Using Machine Learning to Identify Volcanic Ash From Satellite Observations (AO06)

Candidate Number : 1033818

Supervisors : Andrew Prata & Roy Grainger

Abstract

In this report the efficacy of the machine learning method of a 'Random Forest Classifier' is investigated for its potential use of identifying volcanic ash from infrared data from the geostationary satellite 'Himawari-8'. The model's performance is compared to the existing, physics based, 'Brightness Temperature Difference (BTD) Method' [16]. It was found that the model can accurately classify ash on previously seen data even when the training set includes a large amount of false positives, as long as the training set is sufficiently large (order 10^6 pixels). In the training set, the false positive rate was 0.29%.

It was also found that the model has some ability to generalise to completely unseen eruptions; however, it would sometimes have a more mixed accuracy and confidence (classification probabilities closer to 50% for Raikoke, 2019; Karymsky, 2020). In other cases (Suwanosejima, Feb 2022), the model was able to generalise excellently and identified the ash cloud with far fewer false positives than would be obtained using the BTD method.

Contents

1 Introduction

1.1 Motivation and Background

The detection of volcanic ash is important for a range of purposes. It is often ejected high into the atmosphere, where it can be carried by the winds and pose a risk to aviation [18]. In addition, the fall of volcanic material can be hazardous for the communities on the ground due to its effect on agriculture and air quality [6]. By improving on existing methods for detecting ash, we can aid the process of conducting retrievals of physical properties of the ash cloud, such as particle size, optical depth, and cloud-top height. In doing so we, may be able to improve the methods by which we predict the hazards caused by ash.

The advent of more advanced geostationary satellites such as Himawari-8 and the GOES-16 (-17) satellites mean that we can analyse volcanic plumes throughout a volcanic eruption at 10-minute temporal resolution, 1-2 km spatial resolution and using 16 different visible and infrared channels. We can therefore acquire a large and varied dataset with which we can train machine learning algorithms to classify volcanic ash in satellite imagery.

Current methods for detecting ash from geostationary satellite imagery make use of a brightness temperature difference (BTD) between the 11.2 and 12.4 μm channels (i.e. T12 − T11) of the infrared sensor [16]. This method relies on the ash having a greater absorption at 11 μm than it does at 12.4 μm, which results in a negative BTD for optically thin, silica-rich ash plumes. The BTD method is widely used, but there are other situations in the atmosphere that result in a negative BTD, which are not caused by ash, particularly cool convective cloud-tops and temperature inversions over cold land surfaces.

This brightness temperature difference has a number of influences which could affect whether it would reliably identify ash, including: instrument noise, the presence of non-volcanic clouds under/over the volcanic cloud, the surface emissivity, and the viewing geometry. The way these affect the detection varies significantly from scene to scene and so a method which detects ash well at one time in an eruption may begin to detect false positives later in the eruption. It is therefore pertinent to examine a way of removing these false positives.

1.2 Machine Learning Methods

Previous efforts have been made to use machine learning methods to detect volcanic ash [19] where a Support Vector Machine (SVM) was trained on data from the eruption of Nishinoshima in 2020 and was able to generalise to the eruption of Raikoke in 2019. It was also identified that such methods could benefit from having a larger and more varied dataset to learn from.

SVMs work by optimising some decision boundary with respect to the data-points in N dimensional space, where N is the number of features [7]. The training time for SVMs scale with the number of data points N as N^a where a lies between 2 and 3 depending on the kernel used [9]. It therefore cannot benefit from enormous amounts of data, since it will take too long to train.

The efficacy of another machine learning method, the Random Forest Classifier (RFC) is examined in the present study. The RFC is an ensemble of decision trees, which all act as 'weak learners' which might only be moderately better than a random guesser. These 'weak learners' vote based on a subset of the features given to it on which class to assign that data point. The fact that each tree is only given a subset of the features is meant to reduce overfitting, which is something that is examined here by testing how well it generalises to unseen eruptions (see section 3.2 'Classification Accuracy on Unseen Examples'). By having several of these learners vote, we can reduce the effect of the learners which voted incorrectly, this is a concept in machine learning called 'Bagging' which is designed to reduce overfitting and variance. The training time for this method scales linearly with the number of data-points, so can be used on a large data-set and still complete training in reasonable time.

1.3 Objectives

Improvements need to be made in terms of the ability for a method to generalise to an entire volcanic eruption, ideally without having to make tweaks to the model for it to produce a consistent classification. In previous work [5], a different machine learning method based on a convolutional neural network (CNN) was studied and found to be effective for identifying the ash plume of one eruption of Mount Etna. It was identified that an important area of further study was the performance of such algorithms on a variety of different eruptions, which is an objective of this report.

2 Method

2.1 Data Collection

2.1.1 The Himawari8 Satellite

In order to train the RFC, we will need to give it a set of features which it will use to make the classification of 'ash' or 'not ash' for a particular pixel. For this investigation, data from the Advanced Himawari Imager (AHI) on the Himawari-8 satellite was used [4], which takes images of the disk of the Earth with the subfocal point at 140.7 degrees east, which covers much of the 'Pacific rim of fire', containing most of the active volcanoes on the planet. Data for the full disk is provided in 10 minute intervals, and has a digitization of 5500 by 5500 pixels for the infrared channels, resulting in a 2 km resolution at the sub-satellite point.

Available are 16 spectral bands, including 3 visible, 3 near-IR and 10 IR channels. For this report the 10 IR channels will be used, as these measurements are available both day and night, which is not the case for the visible and near-IR channels. We expect the 10.4, 11.2, and 12.4 μm channels to be particularly useful as these are in an 'atmospheric window' affected less by common atmospheric gases. The IR absorption and therefore brightness temperature measured by the satellite in each band is affected other atmospheric substances, as is illustrated in figure 1. The details of the 16 bands that the satellite instrument has are summarised in figure 10 in the Appendix.

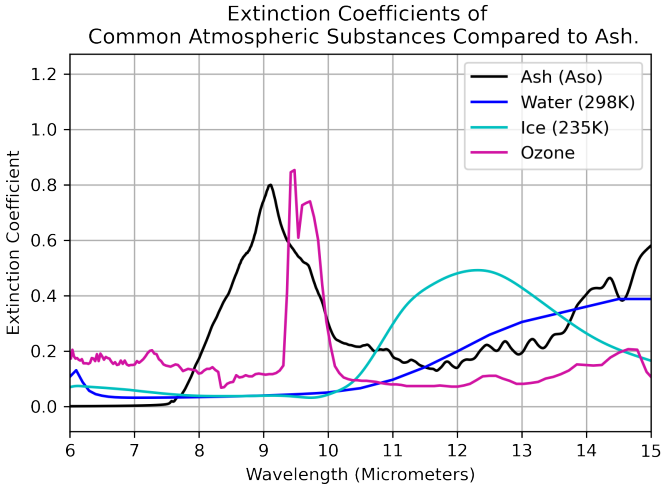


Figure 1: Extinction coefficients of common atmospheric substances including ash from an eruption of mount Aso [2], water [11], ice [1], and ozone [12]. The absorption peak of ozone near 9.5 μm could make it difficult to detect ash in that spectral region. Also note how the absorption of Ash decreases from 10 to 12 μm but for water and ice it increases. This means these substances should produce opposite BTDs.

2.1.2 The Tokyo Volcanic Ash Advisory Centre

The Tokyo Volcano Ash Advisory Centre (VAAC) issues warnings for volcanic ash emitted during eruptions using Himawari-8 data [10]. A polygon roughly enclosing the cloud is described in text and graphic form and is given to aviation authorities. Such a polygon is illustrated in figure 2. The latitude-longitude coordinates can then be extracted from these VAAC warnings. Obviously the polygon is a very coarse description of the ash cloud, so a portion of the polygon will enclose pixels containing no ash. From these we wish to create a labelled data-set where the pixels are assigned a 0 meaning 'not ash' or a 1 meaning 'ash'. For this study a set of slightly less than 1000 VAAC polygons from 2020 and 2021 was used. This means the algorithm will only be able to learn from the volcanoes which erupted in the Tokyo VAAC area in those years. The distribution of training data by volcano and a measure of polygon size is shown in Figure 3.

