, and P. Leinster, "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*, vol. 10, 2018.

[36] P. Baker and B. Kamgar-Parsi, "Using shorelines for autonomous air vehicle guidance," *Computer Vision and Image Understanding*, vol. 114, 2010.

[37] H. Xu, "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," *International Journal of Remote Sensing*, vol. 27, 2006.

[38] G. L. Feyisa, H. Meilby, R. Fensholt, and S. R. Proud, "Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery," *Remote Sensing of Environment*, vol. 140, 2014.

[39] A. Fisher, N. Flood, and T. Danaher, “Comparing Landsat water index methods for automated water classification in eastern Australia,” *Remote Sensing of Environment*, vol. 175, 2016.

[40] T. Kataoka, T. Kaneko, H. Okamoto, and S. Hata, “Crop growth estimation system using machine vision,” in *IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM*, vol. 2, 2003.

[41] D. M. Woebbecke, G. E. Meyer, K. V. Bargen, and D. A. Mortensen, “Color indices for weed identification under various soil, residue, and lighting conditions,” *Transactions of the American Society of Agricultural Engineers*, vol. 38, 1995.

[42] G. E. Meyer, T. W. Hindman, and K. Laksmi, “¡title¿machine vision detection parameters for plant species identification¡/title¿,” in *Precision Agriculture and Biological Quality*, vol. 3543, 1999.

[43] M. Louhaichi, M. M. Borman, and D. E. Johnson, “Spatially located platform and aerial photography for documentation of grazing impacts on wheat,” *Geocarto International*, vol. 16, 2001.

[44] C. J. Tucker, “Red and photographic infrared linear combinations for monitoring vegetation,” *Remote Sensing of Environment*, vol. 8, 1979.

[45] J. Bendig, K. Yu, H. Aasen, A. Bolten, S. Bennertz, J. Broscheit, M. L. Gnyp, and G. Bareth, “Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 39, 2015.

[46] J. Komarkova, I. Cermakova, P. Sedlak, and J. Jech, “Spectral Enhancement of Imagery for Small Inland Water Bodies Monitoring: Utilization of UAV-Based Data,” *Journal of Information Systems Engineering & Management*, vol. 4, 2019.

[47] Z. Wang, J. Liu, J. Li, and D. D. Zhang, “Multi-SpectralWater Index (MuWI): A Native 10-m Multi-SpectralWater Index for accuratewater mapping on sentinel-2,” *Remote Sensing*, vol. 10, 2018.

[48] G. Lefebvre, A. Davranche, L. Willm, J. Campagna, L. Redmond, C. Merle, A. Guelmami, and B. Poulin, “Introducing WIW for detecting the presence of water in wetlands with Landsat and Sentinel satellites,” *Remote Sensing*, vol. 11, 2019.

[49] T. D. Acharya, D. H. Lee, I. T. Yang, and J. K. Lee, “Identification of water bodies in a landsat 8 oli image using a j48 decision tree,” *Sensors (Switzerland)*, vol. 16, 2016.

[50] C. Huang, L. S. Davis, and J. R. Townshend, “An assessment of support vector machines for land cover classification,” *International Journal of Remote Sensing*, vol. 23, 2002.

[51] B. C. Ko, H. H. Kim, and J. Y. Nam, "Classification of potential water bodies using landsat 8 OLI and a combination of two boosted random forest classifiers," *Sensors (Switzerland)*, vol. 15, 2015.

[52] B. D. Vries, C. Huang, M. W. Lang, J. W. Jones, W. Huang, I. F. Creed, and M. L. Carroll, "Automated quantification of surface water inundation in wetlands using optical satellite imagery," *Remote Sensing*, vol. 9, 2017.

[53] S. Mahdavi, B. Salehi, J. Granger, M. Amani, B. Brisco, and W. Huang, "Remote sensing for wetland classification: a comprehensive review," *GIScience and Remote Sensing*, vol. 55, 2018.

[54] F. Isikdogan, A. C. Bovik, and P. Passalacqua, "Surface water mapping by deep learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, 2017.

[55] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 2010.

[56] F. Hu, G. S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised Feature Learning Via Spectral Clustering of Multidimensional Patches for Remotely Sensed Scene Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, 2015.

[57] W. Luo, H. Li, G. Liu, and L. Zeng, "Semantic annotation of satellite images using author-genre-topic model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, 2014.

[58] R. Zhu, L. Yan, N. Mo, and Y. Liu, "Semi-supervised center-based discriminative adversarial learning for cross-domain scene-level land-cover classification of aerial images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 155, 2019.

[59] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, 2021.

[60] G. Hinton, Y. LeCun, and Y. Bengio, "Deep learning," *Nature*, 2015.

[61] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote Sensing Image Scene Classification Using Bag of Convolutional Features," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, 2017.

[62] H. Lin, Z. Shi, and Z. Zou, "Fully Convolutional Network with Task Partitioning for Inshore Ship Detection in Optical Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, 2017.

[63] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[64] X. Wei, Y. Guo, X. Gao, M. Yan, and X. Sun, “A new semantic segmentation model for remote sensing images,” in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017.

[65] Z. Miao, K. Fu, H. Sun, X. Sun, and M. Yan, “Automatic Water-Body Segmentation from High-Resolution Satellite Images via Deep Networks,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, 2018.

[66] W. Fang, C. Wang, X. Chen, W. Wan, H. Li, S. Zhu, Y. Fang, B. Liu, and Y. Hong, “Recognizing global reservoirs from Landsat 8 images: A deep learning approach,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, 2019.

[67] N. Yagmur, N. Musaoglu, and G. Taskin, “Detection of shallow water area with machine learning algorithms,” in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 42, 2019.

[68] A. Ul-Hasan, F. Shafaity, and M. Liwicki, “Curriculum learning for printed text line recognition of ligature-based scripts,” in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2015.

Paper B

Burnt Forest Estimation from Sentinel-2 Imagery of Australia using Unsupervised Deep Learning

Authors:
Nosheen Abid, Muhammad Imran Malik, Muhammad Shahzad, Faisal Shafait, Haider Ali, Muhammad Mohsin Ghaffar, Christian Weis, Norbert Wehn and Marcus Liwicki

Reformatted version of paper accepted for publication in:
International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2021.

