



## Article

# Spectral Analysis to Improve Inputs to Random Forest and other Boosted Ensemble Tree-Based Algorithms for Detecting NYF Pegmatites in Tysfjord, Norway

Douglas Santos <sup>1</sup>, Joana Cardoso-Fernandes <sup>1,2</sup>, Alexandre Lima <sup>1,2</sup>, Axel Müller <sup>3,4</sup>, Marco Brönnner <sup>5</sup> and Ana Cláudia Teodoro <sup>1,2,\*</sup>

<sup>1</sup> Department of Geosciences, Environment and Spatial Planning, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal; douglas.santos@fc.up.pt (D.S.); joana.fernandes@fc.up.pt (J.C.-F.); allima@fc.up.pt (A.L.)

<sup>2</sup> Institute of Earth Sciences, FCUP Pole, 4169-007 Porto, Portugal

<sup>3</sup> Natural History Museum, University of Oslo, 0318 Oslo, Norway; a.b.muller@nhm.uio.no

<sup>4</sup> Natural History Museum, London SW7 5BD, UK

<sup>5</sup> Geological Survey of Norway (NGU), 7040 Trondheim, Norway; marco.bronner@ngu.no

\* Correspondence: amteodor@fc.up.pt

**Abstract:** As an important source of lithium and rare earth elements (REE) and other critical elements, pegmatites are of great strategic economic interest for present and future technological development. Identifying new pegmatite deposits is a strategy adopted by the European Union (EU) to decrease its import dependence on non-European countries for these raw materials. It is in this context that the GREENPEG project was established, an EU project whose main objective is to identify new deposits of pegmatites in Europe in an environmentally friendly way. Remote sensing is a non-contact exploration tool that allows for identifying areas of interest for exploration at the early stage of exploration campaigns. Several RS methods have been developed to identify Li-Cs-Ta (LCT) pegmatites, but in this study, a new methodology was developed to detect Nb-Y-F (NYF) pegmatites in the Tysfjord area in Norway. This methodology is based on spectral analysis to select bands of the Sentinel 2 satellite and adapt RS methods, such as Band Ratios and Principal Component Analysis (PCA), to be used as input in the Random Forest (RF) and other tree-based ensemble algorithms to improve the classification accuracy. The results obtained are encouraging, and the algorithm was able to successfully identify the pegmatite areas already known and new locations of interest for exploration were also defined.

**Keywords:** rare earth elements; critical raw materials; band ratios; principal component analysis; remote sensing; Sentinel 2; exploration; LightGBM; CatBoost



**Citation:** Santos, D.;

Cardoso-Fernandes, J.; Lima, A.; Müller, A.; Brönnner, M.; Teodoro, A.C. Spectral Analysis to Improve Inputs to Random Forest and other Boosted Ensemble Tree-Based Algorithms for Detecting NYF Pegmatites in Tysfjord, Norway. *Remote Sens.* **2022**, *14*, 3532. <https://doi.org/10.3390/rs14153532>

Academic Editor: Amin Beiranvand Pour

Received: 9 June 2022

Accepted: 21 July 2022

Published: 23 July 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In a world that is increasingly seeking to decarbonise its industrial sectors, reduce carbon emissions into the atmosphere and diminish the impact of human activity on nature, there is a growing need for new technologies that respond to this “green” demand and allow for countries to develop without neglecting their environmental and social objectives [1–3]. With the increase in demand for these “green” technologies also comes the demand for raw materials needed for their production and maintenance. Among these commodities are the critical raw minerals (CRM), whose demand is dramatically rising; thus, the need to locate new deposits of these minerals is a major societal task [4]. Granitic pegmatites are an important source of CRM, such as lithium, tantalum, niobium, beryllium, cesium, uranium, and REE, among others [5]. This study occurs in the scope of the Horizon2020 GREENPEG project, an EU project, which has 13 partners, among academic and private sector institutions, in more than six EU countries (<https://www.greenpeg.eu/>,

accessed on 16 July 2021). This project aims to develop innovative, competitive, and more environmentally friendly exploration tools to locate and expand the reserves of pegmatites of the Lithium-Caesium-Tantalum (LCT) and Niobium-Yttrium-Fluorine (NYF) families in Europe [4,5]. Remote sensing is one tool of exploration campaigns developing methodologies for the identification of new possible pegmatite deposits. For this purpose, several image processing techniques were employed, such as RGB combinations, Band Ratios (BR), Principal Component Analysis (PCA), and two machine learning algorithms (MLs)—Support Vector Machine (SVM) and Random Forest (RF) [6,7].

Remote sensing has been applied in the exploration field since the 1970s [8]. Some examples of its application are in the identification of uranium [9], gold deposits [10], REE [11], and hydrothermal alteration zones [12]. Concerning pegmatite exploration, we can cite the work of Cardoso-Fernandes et al. [13] who developed classical RS methods focused on studies of LCT-type pegmatites since 2018. RF and SVM were already used in previous studies by [14] to identify Li-bearing pegmatites in Spain and Portugal and, more recently, to identify NYF pegmatites in Tysfjord in Norway [15]. Hyperspectral data were also applied for mineral mapping of Li-bearing pegmatites at Uis, Namibia [16]. Satellite multispectral data have been applied to detect other types of Li-deposits, namely Li brines in the Salar de Uyuni, Bolivia [17]. Pegmatite exploration can also be complemented by other methods such as soil or stream geochemistry [18,19].

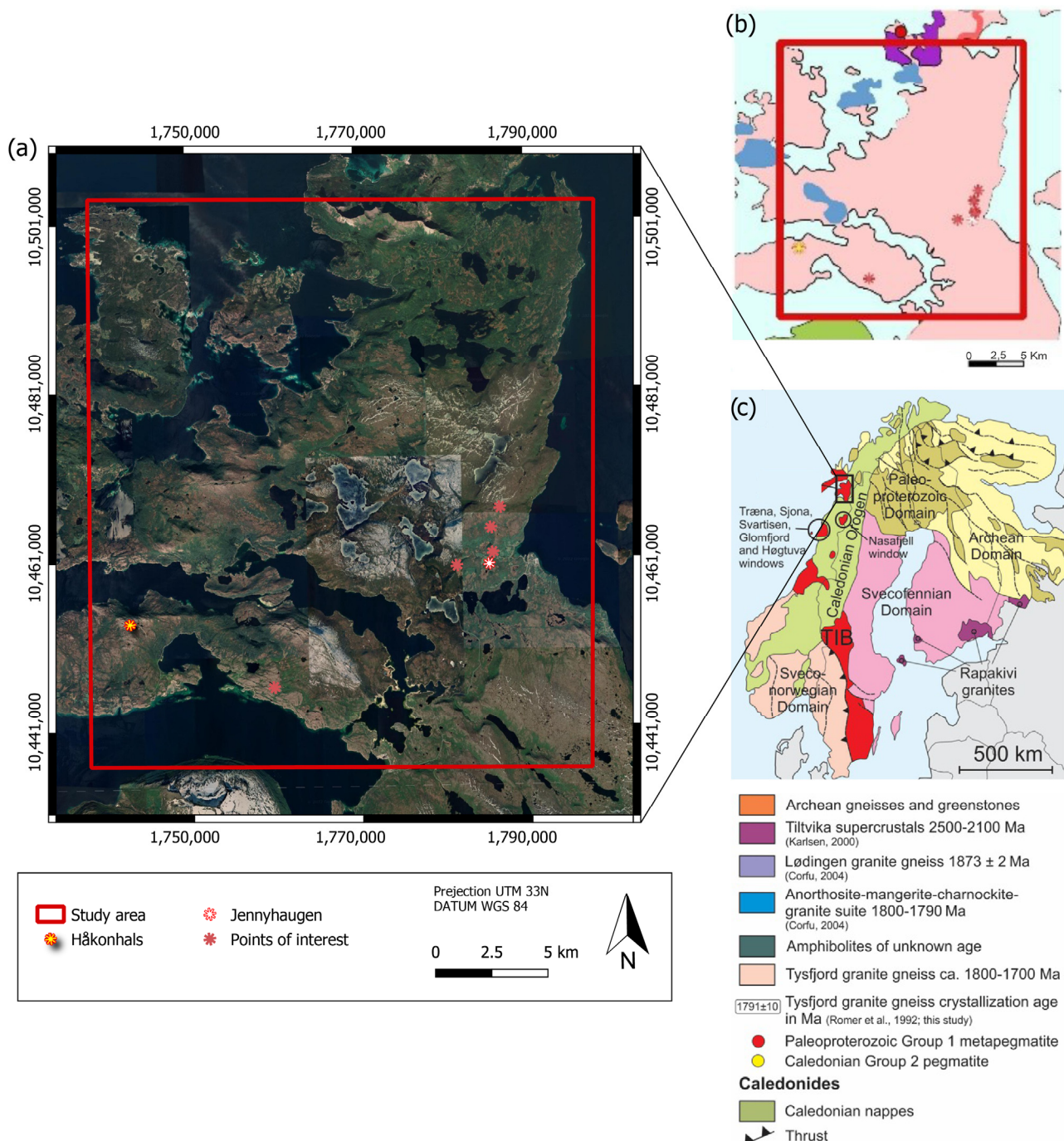
It should be considered that the bands used as inputs in the previous works for the various image processing methods were selected to identify LCT pegmatites. In the Tysfjord area, however, NYF pegmatites occur, and the methodology proposed in this study aims to use spectral analysis (spectroscopy) to select the most appropriate parameters to be used as inputs features for RF and other ensemble tree-based algorithms, thus improving the results obtained in its classification. Besides selecting the most suitable bands according to the spectral response of the NYF pegmatites of Tysfjord, spectroscopy was also used to adapt two image processing methods already known in the literature [7,8,15] and considered “traditional methods” for geological purposes (BR and PCA). These “traditional methods” were used not only for area recognition, but also as inputs for the MLs together with vegetation indexes (Normalised Difference Vegetation Index—NDVI) to improve the sensitivity of the algorithms. The spectral analysis was divided into two steps. The first step focused on spectra extracted directly from the Sentinel 2 bands and the second step was based on spectra collected using a spectroradiometer.

At the end of this study, areas of interest for exploration were evaluated, with new areas identified. The results obtained represent a major contribution not only to remote sensing exploration in the case study of Tysfjord pegmatites, but also in the field of pegmatite exploration in general. Since this method is based on spectral analyses it can be applied in other areas as well. In the future, the application of these new methodological remote sensing approaches to other areas with NYF pegmatites may validate their efficiency in other regions. Although remote sensing only detects surface pegmatite deposits, the GREENPEG project is working on a multi-method toolset to detect sub-surface pegmatites, which includes geophysical and geochemical methods, is categorised according to their respective penetration depth (see Figures 6 and 7 of [5]).

### 1.1. Study Area

The study area is in the Tysfjord district (Figure 1), an area of about 559 km<sup>2</sup> that is located in Northern Norway. The Tysfjord-Hamarøy pegmatitic field includes several NYF pegmatites that are the target of this study [5,20,21]. The shape of the Tysfjord pegmatites ranges from lens-shaped to cigar-shaped pegmatites depending on the length of up to 400 m [5]. First recognised in 1941 by Foslie [22], these granitic pegmatites are known for their great mineral diversity and are genetically linked to the Tysfjord granites [21], which are the host rocks of the pegmatites. Among the 157 identified minerals, the most common accessory minerals are allanite-(Ce), fergusonite-(Y), columbite-(Fe), beryl, various sulphides, and fluorite, besides the major minerals quartz, plagioclase, K-feldspar (variety

‘amazonite’), and biotite [21]. The most important occurrences in the Tysfjord-Hamarøy pegmatite field include: the Jennyhaugen pegmatite, the Nedre Øyvollen pegmatite, and the Håkonhals pegmatite. The Jennyhaugen and Håkonhals mines were chosen as the target of this study as they are large open-pit mines and are not limited by the spatial resolution of Sentinel 2, which has 10 m of spatial resolution.



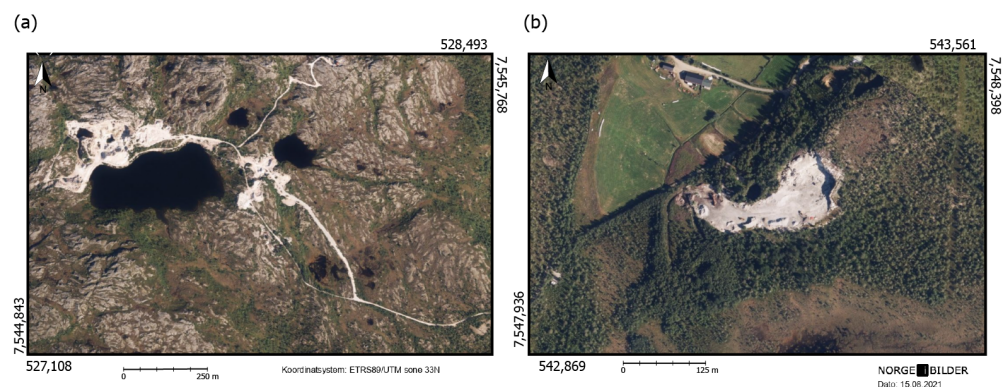
**Figure 1.** Location of the study area. (a) Overview image of the Tysfjord district where the study area is framed by the red square. The Håkonhals pegmatite is marked with a yellow star, Jennyhaugen mine with a white star, and other pegmatites with red stars. (b) Simplified geologic map. (c) Detailed geological map and location of the study area in Norway. TIB: Trans-Scandinavian Igneous Belt. Adapted with permission from Müller et al. [4,5].