2.2 Creating a labelled data-set

In previous work [19] the non-ash pixels inside the polygon were labelled as such by taking a water-vapour corrected BTD [15] inside the polygon and labelling only the pixels with a negative BTD as ash. This would however still identify some pixels as being ash even though they did not contain ash (e.g. due to cold cloud-tops or cold land surfaces). This issue motivated the choice to try training the algorithm by simply labelling all of the pixels inside the VAAC polygon as ash to eliminate any systematic mislabelling in favour of random/statistical mislabelling. The final labelled data-set is created by selecting some quantity of feature-sets corresponding to

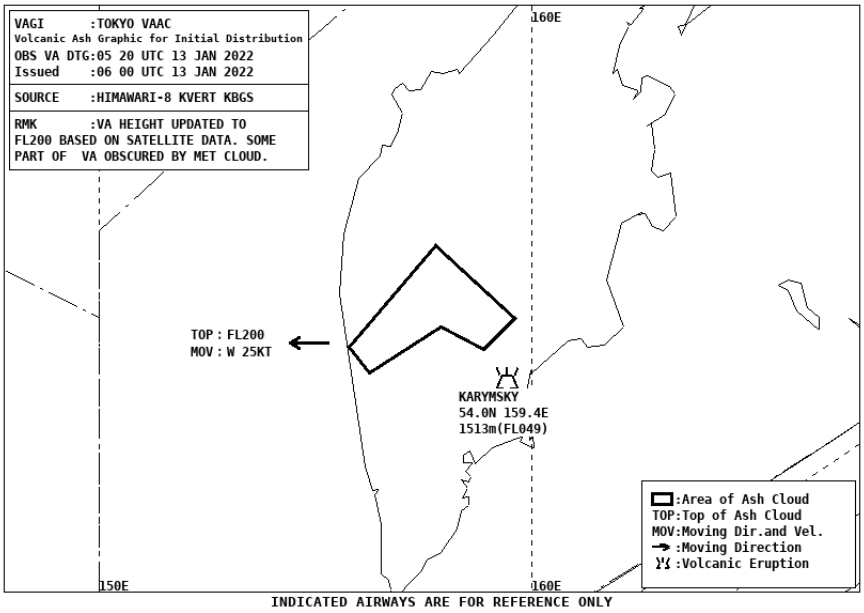


Figure 2: Example of a VAAC graphic

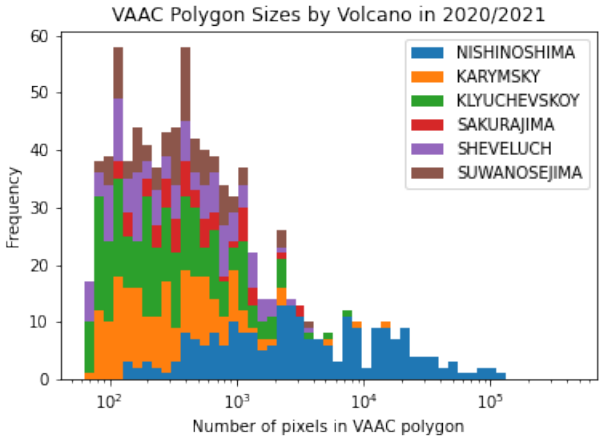


Figure 3: Frequency and size of Tokyo VAAC polygons by volcano (Top 6 shown). Note the prevalence of large polygons from Nishinoshima, which erupted continuously over a period of several weeks in 2020.

pixels inside the VAAC polygon and some much larger quantity of feature sets corresponding to pixels outside the VAAC polygon. This is then repeated for each of the VAAC polygons in the training data.

This will obviously cause the training data to have a large number of false positives, however, it is shown later (Section 3.1) that by vastly outnumbering the false positives with true 'not ash' pixels, this machine learning algorithm can correctly identify the false positives in the training as being 'not ash'. To show the potential issues of systematic mislabelling, an image of an ash cloud partly obscured by water cloud is shown in the $11\mu m$ channel, and also a colour-map showing the 11-12 BTD. As an example, figure 4 shows an image of a Nishinoshima eruption taken by Himawari-8 on 2020/07/30 at 17:20 UTC.

2.3 Training the Random Forest

The RFC used is provided by Scikit-learn [14], as it is the most readily implemented. Others are available e.g. XGBoost , which could potentially be explored in the future. It was found during testing that the RFC's classification accuracy could be improved by including BTDs and contextual information as features in addition to the channel values. The model that was settled on for this study included the IR channel brightness temperatures; the local variance and maximum of the 11 μm BT channel; and the BTDs between the channels at 11.2-12.4μm, 13.3-10.5μm, 10.5-8.6μm, and 11.2-9.6μm. This model was then trained by selecting a random 384 'ash' pixels and a random 3072 'not ash' pixels from a 512 by 512 pixel square centred at the centroid of the VAAC polygon, leading to there being a total of over 3.3 million pixels from 965 different VAAC polygons in the data-set.

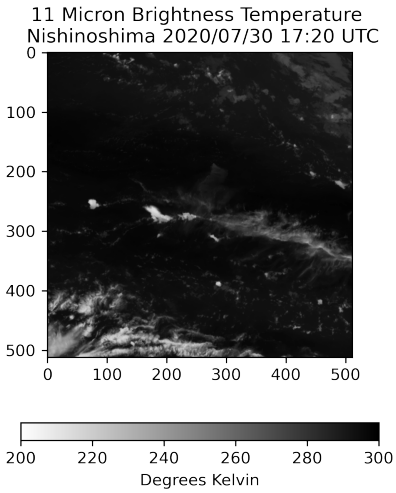


Figure 4: $11\mu m$ channel

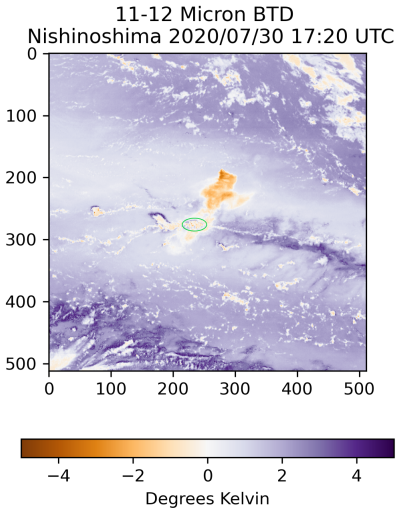


Figure 5: BTD. Highlighted in green is an area of negative BTD which is not ash. In Waugh's SVM, this was both labelled as ash in the training data, and classified as ash by the SVM, which is shown in figure 11 in the appendix.

2.3.1 Choosing hyperparameters

There are a myriad of ways of training the RFC in slightly different ways; a summary of the different variables that can alter the model and how they effect accuracy and usefulness in different scenarios is provided below.

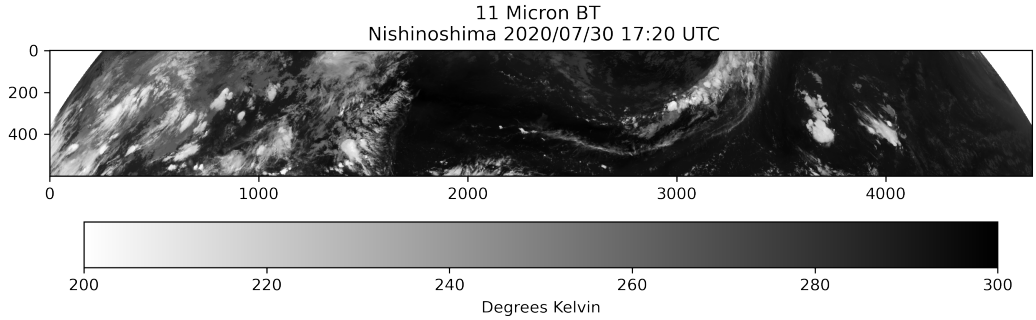
- **Ash:not ash ratio**: Traditionally in this sort of algorithm we desire the number of data-points in each class to be roughly equal. Given the presence of false positives however, the algorithm will benefit from a high number of non-ash pixels. By increasing the number of not-ash pixels in the training, we dilute the false positives in the training data and reduce the false positive rate in the classification. We also however decrease the classification penalty for false negatives, since there is less ash in the training data, which could result in under-classification of ash. In testing it was found that including approximately 10 times as many not ash pixels as ash was generally suitable and consistent with previous work [19].
- **Number of classifiers**: We can select the number of voters to use in the RFC. By including more decision trees we can improve the classification accuracy at a diminishing rate, at the cost of computing time. This computing cost could be reduced by using a more parallelised version of the RFC, such as that included in NVidia's "CuML" package [13]. In testing it was found that any number of classifiers above 10 gave a reasonable accuracy, and 50 was a good balance of speed and accuracy.
- **Training on subsets of data**: We can choose to train on only a subset of the full data-set. For example, by training on only volcanic ash from the arctic volcanoes, the accuracy for predicting ash from future arctic eruptions can be better than when trained on the full data-set, however this would be less suited for predicting ash outside the arctic regions.
- **Using different feature sets**: In testing it was found that including channels which have reflectances (Channels 1-7 on Himawari8) had undesirable effects for ash classification, where the classification of the ash cloud changed significantly between night and day. In the models presented here, the IR channels, the 11 μm variance and maximum, and a selection of BTDs are therefore included. There are many more potential sources of information that the RFC could benefit from, for example: A land/sea mask, the time of day, the solar zenith angle, or the type of volcano. These are not explored in this study but would likely provide some increase in generalisation and accuracy that future work could potentially look at.

