A016: Using machine learning to identify volcanic ash plumes from satellite observations

Candidate number: 1045837
Supervisors: Dr. I. Taylor and Prof R. Grainger

Abstract

In this project the feasibility of training an XGBoost machine-learning classification model on satellite-obtained thermal infrared (TIR) spectra for fast, remote ash plume detection and monitoring is explored. The satellite data used in this project was obtained using the Infrared Atmospheric Sounder Interferometer (IASI) and in order to obtain accurately labelled training data, brightness temperatures (BTS) are simulated using RTTOV (Radiative Transfer for TOVS).

The model is evaluated against both a simulated test spectra and a case study plume: the eruption of Raikoke (Russia) 06/2019. It was found the model was able to well generalise the relationship between ash and its infrared spectra, achieving an $F1$ score of 81% on the test data. The final model has a small bias towards overprediction which is suspected to be a result of the diversity of the 'not-ash' class in the training data. This suggests that machine learning is a suitable tool for satellite-based remote sensing of ash. However, the model was less able to accurately predict the presence of ash when applied to the Raikoke case study which indicates the need for several methodological improvements.

1 Introduction

Volcanic emissions have a considerable impact on the Earth's radiation balance, air quality, and pose an immediate threat to aviation [1]. Explosive eruptions produce large quantities of aerosols and gases, with fine volcanic ash ($< 63\,\mu$m diameter) often accounting for more than 50% of the mass proportion [2]. Correspondingly, fast and accurate detection of ash particles is essential for informing live hazard management protocols [3].

A variety of ground- and satellite-based remote sensing tools form an important part of effective monitoring of ash in the atmosphere, and optimal approaches combine the data from different sources to minimise the error associated with the limitations of each technique [4, 5]. This project focuses on satellite-based remote sensing of ash, for which there continues to be a strong incentive to develop new methods and improve existing techniques [6].

Satellite observations are particularly advantageous as they provide regular views of the same area, facilitating the characterisation of plume evolutions of ash plumes with high spatial and temporal variability [1]. Many characteristic absorption features of volcanic ash are found at thermal infrared (TIR) wavelengths, therefore satellite-based infrared spectrometers are especially useful [7].

The data used in this project was obtained using the Infrared Atmospheric Sounder Interferometer (IASI), a hyperspectral fourier transform spectrometer on the MetOp series of polar orbiters [4]. IASI delivers an exceptional wealth of spectral information, measuring the TIR emitted by the Earth and atmosphere in nadir mode between 645 and 2760 cm^{-1} at a sampling of 0.25 cm^{-1} [8]. Additionally, it offers wide spatial coverage at a reasonable spatial resolution (pixel width of 12 km) and temporal resolution (a global scan once every 12h) [8]. In this report, the feasibility of training a machine learning model on the information-dense spectral data collected by IASI is explored for use in operational, quick ash cloud detection and monitoring.

Most current methods for remotely detecting ash/retrieving ash properties find their origin in the reverse absorption method, also known as the split window technique [9]. This exploits the fact that volcanic ash is typically high in SiO_2 which produces a strong absorption feature in the $9.5\,\mu$m region, whereas