### 1.1.1. The Jennyhaugen Pegmatite

Located at EU89-UTM coordinate Zone 33V 543,313E/7,548,208N, the Jennyhaugen pegmatite has a width of up to 40 m and a length of at least 200 m [21]. When compared to other Tysfjord pegmatites, the Jennyhaugen pegmatite is considered relatively poor in accessory minerals compared to other Tysfjord pegmatites [21]. The minerals exposed in the open pits of Jennyhaugen and found in the mining dumps include the major minerals quartz, plagioclase, K-feldspar (variety ‘amazonite’), and biotite, as well as the accessory minerals garnet, allanite-(Ce), monazite-(Ce), zircon, fluorite, microlite, beryl, tantalite-(Mn), etc. [20,21].

### 1.1.2. The Håkonhals Pegmatite

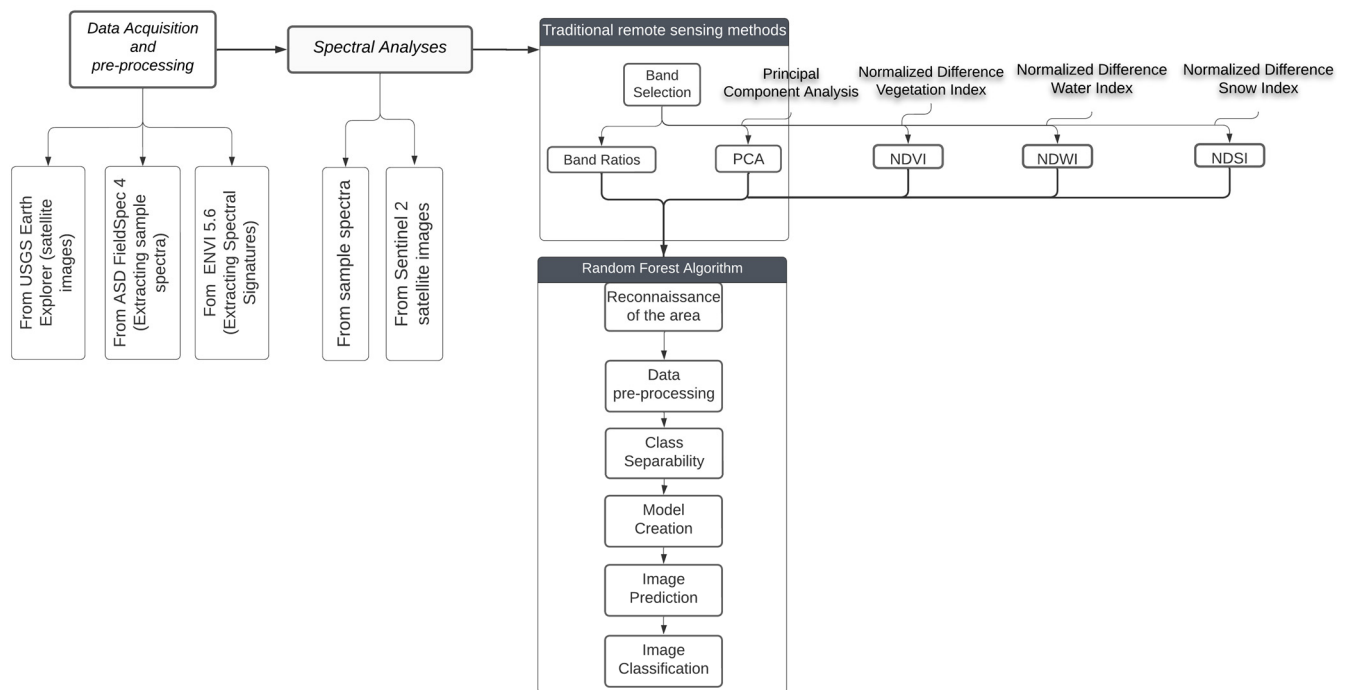
This pegmatite is exposed at coordinates EU89-UTM Zone 33V 527,376E/7,545,507N. The Håkonhals pegmatite is exposed in one large open-pit mine ( $150 \times 150 \text{ m}^2$ ) on its eastern side and two small open pits on its western side. In between, the pegmatite is covered by Tysfjord granite. The mine exposures make the pegmatite an ideal target for this study. Historically this pegmatite was mined for K-feldspar and more recently for high-purity quartz, quartz with less than 50 ppm of contaminating elements [20,22]. At least 28 minerals that can be found in the Håkonhals pegmatite were described by [21]. Among them are magnetite, allanite-(Ce), thorite, bastnäsite-(Ce), monazite-(Y), zircon, beryl, and fluorite. As large open-pit mines, the Håkonhals and Jennyhaugen mines (Figure 2), are ideal for remote sensing studies and in this paper will be referred to as target areas.



**Figure 2.** High-resolution aerial photographs of the (a) Håkonhals and (b) Jennyhaugen mines that were used as target areas in this study. Adapted with permission from Norge I bilder [23].

## 2. Materials and Methods

The methodology applied in this study is based on spectral analysis (spectroscopy) to improve the final result of the RF algorithm classification. As described by the authors of [20], who besides the RF algorithm also applied the SVM algorithm in the identification of NYF pegmatites, the RF was able to identify the locations of known pegmatites. However, this algorithm was not efficient in classifying the other elements of the study area (mainly water and vegetation), which implies that the classification accuracy, in general, can be improved. Overall, SVM classification performed better in Tysfjord when compared to RF classification. Through the spectral analysis of samples from Tysfjord and spectra extracted from all Sentinel 2 bands, the methodology of this study proposes to select the most appropriate bands and to adapt traditional image processing methods (BR and PCA) to be used as input into the algorithm. The workflow of the applied methodology is represented in Figure 3.



**Figure 3.** Workflow of the applied methodology to select and adapt the traditional remote sensing methods.

### 2.1. Data Acquisition and Pre-Processing

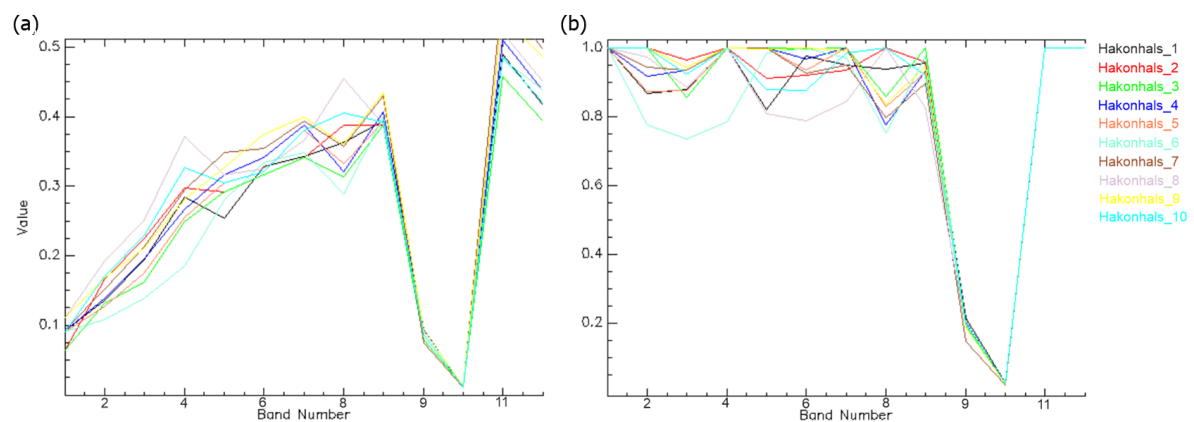
In the scope of the GREENPEG project, we work with three multispectral satellites (Landsat 8 OLI, TERRA ASTER, and Sentinel 2 MSI). Sentinel 2 has the best spatial resolution as well as temporal resolution, allowing for analysing and selecting the best image to apply RS methods in Tysfjord (an area with arctic weather conditions). The image was downloaded from the United States Geological Survey (USGS) website, and all downloaded images have a cloud cover of less than 10%. After downloading the images, the atmospheric correction Dark Object Subtraction (DOS1) algorithm was applied [24]. The atmospheric correction procedure was performed using the Semi-Automatic Classification Plugin (SCP) tool version 6.2.9 available in the QGIS software version 3.2.1. After the atmospheric correction, the Normalised Vegetation Index (NDVI) and Normalised Snow Index (NDSI) were applied to select the images with as little vegetation and ice coverage as possible. The selected image for this study was from 28 September 2019, an autumn image showing less snow cover than winter images and less vegetation when compared with summer and spring images.

The spectral data were obtained in two different ways: (i) Extracted from Sentinel 2 image, where 28 spectra were collected (10 from Håkonhals pegmatite outcrop and 8 from Jennyhaugen mine). The spectra were extracted and analysed in ENVI classic version 5.6 software. (ii) In total, 58 spectra were collected in the laboratory from 21 rock samples from Tysfjord. The spectra were collected with the Analytical Spectral Devices (ASD) FieldSpec 4 spectroradiometer a transportable battery-powered spectrometer with a spectral range of 350–2500 nm, a spectral resolution of 3 nm at 700 nm (VNIR), 10 nm at 1400 nm (SWIR 1), and 10 nm at 2100 nm (SWIR 2), with a scanning time of 100 milliseconds. The equipment was calibrated using a Spectralon plate with reflectance higher than 95% for the 250–2500 nm region and higher than 99% for the 400–1500 nm region [25].

### 2.2. Spectral Analysis

Spectral analysis has been used for several purposes in geology, such as studies involving REE [26], minerals in general [27], gold mining [9], and more recently the work of Cardoso-Fernandes [28] focused on the spectral analysis of Li minerals and pegmatites.

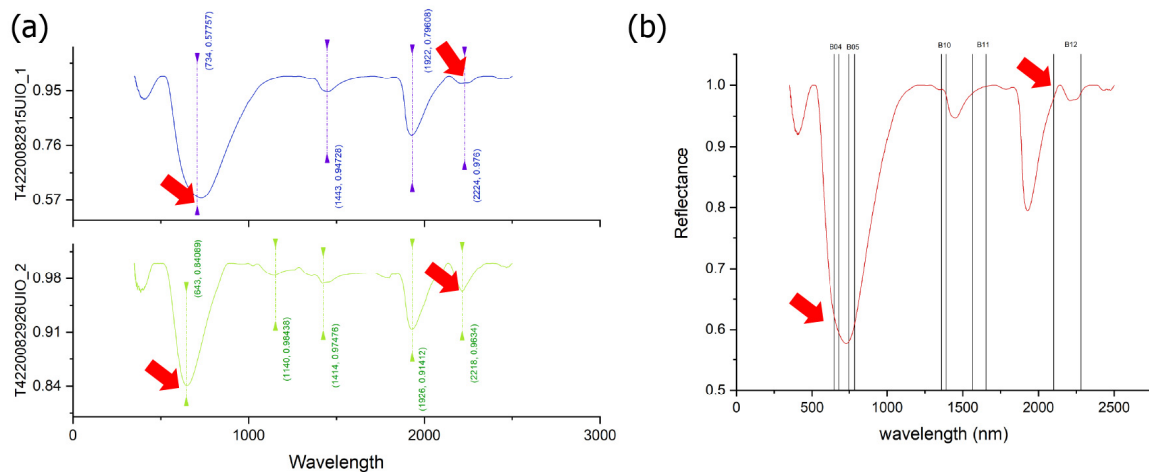
The spectra obtained from the Sentinel 2 satellite bands were extracted and analysed using the ENVI classic software version 5.6. The spectra collected from the Sentinel 2 bands were obtained directly over pixels that are in areas of pegmatite outcrop (these will be called pure spectra). The spectra were then analysed through continuum removal so that we could better evaluate their minimum absorption and reflectance peaks. Taking into account that in a Greenfield exploration area it may not be possible to distinguish pixels of pegmatites from other elements within the target area, averaging the spectra may be an alternative to pure pegmatite spectra. To validate this method, the spectra from the entire mine outlines were compared to the pure spectra and the reflectance analysed. As the spectra were extracted directly from the Sentinel 2 bands, they were provided with information on the band numbers on the X-axis, allowing us to compare the absorption and reflectance zones with the spectral range of the Sentinel 2 bands. The continuum removed spectra were analysed using the Spectral Library Viewer tool (Figure 4), and the bands in the main zones of absorption and reflectance of each spectrum were identified and pointed.



**Figure 4.** Overall comparison between (a) raw and (b) continuum-removed spectra of the Håkonhals mine.

The data collected in the laboratory using the spectrometer were received through the ASD Indico Pro application, an application developed by ASD Inc. to receive and store the spectral data transmitted from the ASD spectrometer [29]. To improve the signal-to-noise ratio, each spectrum collected represents an average of 40 scans [28]. Five spectra were collected for each spot, and an average of these five spectra was made to acquire the final spectra. At least two spectra were collected from each sample, which were arranged according to their mineral composition. The spectra were pre-processed using SpectraGryph software version 1.6 (Friedrich Menges (Oberstdorf, Germany)). The spectra were processed and had their continuum removed in Python programming language using the pysptools library.

In total, 58 spectra were analysed, of which 5 correspond to ‘amazonite’—the green K-feldspar variety, 13 to pink K-feldspar, 9 to plagioclase, 16 to massive quartz from the pegmatite core zone, 6 to Tysfjord granite, 4 to biotite, and 5 were from wall zone of pegmatites, which is a mineral mixture of plagioclase, quartz, biotite, and K-feldspar. The spectra were exported to .txt format and imported to Envi Spectral Library Viewer tools, where they were analysed and one spectrum from each sample was selected as representative, resulting in 21 spectra selected for analysis. The next step was to compare the region of absorptions and reflectance with the spectral range of the Sentinel 2 bands (Figure 5). After this analysis, new bands were selected and assigned in the methods described in the following steps.



**Figure 5.** Spectra of amazonite samples from Tysfjord. (a) It is possible to see—indicated by red arrows—the main absorptions of the spectra collected in the laboratory. (b) Comparison of the absorption and reflectance features—indicated by red arrows—with the spectral range of the Sentinel 2 satellite bands.

### 2.3. Band Ratios

BR is one of the most widely used image processing methods for lithological purposes [30–32] and consists of dividing bands with high reflectance by bands with high absorption to highlight specific spectral differences [33]. The BR tested in this study are presented in Table 1, for BRs derived from spectra collected in the laboratory, and in Table 2, for BRs derived from spectra extracted from Sentinel 2 bands.