For this model, an Not ash:ash ratio of 8 was used, with 64 classifiers, using all polygons, a maximum tree depth of 20, and a maximum of 8 of features for each tree. When training the RFC, a random 20% of the pixels were set aside for testing the classification accuracy.

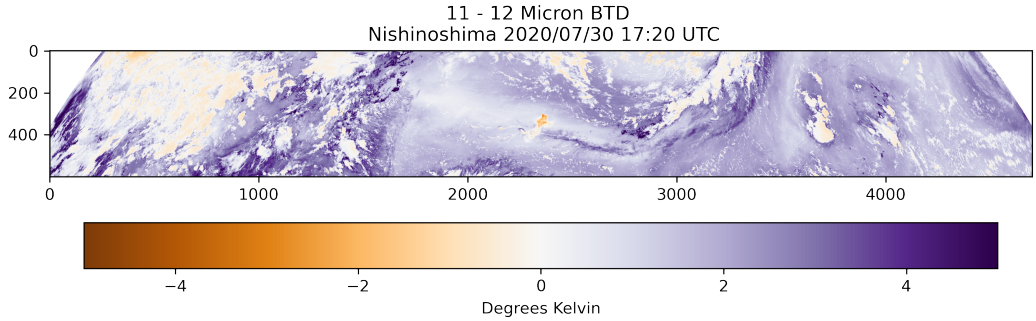
3 Results

3.1 Classification Accuracy on Nishinoshima

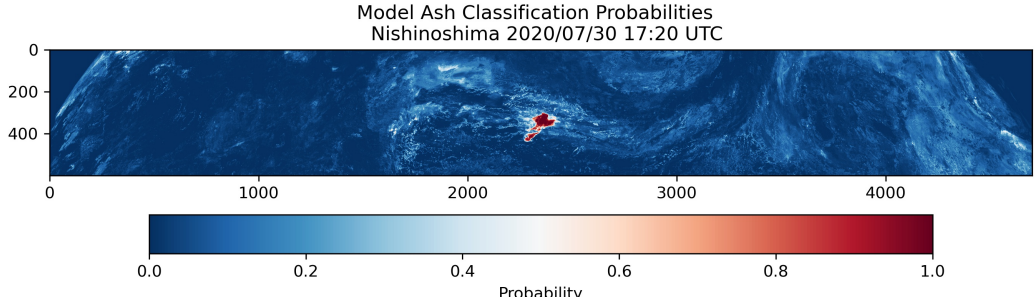
The false negative rate for the RFC described above was 0.29% and the false positive rate was 31%. The false positive rate is quite high since the training data includes an amount of false positives; 31% is not an unreasonable percentage of a normal polygon to contain not-ash. It is more illustrative to show the classification, since the spatial distribution of false positives and negatives in a real image is important, in addition to the rate at which they occur. Shown below is a slice of the full disk image including the Nishinoshima plume. Also shown are images in the 11 μm brightness temperature, the 11-12 BTD, and a colour map showing the ash classification confidence, i.e. the percentage of decision trees which voted 'ash'. Finally, there is a zoomed image of the classification confidence.



(a) 11 μm Brightness Temperature. The ash plume is barely visible in the centre of the image



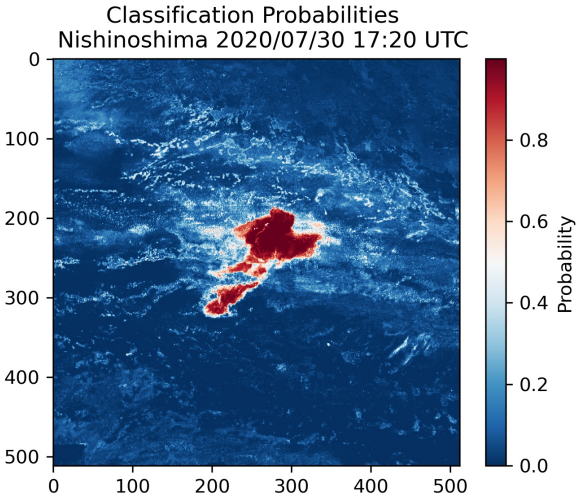
(b) 11-12 μm BTD. Note the similarity in BTD of lots of the convective cloud tops to the ash plume.



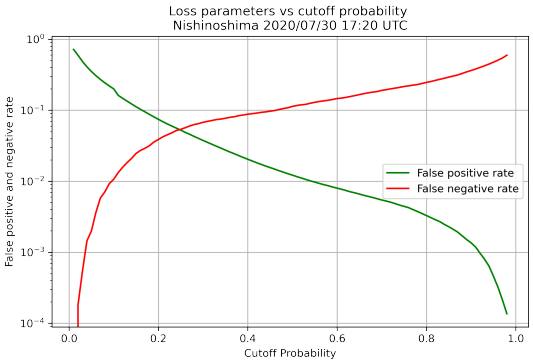
(c) RFC Confidences. Note the fact that there are no regions other than the ash cloud for which the algorithm has confidently identified ash. This indicates that the algorithm is able to generalise to new 'not ash' pixels, meaning there is a low false positive rate.

Figure 6: Comparison of ground truth ash flag to the model classifications for a scene from the eruption of Nishinoshima.

It is very promising that there is a very low false positive rate, as can be seen in the wide image of the RFC confidences. The RFC has correctly identified all of the regions with a negative 11-12 BTD outisde the ash cloud as being not ash, which is already an improvement over using just the BTD method. We should however not be too impressed, as this was included in the training data and could be due to overfitting; we must therefore



(a) RFC Confidences Zoomed In. Notice the correct classification of the obscuring Met Cloud as 'Not Ash'



(b) Graph showing the FPR and FNR of the model compared to the BTD method as a function of the cutoff probability.

Figure 7: 11-12 μm BTD, clearly showing the position of the ash cloud, as well as the loss parameters.

test it on unseen data. For this, the algorithm will be tested on a number of unseen eruptions, including Suwanosejima (Section 3.2). Additional test cases on eruptions of Raikoke and Karymsky are presented in the appendix (Section 5.3).

In order to obtain a binary classification we must take a threshold on the classification probabilities. We might choose 0.5 as our default threshold (i.e. classify any pixel as ash if over half of the decision trees voted 'ash'), but if the classification is very confident, as in this case, we can take a much higher threshold and lower the false positive rate whilst maintaining correct classification of the ash cloud. The ground truth ash flag is generated by labelling pixels as ash if they have a negative water-vapour corrected 11-12 μm BTD (which is determined using the maximum 11 μm BT in the scene), are in the VAAC polygon, and have an 11 μm BT greater than 250K (to exclude the cool met cloud). Figure 7 illustrates the trade-off between increasing the probability threshold, thereby decreasing the FPR, but increasing the FNR. From this we can then compare the false positive and negative rates between the model classification and a BTD threshold method applied to the whole domain instead of just inside the VAAC polygon. Shown in figure 12 in the appendix is a comparison of the ground truth ash flag to the model classification when the probability threshold is taken at 50 and 85%.