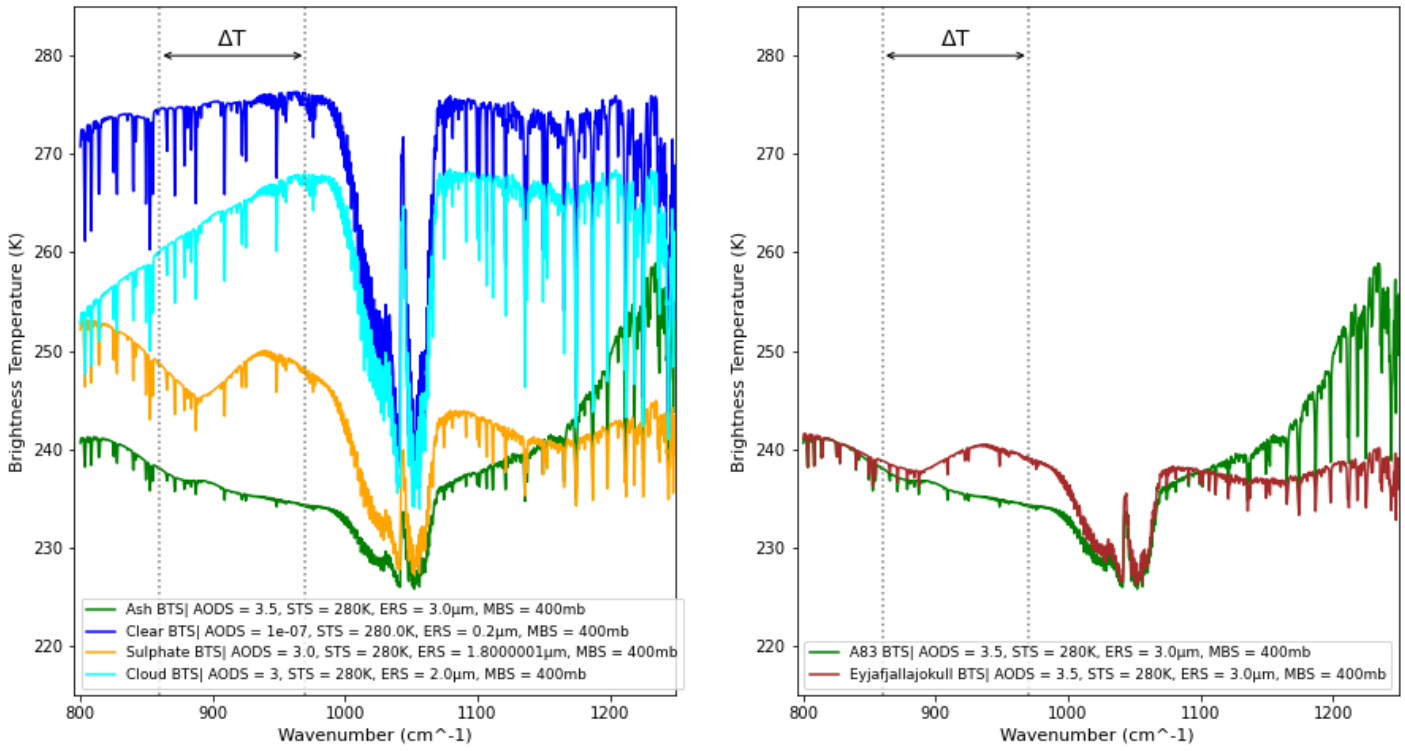


Figure 1: Comparison of top-of-atmosphere brightness temperature spectra as would be measured by IASI (simulated using RTTOV. See Section 2.1.1). The dotted lines indicate the region used for the reverse absorption method. (L) Simulations of Clear, Cloudy, Sulphate and Ash; (R) Simulations of two different ash samples characterised by their refractive index. In the case of Eyjafjallajökul (2010) $\Delta T < 0$ is an insufficient criterion for ash detection. Key: AODS = Ash Optical Depth at $550nm$; STS = Surface Temperature; ERS = Effective Particle Radius; MBS = Atmospheric pressure (proxy for cloud top height).

the ice/water vapour in meteorological clouds absorbs preferentially towards 12 μm [2]. Following Prata (1989), we can define ΔT as T(10.3-11.3 μm) – T(11.5-12.5 μm) where T(λ) are the observed brightness temperatures (BTS) measured in the wavelength range λ [10]. It is suggested that $\Delta T < 0$ is a necessary criterion for the presence of ash, and that for non-volcanic clouds $\Delta T > 0$. Qualitatively, a negative gradient in this region is indicative of the presence of ash, as is illustrated in Figure 1a).

It is well known that this criterion is neither necessary nor sufficient to determine the presence of ash. Windblown mineral dust is one of the most abundant aerosols on Earth [6], and has very similar spectral characteristics to volcanic ash in the TIR, leaving the reverse absorption method vulnerable to over-detection [9]. False alarms can also be caused by a variety of other environmental factors such as temperature inversions near the Earth's surface and stratospheric thunderstorms [5]. Conversely, high concentrations of water or ice in the atmosphere, or the plume itself, and thick volcanic plumes that approach optical saturation, can have a masking effect on the signal leading to $\Delta T \geq 0$ and cause under-detection [9]. This masking effect can be seen in the spectra observed from the eruption of Eyjafjallajökull (2010) in figure 1b), illustrating the need for more sophisticated detection algorithms than the reverse absorption method alone.

There have been numerous adaptations to this technique that aim to address these problems (a survey of several of these advancements can be found in Clarisse and Prata (2016) [11]). Ultimately, these methods all suffer from the fact that 'ashy' atmospheric states are too diverse to contain one clear spectral signature, the exact shape of the spectra depending on many factors (such as particle radii, the optical and geometric thickness of the cloud, and mineral composition) that vary between and within eruptions [12].

In this project, a machine-learning method is selected as an avenue to overcome these challenges. Specifically, an eXtreme Gradient Boosting (XGBoost) model [13] is developed as an approach to satellite-based ash plume detection. XGBoost is a popular machine learning model that has seen great suc-

cess in many research problems across scientific disciplines, but to date there have been few applications in satellite-based aerosol retrievals or classification.

It's important to note that despite the media attention, machine learning is not always the best way to solve research problems. However, in this case, there are many potential advantages to applying machine learning due to the range of possible absorption spectra for ash clouds and corresponding variety of potentially identifying spectral features. Machine learning techniques can typically be divided into 'classical' and 'deep learning'. Classical algorithms build structurally understandable models by learning the parameters directly from the features of the training sample. Deep learning can be advantageous for high complexity problems, but the structure of resultant models is typically uninterpretable [14, 15]. XGBoost is a classical model which facilitates investigation into the specific channels that are driving the classification, and thus can provide scientific insight for future ash retrieval algorithms.

In addition, many current satellite-based retrieval algorithms rely on computationally demanding radiative transfer models that are not suited for real-time ash detection as is required for live hazard mitigation. Correspondingly, faster approximations are adopted with a substantial sacrifice in accuracy (such as the linear retrieval used by Sears et al. (2013) [16]). Advanced machine learning algorithms such as XGBoost are capable of learning the complex relationships encoded in more computationally demanding forward models, yet once trained, take a fraction of the time and processing power to be applied, facilitating real-time detection.

In this paper, a description of the training sample selection and assembly is given in Section 2.1. In Section 2.2, we introduce the XGBoost algorithm and the process of model construction. In Section 2.4 we describe the process of model validation. In Section 3.1, we describe the performance of our model when validated against independently simulated data, as well as analysing the spectral features most used in classification. In Section 3.3, our XGBoost prediction model is applied to a case study eruption: Raikoke (2019)). Section 4 contains a conclusion and discussion of the results.

2 Method

2.1 Data

In order to apply supervised learning, training data has to be provided for the learning phase during which the model 'learns' to recognise the spectral signatures of ash and other volcanic aerosols in order to classify them. For volcanic eruptions, there is very little available data in which the exact location of ash is known, with in situ measurements rare and limited to small parts of the plume at specific times [3]. As an alternative, the established method of simulating BTS is used. This approach offers distinct advantages such as: a) all atmospheric properties are exactly known; b) A variety of ash types, characterised by their refractive index, can be included, increasing the generalisability to plumes of unknown ash-type. The main drawback to this technique is that the full complexity of real observations is challenging to reproduce using radiative calculations alone [3].

In this section the details of the radiative transfer calculations used for the simulation are explained, followed by a description of the training sample selection and assembly.

2.1.1 Radiative Transfer Model

The software package, Radiative Transfer for TOVS (RTTOV), was used to simulate top-of-atmosphere radiances that would be observed by IASI. RTTOV is a fast radiative transfer model that is capable of accurately computing the spectra of a variety of trace gases and atmospheric states. Possible input data relevant to radiative transfer in the thermal range can be varied such as satellite zenith angle, surface properties such as temperature and surface emissivity, vertical profiles of temperature and gas concentrations, cloud properties such as height, water content, and particle effective radius, in addition to aerosol layer properties such as height and optical depth. It is especially advantageous in this time-constrained study as it is very fast compared to line-by-line models [17].

