Investigating large-scale change in volcanic time series data using machine learning analysis



Grace Manley

University College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas Term 2021

Supervisors: Prof. T.A. Mather, Prof. D.M. Pyle, Prof. D.A. Clifton, and Dr M. Rodgers

Declaration

I declare that the contents of this thesis are entirely my own, except where explicitly noted.

Grace Manley

Tuesday 2nd November 2021

Abstract

Investigating large-scale change in volcanic time series data using machine learning analysis

Grace Manley
University College, Oxford
Michaelmas 2021

Volcanic eruptions are characterised by large-scale transitions in behaviour, which include the transitions governing repose to unrest, unrest to eruption, and eruption to quiescence. Identification of these transitions is a fundamental goal in volcanology. This thesis aims to explore the use of machine learning approaches which are established for the monitoring of critical systems, such as healthcare or jet engine monitoring, to better understand the timing of large-scale transitions in volcanic activity. The first research chapter is an application of machine learning approaches to better understand the timing of eruption end, which has no systematic definition. Multi-class machine learning methods were applied to time series derived from volcano-seismic data from two volcanoes with contrasting eruption styles: Nevado del Ruiz, Colombia and Telica, Nicaragua. The resulting eruption end-dates agreed with previous, generic definitions of eruption end and provide an independent constraint on the timescale of eruption cessation processes. The second research chapter is an extension of this methodology to multi-parameter datasets from the 2005 – 2006 unrest and eruption of Augustine, Alaska. Both multi-class and novelty detection approaches, in which models are trained only on non-eruptive data, are applied. The resulting transition from non-eruptive to eruptive identified from the modelling agrees with previous estimates of the timing of dike initiation prior to eruption. Additionally, I explored the use of different types of data and different seismic catalogues for distinguishing between eruptive and non-eruptive activity. The final research chapter is an application of deep active learning to automated volcano-seismic event classification. Active learning is the process by which a machine learning model actively selects the training data which it learns from. This approach has the potential to reduce the quantity of labelled data required to train a successful model. The deep active learning approach is applied to two datasets from Nevado del Ruiz, Colombia and Llaima, Chile. Active learning models performed better on unseen testing data for both datasets and achieved better performance during training for the Nevado del Ruiz dataset.

Acknowledgements

Firstly, thanks to my dream team of supervisors. Thank you to my supervisors in the Earth Sciences department, Tamsin Mather and David Pyle, for allowing me the freedom to explore interesting ideas throughout my DPhil, and for their vast experience which helped me to refine them. I am very grateful for their support and encouragement to write up and present my work throughout. Thank you to David Clifton, for taking me on as a first-year DPhil student and bringing incredible insights to the project at every stage. Thank you to Mel Rodgers, both for our regular discussions and for invaluable technical discussions during critical points of the project.

Additional thanks go to the collaborators who I have worked with from whom I have received considerable feedback and advice, including Glenn Thompson, Diana Roman and John Makario Londoño.

An enormous thanks go to Mel and Rocco for hosting me in Tampa during my first-year trip to the University of South Florida, where I learned much of the fundamentals of seismic data processing. Thanks to the members of the Geosciences department for welcoming me and for useful conversations about my project at this early stage of the work, including Chuck Connor and Steve McNutt.

I have been fortunate to be a part of two research groups during my DPhil, and I thank the Volcanology research group for useful feedback and discussion over the years. Thanks must also be extended to the members of the Computational Health Informatics group in Engineering for being so welcoming to a volcanologist and for many insightful discussions both during and outside meetings.

A great thanks to all my friends at Oxford who have made it a pleasure to spend so many years at this university and I am lucky that there are too many to name here. Thanks to all my friends in department, in Office 40.18 and beyond. Particularly, thanks to Amber for the support as we were both in the final stages. I am grateful to other volcanology DPhils, Anna, David, Annika, and Alice for their friendship and particularly for all the coffee breaks which got me through some cold winter

mornings in lockdown. A huge thanks to all my friends from the Ultimate Frisbee club, who have kept me sane throughout. Special thanks go to SJ and Kay for being such positive influences in the first and final years of my DPhil. Thank you to Kelly for always supporting me from afar.