Classification Method	False Positive Rate (%)	False Negative Rate (%)
BTD Threshold	2.22	0.00
Model, P>0.5	1.60	6.04
Model, P>0.85	0.44	18.94

Table 1: Comparison of error rates between the model and traditional BTD threshold method. A probability threshold of 0.85 is also investigated since this produces the classification which is most visually similar to the physics based BTD method. As such, its false positive rate is very low.

The accuracy of the model for this scene shows that we can reduce the false positive rate by using this method in preference to the traditional BTD method. If we used a sputtering filter (ieėremoving isolated ash pixels), we would widen the gap in the false positive rate and further improve the performance of the model compared to the BTD method. This is because the model seems to be less likely to incorrectly identify large contiguous regions as ash, whereas the BTD method has incorrectly identified large contiguous regions of ash outside of the true ash cloud.

3.2 Additional Test Case: Suwanosejima

In further testing carried out on a wider range of unseen eruptions, it was found that there were some cases where the RFC produced classifications that would have been very difficult to do using the BTD method alone. Such a case was a small explosive eruption of Suwanosejima, one of the Tokara islands just south of the Japanese mainland. The eruption occured at 12:00 UTC on 2022/02/17, the ash from which soon reached a flight level of 9000ft according to the warning issued by the Tokyo VAAC at 12:41 UTC that day. Figure 8 shows the ash cloud at 16:30 UTC on the same day:

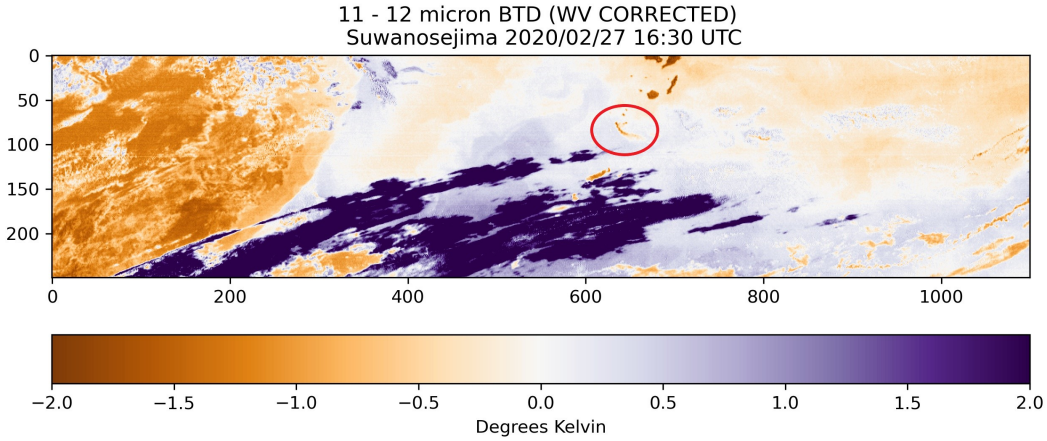


Figure 8: Water-vapour corrected 11-12 μm BTD at 16:30 UTC. The location of the eruption is highlighted in the image by the red ellipse.

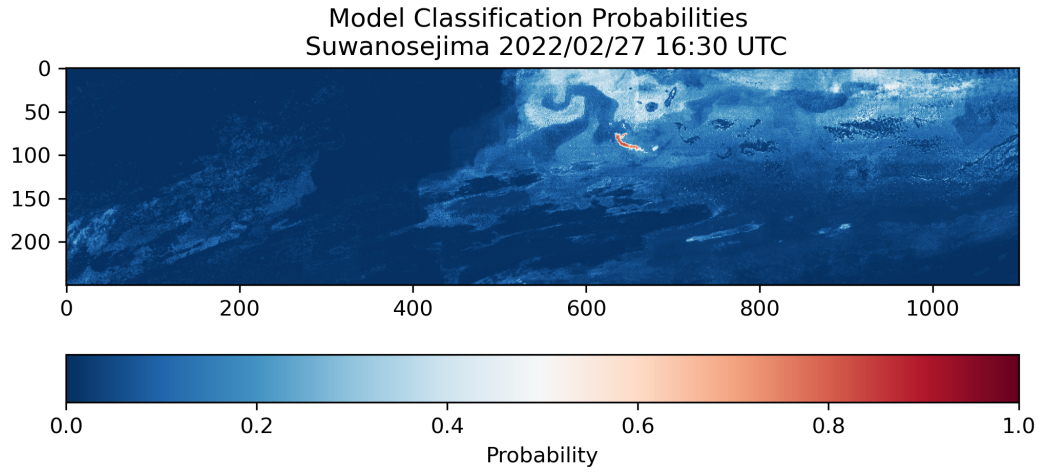


Figure 9: RFC Classification probabilities. Notice that if we were to take a probability threshold of above 0.6, there would be nearly no false positives in the classification within this domain.

Whilst the BTD signal is visible in figure 9 , we could not easily extract the location of the plume automatically without having to restrict our domain significantly, as there would be false positives caused by the islands to the north, and the convective cloud tops to the west. The classifier easily identifies these as being 'not ash', whilst identifying the ash plume very clearly. This presents a clear improvement over the BTD technique for this scenario.

4 Conclusions

In this study we have investigated the performance of the ensemble-based machine learning method, the 'Random Forest Classifier' for classifying volcanic ash from geostationary satellite data. It was found that when the desired data to be classified was from an eruption already in the training data-set, the classification was very good (0.29% FPR). The false positive rates were shown to be less than that of only using the BTD for the eruption of Nishinoshima. We also saw the potential for the model to be used in such cases as the Suwanosejima eruption on 2 Feb 2022. The model detected the ash plume without significant levels of false positives, where using the BTD alone would have detected a huge number of false positives in the vicinity. For these scenarios, the model presents a clear improvement over the BTD method.

In the examples of the Karymsky and Raikoke eruptions (see appendix section 5.3) the model showed less improvement compared to the BTD method, not least because in those scenarios, the nearby atmospheric and ground conditions were such that there were few false positives outside the ash cloud obtained using the BTD method, meaning it would be difficult for any model to improve over the BTD method.

One limitation of the model that has become obvious is that often, in order to get a good classification, we must change the probability threshold from 50% to some other value, depending on how confidently the RFC classifies a particular scene. When a scene (or one very similar e.g. if a scene from the same area at a similar time is included) is in the training data, the classification confidence becomes very high and correspondingly, choosing a higher probability threshold may be required to achieve a low false positive rate. In the case of the Suwanosejima eruption shown, a probability threshold of above 50% was suitable despite it being unseen, which we can explain by the prevalence of such small eruptions from that volcano in the training data.

The fact that the efficacy of the model varies quite significantly from eruption to eruption motivates a number of possible areas of future study or improvement, which might improve such models so that they can be used more generally, a handful of which are presented as follows:

- **Use a sputter filter to remove false positives:** From a quick visual inspection of any of the model ash classifications, we can see that many of the false positives generated by the model are very noisy, and frequently are isolated pixels. It would be simple to remove these isolated pixels and improve the false positive rate.

- **Use a different feature set:** For this model I used the IR channels, the maximum and variance of the 11 μm BT, and a somewhat arbitrary selection of the BTDs between the channels at 11.2-12.4μm, 13.3-10.5μm, 10.5-8.6μm, and 11.2-9.6μm. This could easily be expanded to include all combinations of BTDs, the visible channels, as well as more contextual features, such as land cover or local zenith angle. Adding these would likely make diminishing improvements to the models accuracy, and would allow the study of which features were most important to the classifier for classifying ash.

- **Use different models:** The RFC used in this report was the one provided by Sci-kit Learn. Other similar, more modern models such as XGBoost are popular and might improve accuracy.

- **Use a convolutional neural network:** The current two most popular machine learning techniques for semantic segmentation are CNNs and RFCs. It would therefore be worth examining how a CNN would compare against the RFC; in particular it would have the benefit of being able to use more contextual information. For this task, a CNN incorporating residual layers with a U-like structure would likely be useful. This type of architecture, a 'ResUnet' has already been used for semantic segmentation of satellite data [3].

- **Combine the RFC with a CNN:** The RFC provides very nearly correct classifications in a number of cases. A possible idea to improve its accuracy would be a train a CNN to 'fix' the errors of the RFC, by using contextual information. I expect this would aid in noise reduction, and could even potentially be used to draw in where the ash cloud likely is, even if it is obscured somehow.