The simulations were done as part of a previous student's work [18] and provided at the outset of the project.

2.1.2 The Training Data Set

The training data set was compiled by varying the input parameters of RTTOV to produce 107,361 training

samples. The following atmospheric states were simulated and labelled 'not-ash': Clear Sky, Cloudy Sky, Sulphate, SO_2. These classes were chosen to be representative of the spectra that would be present in a typical scene following a volcanic eruption. Ash was simulated using complex refractive indices measured for ash samples from eruptions of Mount Etna (2019) and Eyjafjallajökull (2010) as measured by Deguine et al. (2020) [19]. These were labelled 'ash'. In total there were 66,321 ash and 41,040 not-ash samples in the training data. A similar number of ash and 'not-ash' samples were simulated in order to avoid the problem of class imbalance. See Table A.1 in Appendix A for details on the exact variation of parameters that enter the simulation of each atmospheric state.

BTS were simulated across the whole interval 800 to 1400 cm^{-1} with a resolution of 0.25 cm^{-1}, leading to 2401 distinct spectral features for each sample.

A further assumption made in the simulation was that all spectra are viewed over seawater, limiting the scope of the project to classifying ash over seawater only. The challenge of accurately simulating the complexity of signals measured by IASI over land, owing to high variability in surface emissivity, fell outside the scope of the computational resources available for this project. The development of a more general ash classifier is a task for future research.

2.2 Model Construction

The model was constructed using XGBoost, an open-source machine learning library for tree boosting, and built in Python using the sci-kit learn wrapper interface [13][20].

XGBoost is an optimised implementation of the gradient boosting framework by Friedman [21] that works exceptionally well for classification problems [14].

The final model is derived from an ensemble of individual tree-based models. The first tree is constructed by application of a loss function to all features and possible split points. A loss function measures the difference between the prediction and target value. By minimizing the loss function, we can select the optimal feature and condition at each root node.

In this study we choose the loss function to be a logistic regression function, as is often used in classification tasks. This is implemented by selecting the `binary:logistic` parameter of the XGBoost model in Sci-Kit Learn.

Features are chosen recursively and further nodes are added to the tree until it reaches the specified height, the ends of branches are leaves holding the results of the tree. These leaves contain the output of the logistic regression calculation and reflect how likely the input is to be ash. The actual structure of the first tree built by the classifier is illustrated in Appendix D.

Tree $i+1$ is built in the same way as tree i, except this tree is built to fit the residuals: the difference between the prediction and the original target of the previous model of i trees. This recursive process ends after a specified number of trees are built.

When the classifier is presented with a test spectra, the data starts at the root node of each tree, and passes through the tree as specified by the conditions. Finally, the results from all the trees are summed and normalised using the following function:

$$P = \frac{1}{1+e^{-S}} \tag{1}$$

where S is the sum of the results of all the trees, and P represents a confidence score in the spectra being ash. If P is above a set threshold, then the sample is predicted to be ash. Otherwise, it is labelled 'not-ash'. This can be seen as a confidence score for the presence of ash. In this study the threshold is set to 95%.

Additionally, machine learning models include hyperparameters such as the number of trees constructed, and the depth of each tree, which are known to have a large effect on predictive model efficacy [22]. GridSearchCV was applied to search for the optimal hyperparameters in this case. Our final model consists of 100 trees (i.e. $n\_estimators = 100$) with heights of 6 ($max\_depth = 6$).

2.3 Data Pre-processing

The data was pre-processed following careful empirical investigation. This involved both theoretical considerations and an iterative process of training a variety of models on differently processed training data and comparing model performances. See section 2.5.2 for detail on how model performance is evaluated. The details of how the data set was modified for the training of the final classifier are provided below.

During the process of sample inspection, it was found that some spectra in the training set were not suitable due to low transmittance of infrared. Optically thick clouds produce information-sparse absorption spectra in TIR without the detail required for ash

classification. Therefore, the data was further cleaned by excluding ash samples with AODS>5. This significantly improved predictive performance on the test data and is further advantageous as it increases the sensitivity of the model to small differences in the features indicative of the presence of ash, as optically thinner ash clouds look much more like clear sky in TIR. This doesn't significantly reduce the applicability of the model, as clouds with AODS>5 are relatively rare (see AODS retrievals in Balis (2016), Spinetti (2008), Ishimoto(2022) for typical values) [23, 24, 25].

Secondly, when evaluating the feature importances of the initial iterations of the classifier, features in range 1200 to 1400 cm^{-1} were found to be significant drivers of classification (see Appendix C for details of the feature importance calculation), which fall outside the typical channels exploited in known algorithms for ash detection in TIR. The spectra in this range are incredibly noisy due to water-vapour absorption and the region surrounding $k = 1370\,\text{cm}^{-1}$ exhibits a strong SO_2 absorption feature [4]. This was suspected to drive under-classification of ash in pixels where ash and SO_2 are collocated as is very often the case in volcanic plumes [16]. Correspondingly the wavenumber range in the training data was reduced to 800 to 1250 cm^{-1}. A further SO_2 absorption feature is found in the region of $k = 1149\,\text{cm}^{-1}$, but any further reduction of the wavenumber range would preclude features key to identifying andesite/basalt heavy ash plumes [9].

Lastly, the brightness temperatures were normalised in order to prevent bias towards features in samples with higher surface temperatures or that are lower in height, which would occur simply as the BTS values are greater. This is a standard pre-processing step in building any machine learning-based classifier [14]. See Appendix B for the details of the normalisation calculation.

2.4 Model Validation

Two approaches are used for the validation of the classifier. First, its performance is evaluated against independently simulated data to test the model's ability to generalise the relationship contained in the training data between input spectra and the presence of ash. Then, the model is evaluated against an existing linear ash retrieval algorithm on real IASI measurements for a case study plume: the eruption of Raikoke (2019).