**Table 1.** BR developed through spectral analysis of spectra collected in the laboratory. The spectra were organised according to mineral samples.

Granite	Plagioclase	Biotite	Quartz
3/6	8A/4	8/5	3/4
12/6	8A/11	xx	3/12
8/6	xx	xx	8/12

**Table 2.** BR developed through spectral analysis of spectra extracted from Sentinel 2 bands. The spectra were organised according to reflectance bands.

Band 4	Band 6	Band 7	Band 8A
4/3	6/3	7/3	8A/3
4/5	6/5	7/5	8A/5
4/8	6/8	7/8	8A/8

### 2.4. Principal Components Analysis

According to Singh and Harrison [34], PCA is a multivariate statistical technique used to enhance and separate certain types of spectral signatures from the background. Over the years, it has been applied in several areas and is now used, for instance, for image enhancement, digital change detection, and determining the underlying statistical dimension of the image data set [34,35]. The formula that expresses the principal components is presented in Equation (1).

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \dots + a_{nj}X_n = a_j^T X \quad (1)$$

where  $T$  denotes the transpose of a matrix and  $a_j^T = [a_{1j}, \dots, a_{nj}]$  are the normalised eigenvectors [i.e.,  $a_j^T a_j = 1$ ] of the variance-covariance matrix [34].

Considering the results of previous studies that have used PCA to identify pegmatites or their minerals of interest [31,32], selective PCA was applied to only two bands' subsets (Table 3).

**Table 3.** PCA on two bands developed through spectral analysis of spectra collected in the laboratory and extracted from Sentinel 2 bands.

PC2 Laboratory Spectra	PC2 Extracted from Sentinel 2 Bands
3 and 6	4 and 8
12 and 6	4 and 5

### 2.5. Random Forest and Other Ensemble Tree-Based Algorithms

As a very robust algorithm that can be used for supervised classification, the RF algorithm consists of an ensemble tree type classifier, capable of improving classification accuracy by combining various decision trees to avoid overfitting [36]. More recently, other ensemble tree-based algorithms have emerged, based on gradient boosting, namely (i) the Light Gradient Boosting Machine (LightGBM) [37] and (ii) the Categorical Boosting (CatBoost) [38] that should present higher training speed, efficiency, and better accuracy than RF, according to the developers.

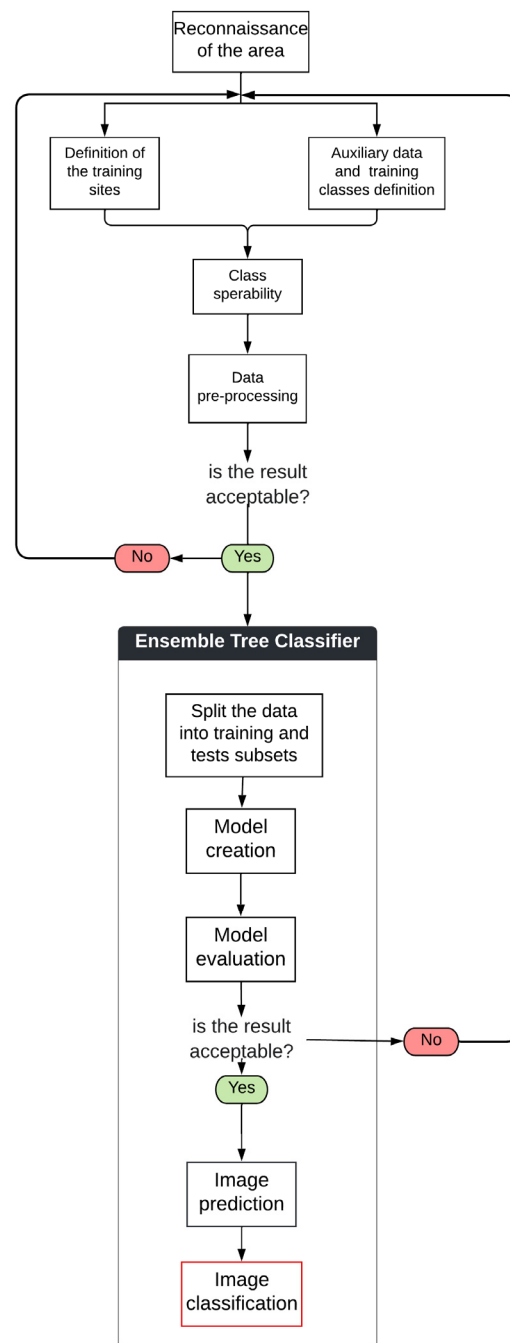
The models proposed have two stops, one after the class separability processes and another after the image prediction. These “stops” were created after important phases of the model to improve the algorithms and minimise errors. These processes can be observed in Figure 6.

The algorithms were implemented in Python programming language using free open-source libraries for ML, namely scikit-learn, lightgbm, and catboost. The Python language allows for the development of an algorithm more interactively than in plugins such as QGIS.

Two RF models were tested for classification. The first (C1) takes into account only the pegmatite pixels that are concentrated within the pegmatite outcrop at Håkonhals. In the second (C2), the sampling polygons are distributed over the mining area of Håkonhals. At Jennyhaugen, the sampling polygons were distributed over the entire mine in both models because, in the geological maps, the entire mining area is considered an open-pit outline. The results of the two models were compared to check if the area east of the outcrop, that may contain traces of pegmatite material (e.g., mining waste dumps and storage sites made of pegmatite material), are also efficient to be used in the classification. The same procedure of Figure 6 was employed for the remaining boosting algorithms and the results were compared with the C1 model for RF.

#### 2.5.1. Reconnaissance of the Area

In this step, a geological map and vegetation (NDVI), snow (NDSI), and water (NDWI) indices were used to understand the elements that make up the study area and, consequently, which classes to consider in the classification stage. The band indices confirm the absence of snow and point out a high vegetation cover. Water is an abundant element throughout the study area. The NDWI, in some areas, does not highlight very well pixels of water bodies with high values, which may indicate high sediment concentration. The study area has two rock classes that may be influential in the classification: the AMCG (anorthositic, mangeritic, charnockitic, and granitic) rocks and the Tysfjord granite gneiss [5]. Due to the spectral similarity between these two rock classes, both were classified in a single class, granite. Built areas are also present but are a very unrepresentative element in the study area. With few villages scattered around the Tysfjord district, this element was masked, after image classification, with a Normalised Difference Built-Up Index (NDBI). After this analysis, four classes were defined for classification: (1) pegmatites, (2) granite, (3) water, and (4) vegetation.



**Figure 6.** Workflow showing the step-by-step method applied to the RF algorithm.

### 2.5.2. Class Separability

To ensure that the classes are well separated and distinguished, the separability of the classes' signatures was performed using Bhattacharyya Distance in PCI Geomatica software. Separability is measured by values between 0 and 2, with values between 0 and 1 ( $0.0 < x < 1.0$ ) representing very poor separability, values between 1 and 1.9 ( $1.0 < x < 1.9$ ) representing poor separability, and values between 1.9 and 2 ( $1.9 < x < 2.0$ ) representing good separability.

### 2.5.3. Data Pre-Processing

To input the data into the algorithm, it is necessary to convert it into a format that can be acceptable in Python language. To do this, a series of procedures were performed in

the ArcMap software. First, each band was clipped only for the training areas. After that, the result was transformed into a point shapefile. The next step was to extract the data from the other rasters using this same point file that allows us to extract the information of all the rasters and organise it in a database (extract multi values to points). Finally, the intersect tool was used to correlate the training areas with the information collected from the pixels of each raster.

#### 2.5.4. Model Creation

To ensure the maximum independence between the training and test subsets, the dataset was split, considering 25% of the pixels for testing and the remaining 75% for training. This method was adopted in previous studies [15]. In this step, the parameters to be optimised and the corresponding parameter range (or variation) to be used in the grid search were defined. For example, in this study, the number of trees to test in the model, defined by the function 'n\_estimators', ranged from 10 to 500, with a steady increase after n\_estimators = 50. Another defined parameter was the 'Class\_weight', which was set to "balanced", thus adjusting the weights inversely, proportional to class frequencies in the input data, to not penalise less frequent data, as is the case of pegmatites.

#### 2.5.5. Model Evaluation

In this step, different metrics are applied to evaluate the best model returned by the grid search of the previous step. The metrics applied in this study were the mean cross-validation score and the Kappa statistics. A confusion matrix was generated considering the test subset.

### 3. Results

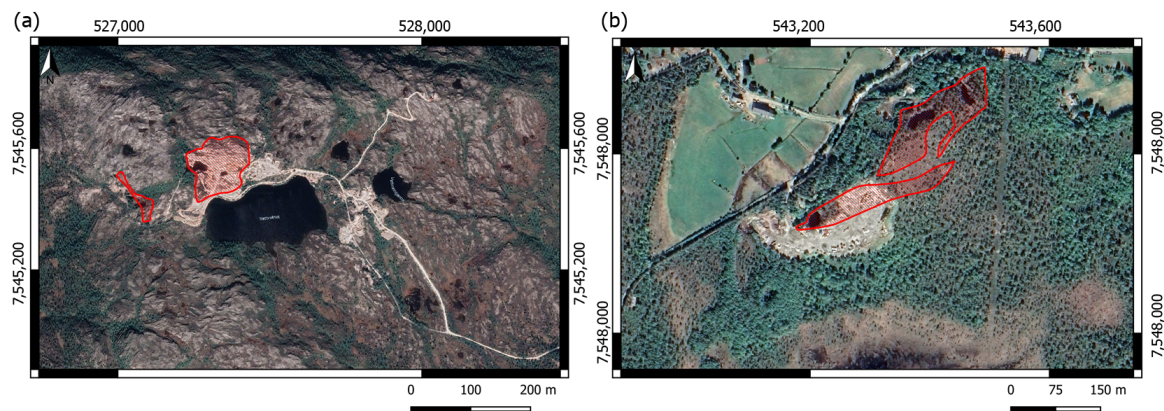
#### 3.1. Spectral Analyses

##### 3.1.1. Spectra from Sentinel 2 Images

Considering the importance of the reflectance peaks for this study, the analysis of the spectra was performed with spectra with continuum removed. In this study, only the main absorptions and reflectance that are in zones of interest for the analysis will be referred to, i.e., zones that fall within the spectral range of the Sentinel 2 bands. The analysis of the spectra directly collected from the 13 bands of Sentinel 2 allowed for the identification and selection of bands of interest for the study. However, these spectra are affected by the water vapour absorption bands (bands 9 and 10), since they are designed for atmospheric correction and cloud detection, and thus were eliminated in this study [39]. This analysis requires previous knowledge of the study area to collect reference spectra correctly while avoiding picking other elements of the image by mistake. The spectra collected on the pegmatite outcrop (Figure 7) were called pure spectra and were further used for the analysis. Ten spectra were collected from Håkonhals and nine from Jennyhaugen.

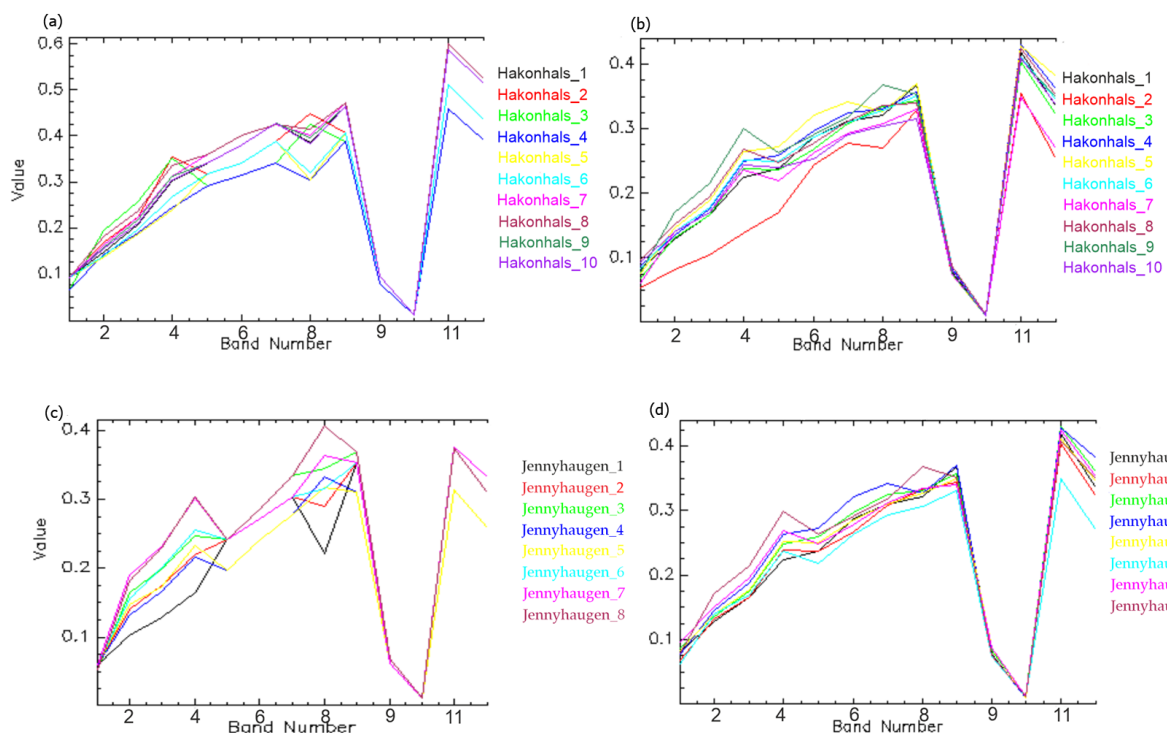
It is known that in the mining area to the east of the Håkonhals pegmatite outcrop there are mining dumps, mining roads, and storage sites made of pegmatite material, but this is not cropping out. As in these areas, there are other geological elements besides pegmatites, such as granites, and to minimise the influence of non-pegmatite spectral pixels on the final results, another pixel extraction method was tested. This method was tested to verify if the average of the spectra can mitigate the impact of other elements in the mining area. In this way, to validate the applicability of this method in Greenfield areas where the exact location of pegmatite outcrops is uncertain, 90 spectra were collected (50 from Håkonhals mine and 40 from Jennyhaugen mine). After the collection, an average was taken from every 5 spectra, resulting in 10 spectra from Håkonhals mine (numbered 1 to 10) and 8 spectra from Jennyhaugen mine (numbered 1 to 8). The average was made using the spectral mathematical tool available in the Envi software. To check whether this method is correct in representing the pegmatite spectra, a comparison was made between the mining spectra and the pure spectra (Figure 8). For this, comparison maps showing the contacts

between the pegmatites and the surrounding rocks were used, with the pegmatite contacts mapped by the authors in the field.