Burnt Forest Estimation from Sentinel-2 Imagery of Australia using Unsupervised Deep Learning

Nosheen Abid, Muhammad Imran Malik, Muhammad Shahzad, Faisal Shafait, Haider Ali, Muhammad Mohsin Ghaffar, Christian Weis, Norbert Wehn and Marcus Liwicki

Abstract

Massive wildfires not only in Australia but also across the globe burn millions of hectares of forests and green land affecting the social, ecological and economical situation worldwide. Widely used indices-based threshold methods like Normalized Burned Ratio (NBR) require a huge amount of data pre-processing and are specific to the data capturing source. The state-of-the-art deep learning models are supervised and require domain experts knowledge for labeling the data in huge quantity. These limitations make the existing models difficult to be adaptable to the new variations in the data and capturing sources. In this work, we have proposed an unsupervised deep learning based architecture to map the burnt regions of forests. The model considers small patches of satellite imagery and classifies them into burnt and not burnt. These small patches are concatenated into binary masks to segment out the burnt region of the forests. The proposed system is composed of two modules: 1) a state-of-the-art deep learning architecture for feature extraction and 2) a clustering algorithm for the generation of pseudo labels to train the deep learning architecture. The proposed method is capable of learning the features progressively in an unsupervised fashion from the data with pseudo labels reducing the exhausting efforts of data labeling that requires expert knowledge. We have used the real-time data of Sentinel-2 for training the model and mapping the burnt regions. The obtained F1-Score of 0.87 demonstrates the effectiveness of the proposed model.

1 Introduction

Australia is, more than any other, a fire continent [1]. It has faced an annihilating beginning of a gigantic fire within the last quarter of 2019, which burnt over 5.8 million hectares of forests, mostly in Victoria(VIC) and New South Wales (NSW). In general, the number of fire alerts in Australia has increased in the past two decades due to an increase in humidity, drought, record heat, and high winds [2]. Similar to Australia, forests in other continents have historically burned up to approximately 5% in the previous decade [3] essentially devastating biodiversity, timberland riches, and human settlements [4].

Considering the severity of the circumstances and disadvantages of the furious blaze [5], the research community has actively worked on the issue. Many methods and solutions have been designed for detecting and monitoring the woodland fire by different sources like Radio Detection and Ranging (RADAR), Light Detection and Ranging (LiDAR), and optical imagery [6]. The primary sources of remote optical imagery are Unmanned Aerial Vehicles (UAV) and satellites. Satellite imagery, captured through multispectral sensors, is particularly used worldwide for forest fire detection and assessment.

Satellite imagery-based burnt area classification algorithms can be generally divided into two major categories; rule-based methods and machine learning methods. Rule-based methods are a combination of the spectral response of burnt region mostly in short wave infrared (SWIR) and near-infrared (NIR) bands of the satellite imagery. This is because fire has a significant reflectance in the SWIR and NIR bands. The spectral responses in these bands are pretty helpful in detecting sound vegetation and burned regions. In burning areas, a significant drop in values is observed in NIR reflectance and a rise in SWIR reflectance after burning. This response is because of the sensitivity of the NIR band to chlorophyll substance of healthy plants, and SWIR captures the moisture of soil and vegetation [7]. These multispectral bands, along with visual bands (red, green, and blue), are commonly used in the indices applied to detect burnt regions in satellite imagery. The most commonly used indices for the purpose are Normalized Burned Ratio (NBR) [8], the Mid-InfraRed Burn Index (MIRBI) [9], and the Modified Burned Area Index (BAIM) [10]. For detection of burnt areas from an aerial view, mostly pre-event imagery of the scene is used along with the post-event to detect the changes with the help of empirically calculated thresholds. Here, an insufficient choice of a pre-event scene may lead to misclassifications. Additionally, these traditional rule-based approaches are sensitive to noise, like cloud cover, and require exhaustive preprocessing of a massive corpus of data–thereby making the task more challenging.

In the recent past, several machine learning-based techniques have been designed to map burnt regions using remotely sensed imagery. The lately designed algorithms MCD64AI at 500 m resolution [11] and FIRECCI51 at 250-meter resolution [12] create temporal composites for capturing the lasting changes, sift low-quality pixels, and combine these processed pixels with active blaze identified using the MODerate resolution Imaging Spectroradiometer sensor (MODIS). For FireCCI51, initially, some candidate pixels for the burnt area are detected. Later the neighboring burnt pixels of the candidate pixels are identified using a pixel growing algorithm. For MCD64A1C6, steps in series are followed, including the region growing procedure. The use of the region growing technique is a very common practice in traditional approaches for mapping burnt regions [13, 14, 15]. Many other algorithms have been presented and used at regional levels. For instance, One-Class Support Vector Machine [16] is used for reducing the omission error produced by the omission of the active blaze. It has minimized the requirement of using the region growing technique to make the approach comparatively easier. However, the proposed method used temporal composite for avoiding cloud cover and cloud shadows leading to discarding some information that could be useful. Furthermore, sensors vary from each other in characteristics. Most of the traditional approaches are sensitive to the particular sensors they are designed and refined for, and adaptation of these algorithms to different sensors becomes a challenging task. Despite improvements over the years in the algorithms for burnt area mapping, there are still some facets that need improvements and/or are outside the limitations of the traditional methods. Specifically, the burnt zone mapping tools would be more useful for larger and steady time-series data, better uncertainty estimation, and mapping blaze areas and combustion completeness [17]. It raises the need to have a method that is scalable as well as adaptable to variations. Deep learning (DL) is capable of addressing the above-stated limitations [18].

Today Deep Learning (DL) techniques are rapidly becoming state-of-the art for learning variant and complex features across various domains [19]. Computer-vision problems of object detection, localization, and recognition are thrived by Convolutional Neural Networks (CNN) [20].

In the domain of remote sensing, CNNs are used in emerging applications of land-cover classification [21], segmentation of buildings, roads [22] and small objects [23], reconstruction of missing information in the data [24], cloud-cover detection [25] and cloud shadows and effective utilization of Spatio-temporal satellite data [26, 27, 28, 29]. On a similar note, burnt land mapping and dating have also been addressed by using deep learning [30]. Pinto et al. [31] combined CNNs and Long-Short-Term-Memory (LSTM) with U-Net based architecture for mapping and dating burnt regions using multispectral imagery. Similarly, Knopp et al. [32] have segmented burnt land from mono-temporal Sentinel-2 using U-Net based architecture.

Even though the deep learning methods are becoming state-of-the-art, they carry a few limitations. 1) They are generally supervised and require a huge amount of labeled data for training the model. Data labeling is time-consuming and an exhaustive task. Furthermore, in many instances, it requires expert knowledge, which is hardly available at scale. 2) They are domain-specific. Their performance diminishes radically when applied to a diverse dataset of the same problem. To bargain with these issues, the concept of Curriculum Learning [33] has been used by a few researchers [34]. In this paper, we have also used a similar concept. We performed burnt forest estimation by combining the state-of-the-art machine learning and computer vision methods with the concept of CL using Sentinel-2 Imagery of Australia. We have performed unsupervised patch-based classification of burnt and unburnt patches in satellite imagery. The process itself covers three stages, including the selection of training examples, computing the discriminative features, and classify the burnt and unburnt regions.