2.4.1 The Test Data Set

The test data set was simulated using the same procedure described in Section 2.1.2 for training data simulation. So that the spectra were independent, the input parameters to RTTOV were varied differently as shown in Table A.2 in Appendix A.

2.4.2 Evaluation Metrics

Appropriate choice of evaluation metrics is integral to assessing the performance of the classifier. The following four metrics are used for evaluation in this study:

$$Accuracy = A_{True} + C_{True}/N_{Total} \tag{2}$$

$$Precision = A_{True}/A_{True} + A_{False} \tag{3}$$

$$Recall = A_{True}/A_{True} + C_{False} \tag{4}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

Here, N_{Total}, is the total number of samples, A_{True} is the number of true ash predictions, A_{False} is the number of false ash predictions, C_{True} is the number of true not-ash predictions, and C_{False} is the number of false not-ash predictions.

The accuracy metric is the obvious choice for assessing overall model performance, however in an imbalanced classification problem like this, it is a weak indicator of success. The proportion of ash pixels in the scene is often very low compared to not-ash, therefore a model that only outputs 'not-ash' would have exceptionally high accuracy, and would be useless as a classifier. Correspondingly, the *Precision*, *Recall* and $F1$ scores also require careful consideration. Maximising precision minimises false positives, while maximising recall minimises false negatives. Both are crucial to high predictive performance, therefore the key metric used to compare model performance is $F1$: the harmonic mean of precision and recall. The harmonic mean is used here as, unlike the arithmetic mean, it punishes extremes; for a model with $Recall = 1$ and $Precision = 0$, $F1 = 0$, appropriately reflecting poor performance.

2.4.3 Linear Ash Retrieval

In order to validate the classifier in real conditions, the linear ash flag developed in Sears et al. (2013) is used [16]. The flag has been found to be very useful as a quick method for remotely predicting the location of

ash clouds, but it should be noted that it is prone to under-detection, especially in low density clouds [16]. Correspondingly, some amount of over-detection from the XGBoost classifier may not be indicative of poor performance.

For the linear flag/XGBoost classifier comparison, the Raikoke (Russia) eruption in June 2019 is used as a case study. Specifically, data collected by IASI at 21:53 on the 22/06/2019 was selected. This was chosen as it featured the largest ash plume over seawater (as detected by the linear flag) of the data that was available for the project.

3 Results

This section contains details of the final model performance on both synthetic and real data. In addition, it contains the results of the feature importance calculations and a discussion of these results.

3.1 Test Data Performance

The results of the XGBoost Classifier on the test data set are shown in the Figure 3. The performance is extremely promising, with high *Accuracy* and $F1$ demonstrating the model's ability to well generalise the relationship between the spectra and class label in the simulated data.

Additionally, in Figure 2 we see the relative confidence the classifier has in these predictions given by the output of (1). A prediction of 1 is indicative of very high confidence in ash, and 0 of very high confidence in not-ash. Only 7284 out of 99939 spectra are predicted in the 20-80% confidence interval (7.2%) which indicates high confidence in the majority of predictions made.

The high *Recall* and lower *Precision* indicates that the model is liable to overprediction; it is producing very few false negatives at the expense of producing some false positives. This trend is also observed in the Raikoke case study.

3.2 Feature Importances

The interpretability of XGBoost's tree structure is extremely useful for: a) Gaining scientific insight into which features of the spectra are driving the predictions of the classifier; b) Verifying that the classifier is identifying features known to be indicative of ash presence.

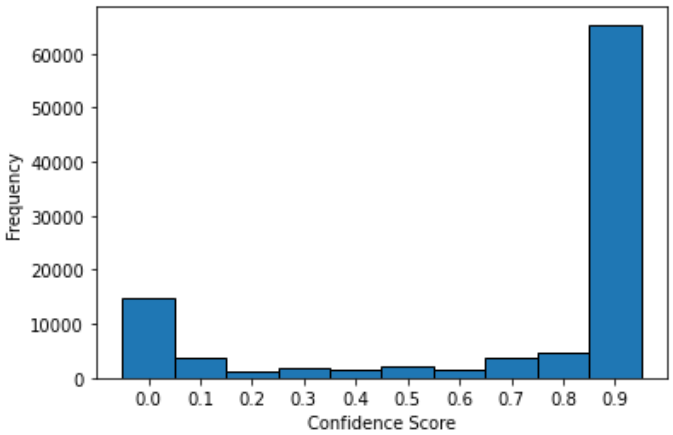
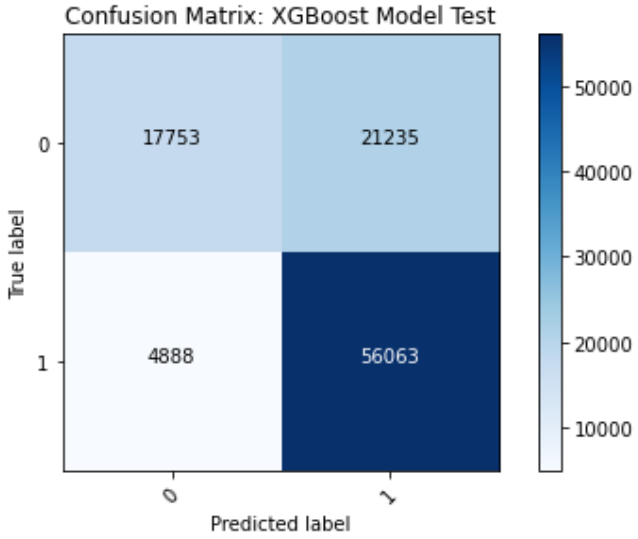


Figure 2: Distribution of the ash prediction confidence scores on the test data. P=1 indicates very high confidence that the spectra is ash, P=0 indicates very high confidence that the spectra is not-ash.