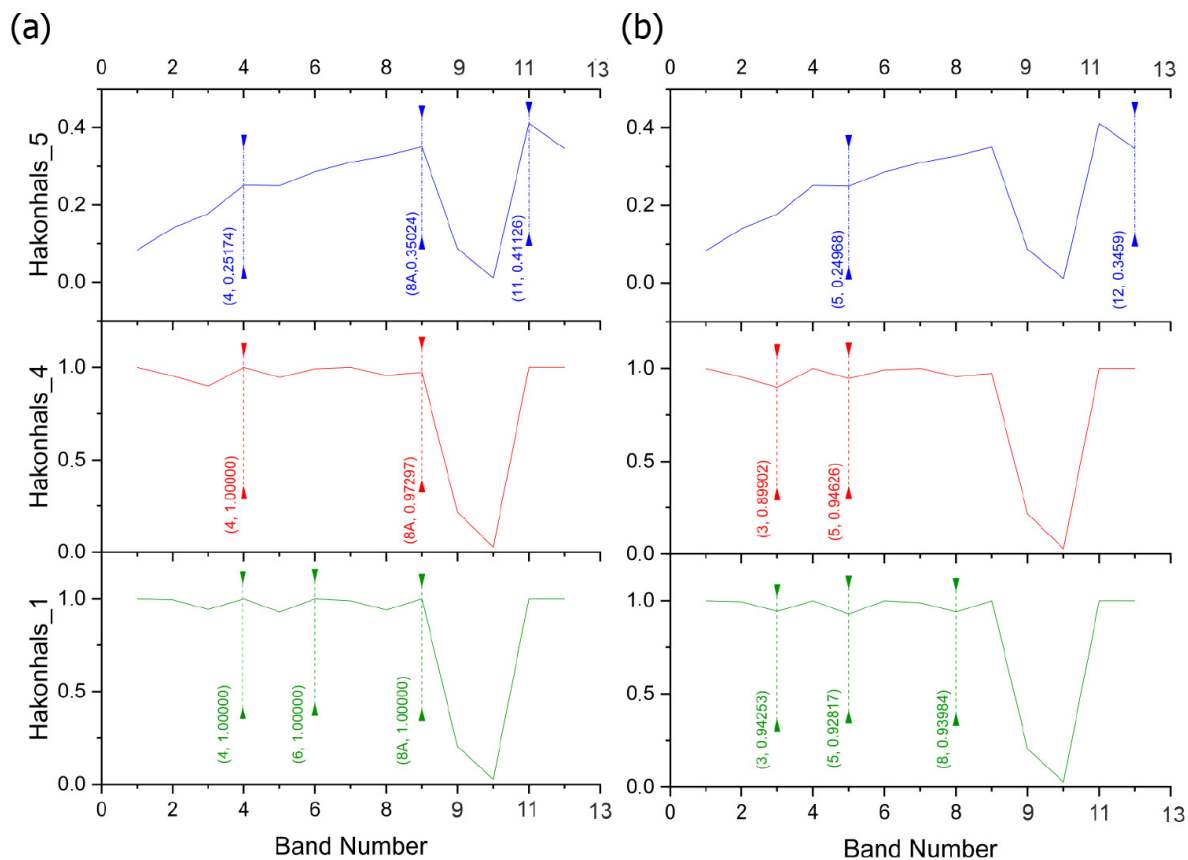
**Figure 7.** Outcrop of pegmatites identified in red polygons. (a) Pegmatite outcrop at Håkonhals. (b) Pegmatite outcrop at Jennyhaugen. The area outside the pegmatite outcrop and within the mining areas may also contain pegmatitic materials.

As shown in Figure 8, the main absorptions and peak reflectance are unchanged from the pure spectra and the mine spectra. The main difference is in the intensity of both the absorptions and the reflectance peaks, which are much more marked in the pure spectra.



**Figure 8.** Sentinel 2 spectral behaviour pure pegmatite spectra (unaveraged) and mining spectra derived from averaging. Håkonhals and Jennyhaugen mines. (a) Håkonhals pure spectra. (b) Håkonhals mining spectra. (c) Jennyhaugen pure spectra. (d) Jennyhaugen mining spectra.

Analysing the relationship between the reflectance of the pegmatites and the spectral range that each Sentinel 2 band covers in the continuum spectrum (Figure 9), it is possible to see that the main absorption bands are 3, 5, 8, and 12, while the most relevant bands in reflectance peaks are 4, 6, 7, and 8A.

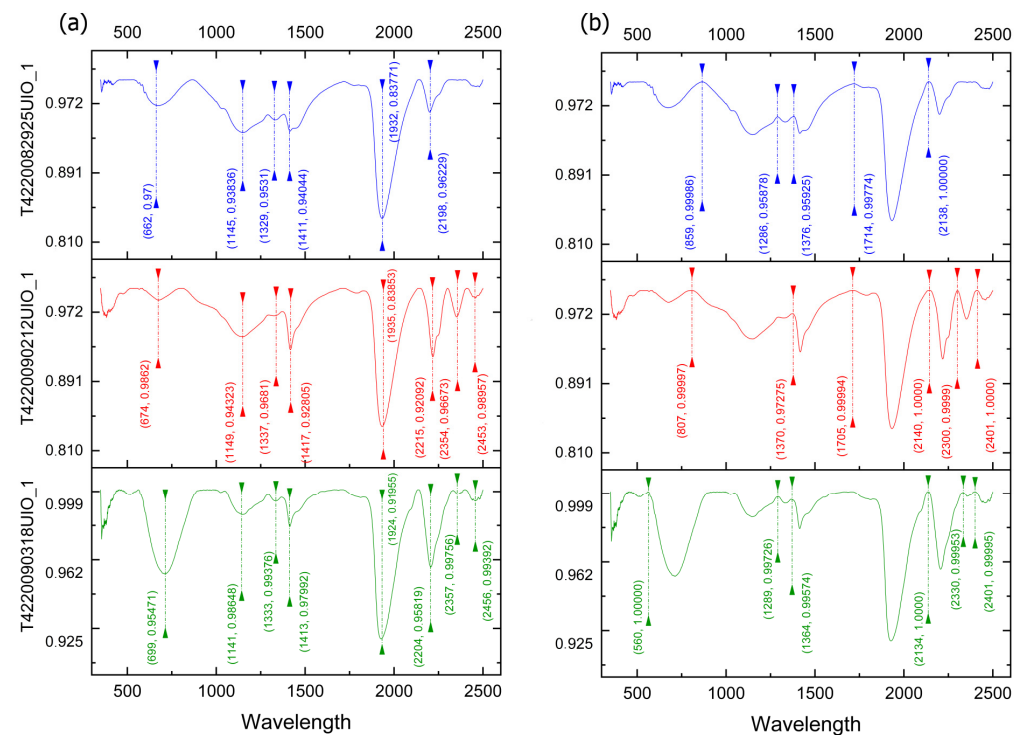


**Figure 9.** Sentinel 2 spectral behaviour from Håkonhals mine (mining spectra). (a) Bands in absorption features. (b) Bands in reflectance peaks. For each spectrum, in either absorption or reflectance regions, is presented, between brackets, the number of the Sentinel 2 bands, and the respective reflectance value.

Comparing the reflectance characteristic of the spectra with the spectral range of the Sentinel 2 bands, we see that there is no difference between the pure spectra and the entire mine spectra. However, the extraction of spectra where it is certain that there is a pegmatitic outcrop is the best option for spectral studies.

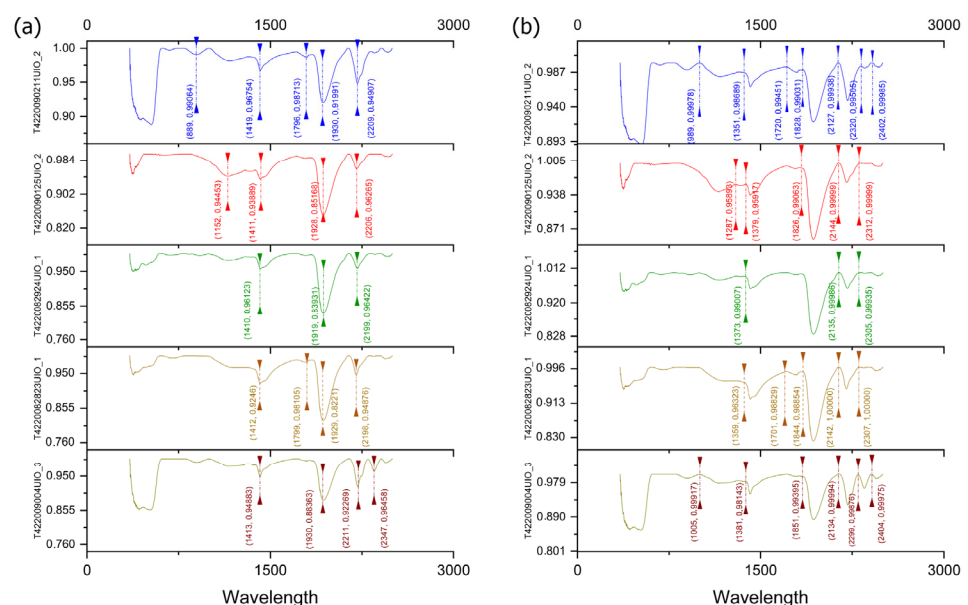
### 3.1.2. Spectra Collected in the Laboratory

The plagioclase spectra (Figure 10) show absorptions at around 670 nm (covered by band 4) and at 2200 nm (covered by band 12). The reflectance peaks start at around 850 nm (band 8A) and finish at around 1700 nm (band 11). The strong, symmetric OH- and water absorptions (around 1411 nm and 1904 nm, respectively) and a single AlOH absorption (around 2200 nm) indicate the presence of montmorillonite in the three samples. Illite is probably mixed with montmorillonite in the red spectra of Figure 10 due to the AlOH secondary absorptions at 2354 nm and 2453 nm. The double absorptions' features at 2215 nm and 2354 nm in the same spectrum are diagnostic of the presence of FeOH and MgOH, respectively. These two absorptions together with a peak between them (at 2300 nm), and diagnostic iron features in VNIR indicate the presence of biotite in sample T4220090212UIO\_1.



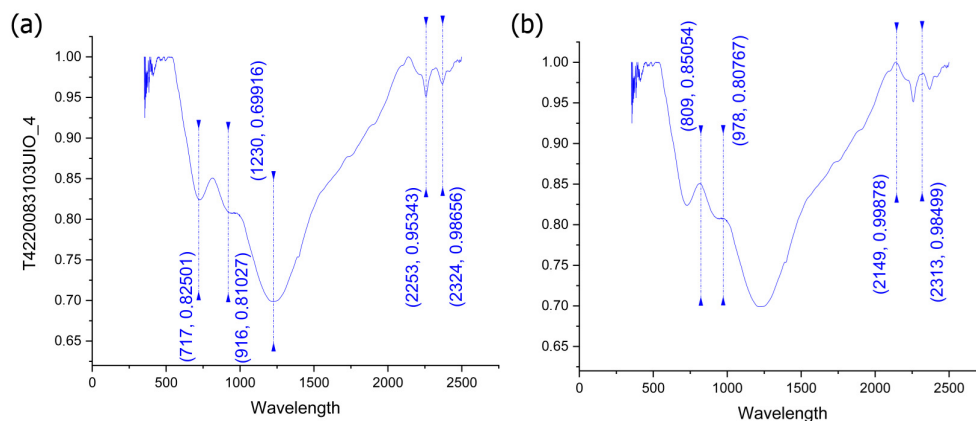
**Figure 10.** Absorption and reflectance peaks from spectra collected from samples of Tysfjord plagioclases. (a) Main absorption peaks (minimum). (b) Main reflectance peaks.

As shown in Figure 11, the spectral bands of the Sentinel 2 do not cover the absorption features of the spectra collected from the K-feldspar samples (around 890, 1100, 1400, 1800, and 1900 nm). The exception is an absorption around 2200 nm (covered by band 12). It is worth mentioning that band 12 also covers the reflectance peaks for these samples (which occur around 2130 nm). The absorption features of the K-feldspar samples point to the possible presence of montmorillonite and  $\text{Fe}^{2+}$ . The extremely pronounced and rounded water features could be due to aqueous fluid inclusions.