The proposed unsupervised method is composed of deep learning architecture, clustering algorithm, and selection operation based upon curriculum learning for burnt region classification. The model takes satellite image patches as input and classify them into ”burnt” and ”not burnt” class. It is assumed that the input patch is not labeled. Therefore, a clustering algorithm is used to tackle the issue of labeling. Firstly, pre-trained deep learning architecture is used to extract feature vectors of patches. Secondly, these feature vectors are clustered into two categories using state-of-the-art clustering techniques to generate pseudo-labels, assuming them to be burnt and not burnt. The pseudo-labels are used as a new identity of the patch. Initially, the pre-trained deep learning model is trained on the ImageNet dataset, which does not contain aerial imagery. Hence, the pre-trained model may not extract good features vector resembling burnt and not burnt region of aerial imagery, resulting in loose clusters in the feature space. Our hypothesis is that clustering will give better distribution of burnt and non-burnt patches than random division. The main idea of the approach is to iteratively improve the feature extractor by fine-tuning the deep learning model on representative samples from each cluster. A selection operation is used for selecting the samples from generated clusters for fine-tuning. The selection operation selects the samples present near the centroids of the clusters. These samples indicate the prominent features of the respected cluster. Fine-tuning of the model with selected samples and respected pseudo-labels makes the model learn the discriminative features between the two clusters. As a result, better discrimination is achieved between burnt and not burnt regions. When the model converges, one of the clusters will belong to the ”burnt” and the other to the ”not burnt” class. The major contributions of our work include:

1. A progressive deep learning model for burnt region classification from multispectral aerial imagery.

2. An unsupervised architecture removes the need for data labeling, which is a primary requirement of state-of-the-art supervised deep learning based methods.

3. A patch-based Sentinel-2 imagery dataset of burnt Australian regions from the 2020 fire incident has been developed and will be publicly available.

2 Materials and Methods

2.1 Study Area

In this study, different regions of Victoria (VIC) and New South Wales (NSW) have been considered for analyzing the wild bushfire. Figure 1 graphically shows the regions selected for this analysis while Table 1 lists down the precise geographical locations in the standard latitude/longitude (WGS84) coordinates and surrounding cities. The NSW and Victoria regions are chosen as these were the worst-hit states of Australia affected by the massive fire of 2019-2020. It has burnt more than 5 million hectares of bush, forests, and parks across the state and destroyed more than 2,000 houses. The wildfire was mostly located at the coast of the Tasman Sea in NSW. The windy conditions and hot weather added fuel to the fire and resulted in an uncontrollable situation. The massive bushfire raged the area, including the Australian capital Canberra, for weeks and months. In Victoria, the fire affected 1.2 million hectares by early January 2020 [35]. The generated smoke had drastically polluted the environment, air quality, and satellite imagery. According to Swiss-based group AirVisual [36], the quality of the polluted air in Canberra (capital) was rated as the 3rd worst of all major global cities on January 3, 2020. The satellite images from early January 2020 manifested significant dissemination of smoke from firestorms in NSW and Victoria and spread far away to New Zealand as reported in the BBC report [37].

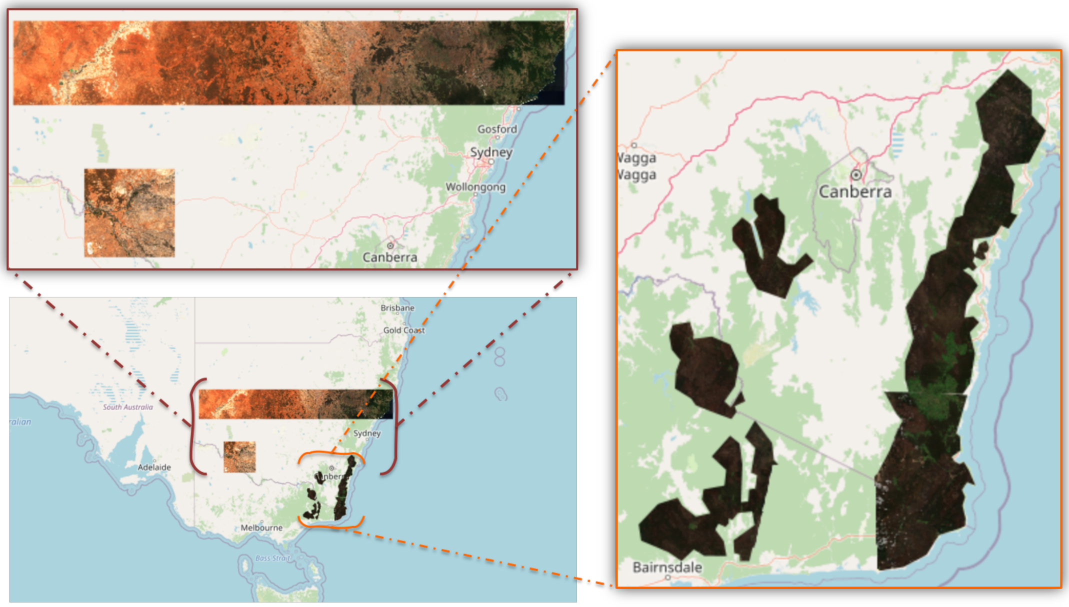


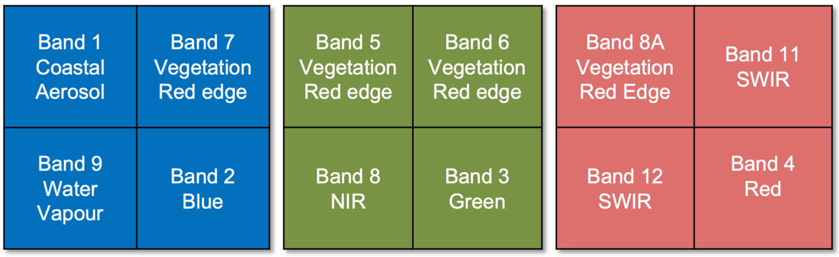
Figure 1: An illustration of the regions of New South Wales (NSW) and Victoria is considered in this study for analyzing and training the deep learning model. The rectangular regions indicate the area of early 2019 and are considered as not burnt. Whereas the other polygons in NSW and Victoria are of the 1st quarter of 2020 and considered as burnt regions as a result of the massive fire.

Class	**Region**	**Latitude**	**Longitude**	Cities			
				East	**West**	**North**	**South**
Not Burnt 2019	Mymagee	-32.14	146.72	Crowdy Head	Broken Hill	Gowang	Mount Hope
	Balrang	-34.70	143.64	Maude	Rabinvale	Corrong	Winlatin
Burnt 2020	-	-35.76	148.43	Yarrangobily	Buddong	-	Cabramurra
	Torn Groggin	-36.53	1447.92	Murray Gorge	-	Nariel Vally	-
	East-South Coast NSW	-36.40	149.59	Narooma	-	Bundanoon	Tamboon

Table 1: The table shows the details of the considered NSW regions of 2019 and 2020. Two polygons are considered from 2019 for "Not Burnt" class, whereas three polygons are considered from 2020 for "Burnt" class. These regions are used for training and testing the unsupervised deep learning model.