Accuracy	Precision	Recall	F1
0.74	0.73	0.92	0.81

Figure 3: Model Evaluation on Test Data

The more times a feature is used in the classification algorithm, the more important it is. For each feature, k, we calculate a relative feature importance, FI, using the `feature_importance_` attribute of the `XGBClassifier` object in Sci-kit Learn. This calculates a relativised score using the number of times a feature is used to split the data across all trees [13]. See Appendix C for the full details of the calculation.

Once the final classification model was trained, FI scores were calculated for each wavenumber and are displayed in Figure 4. Only 955 features out of 1441 contributed at all to the classification algorithm, and merely 36 features score $FI \geq 0.005$.

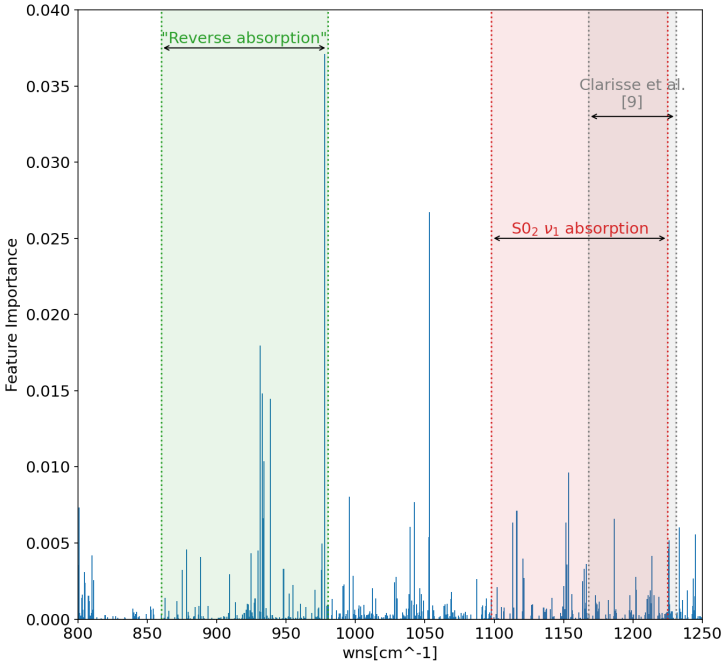


Figure 4: Relative feature importances calculated for the XGBoost model. The dotted lines mark out the regions expected to contribute to classification based on existing classification methods.

These results are generally very consistent with the empirical features. Most ash detection algorithms, like the reverse absorption method, use very few features in this range to determine the presence of ash, so it is unsurprising that the majority have $FI \approx 0$. It is likely that the prevalence of uniformly very small feature importance in unexpected channels is a symptom of a small amount of overfitting rather than significance in ash classification. Conversely, the 36 most significant features do span the range of the wavenumber interval, confirming the complexity of possible ash spectral signatures which motivated the machine learning approach.

Some further comments about the consistency of these results with the empirical features:

1. Reverse Absorption: 5/7 of the most important features are found in the range 860-980 cm^{-1} in which a negative gradient is known to be a strong indicator of ash. This region is by far the most important in classification as is expected.
2. [1168, 1231] cm^{-1}: There is a small cluster of important features in this wavenumber range in which a positive gradient is known to distinguish ash from other mineral heavy aerosols such as sand [9].
3. SO_2 Absorption: We see a small cluster of important features around 1150cm^{-1} which contains the ν_1 SO_2 absorption feature. Use of these features to identify a sample as 'not-ash' in cases where ash and SO_2 are colocated is likely to drive some misclassification and reduce the *Recall* when applied to real scenes.
4. [985, 1075] cm^{-1}: Several occurrences of high feature importances in this range are unexpected. This region is strongly associated with tropospheric ozone absorption [26] resulting in the characterstic 'v-shape' of absorption spectra in the TIR as seen in Figure 1. The high value of $k = 1053.25 cm^{-1}$ may be indication that there is an unphysical relationship in the training data that is weakening model performance.

3.3 Case Study: Raikoke (2019)

The model yielded worse results when applied to the Raikoke plume. The full classifier performance metrics were: $Accuracy = 0.87$, $Precision = 0.49$, $Recall = 0.56$, $F1 = 0.52$.

Consistent with the test data, *Recall* continues to be higher than *Precision*, highlighting that the model is more liable to overprediction than underprediction. However, the side-by-side comparison in Figure 5 shows that there is also substantial under-detection in places, leading to the low $F1$ score. The shape of the upper section of the plume can be seen, yet the majority of the under-section of the plume is missed. In addition, there is substantial false detection in the lower and upper right sections of the region.

4 Discussion and Conclusion

4.1 Discussion

The strong performance of the XGBoost model on the test data set strongly suggests that it is a capable framework for generalising the relationship between ash and its infrared absorption spectra, and therefore confirms that a machine learning method is a promising approach to ash classification using hyperspectral satellite measurements. The consistency of the features most used by the model and empirical features provides further support that the model is capable of 'learning' the patterns required for ash classification.

The serious disparity between the model efficacy when applied to the Raikoke plume suggests that the poor performance is a result of an unphysically repre-