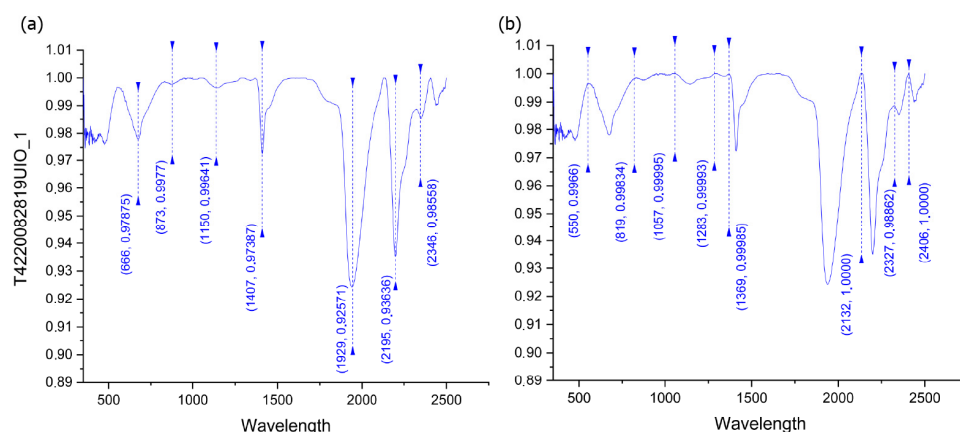
**Figure 11.** Absorption and reflectance peaks from spectra collected from samples of Tysfjord K-feldspar. (a) Main absorption peaks (minimum). (b) Main reflectance peaks.

The first absorption feature of biotite (Figure 12) is at around 710 nm (band 5), which, together with the absorptions at 717 nm and 916 nm, indicates that chlorite is also present in the mixture. Another significant absorption feature is at around 2250 nm (covered by band 12). As for the reflectance, there is a peak around 810 nm (band 8) and another at 2100 nm (band 12). In addition to the two absorptions at 2253 nm–2324 nm, which are due to either chlorite or biotite, the peak reflectance between them is relatively high compared to the reflectance that is diagnostic of biotite.



**Figure 12.** Absorption and reflectance peaks from spectra collected from samples of Tysfjord biotite. (a) Main absorption peaks (minimum). (b) Main reflectance peaks.

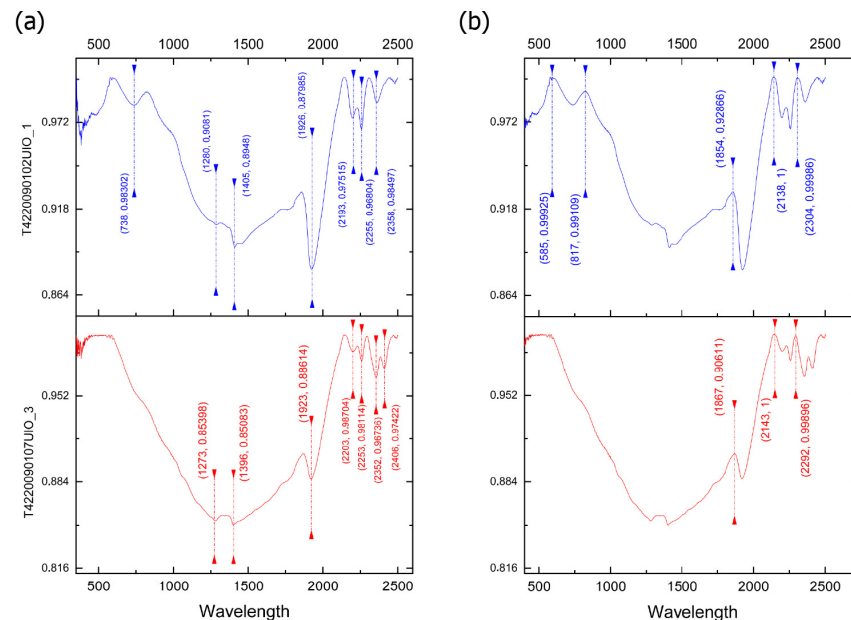
As the quartz samples from the pegmatite core zone have similar absorption and reflectance zones. The spectrum of sample T4220082819UIO\_1 was selective as a representative spectrum. As shown in Figure 13, absorptions occurring at 666 nm and 673 nm are covered by band 4 and another significant absorption feature is at around 2195 nm (band 12). The main peaks of reflectance are at 550 nm (covered by band 3) and 819 nm (covered by band 8) and the last minimum is around 2130 nm (band 12). A single, asymmetric OH absorption at 1407 nm together with a strong water absorption around 1920 nm and a single sharp, asymmetric AIOH absorption feature at 2190 nm indicates the presence of montmorillonite. The broad water adsorption feature indicates the presence of aqueous fluid inclusions in the quartz samples. The two absorptions around 2350 nm and 2440 nm suggest that illite also occurs in this sample.



**Figure 13.** Absorption and reflectance peaks from spectra collected from pegmatite quartz samples of Tysfjord. (a) Main absorption peaks (minimum). (b) Main reflectance peaks.

As shown in Figure 14, the main absorption wavelengths of significance to the study for Tysfjord granite samples are at around 738 nm (covered by band 6) and between 2200 and 2400 nm (covered by band 12). The reflectance zones of interest are at 580 nm (covered by

band 3), 817 nm (covered by band 8), and around 2300 nm (covered by band 12). Band 12 covers both absorption features and reflectance peaks. Both spectra show absorptions around 2200 nm, 2250 nm, and 2360 nm, but the absorptions in VNIR and the water absorption indicate that the top spectrum (in blue) is characterised by the occurrence of chlorite mixed with biotite and possibly montmorillonite, while the bottom spectra (in red) indicates the existence of biotite mixed with white mica and minor chlorite.



**Figure 14.** Absorption and reflectance peaks from spectra collected from samples of Tysfjord granite. (a) Main absorption peaks (minimum). (b) Main reflectance peaks.

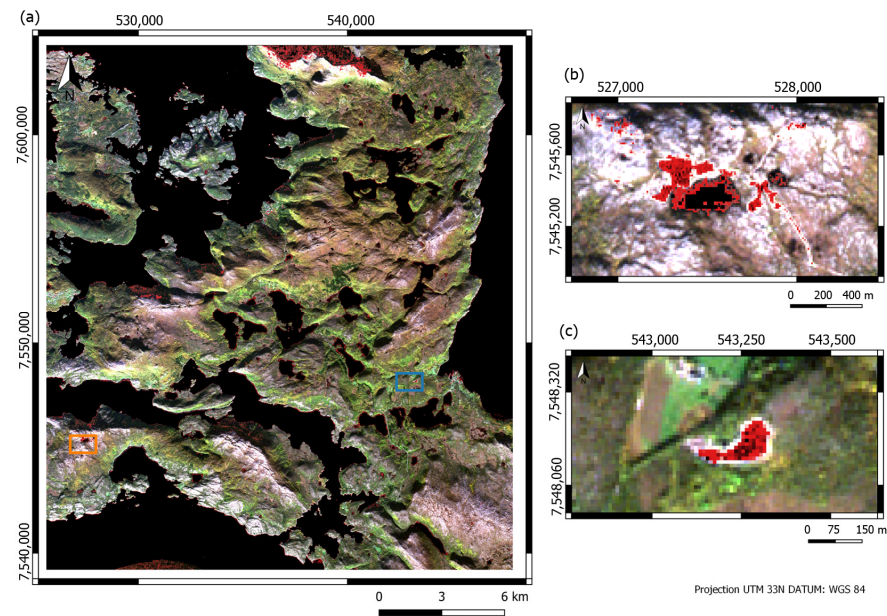
### 3.2. Traditional Methods

#### 3.2.1. Band Ratios

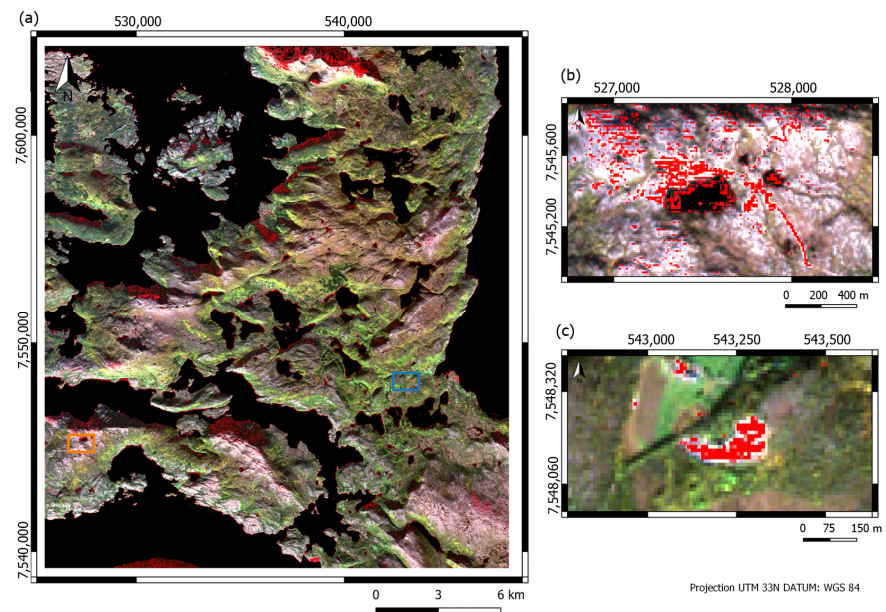
Knowing the bands in absorption and reflectance zones, it is then possible to propose a new BR. Among the BR tested from the analysis of the spectra of the Sentinel 2 bands (Table 2), the BR 4/8 and BR 4/5 were able to successfully highlight the target areas. BR 4/8 was able to highlight the target areas with values between 0.70 and 0.90 (on a scale of 0 to 2.94), which are represented in Figure 15, in red pixels. This BR also highlighted water bodies with high values making it necessary to include a water mask.

BR 4/5 highlights areas with known pegmatites with values between 0.90 and 1.51 (on a scale of 0 to 2.94). Even though it highlights the target areas well, the number of false positives in granite areas is much higher in comparison to BR 4/8. A water mask was also applied (Figure 16).

Regarding the ratios derived from the analysis of the rock samples, and starting with the BR from granite samples, the BR 3/6 highlights the pixels of the pegmatite areas at values around 0.50 (at Håkonhals) and 0.73 (at Jennyhaugen). Water was highlighted with high values for this BR with values greater than 1.0. Granite zones were not well identified (0.30). BR 12/6 cannot distinguish granite from pegmatite. Both present pixels with values around 0.50. BR 8/6 could not distinguish either granite or pegmatite. As these ratios were developed from granite samples, it was expected that they would highlight granite. BR 8A/4 and 8A/11 (from plagioclase) obtained similar results. Both highlighted vegetation better (with values between 1.0 and 2.0) than pegmatites (0.80). As with the plagioclase BRs, the BR 8/5 (from biotite), also highlighted vegetation pixels better (around 4.0) than pegmatites. The BR 8/12 (from quartz) was able to highlight the vegetation better. However, it cannot distinguish pegmatite from granite and assigns high values for water.



**Figure 15.** Result for BR 4/8. Pixels higher than 0.70 are represented in red colour. (a) Overview of the study area where Håkonhals mine is highlighted by the orange rectangle and Jennyhaugen mine by the blue rectangle. (b) Håkonhals mine in focus. (c) Jennyhaugen mine in focus.

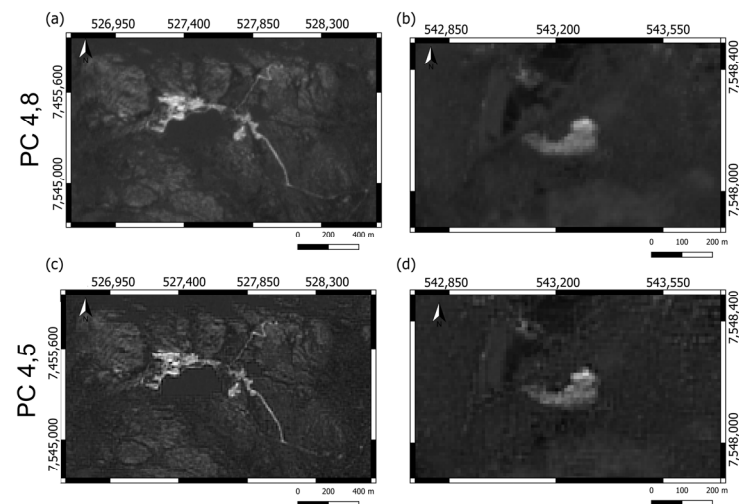


**Figure 16.** Result for BR 4/5. (a) Overview of the study area where Håkonhals mine is highlighted by the orange rectangle and Jennyhaugen mine by the blue rectangle. (b) Håkonhals mine in focus. Here it can be seen that this BR has also highlighted many granite pixels with elevated values. (c) Jennyhaugen mine in focus.

In general, the BRs elaborated from the analysis of the spectra collected in the laboratory are not suitable to be used as input for the RF algorithm, as they highlight other elements, such as vegetation, instead of the pegmatite that is the target of this study. Although BR 3/6 can highlight the pegmatite areas, its result is much inferior to BR 4/5 and 4/8 and, because of that, it was not selected as input.

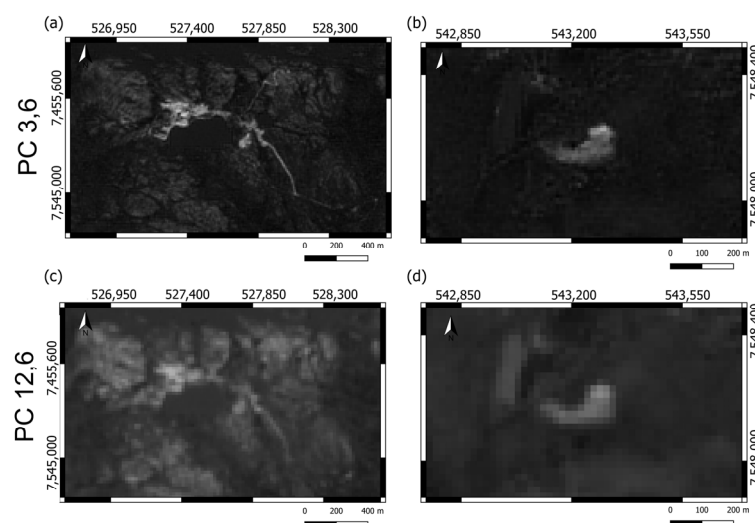
### 3.2.2. Principal Components Analyses

The PCA of bands 4, 5 and 4, 8 obtained very good results. Both were able to highlight the target areas very well and without signal confusion with the surrounding elements, as occurred with the BR. As can be seen in Figure 17, both obtained very similar results. The PCA of the 4 And 5 bands highlighted the target areas with pixels at high values around 1.75 (on a scale of  $-0.03$  to  $1.8$ ), while the PCA 4, 8 highlighted the target area pixels at values around  $0.20$  (scale of  $-0.07$  to  $0.308$ ).