2.2 Data and Pre-Processing

The Sentinel-2 multispectral imagery of the selected regions of NSW and Victoria is used to generate the required dataset. We have visualized the Sentinel-2 imagery and accordingly labeled the unaffected and affected regions from fire. The rectangular regions, as shown in Figure 1, belong to the 1st quarter of 2019 and are considered as unaffected from the wildfire. Whereas the randomly shaped polygon corresponds to the 1st quarter of 2020 and is the affected burnt regions. We have extracted the respective bundle of images considering only equal to or less than 1% of the clouds. From both bundles, the two respected median images are computed. The 2019 median image of the rectangular region is used as an unaffected class, whereas the 2020 median image of irregular polygon regions is used as the burnt forest class. These two images are divided into small patches of size 64x64. We have considered the 12 bands of Sentinel-2, that are, Band 1 – Coastal aerosol, Band 2 – Blue, Band 3 – Green, Band 4 – Red, Band 5 to 7 – Vegetation red edge, Band 8 – NIR, Band 8A – Narrow NIR, Band 9 – Water vapor, Band 11 – SWIR, Band 12 – SWIR (i.e., all the bands are used except the Band – SWIR – Cirrus – 10 as it does not provide the surface information).



Band 1 Coastal Aerosol	Band 7 Vegetation Red edge	Band 5 Vegetation Red edge	Band 6 Vegetation Red edge	Band 8A Vegetation Red Edge	Band 11 SWIR
Band 9 Water Vapour	Band 2 Blue	Band 8 NIR	Band 3 Green	Band 12 SWIR	Band 4 Red

Figure 2: Shows the considered 12 bands of multispectral satellite imagery of Sentinel-2 into three-channel input. Each channel contains four bands, 1 in each quarter. Each of the red, green, and blue bands is kept in each channel, considering the input configuration of the CNN model. All the patches are preprocessed in this way to make them suitable for the input of the deep learning model.

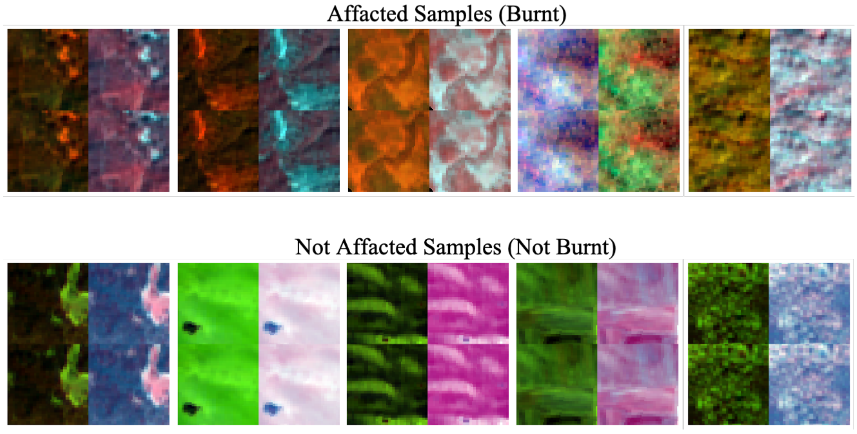


Figure 3: Shows the visuals of 10 samples of burnt and not burnt from the dataset where each sample is composed of 12 bands of Sentinel-2 imagery concatenated into three channels.

These bands are concatenated to make them suitable to feed as input to the deep learning network. The four of these bands are concatenated together to form one channel of size (128x128). Similarly, three channels sample size (128x128x3) is created out of the chosen twelve bands. Figure 2 graphically shows the bands concatenation. Among the generated data, we have randomly selected 12,000 samples for training and evaluating the employed unsupervised deep learning model. The considered dataset of 12,000 samples includes 6,000 affected (i.e., burnt forest) and 6,000 not affected regions‘ samples, out of which 8,000 were used for training and 4,000 for testing. Some affected and unaffected samples from the generated dataset are shown in Figure 3.

2.3 Deep Unsupervised Burnt Forest Learning Scheme

The proposed methodology essentially frames the burnt forest monitoring process in an unsupervised learning manner. It does so by adopting a two-phase procedure: In the first phase, a base deep convolutional neural network (CNN) is patch-wise trained (as an initializer for the subsequent phase) on the relevant dataset to learn robust and distinctive burnt forest features. Subsequently, in the second phase, these features are then input to an unsupervised clustering scheme to perform the grouping of image patches that share similar appearance characteristics. The fundamental underlying idea is to *iteratively* fine-tune this whole feature extraction and clustering scheme in an unsupervised way. The idea has been adopted from computer vision community (e.g., [38] [39] [40]) which combines the strengths of transfer learning and latent space representation to enable cross-domain adaptation. The clustering results are treated as *pseudo* labels and are fed back to the network to further fine-tune the base model. The process then continues with increasingly growing training samples with pseudo labels until convergence. Following are the individual steps outlined in a sequential manner:

1. Perform feature extraction using a pre-trained base model to extract robust burnt forest feature representations;

2. Feed the extracted feature representations to an unsupervised clustering to cluster burnt forests from the rest image patches;