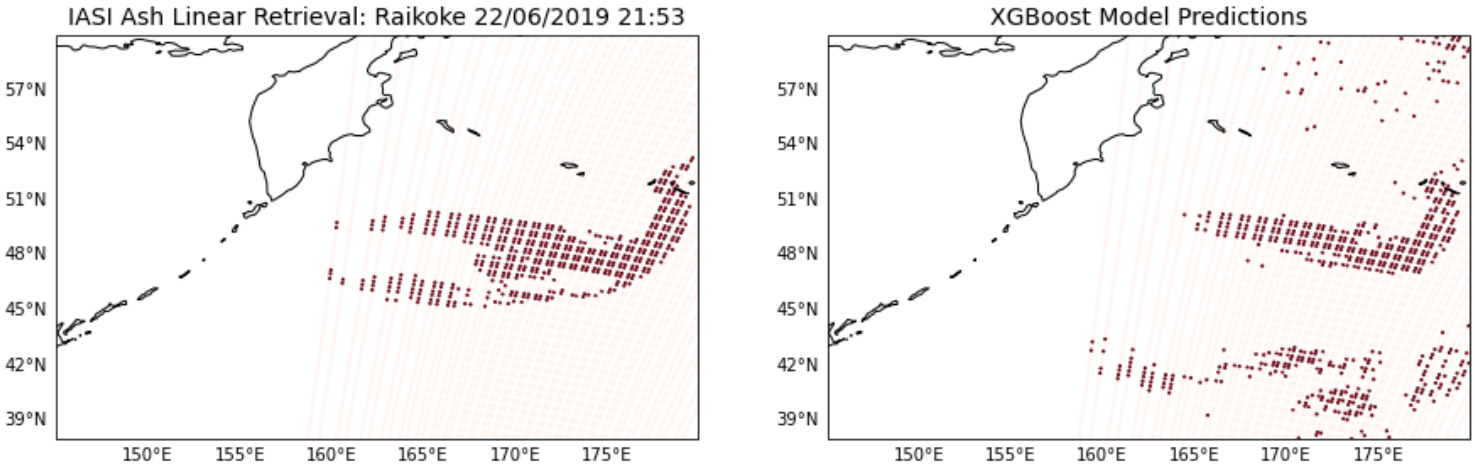


Figure 5: Comparison of the XGBoost model predictions with the linear ash retrieval. The spacing of the dots indicates the resolution of the IASI instrument. Red dots indicate an ash prediction. Both significant over- and under-prediction drive a low $F1 = 0.52$

sentative training data set as opposed to the unsuitability of the XGBoost model in spectral ash classification. If the training data is insufficiently similar to a real scene, the relationship between spectra and class will often be different in the real case and the model will be unable to make correct predictions where they are different.

As described in section 2.5.1, the spectra in the test data are simulated as independent species and labelled 'ash' or 'not-ash', therefore each spectra is 'pure'; they are as would be seen by IASI if that class was all that was contained in the $12km$ pixel radius. Of course, in reality this is a crude approximation and typically many different classes, such as 'cloudy' and 'sulphate', will be collocated within a pixel, producing spectra containing the absorption features of both. When ash is collocated with another species this can be particularly damaging to model performance.

In the case of the Raikoke eruption, significant collocation of ash and SO_2 is suspected to be driving the underprediction of ash in many cases. SO_2 has a strong absorption feature at $1149cm^{-1}$ which is not present in any of the simulated ash spectra, yet it has a high FI score as can be seen on Figure 4, and correspondingly any absorption at this value is very likely to yield a 'not-ash' class prediction. Additionally, the 'moist' atmosphere and presence of any airborne particles not present in the training data will also be contributing to the low predictive accuracy. The lack of aerosols other than sulphate and ash in the training data is assumed to be driving the high false positive rate in clusters in the upper and lower regions of the Raikoke scene, as the classifier will be very sensitive to the spectra of mineral containing particles.

To improve performance, higher quality training data must be used that better physically represents real scenes. One suggestion for this is to use a more sophisticated method for data simulation similar to the approach taken in Bugliaro (2022) [3]. This involves simulating a greater variety of 'not-ash' scenes, and then 'injecting a layer of ash' into the scene and calculating the resultant spectra in the ash and not ash cases. In this way, the reality of species collocation is more adequately reflected in the training data. Additionally, the use of other radiative transfer models, such as the Reference Forward Model (RFM) [27] could be explored and compared. This has a much greater computation time but this is irrelevant once the model has been trained, and may produce more accurate absorption spectra.

It is worth mentioning that the most accurate way of assembling labelled physically realistic training data would be to use a variety of existing ground-based and satellite-based detection methods on real scenes, and then train the model on real labelled data. Since many existing ash detection methods are flawed in certain environments (see Section 1), highly accurate data-labelling would require expert classification of each pixel as done in Taylor (2015) [6]. This is, of course, incredibly time and labour intensive, which partially motivates the simulation approach as taken in this project.

An additional consideration is that the relatively high false positive rate observed on even the test data

($Precision = 73\%$ whereas $Recall = 92\%$) may be attributable to the large diversity of the 'not-ash' spectra. For example, 'not-ash' refers to clear skies, dense SO_2 clouds, and sulphate-rich atmospheres. Therefore a further methodological change that is likely to improve model performance is to develop from a binary to a multi-class classification model in which each atmospheric state is designated its own class. The drawback of this is that far more training data would need to be included for XGBoost to accurately identify the relevant spectral class relationships, as the amount of data required increases with the number of classes. Although careful removal of irrelevant features through Principle Component Analysis may help reduce the size of the data set needed. Additionally, without improvements to the training data to be more physically representative as described above, model performance is unlikely to improve significantly on real plumes.

4.2 Conclusion

This report investigated the suitability of the application of an XGBoost classification model to predict the presence of ash using hyperspectral observations from IASI. The model was trained and tested on a synthetic dataset produced by varying input parameters to the fast radiative transfer model, RTTOV. On the synthetic test dataset, the model performed well with an $F1$ score of 81%, demonstrating the suitability of XGBoost as a method for the spectral classification of ash. This is further validated by the empirical consistency of the features used by the model in prediction and a variety of known spectral signatures of ash.

The model performed significantly less well when applied to a case study of the eruption of Raikoke (2019), with both significant over-detection and under-detection of ash across the scene. The strong performance on synthetic test data yet much weaker performance on case study plumes suggests that the training data used doesn't sufficiently replicate the complexity of a real scene. Several options for obtaining more accurate training data have been suggested for future research, most significantly including accurate simulations of IASI observations of pixels that include collocated ash, SO_2 and/or clouds, and with a greater variety of atmospheric conditions. In addition, once greater quality training data is obtained, a multi-class classification approach is suggested.

Acknowledgements

Many thanks must go to my supervisors, Dr Isabelle Taylor and Prof Roy Grainger, for their exceptional advice and encouragement throughout the project. Additional thanks must go to Dr Duncan Watson-Parris for sharing much valued expertise on machine learning and classification algorithms.

References

[1] L. J. Ventress et al. Retrieval of Ash Properties from IASI measurements. *Atmos. Meas. Tech.*, 9:5407–5422, 2016.