Finally, thanks to my family. Thank you to my parents, who have been unbelievably supportive of my research even when it strays so far outside the bounds of their areas of expertise. Helen, Kitty, Nay, Alex, John, I am so lucky to have you all as such wise and caring (and numerous!) older siblings. A great thanks to my nieces and nephews, though they probably do not realise their contribution. Finally, my deepest thanks go to Tom without whose love and support I would not have got this far.

Extended abstract

Investigating large-scale change in volcanic time series data using machine learning analysis

Grace Manley
University College, Oxford
Michaelmas 2021

Volcanic eruptions are characterised by large-scale transitions in behaviour, which include the transitions governing repose to unrest, unrest to eruption, and eruption to quiescence. Identification of these transitions is a fundamental goal in volcanology. Machine learning approaches are well-established in the field of condition monitoring for critical systems, such as healthcare or jet engine monitoring. Volcanic and healthcare systems share fundamental conceptual similarities: observation of failure (e.g., for volcanic eruptions) is typically rare in comparison with stable behaviour, and the number of failure modes are not known or not well characterised. Moreover, due to the nature of surface-based monitoring, data from both systems is subject to noise from external factors and can be irregularly sampled. The principal goal of this thesis is to use knowledge and techniques which are well-developed in a healthcare context, and apply them to better understand large-scale changes in volcanic activity.

Chapter 3 is an application of machine learning approaches to better understand the timing of eruption end. The timing and processes that govern the end of volcanic eruptions are not yet fully understood, and there currently exists no systematic definition for the end of a volcanic eruption. Currently, end of eruption is established either by generic criteria (typically 90 days after the end of visual signals of eruption) or criteria specific to a given volcano. I applied supervised machine learning classification methods, in which models are trained on both non-eruptive and eruptive data, and defined a decisiveness index D to evaluate the consistency of the classifications obtained by these models. These multi-class methods were applied to seismic time series from two volcanoes chosen because they display contrasting styles of eruption: Telica (Nicaragua) and Nevado del Ruiz (Colombia). For both volcanic systems, the end-date obtained by classification of seismic data is 2 - 4 months later than end-dates defined by the last occurrence of visual eruption (such as ash emission). This finding agrees with previous, general definitions of eruption end and is consistent across models. The

classifications have a higher correspondence of eruptive activity with visual activity than with database records of eruption start and end. Additional analysis on the relative importance of the different features of seismic activity used in our models (e.g., peak event amplitude, daily event counts) found little consistency between the two volcanic systems in terms of the most important features which determine whether activity is eruptive or non-eruptive.

Chapter 4 is an extension of the methodology from seismic data, to also include geodetic and gas data. Some traditional approaches to forecasting transitions prior to volcanic eruptions have used models motivated by the underlying physics of eruption onset, which assume that geophysical precursors will consistently display similar patterns prior to transition in volcanic state. A machine learning classification approach for detecting significant changes in patterns of volcanic activity, potentially signalling transitions during the onset or end of volcanic activity, does not require a model of the physical processes underlying critical changes. I explore the application of a novelty detection approach, where models are trained only on data prior to eruption, to the precursory unrest at Augustine Volcano, Alaska in 2005. This approach could be useful for geophysically-monitored volcanic systems which have been in repose for some time, as no eruptive data is required for model training. The novelty detection results are compared with multi-class classification, where models are trained on examples of both non-eruptive and eruptive data. The results of these classification models are contextualised using constraints from petrological, satellite and visual observations from the 2006 eruption of Augustine Volcano. The transition from non-eruptive to eruptive behaviour we identify in mid-November 2005 agrees with previous estimates of the initiation of dike intrusion prior to the 2006 eruption. Models which include multiple types of data (seismic, deformation and gas emissions) can better distinguish between non-eruptive and eruptive data than models formulated on single data types.

The final part of this work explores the use of active learning to reduce the training set size in volcano-seismic event classification. Volcano-seismic event classification represents a fundamental component of volcanic monitoring and recent advances in techniques for the automatic classification of volcano-seismic events using supervised deep learning models achieve high accuracy. However,

these deep learning models require a large, labelled training dataset to successfully train a generalisable model. I develop an approach to volcano-seismic event classification making use of a method of active learning, wherein a machine learning model actively selects the training data which it learns from. This active learning approach is a diversity-based method, which works by selecting new training points which are most dissimilar from points already in the model according to a distance-based calculation applied to the model features. The active learning is combined with an existing volcano-seismic event classifier and applied to data from two volcanoes: Nevado del Ruiz, Colombia and Llaima, Chile. Models with data selected using an active learning approach achieve better testing accuracy and Area Under the (Receiver Operating Characteristic) Curve (AUC) than models where the next training data is selected via random sampling. Active learning decreases the labelling burden for the Nevado del Ruiz dataset but offers no increase in performance for the Llaima dataset. I explore the reasons for the variable outcomes using an unsupervised visualisation of the data, and suggest that active learning has potential to reduce the amount of training data required for datasets where the event types are less separable.