3. Refine the obtained clusters to probabilistically retain the representative image patches;

4. The cluster IDs are used to assign pseudo labels to the unlabeled refined image patches;

5. Retrain (i.e., fine-tune) the deep learning module with each refined image patch of every cluster;

6. Extract features of the whole unlabeled training corpus using the fine-tuned model obtained from Step 5.

7. Repeat Step 2 to 6 until the deep learning model is converged.

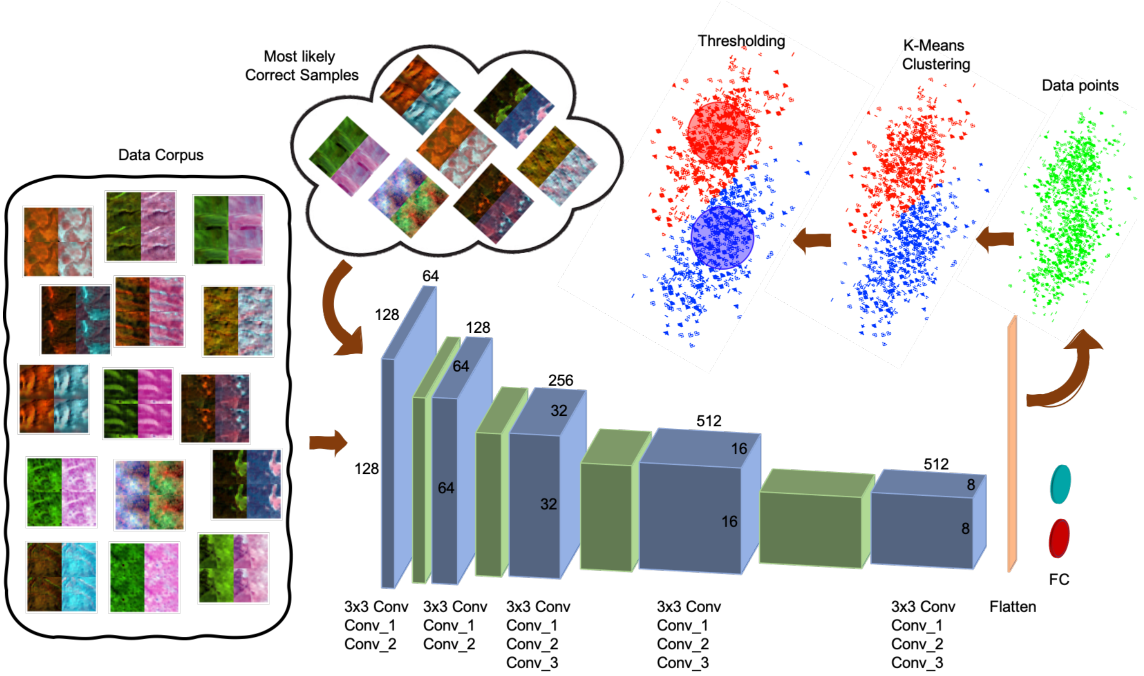


Figure 4: Shows the three prominent steps of the technique; 1) Deep learning model (VGG-16) to learn and extract the features, 2) Clustering to generate pseudo labels, and 3) Thresholding the samples present near the centroids of the clusters and declaring them as reliable samples.

For Step 1, the adopted base model is the VGG16 model pre-trained on a large-scale ImageNet dataset which is employed for extracting features from the training input image patches. The output of the last convolutional layer is extracted to get feature maps of each sample in the dataset. The extracted features are flattened to get the feature vectors. To cluster these feature representations, a well-known unsupervised k-means clustering algorithm is adopted. The input layer of the VGG-16 is adapted according to remote sensing image patch size and the output layer to the number of clusters that are generated. If we suppose that the features extracted from the training image patches $\{x_i\}_{i=1}^{N}$ are represented by $\{f_i\}_{i=1}^{N}$, then in Step 2, these features are clustered using k-means objective function: $\{y_i\}_{i=1}^{N} \leftarrow \min \sum_{i=1}^{N} \sum_{k=1}^{2} |f_i - c_k|$ where each feature vector is assigned a cluster label $\{y_i\}_{i=1}^{N}$ on the basis of its minimum distance from the particular centroid c_k, where c is the centroid of the kth cluster. In the current scenario, we have set the value of k to be two so that all the image patches are clustered into two groups, namely burnt forest and the other category. In Step 3 and 4, the obtained clustering IDs are used to assign the pseudo labels that are later used for fine-tuning the CNN model.

Since the employed VGG16 model use model weights that have been trained on a completely irrelevant (natural) dataset, therefore the obtained clusters are quite noisy and cannot be directly used to fine-tune remote sensing images to recognize burnt forest image patches. To cope with this issue, the obtained clusters are passed through a filtering mechanism to prune individual clusters. For this purpose, only those features

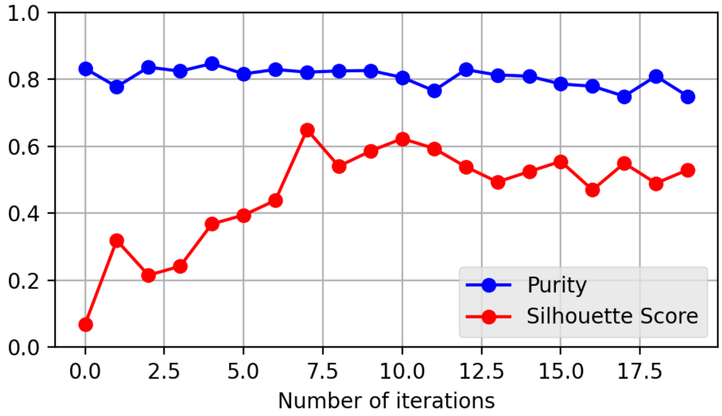


Figure 5: Graph showing the purity and Silhouette Score of generated clusters for burnt forest and other regions over every iteration of fine-tuning of VGG16.

are retained whose distance to the centroid of the pixel is less than a certain threshold. The refinement process keeps only the feature points near the cluster centroids and thus restricts the CNN to learn only the prominent features and avoid unnecessary noisiness. The image patches belonging to the refined clusters are then subsequently used to retrain the whole network in an unsupervised manner, i.e., with the cluster IDs as pseudo labels, in Step 5. In the next iteration, the updated (fine-tuned) model is used to extract the features from the image patches. With every iteration, the model learns the image features using the pseudo labels of clusters resulting in comparatively better clusters than the previous iteration. The process thus iterates until the loss of the deep model is converged. See Figure 4 for model visualization.

3 Results and Discussion

3.1 Clustering

Initially, the clusters are generated out of features extracted from a pre-trained deep learning model. The model is trained on an irrelevant domain (ImageNet) which does not contain the satellite imagery, more specifically, burnt forests in Sentinel-2 Imagery. As a result, the clusters are not compact and loosely packed for our input of Sentinel-2 imagery. To evaluate the compactness of the clusters, purity and the Silhouette Score are computed (see Figure 5). Purity is a supervised measure that calculates the ratio of correctly classified samples to the total number of samples for all clusters. Silhouette Score is an unsupervised measure that calculates the ratio on the basis of the distance between each sample within-cluster and the neighboring clusters.

It can be seen in the graph that, in the beginning, the Silhouette Score is a small number, which is 0.07, indicating the lack of compactness in the clusters. With every iteration of fine-tuning of the model, the compactness in the clusters increases (at max to

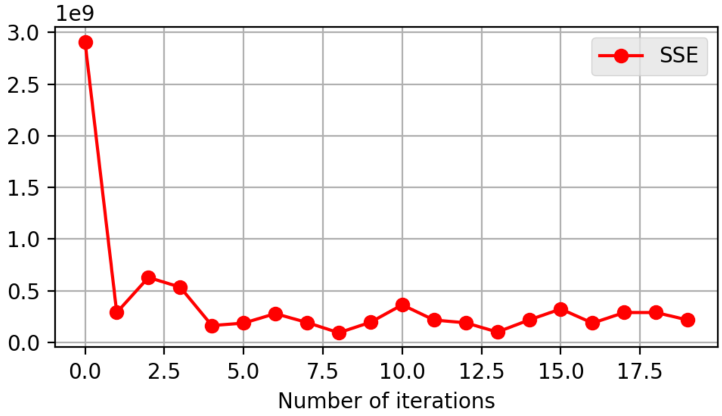


Figure 6: Graph showing Sum of Squared Error of generated clusters for burnt forest and other regions over every iteration of fine-tuning of VGG16.

0.64 at the 7th iteration). After a few iterations, the compactness is saturated. Whereas purity remains consistent throughout between the interval (0.75 - 0.85), indicating the ratio of correctly classified samples.