In summary, this project has demonstrated the promise of AI in advancing the methods for detecting atmospheric volcanic ash beyond the current state-of-the-art.

References

[1] A. Y. Zasetsky et. al. "Frequency Dependent Complex Refractive Indices of Supercooled Liquid Water and Ice Determined from Aerosol Extinction Spectra." In: *Journal of Physical Chemistry* 109 (12 2005), pp. 2760–2764. URL: https://doi.org/10.1021/jp044823c.

[2] B.E Reed et. al. "The complex refractive index of volcanic ash aerosol retrieved from spectral mass extinction". In: *Journal of Geophysical Research* 123 (2018), pp. 1339–1350. URL: https://doi.org/10.1002/2017JD027362.

[3] D. Watson-Parris et. al. "A Large-Scale Analysis of Pockets of Open Cells and Their Radiative Impact". In: *Geophysical Research Letters* 48 (2021). URL: https://doi.org/10.1029/2020GL092213.

[4] K. Bessho et. al. "An Introduction to Himawari-8/9 — Japan's New-Generation Geostationary Meteorological Satellites". In: *Journal of the Meteorological Society of Japan* 94 (2 2016), pp. 151–183. URL: https://doi.org/10.2151/jmsj.2016-009.

[5] L.Liu Et. Al. "Monitoring of volcanic ash cloud from heterogeneous data using feature fusion and convolutional neural networks–long short-term memory". In: *Neural Computing and Applications* 33 (2021), pp. 667–678. URL: https://doi.org/10.1007/s00521-020-05050-y.

[6] T.M. Wilson et. al. "Ash storms: impacts of wind-remobilised volcanic ash on rural communities and agriculture following the 1991 Hudson eruption, southern Patagonia, Chile". In: *Bull Volcanol* (2010).

[7] Himani Bhavsar. "A Review on Support Vector Machine for Data Classification". In: *IJARCET* 1 (10 2012).

[8] A. Bokhovkin and E. Burnaev. "Boundary Loss for Remote Sensing Imagery Semantic Segmentation". In: *Advances in Neural Network*. ISNN 2019 (2019).

[9] A. Bordes. "Fast Kernel Classifier with Online and Active Learning". In: *Journal of Machine Learning Research* 6 (2005), pp. 1579–1619.

[10] Tokyo Volcanic Ash Advisory Center. *Volcanic Ash Advisory*. 2022. URL: https://ds.data.jma.go.jp/svd/vaac/data/Inquiry/graphic_and_dispersion.htm.

[11] George M. Hale and Marvin R. Querry. "Optical Constants of Water in the 200-nm to 200-m Wavelength Region". In: *Journal of Applied Optics* 12 (1973), pp. 555–563. URL: https://doi.org/10.1364/AO.12.000555.

[12] *IR Spectrum of Ozone*. 1969. URL: https://webbook.nist.gov/cgi/cbook.cgi?ID=C10028156&Mask=80#IR-Spec.

[13] NVidia. *CuML Documentation*. 2020. URL: https://docs.rapids.ai/api/cuml/stable/.

[14] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[15] A. Prata and I. Grant. "Determination of mass loadings and plume heights of volcanic clouds from satellite data". In: *CSIRO Atmospheric Research* 48 (2001).

[16] A.J. Prata. "Observations of volcanic ash clouds in the 10-12 µm window using AVHRR/2 data". In: *Int. J of Remote Sensing* (1989).

[17] Adam Voiland. *Raikoke Erupts*. 2019. URL: https://www.earthobservatory.nasa.gov/images/145226/raikoke-erupts.

[18] W.R.Chen. "Volcanic ash and its influence on aircraft engine components". In: *Procedia Engineering* (2014).

[19] Michael Waugh. "Using AI to Classify Volcanic Ash in Satellite Observations". In: (2021).

5 Appendix

5.1 Himawari8 Information

Band	Central wavelength	Bandwidth	SNR or NEΔT @ specified input	Resolution at SSP (Sub Satellite Point)	Prime measurement objectives and use of sample data
1	455 nm	50 nm	≤ 300 @ 100 % albedo	1.0 km	Daytime aerosol over land, coastal water mapping
2	510 nm	20 nm	≤ 300 @ 100 % albedo	1.0 km	Green band – to produce color composite imagery
3	645 nm	30 nm	≤ 300 @ 100 % albedo	**0.5 km**	Daytime vegetation/burn scar and aerosols over water, winds
4	860 nm	20 nm	≤ 300 @ 100 % albedo	1.0 km	Daytime cirrus cloud
5	1610 nm	20 nm	≤ 300 @ 100 % albedo	2.0 km	Daytime cloud-top phase and particle size, snow
6	2260 nm	20 nm	≤ 300 @ 100 % albedo	2.0 km	Daytime land/cloud properties, particle size, vegetation, snow
7	3.85 μm	0.22 μm	≤ 0.16 @ 300 K	2.0 km	Surface and cloud, fog at night, fire, winds
8	6.25 μm	0.37 μm	≤ 0.40 @ 240 K	2.0 km	High-level atmospheric water vapor, winds, rainfall
9	6.95 μm	0.12 μm	≤ 0.10 @ 300 K	2.0 km	Mid-level atmospheric water vapor, winds, rainfall
10	7.35 μm	0.17 μm	≤ 0.32 @ 240 K	2.0 km	Lower-level water vapor, winds and SO_2
11	8.60 μm	0.32 μm	≤ 0.10 @ 300 K	2.0 km	Total water for stability, cloud phase, dust, SO_2, rainfall
12	9.63 μm	0.18 μm	≤ 0.10 @ 300 K	2.0 km	Total ozone, turbulence, winds
13	10.45 μm	0.30 μm	≤ 0.10 @ 300 K	2.0 km	Surface and cloud
14	11.20 μm	0.20 μm	≤ 0.10 @ 300 K	2.0 km	Imagery, SST, clouds, rainfall
15	12.35 μm	0.30 μm	≤ 0.10 @ 300 K	2.0 km	Total water, ash, SST
16	13.30 μm	0.20 μm	≤ 0.30 @ 300 K	2.0 km	Air temperature, cloud heights and amounts

Figure 10: Summary of the Himawari8 Bands [4]

5.2 Figures

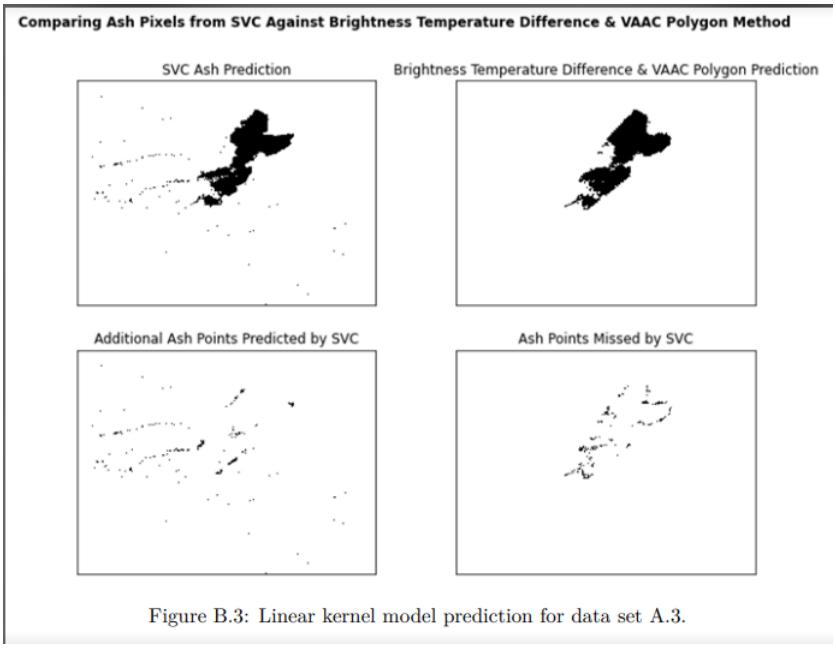
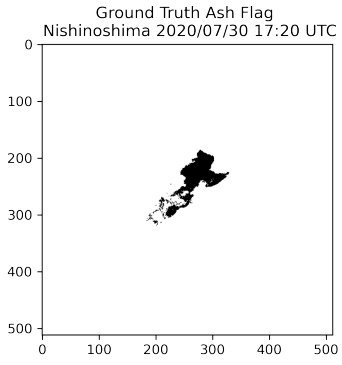
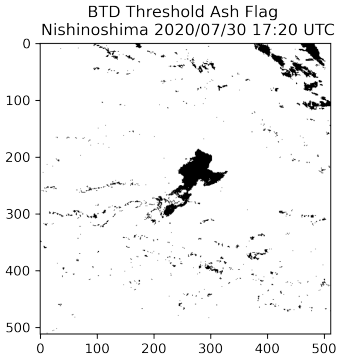


Figure B.3: Linear kernel model prediction for data set A.3.