Finally, I discuss the implications of the work presented in this thesis for using machine learning approaches in a monitoring framework. This discussion includes the adaptations which would be required for a real-time monitoring approach, the considerations on the completeness of seismic catalogues, and the extension of these methods to other types of data. Additionally, I discuss several avenues of future work which could extend the methods used in this thesis to better understand eruptive transitions and the nature of large-scale change in volcanic systems.

Commonly-used abbreviations

ANN . . . . . Artificial Neural Network

AVO . . . . . Alaska Volcano Observatory

BNN . . . . . Bayesian Neural Network

CLVF . . . . . Carry Last Value Forward

CNN . . . . . Convolutional Neural Network

FFM . . . . . Failure Forecast Method

GPC. . . . . Gaussian Process Classification

GPR. . . . . Gaussian Process Regression

GPS . . . . . Global Positioning System

GVP . . . . . Global Volcanism Program

InSAR . . . . . Interferometric Synthetic Aperture Radar

k-NN . . . . . k-Nearest Neighbour

LaMEVE . . . . . Large Magnitude Explosive Volcanic Eruption

LR . . . . . Logistic Regression

ML . . . . . Machine Learning

NDR . . . . . Nevado del Ruiz

RF . . . . . Random Forest

RSAM . . . . . Real-time Seismic Amplitude Measurement

SO_2 . . . . . Sulphur dioxide

SVM . . . . . Support Vector Machine

t-SNE . . . . . .t-Distributed Stochastic Nearest Neighbour Embedding

USGS . . . . . United States Geological Survey

VEI . . . . . .Volcanic Explosivity Index

Contents

Declaration ..... **iii**

Abstract ..... **v**

Acknowledgements ..... **vii**

Extended abstract ..... **ix**

Commonly-used abbreviations ..... **xii**

Contents ..... **xiii**

List of Figures ..... **xvii**

List of Tables ..... **xviii**

Chapter 1 - Introduction **1**

1.1) Eruption onset 3

1.2) Eruption end 6

1.3) Machine learning 11

1.4) Thesis structure 15

1.5) Related references 17

Chapter 2 - Data processing **19**

2.1) Seismic data 19

2.1.1) Data collection 19

2.1.2) Continuous data processing 19

2.1.3) Waveform processing 21

2.1.4) Feature extraction 24

2.2) Gas data 25

2.2.1) Data collection 25

2.2.2) Data processing 26

2.2.3) Feature extraction 28

2.3) Geodetic data 28

2.3.1) Data collection 28

2.3.2) Data processing 29

2.3.3) Feature extraction 29

Chapter 3 - Understanding the timing of eruption end using a machine learning approach to classification of seismic time series.....30

3.1) Abstract.....30

3.2) Introduction.....31

3.3) Methods.....34

3.3.1) Machine learning methods used.....34

3.3.2) Model assumptions.....37

3.3.3) Nevado del Ruiz and Telica volcanoes.....38

3.3.4) Classification scheme.....39

3.3.5) Feature extraction.....41

3.3.6) Eruption classification.....44

3.3.7) Quality assessment: Decisiveness index.....44

3.4) Results.....46

3.4.1) Nevado del Ruiz.....46

3.4.2) Telica.....52

3.4.3) Training models with data from the end of volcanic eruption.....57

3.5) Discussion.....58

3.5.1) Classification compared to visual observations.....58

3.5.2) Possible application of methods and future work.....60

3.5.3) Failure of logistic regression.....62

3.5.4) Feature importance: feature ranking.....62

3.5.5) Verification of model assumptions.....63

3.6) Conclusions.....65

Chapter 4 - Machine learning approaches to identifying changes in eruptive state using multi-parameter datasets from the 2006 eruption of Augustine Volcano, Alaska.....66