For further analysis, the Sum of Squared Error (SSE) is calculated, which is also the objective function of K-Means clustering. It can be seen in Figure 6. Initially, the value of SSE was quite high when clustering was done using pre-trained VGG16. As soon as the model is fine-tuned on a few images of Sentinel-2, the SSE decreased significantly by the one-degree exponent. After that it remains consistent at mean 3.9xe^8.

As the clusters are loosely packed in the beginning, it is better to use those samples present near the centroids of the clusters for fine-tuning of VGG16. It restricts the model from learning random features and ignores the noisy samples from the clusters. To do so, the dot product is used to find the similarity of every sample within a cluster with its respective centroid. Its value ranges from 0 to 1. If the dot product is greater than or equal to a pre-defined threshold, it is counted as the sample present near the centroid and declared as a reliable sample.

It can be seen in Figure 7 that at the start, only a few, i.e., 22 reliable samples, are extracted using the pre-trained VGG16. It indicates that the clusters are loosely packed. As the VGG-16 gets fine-tuned iteratively, the count of a reliable sample grows. The growth of reliable samples gets saturated after some iterations. After the 7th iteration, the graph remains quite consistent, with a count close to 8,000. It indicates that the clusters contain the majority of the samples from the training corpus of 9,000 samples as a reliable set and declaring only a small fraction of about 1,000 as the noisy ones.

Considering the purity, Silhouette Score and SSE measures, and count of reliable samples, it can be seen that the clusters generated at the 7th and 10th iteration are the best ones. We have fine-tuned the model for 20 interactions. After this, no more improvement in the clustering and fine-tuning is observed.

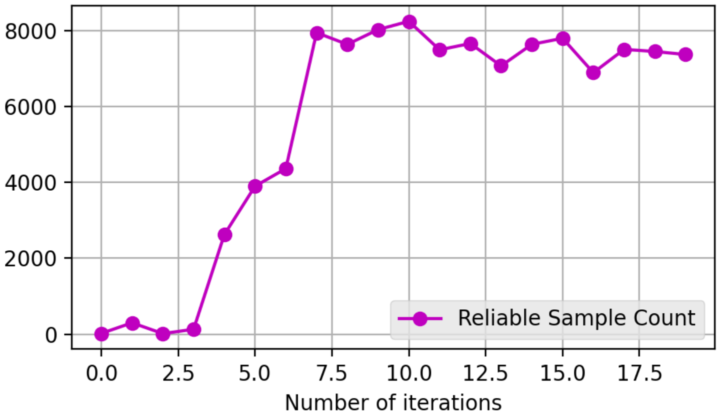


Figure 7: Graph showing the count of reliable samples over every iteration of fine-tuning of VGG16.

3.2 Fine-tuning VGG16

VGG16 is a supervised deep learning architecture that requires labels along with images to train the model. The pseudo labels of generated clusters are used to train VGG16. It can be seen in Figure 8 that the model is reporting almost a very small cross-validation loss on every iteration of fine-tuning VGG-16, considering the pseudo labels as the labels of the patches. It shows that the model is effectively learning the generated labels of the

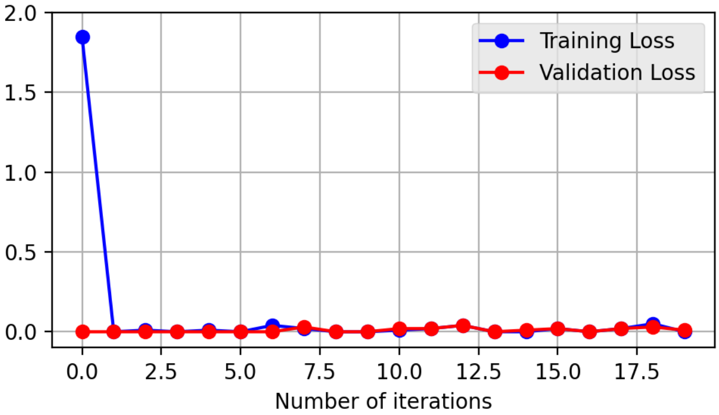


Figure 8: Graph showing the training and cross-validation loss with pseudo labels over every iteration of fine-tuning of VGG16.

patches. The model is fine-tuned end-to-end till the classification layer. This fine-tuned model is used for feature extraction in the next iteration, where the last max-pooling layer's output is used to generate the feature vectors for the clustering step.

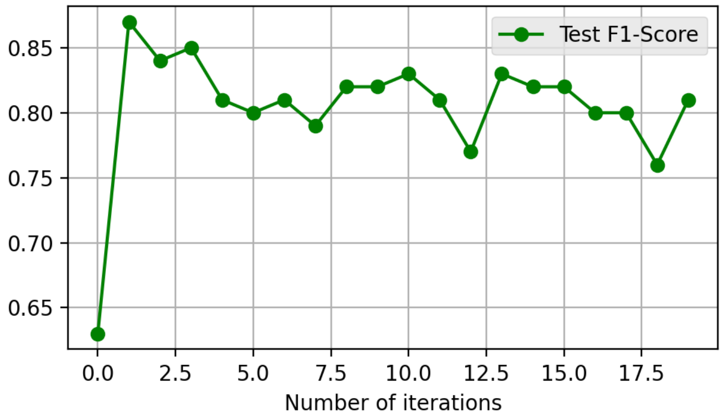


Figure 9: Graph showing F1-Score on the test corpus over every iteration of fine-tuning of VGG16.

The fine-tuned models generated over 20 iterations are evaluated on a real-time dataset. Figure 9 shows the calculated F1-Score. It can be seen that the model at 1st, 2nd, 3rd, 10th, and 13th gave the top 5 F1-Scores. Considering the top 5 results on test corpus over the 20 iterations, a mean of precision, recall, F1-Score, and accuracy are reported in Table 2.

	Precision	**Recall**	**F1-Score**
Class 0	0.86	0.83	0.84
Class 1	0.83	0.86	0.85
Accuracy	-	-	0.85
Macro Avg	0.85	0.85	0.85
Weighted Avg	0.85	0.85	0.85

Table 2: The table shows the average precision, recall, F1-Score and accuracy for best 5 iterations on the test corpus.