**Figure 17.** Comparison of PCA results of bands 4 and 5 and 4 and 8. (a) Håkonhals mine in focus for PCA 4, 8. (b) Jennyhaugen mine in focus for PCA 4 And 8. (c) Håkonhals mine in focus for PCA 4, 5. (d) Jennyhaugen mine in focus for PCA 4, 5.

For the PCA based on the spectra collected in the laboratory, the PCA of bands 3, 6 shows a good result, highlighting pixels in the white target areas and with values around  $0.22$  (on a scale of  $-0.028$  and  $0.26$ ). Despite highlighting the areas of the mines, the PCA of the bands 12, 6, as well as the BR made of the same bands, also highlights the granite, which is not ideal for this study (Figure 18).



**Figure 18.** Comparison of PCA results of bands 3, 6 and 12, 6. (a) Håkonhals mine in focus for PCA 3, 6. (b) Jennyhaugen mine in focus for PCA 3, 6. (c) Håkonhals mine in focus for PCA 12, 6. (d) Jennyhaugen mine in focus for PCA 12, 6.

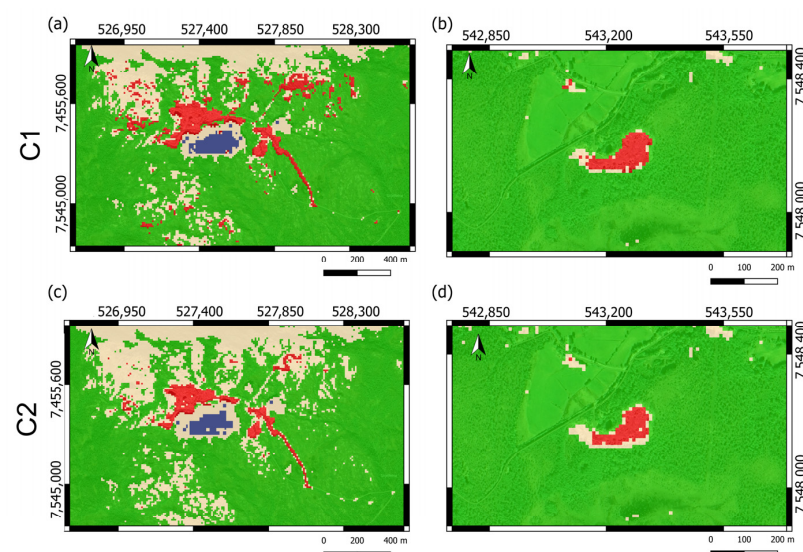
### 3.3. Random Forest and Other Ensemble Tree-Based Algorithms

Based on the previous results, the inputs that were selected to be introduced in the RF algorithm were (i) bands 4, 6, 7, and 8A; (ii) band ratios 4/5 and 4/8; (iii) PCA 4/5 and 4/8; and, taking into consideration a greater inefficiency of the algorithm in classifying vegetation areas in previous works [15], it was decided to add the (iv) NDVI to improve the classification of vegetated areas. The classification occurs with all inputs previously described. Two classifications were made. One (C1) where the pegmatite training areas at Håkonhals only include certain pegmatite outcrop pixels. The other (C2) where the training areas include the outcrops and also other parts of the mining area where the presence of pegmatite material may be present, but it is uncertain. Table 4 shows the scores for the C1 and C2 models.

**Table 4.** Scores for RF model.

Kappa Statistics: C1 = 0.95/C2 = 0.96						
Mean Cross-Validation Score (Accuracy): C1 = 0.96/C2 = 0.97						
	Precision		Recall		F1 Score	
	C1	C2	C1	C2	C1	C2
Granite	0.95	0.92	0.90	0.97	0.93	0.95
Pegmatite	0.95	0.91	0.97	1.00	0.96	0.95
Vegetation	0.98	1.00	0.99	1.00	0.99	1.00
Water	1.00	1.00	1.00	0.97	1.00	0.98

Both models' scores are similar. Concerning the overall classification metrics, the Kappa Statistics and the accuracy are similar. When looking at the classification of the class level separately (Recall, Precision, and F1 Score), there is also little difference in the results. The biggest difference between the two models for the pegmatite class is the precision, where the model with the highest precision is C1 with 0.95, while model C2 shows 0.91. A difference of 0.04 values. Regarding image classification (Figure 19). Model C1 classifies more pixels as pegmatites and, consequently, more false positives with granite and coastal areas. Both models identified pixels as pegmatites outside the outcrop area at Håkonhals. This shows that there are pegmatite materials in this region and that the algorithm is sensitive to them.



**Figure 19.** RF classifier for C1 and C2 models. Pegmatites are classified as red colour, granite as beige, water as blue, and vegetation as green (a) C1 model classification for Håkonhals mine. (b) C1 model classification for Jennyhaugen mine. (c) C2 model classification for Håkonhals mine. (d) C2 model classification for Jennyhaugen mine.

Model C1, despite having more false positives, uses more enhanced sample data with more pixels that are sure to be pegmatites. Therefore, model C1 is the most appropriate for the study area. The results presented hereafter refer to this model.

When the confusion matrix was analysed (Table 5), it was noticed that many granite pixels were misclassified as vegetation. As expected, some pegmatite pixels were also wrongly classified as granite and vice versa. The confusion matrix also indicates misclassification between the vegetation and pegmatite classes and between the granite and vegetation classes.

**Table 5.** C1 model confusion matrix.

	Predicted			
	Granite	Pegmatite	Vegetation	Water
Granite	37	2	3	0
Pegmatite	2	58	0	0
Vegetation	0	1	99	0
Water	0	0	0	39

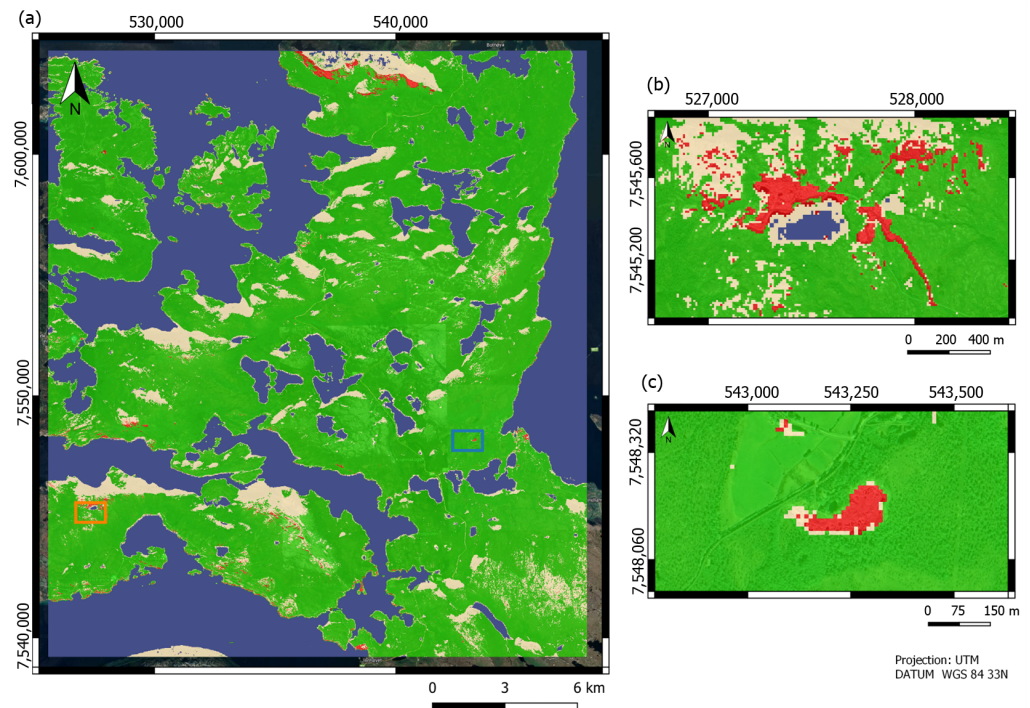
The distinction between the granite and pegmatite classes was a challenge, but sampling was redone until good separability was obtained. As exemplified in Table 6, all classes have good separability, which is encouraging especially concerning the separability of the pegmatite and granite classes, which was low in previous studies in Tysfjord [15].

**Table 6.** Signature separability for training classes.

	Pegmatite	Granite	Water	Vegetation
Pegmatite				
Granite	1.953983			
Water	2.000000	1.999965		
Vegetation	1.993627	1.964088	1.999998	

Still, a lot of false positives are concentrated in the upper right corner of the study area (Figure 20). When comparing the classification with high-resolution images, it is noticeable that areas that are in the shadow of tall landforms, such as mountains or slopes, were mostly classified as granite, even when vegetation is present in these areas. The pixels in these zones have extremely low reflectance, around 0.020, which may have negatively influenced the classification.

The importance of the inputs was also analysed (Table 7). Regarding C1 model, the most relevant inputs for the analysis were, in order of importance, NDVI, band 7, PCA 4, 8, and band 4. However, all inputs were relevant in the image classification, it is evident that, among the traditional image processing methods, PCA 4, 8 and BR 4/8 have the highest importance. Among the bands, band 7 and band 4 are the ones that stood out the most. While in model C1, there was a great discrepancy between the importance of band 8A and the others, in the Boosting algorithms the importance of the bands is more balanced. In the LightGBM algorithm, band 6 had the highest contribution in the classification and band 4 the lowest. In contrast to LightGBM, band 4 is the most important band for the Catboost algorithm. In both Boosting algorithms, BR 4/8 had the highest importance, followed by PCA 4, 8. While BR 4/5 and PCA 4, 5 have the lowest importance in both Boosting algorithms. The BR 4/8 was the most important input for LightGBM and BR 4/5 was the least important. NDVI was the most important input for Catboost, while BR 4/5 was the least important for the classification.

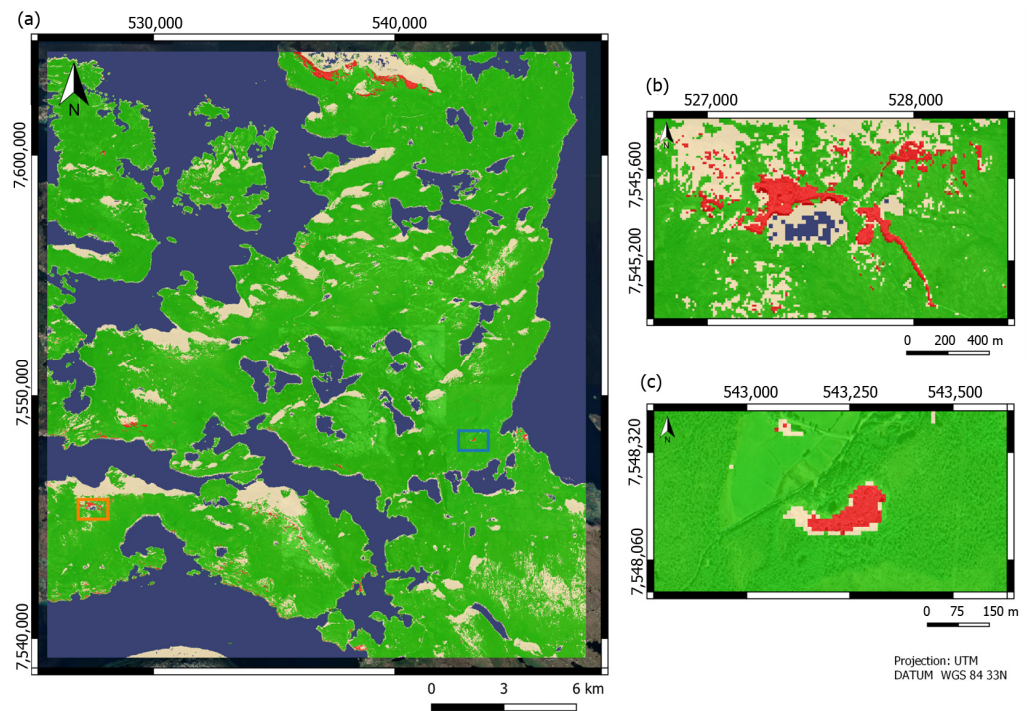


**Figure 20.** RF classifier. Pegmatites are classified as red colour, granite as beige, water as blue, and vegetation as green. (a) Overview of the study area where Håkonhals mine is highlighted by the orange rectangle and Jennyhaugen mine by the blue rectangle. (b) Håkonhals mine in focus. (c) Jennyhaugen mine in focus.

**Table 7.** Importance of inputs in image classification, measured in percentage.

Input	Importance C1	Importance LightGBM	Importance Catboost
Band 04	16.46	9.4	13.25
Band 06	12.24	13.18	12.27
Band 07	15.06	11.77	11.35
Band 8A	5.22	10.93	11.54
PCA 4, 5	1.20	5.6	4.8
PCA 4, 8	13.35	11.79	11.24
BR_4/5	8.03	4.3	3.8
BR_4/8	8.53	18.57	14.69
NDVI	19.87	14.25	16.95

Both boosting models obtained the same result as model C1 for accuracy (0.96), while the kappa statistic was 0.97 for the LightGBM algorithm and 0.96 for Catboost. Regarding the classification output, both boosting models were able to classify pegmatite. When comparing the results with model C1, it is possible to see that the boosting models classified fewer pixels as pegmatite, but the difference is so small that it cannot be said that this has reduced false positives. Despite being able to identify known pegmatites as effectively as the other models, Catboost did not correctly classify water bodies and vegetation. Among the boosting algorithms, the LightGBM algorithm obtained the best result, being able to identify pegmatites in target areas and classify other elements of the study area (Figure 21).