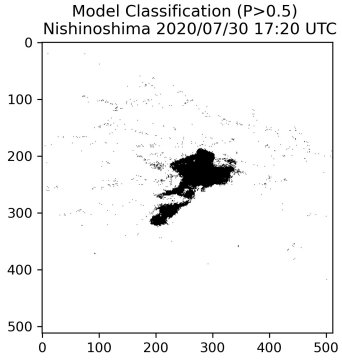
Figure 11: Figure taken from [Waugh 2021]



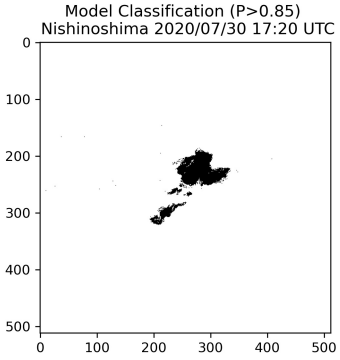
(a) Ground Truth (Note the difference to the ground truth in Waugh's paper where the met cloud is included in the ground truth)



(b) Classification of scene using BTD threshold combined with a BT threshold to remove the met clouds.



(c) Model classification if the probability threshold is set to 50%. Note that there are a lot of noisy false positives. These might be removed in post-processing by a sputter filter.



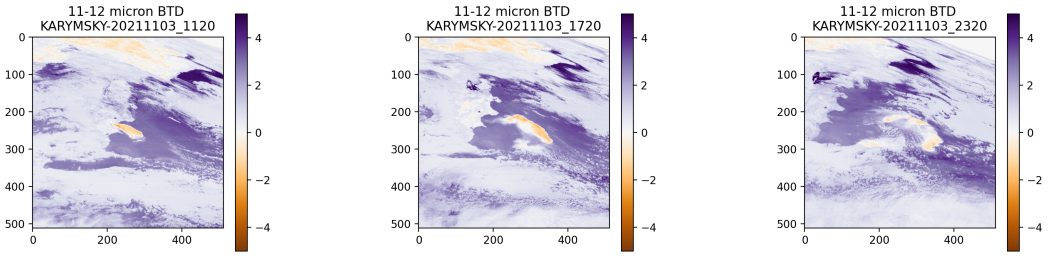
(d) Model classification if the probability threshold is set to 85%. The number of false positives has dropped significantly at the cost of a few false negatives.

Figure 12: Comparison of ground truth ash flag to the model classifications for a scene from the eruption of Nishinoshima.

5.3 Analysis of accuracy on additional eruptions

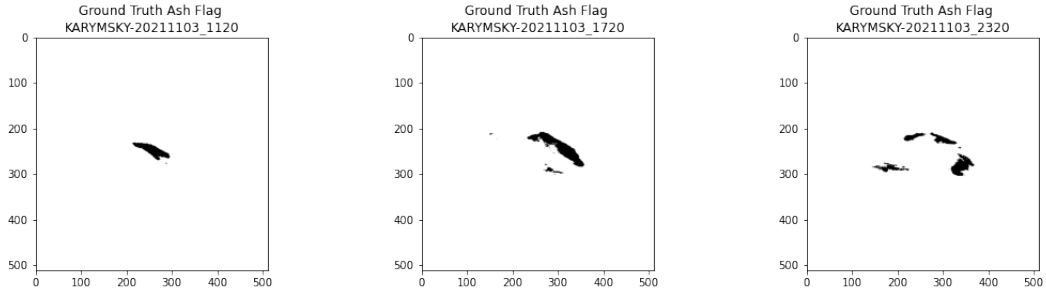
5.3.1 Karymsky

On 2021/11/03 at 07:30 UTC, Karymsky erupted explosively, with the ash cloud reaching an altitude of 28,000ft and extending over around 400km by 23:20 UTC according to the Tokyo VAAC warning. Initially this was within the training data-set. In order to see how the classifier performs were this to have been unseen, all of the training data pertaining to Karysmky eruptions after this time were removed and the RFC was retrained. The classifier was then run on 3 scenes where VAAC polygons were issued, at 11:20 UTC, 17:20 UTC and at 23:20 UTC. The 11-12 BTD signal clearly identified the ash in this eruption, so to validate it, a BTD threshold within the VAAC polygon as the 'Ground Truth' is used. The 11-12 BTD is shown for the 3 scenes described in figure 13 and the corresponding 'Ground Truths' in figure 14.



(a) Scene 1, 11:20 UTC **(b)** Scene 2, 17:20 UTC **(c)** Scene 3, 23:20 UTC

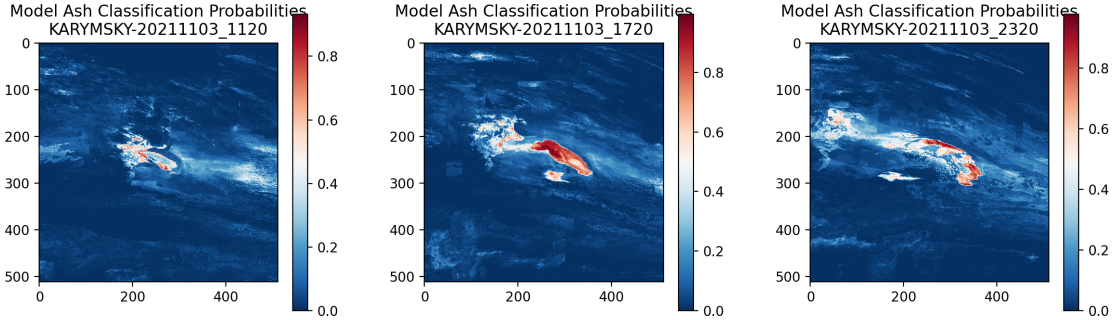
Figure 13: 11-12 μm BTD, clearly showing the position of the ash cloud.



(a) Scene 1, 11:20 UTC **(b)** Scene 2, 17:20 UTC **(c)** Scene 3, 23:20 UTC

Figure 14: Ground truth ash classification for the 3 scenes

These can then be compared against the model classification probabilities, and the classification given some threshold. If we change this threshold from the default value of 0.5, we can obtain more or less sensitive detections than what would normally be output by the classifier.



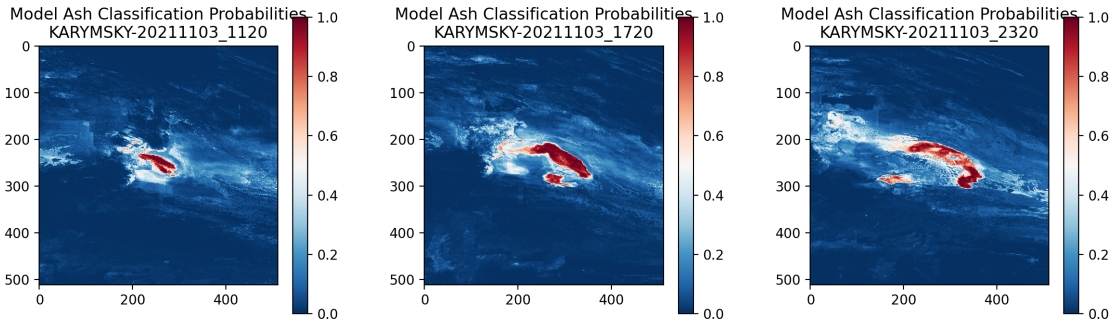
(a) Scene 1, 11:20 UTC **(b)** Scene 2, 17:20 UTC **(c)** Scene 3, 23:20 UTC

Figure 15: RFC ash classification for the 3 scenes when those scenes were excluded from the data-set

Visually we can see in figure 15 that the classifier seems to over-classify land pixels as being ash during the

nighttime, and seems to under-classify the ash-cloud early in the eruption. It does however perform better once the ash cloud has thinned out in scenes 2 and 3.