3.3 Analysis on Sentinel-2 Imagery

We have considered the median Sentinel-2 imagery of three months (Feb 2020 - Apr 2020) for the region, Australian Capital Territory, and South of it, see Figure 10-(a). This area is the worst affected region by the massive wildfire. The top 5 iterations of fine-tuning the model reporting the highest performance on test corpus were deployed to analyze their performance on the considered region. The results can be seen in Figure 10-(b-f). Image (b) of the figure shows the result of 1st fine-tune iteration, reporting the highest accuracy of 0.87 on the test corpus. Similarly, (c) shows the result of the 3rd iteration reporting

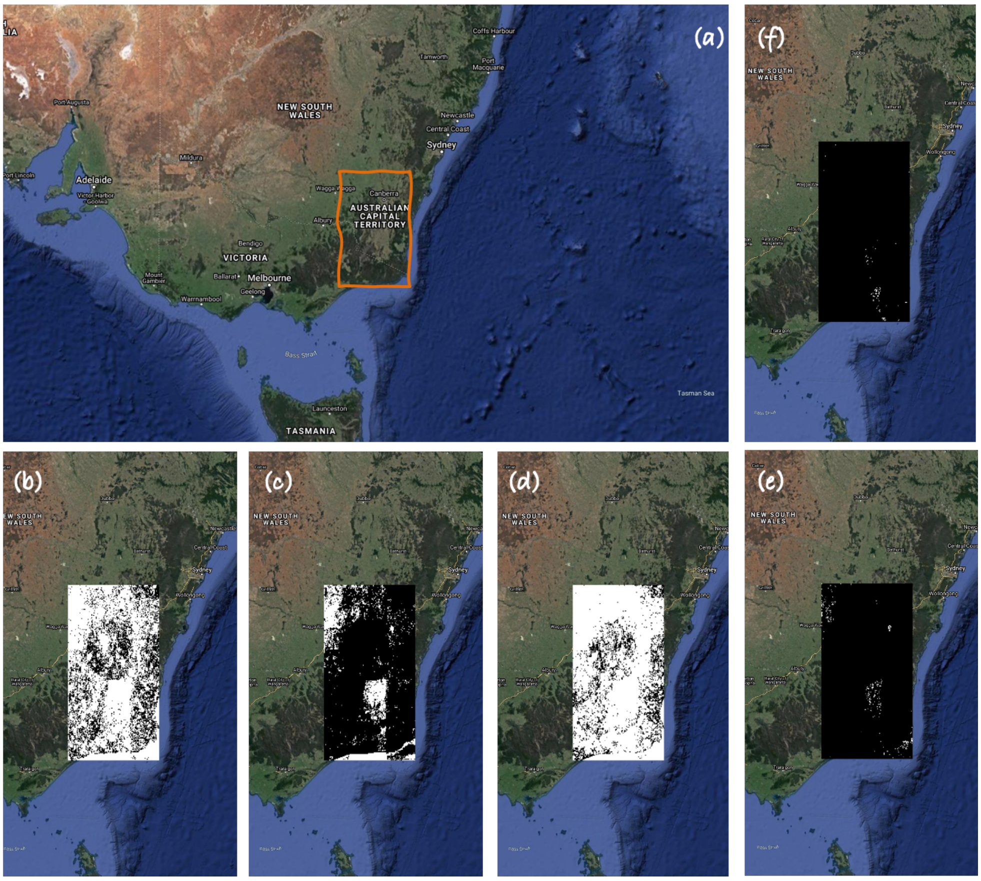


Figure 10: **(a)** Shows the region considered, covering Australian Capital Territory and South of it in an orange polygon, for testing the fine-tuned models on Sentinel-2 median image of three months (Feb 2020 - Apr 2020). **(b)** Shows the prediction results for the fine-tune iteration of the deep learning model, reporting the highest accuracy on test corpus. **(c)** Shows the prediction results for iteration reporting the 2nd highest accuracy on test corpus. **(d)** Shows the prediction results for iteration reporting the 3rd highest accuracy on test corpus.. **(e)** Shows the prediction results for iteration reporting the 4th highest accuracy on test corpus. **(f)** Shows the prediction results for iteration reporting the 5th highest accuracy on test corpus.

the 0.85 of accuracy, (d) shows the result of 2nd iteration reporting 0.84 of accuracy, and (e-f) results of 10th and 13th iterations, each reporting 0.83 of accuracy. The accuracy on the test corpus for 1st, 2nd, and 3rd is comparatively higher, but the generated clusters are loosely packed compared to later iterations of fine-tuning, and the count of reliable samples is less than 50 (see sub-session 3.1). Considering the compactness and good count of reliable samples, the model at iteration 10 and 13 generate better clusters but performance decreases on the test set. This might be because of multiple iterations of fine-tuning leading the models towards overfitting.

4 Conclusions

In this paper, we have proposed an unsupervised deep learning technique for mapping the burnt regions of Australia. The method is capable of learning the features progressively from the data without expert knowledge. The proposed solution provides the advantages of supervised deep learning models along with removing the tedious step of data labeling. We are able to achieve the F1-Score of 0.85 with the progressive learning behavior of the model in an unsupervised manner. The real-time Sentinel-2 Imagery is used for training the deep learning architecture and mapping the burnt region of Australia. The method can be applied without any modifications to estimate the burnt forest region in any other region of the world due to its unsupervised nature.

References

[1] S. J. Pyne, *World fire: the culture of fire on earth.* University of Washington press, 1997.

[2] D. Bowman, G. Williamson, M. Yebra, J. Lizundia-Loiola, M. L. Pettinari, S. Shah, R. Bradstock, and E. Chuvieco, “Wildfires: Australia needs national monitoring agency,” 2020.

[3] M. M. Boer, V. R. de Dios, and R. A. Bradstock, “Unprecedented burn area of australian mega forest fires,” *Nature Climate Change*, vol. 10, no. 3, pp. 171–172, 2020.

[4] A. Singh, “Case study on 2019 australian bushfire,” 04 2020.

[5] G. R. Van Der Werf, J. T. Randerson, L. Giglio, T. T. Van Leeuwen, Y. Chen, B. M. Rogers, M. Mu, M. J. Van Marle, D. C. Morton, G. J. Collatz *et al.*, “Global fire emissions estimates during 1997-2016,” *Earth System Science Data*, vol. 9, no. 2, pp. 697–720, 2017.

[6] M. C. Stambaugh, L. D. Hammer, and R. Godfrey, “Performance of burn-severity metrics and classification in oak woodlands and grasslands,” *Remote Sensing*, vol. 7, no. 8, pp. 10 501–10 522, 2015.

[7] E. Chuvieco, F. Mouillot, G. R. van der Werf, J. San Miguel, M. Tanase, N. Koutsias, M. García, M. Yebra, M. Padilla, I. Gitas *et al.*, “Historical background and current developments for mapping burned area from satellite Earth observation,” *Remote Sensing of Environment*, 2019.

[8] C. H. Key and N. C. Benson, “Measuring and remote sensing of burn severity,” in *Proceedings joint fire science conference and workshop*, vol. 2. University of Idaho and International Association of Wildland Fire Moscow, ID, 1999, p. 284.

[9] S. Trigg and S. Flasse, "An evaluation of different bi-spectral spaces for discriminating burned shrub-savannah," *International Journal of Remote Sensing*, vol. 22, no. 13, pp. 2641–2647, 2001.

[10] M. Martín, I. Gómez, and E. Chuvieco, "Performance of a burned-area index (BAIM) for mapping Mediterranean burned scars from MODIS data," in *Proceedings of the 5th International Workshop on Remote Sensing and GIS Applications to forest fire management: fire effects assessment.* Universidad de Zaragoza, GOFC GOLD, EARSeL, 2005.