**Figure 21.** LGB classifier. Pegmatites are classified as red colour, granite as beige, water as blue, and vegetation as green (a) Overview of the study area where Håkonhals mine is highlighted by the orange rectangle and Jennyhaugen mine by the blue rectangle. (b) Håkonhals mine in focus. (c) Jennyhaugen mine in focus.

## 4. Discussion

### 4.1. Spectral Analyses

#### 4.1.1. Spectra from Sentinel 2 Images

Although the spectra from the Jennyhaugen mine have a more prominent marked minimum in band 8 when compared to the Håkonhals spectra, the spectra extracted from both study areas have, in general, similar absorption features and reflectance behaviour. The results shown in Figure 8, confirm that the spectral average for spectra obtained from multispectral satellite images can be an effective method in Greenfield areas where the exact location of the pegmatite to be exploited is not known. On the other hand, the results of this method may be impacted if the pixels present in the mining area are mostly represented by pixels of the bedrock. At Håkonhals the pegmatite occupies a considerable part of the mine area and as shown in Figure 8a,b, the differences in intensity between the absorptions and reflectance peaks are minor. At Jennyhaugen the pegmatite is found in a small part of the mine area, as illustrated in Figure 8c,d, and the absorptions, in particular, are much stronger in the mining spectra. It can be said that the pegmatites have strong absorption in band 8 (NIR) in the individual pixels and this absorption was attenuated when the spectra from Jennyhaugen were averaged. This did not happen for Håkonhals, where the average spectra still show, even if timidly, absorption in band 8. The main bands for this set of spectra are the same for pure and mining spectra. The difference in absorption intensity and reflectance are important elements in distinguishing between minerals. However, in the absence of such information, the average of the spectra can be used in Greenfields. We should point out that averaging the spectra may not have much impact on multispectral bands, but this is different for hyperspectral images. With much more robust spectra carrying much more information, averaging can negatively impact the analysis of the results. Therefore, it should be noted that the same approach would not be viable when using hyperspectral data, as seen by some spectral changes around band 8 between the pure and the mine spectra.

#### 4.1.2. Spectra Collected in the Laboratory

Analysis of the spectra collected in the laboratory shows that most of the absorptions of the samples studied take place in the SWIR region (1100–2500 nm). However, only a few of these absorptions are covered by the spectral range of the Sentinel 2 bands, except for band 12, which covers the absorptions in almost all samples around 2200 nm. This absorption is characteristic of the presence of AlOH, which indicates that most samples contain this group [16]. However, band 12 also covers the FeOH and MgOH features that, in the analysed samples, are related to the presence of chlorite or biotite in the samples. Regarding reflectance, band 10 (1360–1390 nm) comprises reflectance peaks in the quartz, K-feldspar, plagioclase, and granite samples and could be the most representative band regarding reflectance. However, this band has no surface information and was omitted from this study. As pointed out by [16], it is evident that the SWIR region has a strong contribution to lithological studies. However, as the results of this study show, Sentinel 2 does not have an important band in this region for detecting NYF pegmatites.

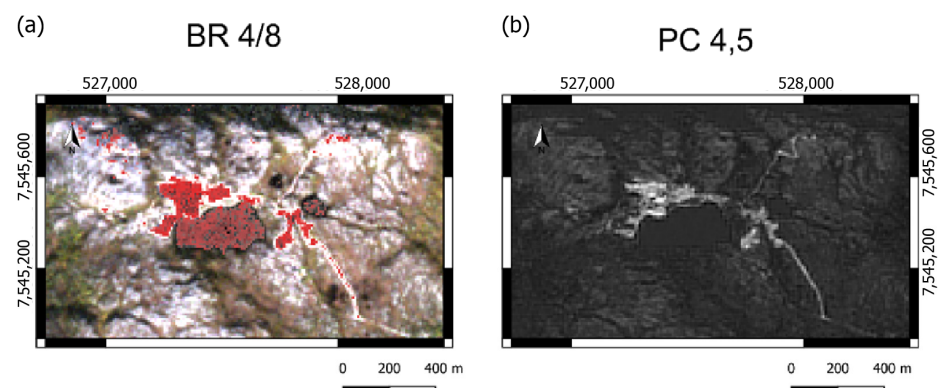
#### 4.2. Traditional Methods

##### 4.2.1. Band Ratios

In general, the BRs elaborated from the analysis of the spectra collected in the laboratory are not suitable to be used as input for the RF algorithm, as they highlight other elements, such as vegetation, instead of the pegmatite that is the target of this study. Although BR 3/6 can highlight the pegmatite areas, its result is much inferior to BR 4/5 and 4/8 and, because of that, it was not selected as an input.

##### 4.2.2. Principal Components Analyses

The results obtained indicate that PCA is more efficient in highlighting NYF pegmatites from Tysfjord than BR. On the other hand, BRs were able to highlight pegmatite areas, but the signal confusion with water and granite bodies (in the case of BR 4/5) makes their result inferior to PCAs that, in turn, highlight pegmatites with fewer false positives (Figure 22). Regardless of the false positives presented by BR, they were able to successfully identify the zones of interest, and together with selected bands, PCA and NDVI were used as input in the algorithms.



**Figure 22.** Comparison of the best BR and PCA results with the Håkonhals mine in focus. (a) BR 4/8. It is possible to see that the ratio also highlights water bodies and granite. (b) PCA of bands 4, 5. Note that it identifies the target area perfectly with no signal confusion with water or granite.

The traditional methods, which were adapted from the analysis of the spectra extracted from the Sentinel 2 bands, proved to be more efficient in highlighting the target areas when compared to the methods proposed from the spectral analysis of the spectra collected in the laboratory. This suggests that extracting spectra directly from satellite bands is a better method for selecting the most appropriate bands for identifying the target. It should be noted that the target of this study is not individual minerals, but the whole pegmatite rock;

therefore, analysing the spectra of specific mineral samples is not the most efficient method of identifying NYF pegmatites.

In the case of the methods adapted from the spectra collected in the laboratory, except for granite, whose ratios of 3/6 were able to discard the target areas, none of the other ratios proposed achieved the same effect so well. On the other hand, the PCA achieved good results, especially the PC 3, 6, which highlighted the target areas very well.

#### 4.3. Random Forest Algorithm

Despite the good results, the C2 model was developed with the sole intention of validating the average approach to be used in Greenfield exploration areas with little information on the pegmatite outcrop. In such cases, pixels across the entire open pit mining area would be used to sample. When comparing satellite images in true colour, we noticed high reflectance in this area, which could indicate snow, but the NDSI does not identify snow cover in this same location. The reasons for this false positive will be investigated in the future. With exception of “shadow” zones, the classification of the other classes seems correct as to the reality of the study area, which demonstrates an improvement in the classification of vegetation and water bodies in relation to previous studies [15].

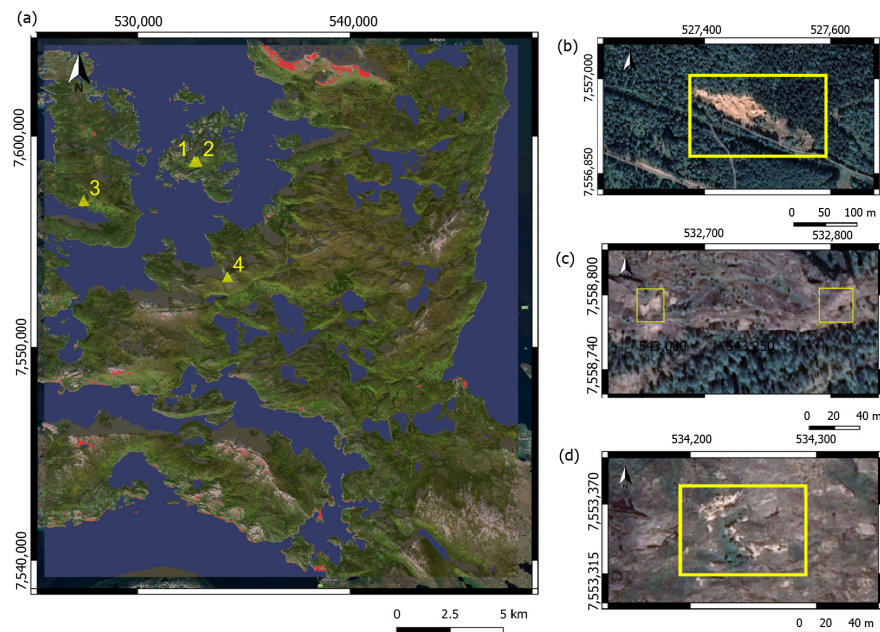
The RF C1 model developed was shown to be effective in detecting NYF pegmatites, indicating that selecting bands for their ability to highlight (reflect) the target is an accurate method. Even so, when compared to previous works, an improvement in the classification of other elements in the study area such as granite and vegetation was seen. However, the number of false positives is an issue that should be investigated in the future. In terms of relevance for the classification, it can be said that the Bands with higher reflectance (B4, B6, B7 and B8A) and the PCA (4, 8) are the most important for the study. In an area with a strong presence of vegetation, NDVI is an important contribution to improving the classification of this class.

#### 4.4. Boosting Algorithms

Both algorithms were able to highlight pegmatites pixels from the target areas. In relation to the classification, in general, the LightGBM algorithm was shown to be superior to the Catboost, managing to classify all the elements of the study area. The Catboost algorithm was not effective in classifying water and vegetation, misclassifying as vegetation pixels that correspond to water bodies. The confusion in the classification between these two classes is already known from previous works [15], where RF also misclassified vegetation pixels as belonging to the water class. In this work, this was avoided using the NDVI to increase the sensitivity of the algorithms in the vegetation classification. Despite the misclassification between these classes, Catboost was able to identify pegmatites in the target areas, which proves that using inputs with high pixel values for pegmatites is an accurate method of identifying this target. In terms of relevance for the classification BR 4/8 and the bands with higher reflectance were the most important for both boosting algorithms.

### 5. New Possible Areas of Interest for Exploration

The applied RF C1 model shows a robust classification of pegmatites having classified more areas as pegmatites than previous models [15], and this demonstrates the possibility that new areas of interest for exploration can be found. For this analysis, the areas indicated by the RF algorithm that were also highlighted by traditional methods were compared with high-resolution photographs. Three points were selected (Figure 23). Points 1 and 2 are very closely located at the WGS 84-UTM coordinate Zone 33N 15,786N/68,140Y and 15,789X/68,140Y, respectively. The sites are located at the flank of a hilltop on the island of Tannøya. The pictures show greyish to whitish hard rock exposures. However, the RF model most likely has probably aimed at the whitish areas, which could be either quartz pegmatite or Tysfjord granite. A follow-up in the field is necessary.



**Figure 23.** Points of interest for exploration. (a) Overview of the study area where the points of interest were identified by the yellow rectangle. (b) Points 1 and 2 in focus. (c) Point 3 in focus. (d) Point 4 in focus.

Point 3 (Figure 23c) is located at 15660N/68125Y on a mountain flank in the eastern part of Hamarøy and north of Brennvikvatnet. The area of interest shows an elongated area of yellowish-coloured rock exposure surrounded by forest. The area seems to be used as an open pit for gravel and sand, but also sections with hard rock of the same colour can be seen. In Northern Norway and, in particular, on Hamarøy, deep weathering of the basement rocks and chemical alteration to clay minerals like illite and smectites is well known and widespread. Since quartz is very hard to weather, the site shows potentially original feldspar pegmatite occurrence with an abundance of feldspar weathering. Locals use the saprolite preferable to build and maintain their gravel roads, which could explain this open pit. Furthermore, a reported minor location with feldspar pegmatite just south of the Brennvikvatnet [40] increases the possibility for another feldspar pegmatite location here.

Finally, point 4 is located at 15821X/68125Y at the top of Jørenvik mountain. The site shows bright, whitish areas within the Tysfjord granite, which could represent quartz-pegmatitic quartz, but it could also be loose gravel or sand derived from physical weathering, enriched in quartz from the underlying Tysfjord granite. This needs to be followed up in the field. In any case, the points identified are promising and show that the approach works. Having a more comprehensive spectral database, including spectra from altered and weathered bedrock, will help to fine-tune and improve it.

## 6. Conclusions

This study was based on spectral analysis to select the most adequate bands for image processing methods and then the most suitable inputs for tree-based ensemble classifiers.

Extracting spectra from areas where it is certain that they are pegmatite outcrops (pure spectra) is the best option for spectral studies, but in the absence of such information, the average spectra can be used in Greenfields areas.

The traditional methods, which were adapted from the analysis of the spectra extracted from the Sentinel 2 bands, proved to be more efficient in highlighting the target areas when compared to the methods proposed from the spectra collected in the laboratory.

As for the methods adopted from the analysis of the spectra extracted from the Sentinel 2 bands, we can say that the BR could highlight the target areas, despite the

spectral confusion with water bodies. Nonetheless, adding a water mask is still a very useful tool in the application of such methods. The PCA was the method that best allowed us to attain the objectives of this work, correctly identifying the target areas.

Regarding the identification of pegmatites, all tree-based ensemble algorithms obtained a similar result. As for the other elements of the study area, Catboost obtained the lowest performance, showing less sensitivity in the spectral differences between water and vegetation classes.

Despite this, the results still need to be improved, especially to decrease false positives. Overall, the applied method proved to be a powerful and robust tool for the exploration of NYF pegmatites in areas with poor vegetation coverage.

The spectral analysis allows for the method to be adapted according to the characteristics of the target under study, and it can be adapted to identify LCT pegmatites. In the future, other algorithms such as SVM and Convolutional Neural Networks (CNN) will be applied that, together with the products generated in this research, will be used as parameters to perform a robust spatial analysis and generate prospect maps that can be used to support decision making regarding the exploration of NYF pegmatites in Tysfjord.