For comparison the classifier is also trained including these scenes, and the difference is shown in figure 16.



(a) Scene 1, 11:20 UTC **(b)** Scene 2, 17:20 UTC **(c)** Scene 3, 23:20 UTC

Figure 16: RFC ash classification for the 3 scenes when those scenes were not excluded from the data-set. Note the increase in classification confidence and decrease in false positives over the coast of Kamchatka

It is clear to see that once these are included in the training data, the classifier becomes much more confident and accurate in identifying ash. This is despite the fact that the training data given was the VAAC polygon and not the BTD threshold. It can therefore be concluded that the RFC is able to identify the false positives within the coarse VAAC polygon description of the ash cloud. There is a subtlety by what we mean by 'accurate' however. In addition to the false positive and false negative rate, we can also use a loss function called the 'Intersection Over Union Loss' ,which is often used in computer-vision semantic segmentation [8]. An IOU loss of 0 corresponds to a perfect match and an IOU loss of 1 is a total mismatch. How this loss function works is shown in figure 17:

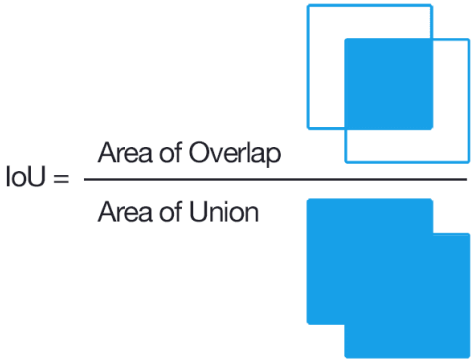
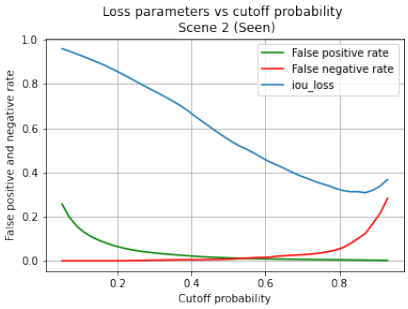


Figure 17: The IOU is the area of overlap divided by the total area. The IOU loss is 1 - IOU.

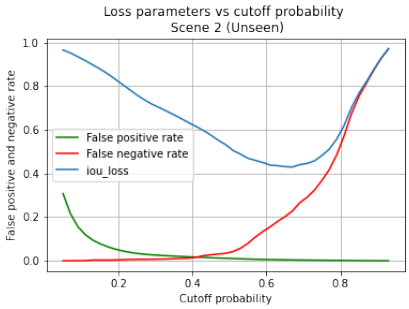
How the IOU loss, false positive rate (FPR), and false negative rate (FNR) vary with cutoff probability is then plotted for scene 2 when the scene is given to, and withheld from the algorithm during training (Figure 18). Notable is the fact that, at the point where the IOU loss is lowest, the false positive rate is extremely low, which presents a possible advantage over the BTD threshold method.

We can compare the false positive and negative rates for both of these models with the traditional BTD threshold, as was done for the Nishinoshima eruption.

The low false positive rates of the BTD method indicate this was an eruption for which using the raw BTD method would be very suitable, and using the RFC would likely be unnecessary. For scenes 1 and 2, the model could produce a better false positive rate than the raw BTD method if it had seen the data before, but at the cost of a fairly large false negative rate.



(a) Scene 2 (Karymsky 17:20 UTC), Data seen.



(b) Scene 2 (Karymsky 17:20 UTC), Data unseen.

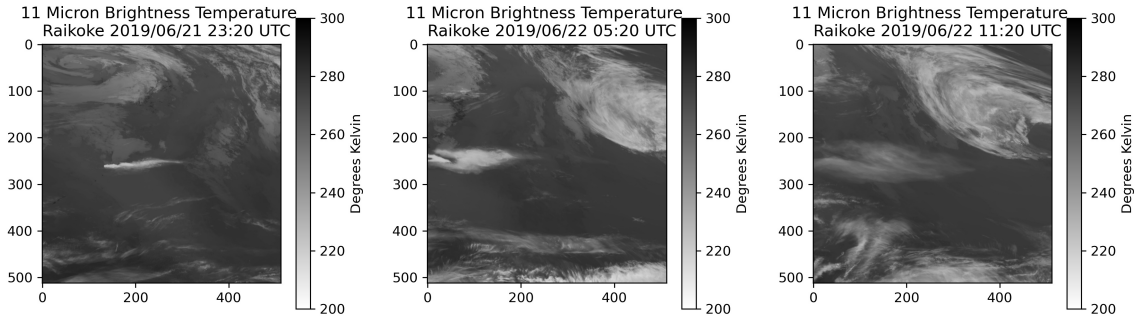
Figure 18: Loss parameters as a function of cutoff probability. Notice that the minimum IOU loss is achieved in both cases for probabilities above 50%.

Scene	Classification Method	False Positive Rate (%)	False Negative Rate (%)
Scene 1	BTD Threshold	0.03	0.00
//	Unseen Model, P>0.5	0.52	72.94
//	Unseen Model, P>0.65	0.08	87.87
//	Seen Model, P>0.5	0.51	1.71
//	Seen Model, P>0.85	0.04	22.08
Scene 2	BTD Threshold	0.43	0.00
//	Unseen Model, P>0.5	1.50	6.48
//	Unseen Model, P>0.65	0.40	12.20
//	Seen Model, P>0.5	1.29	0.72
//	Seen Model, P>0.85	0.33	10.05
Scene 3	BTD Threshold	0.00	0.00
//	Unseen Model, P>0.5	0.88	27.97
//	Unseen Model, P>0.65	0.28	50.74
//	Seen Model, P>0.5	2.15	1.98
//	Seen Model, P>0.85	0.26	33.76

Table 2: Comparison of error rates between the model and traditional BTD threshold method for the Karymsky eruption. The error rates using the BTD threshold method also incorporated a geographical filter to remove the large number of false positives in the very north of the images.

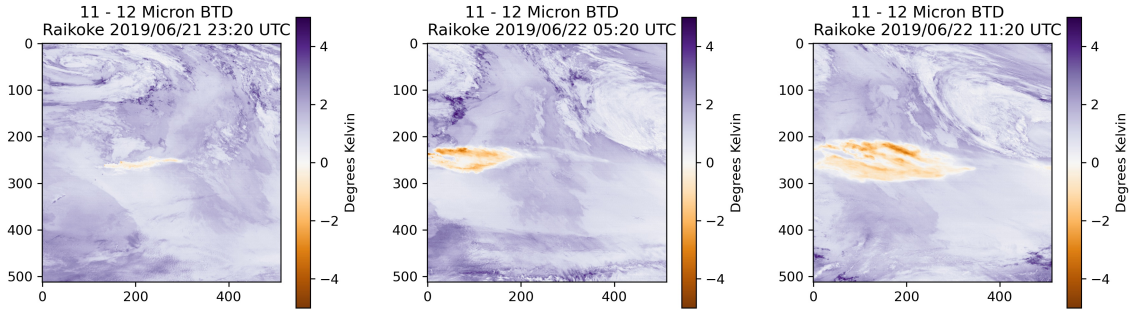
5.3.2 Classification Accuracy on Raikoke Eruption

The Raikoke Volcano is located among the Kuril islands at 48.3^{o}N 152.3^{o}E. The volcano began erupting at 18:05 UTC on 2019/06/21 and continued to erupt until 05:40 UTC on 2019/06/22. This eruption released a very significant quantity of ash which reached between 13 and 17km in height, well into the stratosphere [17]. The same procedure of running the classifier when the data is unseen was carried out for 3 scenes. The 11-12 μm BTD and the 11 μm BT for these scenes are also shown for comparison in figure 20 and figure 19 respectively. The loss parameters versus cutoff probability for the model classifications of these scenes are plotted in figure 21.