[2] G. S. Prata et al. A New Parameterization of Volcanic Ash Complex Refractive Index Based on NBO/T and SiO_2 Content. *Journal of Geophysical Research: Atmospheres*, 124:1179–1797, 2019.

[3] L. Bugliaro et al. VADUGS: a neural network for the remote sensing of volcanic ash with MSG/SEVIRI trained with synthetic thermal satellite observations simulated with a radiative transfer model. *Nat. Hazards Earth Syst. Sci.*, 22:1029–1054, 2022.

[4] H. E. Thomas and I. M. Watson. Observations of volcanic emissions from space: current and future perspectives. *Nat Hazards*, 54:323–354, 2010.

[5] A. Tupper et al. An evaluation of volcanic cloud detection techniques during recent significant eruptions in the western 'ring of fire'. *Remote Sensing of Environment*, 91(1):27–46, 2004.

[6] I. Taylor et al. Investigating the use of the saharan dust index as a tool for the detection of volcanic ash in seviri imagery. *Journal of Volcanology and Geothermal Research*, 304:126–141, 2015.

[7] A. T. Prata et al. Uncertainty-bounded estimates of ash cloud properties using the ORAC algorithm: application to the 2019 Raikoke eruption. *Atmos. Meas. Tech.*, 15:5985–6010, 2022.

[8] IASI documentation. `https://www.eumetsat.int/iasi`. Accessed: 14/11/22.

[9] L. Clarisse et al. A correlation method for volcanic ash detection using hyperspectral infrared measurements. *Geophys. Res. Lett.*, 37:L19806, 2010.

[10] A. J Prata. Observations of volcanic ash clouds in the 10-12 µm window using AVHRR/2 data. *International Journal of Remote Sensing*, 10(4–5):751–761, 1989.

[11] L. Clarisse and F. Prata. Chapter 11 - infrared sounding of volcanic ash. In Shona Mackie, Katharine Cashman, Hugo Ricketts, Alison Rust, and Matt Watson, editors, *Volcanic Ash*, pages 189–215. Elsevier, 2016.

[12] S. Mackie et al. How assumed composition affects the interpretation of satellite observations of volcanic ash. *Met. Apps*, 21:20–29, 2014.

[13] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[14] K. P. Murphy. *Machine Learning a Probabilistic Perspective*. The MIT Press, Cambridge, Massachusetts, 2012.

[15] Y. Zhenping et al. An efficient spectral selection of m giants using xgboost. *The Astrophysical Journal*, 887(2):241, dec 2019.

[16] T. M. Sears et al. So2 as a possible proxy for volcanic ash in aviation hazard avoidance. *Journal of Geophysical Research: Atmospheres*, 118(11):5698–5709, 2013.

[17] R. Saunders et al. An update on the rttov fast radiative transfer model (currently at version 12). *Geoscientific Model Development*, 11(7):2717–2737, 2018.

[18] M. Okanik. Classification of ash spectra in various environmental settings. *EODG Vacation Reports*, 2021.

[19] A. Deguine et al. Complex refractive index of volcanic ash aerosol in the infrared, visible, and ultraviolet. *Applied Optics*, 59:884, 02 2020.

[20] F. Pedregosa et al. Scikit-learn: Machine learning in python. *CoRR*, abs/1201.0490, 2012.

[21] J. H Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001.

[22] S. Kapoor and V. Perrone. A simple and fast baseline for tuning large xgboost models. *CoRR*, abs/2111.06924, 2021.

[23] D. Balis et al. Validation of ash optical depth and layer height retrieved from passive satellite sensors using earlinet and airborne lidar data: The case of the eyjafjallajökull eruption. *Atmospheric Chemistry And Physics*, 01 2016.

[24] C. Spinetti et al. Mt. etna volcanic aerosol and ash retrievals using meris and aatsr data. 09 2008.

[25] H. Ishimoto et al. Ash particle refractive index model for simulating the brightness temperature spectrum of volcanic ash clouds from satellite infrared sounder measurements. *Atmospheric Measurement Techniques*, 15(2):435–458, 2022.

[26] J. Landgraf and O. P. Hasekamp. Retrieval of tropospheric ozone: The synergistic use of thermal infrared emission and ultraviolet reflectivity measurements from space. *Journal of Geophysical Research: Atmospheres*, 112(D8), 2007.

[27] A Dudhia. The reference forward model. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 186:243–253, 2017.

A RTTOV Input Parameters

Table 1: RTTOV Input Parameters (Training)				
Class	Optical Depth at $550nm$ (AODS)	Height/Mb (MBS)	Particle Effective Radius/μm (ERS)	Surface Temperature/K (STS)
Ash	0.5 to 5.0 Step: 0.25	50 to 900 Step: 50	0.2 to 8.0 Step: 0.4	260 to 300 Step: 20
Clear	0*	400	0.2	240 to 320 Step: 0.1
Cloudy	1 to 29 Step: 2	100 to 900 Step: 100	0.2, 1.0, 2.0 to 62.0 Step: 4	260 to 300 Step: 20
Sulphate	0.5 to 10.0 Step: 0.5	50 to 900 Step: 50	0.2 to 8.0 Step: 0.4	260 to 300 Step: 20
	Column Amount (DUS)	Height/Mb (MBS)	Plume Thickness/μm (THICK)	Surface Temperature/K (STS)
SO_2	0.5 to 240.5 Step: 10.0	10 to 940 Step: 15	30, 50, 100	260 to 300 Step: 20
*Note: RTTOV cannot take a zero input so AODS was set to 10^{-7}				

Table 2: RTTOV Input Parameters (Testing)				
Class	Optical Depth at $550nm$ (AODS)	Height/Mb (MBS)	Particle Effective Radius/μm (ERS)	Surface Temperature/K (STS)
Ash	0.51 to 5.01 Step: 0.25	60 to 910 Step: 50	0.3 to 7.5 Step: 0.4	265 to 305 Step: 20
Clear	0*	400	0.2	260.05 to 310.05 Step: 0.1
Cloudy	1.5 to 27.5 Step: 2	110 to 910 Step: 100	0.4, 1.4, 2.5 to 58.5 Step: 4	265 to 305 Step: 20
Sulphate	0.6 to 9.6 Step: 0.5	60 to 910 Step: 50	0.3 to 7.5 Step: 0.4	265 to 305 Step: 20
	Column Amount (DUS)	Height/Mb (MBS)	Plume Thickness/μm (THICK)	Surface Temperature/K (STS)
SO_2	2 to 242 Step: 10.0	15 to 945 Step: 15	35, 55, 105	265 to 305 Step: 20
*Note: RTTOV cannot take a zero input so AODS was set to 10^{-7}				

B Normalisation Process

The feature normalisation procedure used is known as *standardising*, which ensures the variance of the data along each dimension is 1, while preserving the relationship between data points. Standardisation prevents sensitivity to the scale of the input features, improving algorithmic performance [14].