[11] L. Giglio, L. Boschetti, D. P. Roy, M. L. Humber, and C. O. Justice, "The Collection 6 MODIS burned area mapping algorithm and product," *Remote sensing of environment*, vol. 217, pp. 72–85, 2018.

[12] E. Chuvieco, J. Lizundia-Loiola, M. L. Pettinari, R. Ramo, M. Padilla, K. Tansey, F. Mouillot, P. Laurent, T. Storm, A. Heil *et al.*, "Generation and analysis of a new global burned area product based on MODIS 250 m reflectance bands and thermal anomalies," *Earth System Science Data*, vol. 10, no. 4, pp. 2015–2031, 2018.

[13] A. Bastarrika, E. Chuvieco, and M. P. Martin, "Mapping burned areas from Landsat TM/ETM+ data with a two-phase algorithm: Balancing omission and commission errors," *Remote Sensing of Environment*, vol. 115, no. 4, pp. 1003–1012, 2011.

[14] D. Stroppiana, G. Bordogna, P. Carrara, M. Boschetti, L. Boschetti, and P. Brivio, "A method for extracting burned areas from Landsat TM/ETM+ images by soft aggregation of multiple Spectral Indices and a region growing algorithm," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 69, pp. 88–102, 2012.

[15] I. Alonso-Canas and E. Chuvieco, "Global burned area mapping from ENVISAT-MERIS and MODIS active fire data," *Remote Sensing of Environment*, vol. 163, pp. 140–152, 2015.

[16] A. A. Pereira, J. Pereira, R. Libonati, D. Oom, A. W. Setzer, F. Morelli, F. Machado-Silva, and L. M. T. De Carvalho, "Burned area mapping in the Brazilian Savanna using a one-class support vector machine trained by active fires," *Remote Sensing*, vol. 9, no. 11, p. 1161, 2017.

[17] F. Mouillot, M. G. Schultz, C. Yue, P. Cadule, K. Tansey, P. Ciais, and E. Chuvieco, "Ten years of global burned area products from spaceborne remote sensing—a review: Analysis of user needs and recommendations for future developments," *International Journal of Applied Earth Observation and Geoinformation*, vol. 26, 2014.

[18] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *International conference on artificial neural networks.* Springer, 2018, pp. 270–279.

[19] G. Hinton, Y. LeCun, and Y. Bengio, "Deep learning," *Nature*, 2015.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[21] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 96–107, 2018.

[22] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. Dalla Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 139–149, 2017.

[23] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 1–9.

[24] F. Zhang, J. Li, B. Zhang, Q. Shen, H. Ye, S. Wang, and Z. Lu, "A simple automated dynamic threshold extraction method for the classification of large water bodies from landsat-8 OLI water index images," *International Journal of Remote Sensing*, vol. 39, 2018.

[25] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote sensing of environment*, vol. 229, pp. 247–259, 2019.

[26] K. He, X. Zhang, S. Ren, and J. Sun, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

[27] M. Rußwurm and M. Körner, "Multi-temporal land cover classification with sequential recurrent encoders," *ISPRS International Journal of Geo-Information*, vol. 7, no. 4, p. 129, 2018.

[28] C. Pelletier, G. I. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote Sensing*, vol. 11, no. 5, p. 523, 2019.

[29] A. Zulfiqar, M. M. Ghaffar, M. Shahzad, C. Weis, M. I. Malik, F. Shafait, and N. Wehn, "Ai-forestwatch: semantic segmentation based end-to-end framework for forest estimation and change detection using multi-spectral remote sensing imagery," *Journal of Applied Remote Sensing*, vol. 15, 2021.

[30] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais *et al.*, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, pp. 195–204, 2019.

[31] M. M. Pinto, R. Libonati, R. M. Trigo, I. F. Trigo, and C. C. DaCamara, "A deep learning approach for mapping and dating burned areas using temporal sequences of satellite images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 160, pp. 260–274, 2020.

[32] L. Knopp, M. Wieland, M. Rättich, and S. Martinis, "A Deep Learning Approach for Burned Area Segmentation with Sentinel-2 Data," *Remote Sensing*, vol. 12, no. 15, p. 2422, 2020.

[33] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ACM International Conference Proceeding Series*, vol. 382, 2009.

[34] A. Ul-Hasan, F. Shafaity, and M. Liwicki, "Curriculum learning for printed text line recognition of ligature-based scripts," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2015.

[35] "Fires in victoria destroy estimated 300 homes, former police chief to lead bushfire recovery victoria - abc news," https://www.abc.net.au/news/2020-01-06/bushfires-in-victoria-destroy-at-least-200-homes/11844292, Januray 2020.

[36] "Australia bushfires continue to wreak havoc – the state times," https://thestatetimes.com/2020/02/07/australia-bushfires-continue-to-wreak-havoc/, February 2020.

[37] "Australia fires: A visual guide to the bushfire crisis - bbc news," https://www.bbc.com/news/world-australia-50951043, January 2020.

[38] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[39] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 4, 2018. [Online]. Available: https://doi.org/10.1145/3243316

[40] R. M. S. Bashir, M. Shahzad, and M. M. Fraz, "VR-PROUD: Vehicle Re-identification using PROgressive Unsupervised Deep architecture," *Pattern Recognition*, vol. 90, pp. 52–65, 2019.

Acronyms

AWEI Automated Water Extraction Index. 29

BAIM Modified Burned Area Index. 6, 29

CL Curriculum Learning. 8, 12, 14

CNN Convolutional Neural Networks. 8, 12–15, 23, 27

DL Deep Learning. v, 8, 11–15, 23, 28, 29

DT Decision Tree. 6

EO Earth Observation. v, 3, 5, 6, 29–31

ESA European Space Agency. 6

FCM Fuzzy C-Means. 15

GAN Generative Adversarial Neural network. 8

GSD Ground Sample Distance. 5

LiDAR Light Detection and Ranging. 3

LSTM Long-Short Term Memorys. 8

MIRBI Mid-InfraRed Burn Index. 6

MNDWI Modified Normalized Difference Water Index. 6

NASA Aeronautics and Space Administration. 4, 6

NBR Normalized Burned Ratio. 6, 28

NDVI Normalized Difference Vegetation Index. 6

NDWI Normalized Difference Water Index. 6

NIR Near Infrared. 5–7

NLP Natural Language Processing. v, 31

RADAR Radio Detection and Ranging. 3

RF Random Forest. 6

RS Remote Sensing. 3, 5, 6, 8, 9, 12, 13, 17, 20, 21, 23

SGD Stochastic Gradient Descent. 16

SSE Sum of Squared Error. 17, 22

SVM Support Vector Machine. 6

SWIR Short Wavelength Infrared. 6, 7

TIRS Thermal Infrared. 7

UAV Unmanned Aerial Vehicles. 3, 5, 17

UCL Unsupervised Curriculum Learning. v, 9, 14–17, 19–24, 27, 30, 31

WWF World Wildlife Fund for Nature. 4, 5