The results of the techniques applied in this study are very promising, being able to accurately identify the occurrence of pegmatites. Furthermore, this method is highly adaptable and can be applied to other study areas and tested with other satellite products such as Landsat 9 OLI 2, Landsat 8 OLI, ASTER TERRA, WORLDVIEW-3, and hyper-spectral images. Thus, this approach is highly valuable to the mining industry and is a strong contribution to the field of pegmatite prospecting, providing information that can be used by other researchers and mining professionals to decrease the impacts of early-stage exploration.

**Author Contributions:** Conceptualisation, D.S., J.C.-F., A.L. and A.C.T.; methodology, D.S., J.C.-F., A.L. and A.C.T.; software, D.S. and J.C.-F.; formal analysis D.S., J.C.-F., M.B. and A.M.; investigation, D.S., J.C.-F., A.L., A.C.T., M.B. and A.M.; resources A.L. and A.M.; data curation, D.S. and J.C.-F.; writing—original draft preparation, D.S.; writing—review and editing, all remaining authors; visualisation, D.S. and J.C.-F.; supervision, A.C.T., J.C.-F. and A.L.; project administration D.S., A.C.T., J.C.-F., A.L. and A.M.; funding acquisition, A.M. and A.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study is funded by European Union's Horizon 2020 innovation programme under grant agreement No 869274, project GREENPEG: New Exploration Tools for European Pegmatite Green-Tech Resources. The work was also supported by Portuguese National Funds through the FCT-Fundação para a Ciência e a Tecnologia, I.P., with projects UIDB/04683/2020 and UIDP/04683/2020-ICT (Institute of Earth Sciences).

**Data Availability Statement:** Cardoso-Fernandes, J., Teodoro, A.C., Santos, D., and Lima, Alexandre. (2022). Spectral Library of European Pegmatites, Pegmatite Minerals and Pegmatite Host-Rocks—The Greenpeg Database (Version 1) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.6518319>.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. European Commission, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs; Bobba, S.; Carrara, S.; Huisman, J.; Mathieux, F.; Pavel, C. *Critical Raw Materials for Strategic Technologies and Sectors in the EU: A Foresight Study*; Publications Office: Luxembourg, 2020; Available online: <https://data.europa.eu/doi/10.2873/58081> (accessed on 8 June 2022).
2. Haxel, G. *Rare Earth Elements—Critical Resources for High Technology*; U.S. Department of the Interior: Washington, DC, USA, 2002; pp. 1–11.
3. Kesler, S.E.; Gruber, P.W.; Medina, P.A.; Keoleian, G.A.; Everson, M.P.; Wallington, T.J. Global lithium resources: Relative importance of pegmatite, brine and other deposits. *Ore Geol. Rev.* **2012**, *48*, 55–69. [\[CrossRef\]](#)
4. Müller, A.; Reimer, W.; Wall, F.; Williamson, B.; Menuge, J.; Brönnner, M.; Haase, C.; Brauch, K.; Pohl, C.; Lima, A.; et al. GREENPEG—exploration for pegmatite minerals to feed the energy transition: First steps towards the Green Stone Age.

- In *The Green Stone Age: Exploration and Exploitation of Minerals for Green Technologies*; Smelror, M., Hanghøj, K., Schiellerup, H., Eds.; Special Publications; Geological Society of London: London, UK, 2022; p. 526. [CrossRef]
5. Müller, A.; Romer, R.L.; Augland, L.E.; Zhou, H.; Rosing-Schow, N.; Spratt, J.; Husdal, T. Two-Stage Regional Rare-Element Pegmatite Formation at Tysfjord, Norway: Implications for the Timing of Late Svecofennian and Late Caledonian High-Temperature Events. *Int. J. Earth Sci.* **2022**, *111*, 987–1007. [CrossRef]
  6. Cardoso-Fernandes, J.; Teodoro, A.C.; Lima, A. Remote sensing data in lithium (Li) exploration: A new approach for the detection of Li-bearing pegmatites. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *76*, 10–25. [CrossRef]
  7. Cardoso-Fernandes, J.; Teodoro, A.C.; Lima, A.; Roda-Robles, E. Semi-Automatization of Support Vector Machines to Map Lithium (Li) Bearing Pegmatites. *Remote Sens.* **2020**, *12*, 2319. [CrossRef]
  8. Cardoso-Fernandes, J.; Teodoro, A.C.; Lima, A.; Perrotta, M.; Roda-Robles, E. Detecting Lithium (Li) Mineralizations from Space: Current Research and Future Perspectives. *Appl. Sci.* **2020**, *10*, 1785. [CrossRef]
  9. Salles, R.R.; Filho, C.R.S.; Cudahy, T.; Vicente, L.E.; Monteiro, L.V.S. Hyperspectral remote sensing applied to uranium exploration: A case study at the Mary Kathleen metamorphic-hydrothermal U-REE deposit, NW, Queensland, Australia. *J. Geochemical Explor.* **2017**, *179*, 36–50. [CrossRef]
  10. Amer, R.; El Mezayen, A.; Hasanein, M. ASTER spectral analysis for alteration minerals associated with gold mineralization. *Ore Geol. Rev.* **2016**, *75*, 239–251. [CrossRef]
  11. Zimmermann, R.; Brandmeier, M.; Andreani, L.; Mhopjeni, K.; Gloaguen, R. Remote sensing exploration of Nb-Ta-LREE-enriched carbonatite (Epembe/Namibia). *Remote Sens.* **2016**, *8*, 620. [CrossRef]
  12. Pour, A.B.; Hashim, M. Hydrothermal alteration mapping from Landsat-8 data, Sar Cheshmeh copper mining district, south-eastern Islamic Republic of Iran. *J. Taibah Univ. Sci.* **2015**, *9*, 155–166. [CrossRef]
  13. Cardoso-Fernandes, J.; Lima, A.; Teodoro, A.C. Potential of Sentinel-2 data in the detection of lithium (Li)-bearing pegmatites: A study case. In Proceedings of the Earth Resources and Environmental Remote Sensing/GIS Applications IX, Berlin, Germany, 11–13 September 2018; p. 10790. [CrossRef]
  14. Cardoso-Fernandes, J.; Teodoro, A.C.; Lima, A.; Roda-Robles, E. Evaluating the performance of support vector machines (SVMs) and random forest (RF) in Li-pegmatite mapping: Preliminary results. In Proceedings of the Earth Resources and Environmental Remote Sensing/GIS Applications X., Strasbourg, France, 10–12 September 2019; Schulz, K., Michel, U., Nikolakopoulos, K.G., Eds.; SPIE: Bellingham, WA, USA, 2019; Volume 11156, pp. 146–157. [CrossRef]
  15. Teodoro, A.C.M.; Santos, D.; Cardoso-Fernandes, J.; Lima, A.; Brönnner, M. Identification of pegmatite bodies, at a province scale, using machine learning algorithms: Preliminary results. In Proceedings of the Earth Resources and Environmental Remote Sensing/GIS Applications XII, Online, 13–17 September 2021; Volume 11863. [CrossRef]
  16. Booyesen, R.; Lorenz, S.; Thiele, S.T.; Fuchsloch, W.C.; Marais, T.; Nex, P.A.M.; Gloaguen, R. Accurate Hyperspectral Imaging of Mineralised Outcrops: An Example from Lithium-Bearing Pegmatites at Uis, Namibia. *Remote Sens. Environ.* **2022**, *269*, 112790. [CrossRef]
  17. Rossi, C.; Bateson, L.; Bayarar, M.; Butcher, A.; Ford, J.; Hughes, A. Framework for Remote Sensing and Modelling of Lithium-Brine Deposit Formation. *Remote Sens.* **2022**, *14*, 1383. [CrossRef]
  18. Cardoso-Fernandes, J.; Lima, J.; Lima, A.; Roda-Robles, E.; Köhler, M.; Schaefer, S.; Barth, A.; Knobloch, A.; Gonçalves, M.A.; Gonçalves, F.; et al. Stream Sediment Analysis for Lithium (Li) Exploration in the Douro Region (Portugal): A Comparative Study of the Spatial Interpolation and Catchment Basin Approaches. *J. Geochemical Explor.* **2022**, *236*, 106978. [CrossRef]
  19. Galeschuk, C.R.; Vanstone, P.J. Exploration techniques for rare-element pegmatite in the Bird River greenstone belt, southeastern Manitoba. In Proceedings of the Exploration 07: Fifth Decennial International Conference on Mineral Exploration, Toronto, ON, Canada, 9–12 September 2007; Milkereit, B., Ed.; pp. 823–839.
  20. Hetherington, C.J.; Mailloux, G.A.; Miller, B.V. A multi-mineral U-(Th)-Pb dating study of the Stetind pegmatite of the Tysfjord region, Norway, and implications for production of NYF-rare element pegmatites during orogenic collapse. *Lithos* **2021**, *398–399*, 106257. [CrossRef]
  21. Müller, A.; Husdal, T.; Sunde, Ø.; Friis, H.; Andersen, T.; Johansen, T.S.; Werner, R.; Thoresen, Ø.; Olerud, S. *Norwegian Pegmatites I: Tysfjord-Hamarøy, Evje-Iveland, Langesundsfjord*, 6th ed.; Norsk Geologisk Forening: Trondheim, Norway, 2017; ISBN 9788283470208.
  22. Foslie, S. Steinar Foslie Tysfjords geologi. In *Beskrivelse til det Geologiske Gradteigskart Tysfjord.*; H. Aschehoug & Company: Oslo, Norway, 1941; Volume 149.
  23. Norge i Bilder. Available online: <https://www.norgebilder.no/> (accessed on 9 March 2021).
  24. Chavez, P., Jr. Image-Based Atmospheric Corrections-Revisited and Improved. *Photogramm. Eng. Remote Sens.* **1996**, *62*, 1025–1036.
  25. Cardoso-Fernandes, J.; Silva, J.; Dias, F.; Lima, A.; Teodoro, A.C.; Barrès, O.; Cauzid, J.; Perrotta, M.; Roda-robles, E.; Ribeiro, M.A. Tools for remote exploration: A lithium (li) dedicated spectral library of the fregeneda–almendra aplite–pegmatite field. *Data* **2021**, *6*, 33. [CrossRef]
  26. Martin, M.Z.; Fox, R.V.; Miziolek, A.W.; DeLucia, F.C., Jr.; André, N. Spectral analysis of rare earth elements using laser-induced breakdown spectroscopy. In *Next-Generation Spectroscopic Technologies VIII*; SPIE: Bellingham, WA, USA, 2015; Volume 9482. [CrossRef]
  27. Patel, C.M.; Patel, C.D.; Rami, J.M.; Patel, K.R. Optical spectroscopic study of natural rock's minerals. *Mater. Today Proc.* **2020**, *43*, 497–501. [CrossRef]

28. Cardoso-Fernandes, J.; Silva, J.; Perrotta, M.M.; Lima, A.; Teodoro, A.C.; Ribeiro, M.A.; Dias, F.; Barrès, O.; Cauzid, J.; Roda-Robles, E. Interpretation of the Reflectance Spectra of Lithium (Li) Minerals and Pegmatites: A Case Study for Mineralogical and Lithological Identification in the Fregeneda-Almendra Area. *Remote Sens.* **2021**, *13*, 3688. [\[CrossRef\]](#)
29. ASD Inc Indico Pro User's Guide. Available online: <https://www.malvernpanalytical.com/en/learn/knowledge-center/user-manuals/indico-pro-users-guide.html> (accessed on 7 February 2022).
30. Lyu, P.; He, L.; He, Z.; Liu, Y.; Deng, H.; Qu, R.; Wang, J.; Zhao, Y.; Wei, Y. Research on remote sensing prospecting technology based on multi-source data fusion in deep-cutting areas. *Ore Geol. Rev.* **2021**, *138*, 104359. [\[CrossRef\]](#)
31. Santos, D.; Teodoro, A.C.M.; Lima, A.; Cardoso-Fernandes, J. Remote sensing techniques to detect areas with potential for lithium exploration in Minas Gerais, Brazil. In Proceedings of the Earth Resources and Environmental Remote Sensing/GIS Applications X, Strasbourg, France, 10–12 September 2019; Volume 111561f. [\[CrossRef\]](#)
32. Cardoso-Fernandes, J.; Lima, A.; Roda-Robles, E.; Teodoro, A.C. Constraints and potentials of remote sensing data/techniques applied to lithium (Li)-pegmatites. *Can. Mineral.* **2019**, *57*, 723–725. [\[CrossRef\]](#)
33. Yazdi, M.; Taheri, M.; Navi, P.; Sadati, S. Landsat ETM+ imaging for mineral potential mapping: Application to Avaj area, Qazvin, Iran. *Int. J. Remote Sens.* **2013**, *34*, 5778–5795. [\[CrossRef\]](#)
34. Singh, A.; Harrison, A. Standardized principal components. *Int. J. Remote Sens.* **1985**, *6*, 883–896. [\[CrossRef\]](#)
35. Byrne, G.F.; Crapper, P.F.; Mayo, K.K. Monitoring land-cover change by principal component analysis of multitemporal landsat data. *Remote Sens. Environ.* **1980**, *10*, 175–184. [\[CrossRef\]](#)
36. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
37. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3147–3155.
38. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient boosting with categorical features support. *arXiv* **2018**, arXiv:1810.11363.
39. Ge, W.; Cheng, Q.; Jing, L.; Wang, F.; Zhao, M.; Ding, H. Assessment of the capability of sentinel-2 imagery for iron-bearing minerals mapping: A case study in the cuprite area, Nevada. *Remote Sens.* **2020**, *12*, 3028. [\[CrossRef\]](#)
40. Øvereng, O. Kvarts-feltspat-undersøkelser i Hamarøy kommune, Nordland fylke. In *NGU-Rapport (1164/15)*; Norges Geologiske Undersøkelse: Trondheim, Norway, 1974; p. 45.