The standardisation procedure was carried out as follows: For each feature vector, k, the distribution mean, $\bar{k}$, and standard deviation, σ, are calculated. Then the mean is subtracted from each feature, and this is divided by the standard deviation:

$$k' = \frac{k - \bar{k}}{\sigma} \tag{6}$$

where k' is the standardised feature vector.

C Feature Importance

For each feature, k, we calculate a relative feature importance, FI, using the `feature_importance_` attribute of the `XGBClassifier` object in Sci-kit Learn. The `importance_type_` parameter is set to `weight`. See the xgboost documentation for additional details and the potential limitations of this method [13]. We define FI for each feature k as follows:

$$FI_k = \frac{N_k}{\sum_k N_k} \tag{7}$$

where FI_k is the relative feature importance of feature k, and N_k is the number of times a feature is used for splitting at a root node. Notice that the zero importance features (those features that have not been used in any split conditions) do not contribute to the sum, therefore they do not contribute to the relativised score.

D XGBoost Structure Illustration

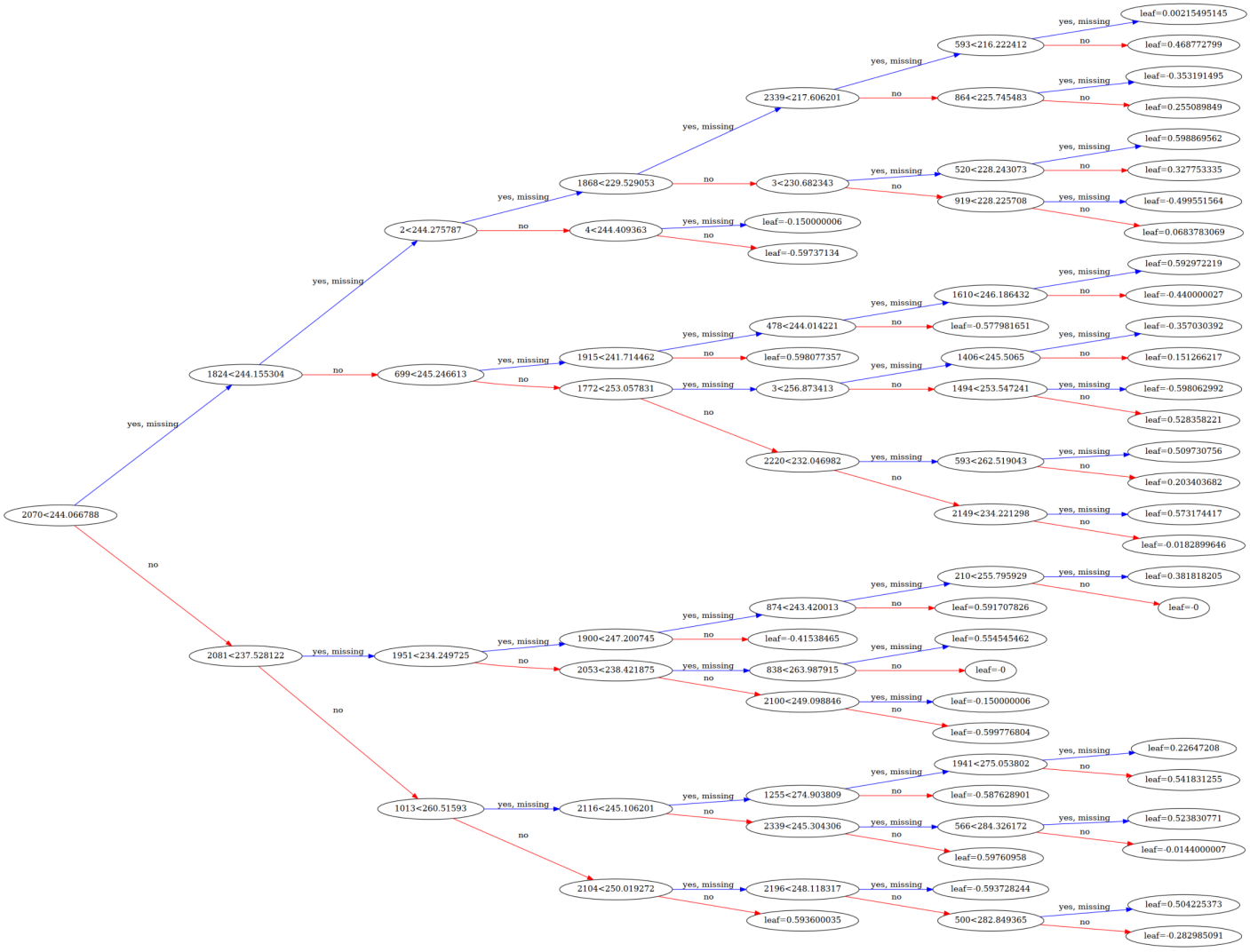


Figure 6: An illustration of the first tree of the XGBoost Classifier. Within each ellipse is the condition at the root node: $983 < 245$ is the condition that the BTS value of the 983rd feature is less than 245K. $max\_depth = 6$ is clearly illustrated by the maximum number of nodes an input value passes through before being assigned a P score at a leaf node.