(a) Scene 1, 2019/07/21 23:20 UTC **(b)** Scene 2, 2019/07/22 05:20 UTC **(c)** Scene 3, 2019/07/22 11:20 UTC

Figure 19: 11 μm BT at 3 times for the 2019 eruption of Raikoke

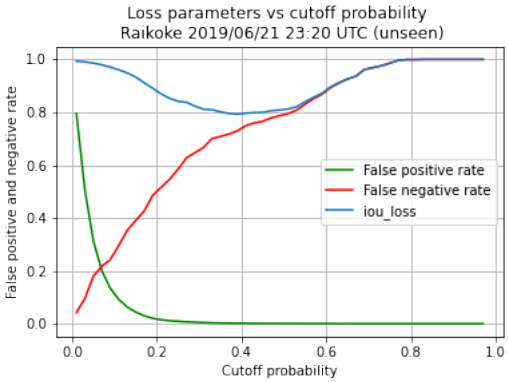


(a) Scene 1, 2019/07/21 23:20 UTC **(b)** Scene 2, 2019/07/22 05:20 UTC **(c)** Scene 3, 2019/07/22 11:20 UTC

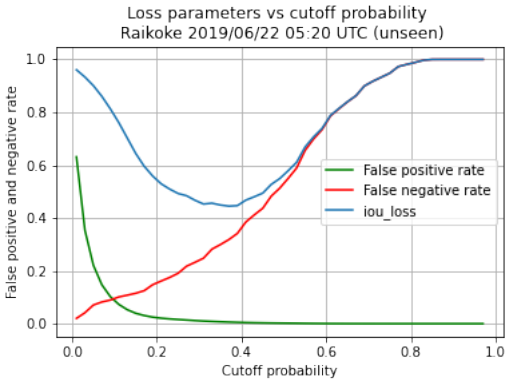
Figure 20: 11 - 12 μm BTD at 3 times for the 2019 eruption of Raikoke

Using a probability cutoff of 0.4, we can compare the model's classification to the ground truth, which is obtained by taking the pixels with a negative 11-12 μm BTD inside the Tokyo VAAC polygons at each respective time.

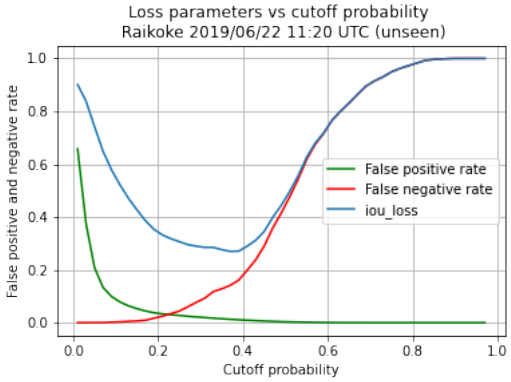
Comparison of these images show that the RFC does have some capacity to generalise to a completely unseen eruption, however its confidence can suffer significantly as a result. In the case of the 1st scene, it fails to identify a significant portion of the ash. For the 3rd scene however, the classification is visually similar, and there are almost no visible false positives outside of the plume. The false positives that do occur are frequently isolated pixels, which might be removed in post-processing with a 'sputter filter'. We can again compare the BTD method to the model classification in a table. We also include the model accuracy when the model is given these 3 scenes in its training data-set for comparison.



(a) Scene 1

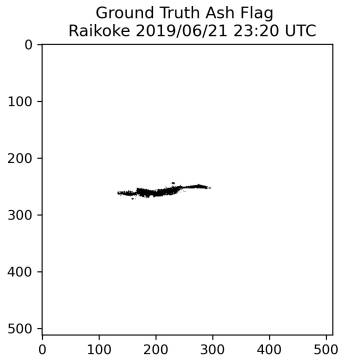


(b) Scene 2

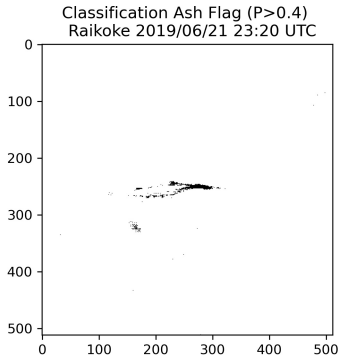


(c) Scene 3

Figure 21: Loss parameters (FNR, FPR, IOU loss) for 3 Raikoke scenes. Note that the best IOU loss is obtained for P>0.4. This indicates a low overall classification confidence.

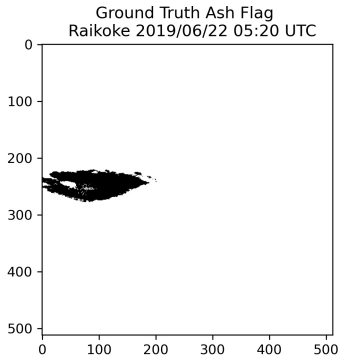


(a) Scene 1

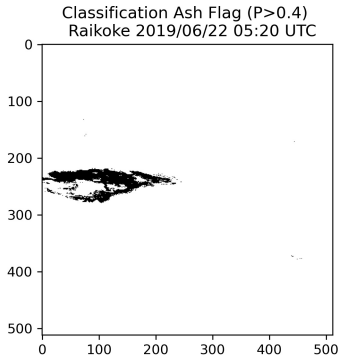


(b) Scene 1

Figure 22: Comparison of BTD threshold inside VAAC polygon to model classification at 23:20 UTC on 2019/06/21. The classification is very poor in the optically thick centre of the plume even though the BTD signal is still clear.

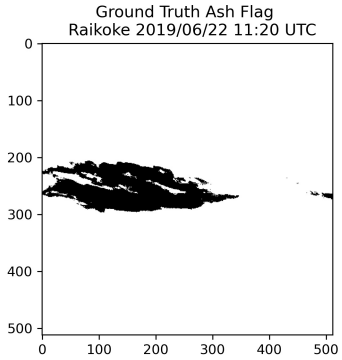


(a) Scene 2

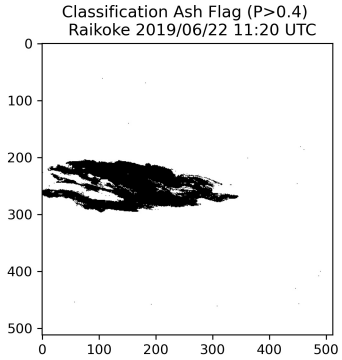


(b) Scene 2

Figure 23: Comparison of BTD threshold inside VAAC polygon to model classification at 05:20 UTC on 2019/06/22. The ash has begun the thin out by this point, and a greater region of the plume is now accurately identified. There is still a portion being missed in the centre of the plume however.



(a) Scene 3



(b) Scene 3

Figure 24: Comparison of BTD threshold inside VAAC polygon to model classification at 11:20 UTC on 2019/07/21. The ash classification is best here once the ash has become very thin, perhaps this is because there are few examples of optically thick ash in the training data.

Scene	Classification Method	False Positive Rate (%)	False Negative Rate (%)
Scene 1	BTD Threshold	0.02	0.00
//	Unseen Model, P>0.5	0.06	79.19
//	Unseen Model, P>0.4	0.15	74.47
//	Seen Model, P>0.5	0.82	18.42
//	Seen Model, P>0.85	0.21	63.38
Scene 2	BTD Threshold	0.08	0.00
//	Unseen Model, P>0.5	0.21	51.92
//	Unseen Model, P>0.4	0.47	35.97
//	Seen Model, P>0.5	2.30	8.12
//	Seen Model, P>0.85	1.26	28.35
Scene 3	BTD Threshold	0.04	0.00
//	Unseen Model, P>0.5	0.35	44.23
//	Unseen Model, P>0.4	1.07	17.37
//	Seen Model, P>0.5	4.58	2.32
//	Seen Model, P>0.85	2.28	21.97

Table 3: Comparison of error rates between the model and traditional BTD threshold method for the Raikoke Eruption.

Again, like the Karymsky eruption, the BTD method is already very good at picking out the ash in the eruption, at least for these 3 scenes. The model is still able to produce good false positive rates if we select the threshold probability carefully. Interestingly, when we include these scenes in the training data-set, the false positive rate actually increases. Some investigation into this revealed that this was due to the algorithm identifying a large SO2 cloud to the east of the ash cloud, which was included in the VAAC polygon but had little to no identifiable ash.

5.4 Acronyms

- **BTD**: Brightness Temperature Difference - The difference in brightness temperatures between two infrared channels.
- **BT**: Brightness Temperature - The radiating temperature for a particular wavelength band.
- **SVM**: Support Vector Machine.
- **RFC**: Random Forest Classifier.
- **VAAC**: Volcanic Ash Advisory Committee.
- **CNN**: Convolutional Neural Network