4.1) Abstract.....66

4.2) Introduction.....67

4.3) Methods.....74

4.3.1) Augustine Volcano.....74

4.3.2) Seismic time series.....77

4.3.3) Gas time series.....80

4.3.4) Geodetic measurements.....84

4.3.5) Machine learning methods.....84

4.3.6) Assumptions.....87

4.4) Results 88
4.4.1) Novelty detection results 88
4.4.2) Multi-class results 89
4.5) Discussion 93
4.5.1) Physical interpretation of model results 93
4.5.2) Importance of feature groups 95
4.5.3) Applications of the method 98
4.6) Conclusions 101
Chapter 5 - A deep active learning approach to the automatic classification of volcano-seismic events 102
5.1) Abstract 102
5.2) Introduction 103
5.3) Methods 106
5.3.1) Data 106
5.3.2) Machine learning classification 109
5.3.3) Active learning 110
5.4) Results 116
5.4.1) Nevado del Ruiz dataset 116
5.4.2) Llaima dataset 117
5.5) Discussion 120
5.5.1) Model comparison 120
..... 122
5.5.2) Applications for active learning 122
5.5.3) Limitations 123
5.6) Conclusions 125
Chapter 6 - Conclusions 126
6.1) Thesis summary of science chapters 126
6.2) Implications for monitoring 128
6.3) Future perspectives 131
6.3.1) Uniting quantitative and qualitative information in models of volcanic state 131
6.3.2) Repose on long timescales 135
6.3.3) Understanding the physical processes governing eruption end 137
6.3.4) Intra-eruptive transitions 138

References **141**

Appendices **179**

Appendix A: 179

A.1) Nevado del Ruiz features 179

A.2) Telica features 180

Appendix B 183

B.1) Confirmed ash emission dates from Nevado del Ruiz 183

Appendix C 186

C.1) List of seismic stations used in analysis 186

C.2) List of GPS stations used in analysis 186

C.3) List of seismic features used as input variables. 187

C.4) List of geodetic features used as input variables 188

C.5) List of gas features used as input variables 188

C.6) Summary of classification results 189

C.7) Three-dimensional visualisation of the seismic, gas flux and geodetic features. 190

C.8) Single-parameter novelty detection models 191

List of Figures

Figure 1.1: Illustration of a generic eruptive cycle and remote sensing techniques which can be used to monitor volcanic activity. 2

Figure 1.2: Schematic illustration of the asymmetric distribution of lava discharge rate time-series proposed by Wadge (1981). 7

Figure 2.1: Map of seismic network for the seismic data used in this thesis. (a) Nevado del Ruiz seismic station map, modified from Castaño *et al*. (2020). (b) Telica seismic station map, from Roman *et al*. (2019). (c) Augustine seismic station map, from Chapter 4 of this thesis. (d) Llaima seismic station map, from Canario *et al*. (2020b) 23

Figure 2.2: Illustration of the noise reduction process for the seismic data. 24

Figure 3.1: Visualisation of how boundaries are defined within the four machine learning methods used in this paper. 37

Figure 3.2: The framework of training and testing supervised multi-class classification models. 41

Figure 3.3: Diagram of the framework for extracting features from raw seismic waveform data. 44

Figure 3.4: Results from SVM classification on Nevado del Ruiz data over the study period. 48

Figure 3.5: Receiver Operating Characteristic (ROC) curve for all the methods applied to the Nevado del Ruiz data. 50

Figure 3.6: Results from feature importance analysis methods on Nevado del Ruiz data. 51

Figure 3.7: Results from GPC classification on Telica data over the study period. 54

Figure 3.8: Receiver Operating Characteristic (ROC) curve for all the methods applied to the Telica data. 54

Figure 3.9: Results from feature importance analysis methods on Telica data. 56

Figure 3.10: Comparison of the event rate and alert level (top panel) and SVM classification and recorded ash emissions (bottom panel) at Nevado del Ruiz between the dates of 28th June 2010 to 25th February 2015. 60

Figure 3.11: Plot of the top 10 ranked features from each feature importance method on Nevado del Ruiz (top) and Telica (bottom). 64

Figure 4.1: Illustration of the range and relative depths of physical processes associated with observation of seismicity, gas and geodetic measurements. 71

Figure 4.2: Seismic and GPS station map for Augustine Island, inset in a map of the Cook Inlet, Alaska. 73

Figure 4.3: Timeline of the data period from April 2005 – May 2006 used in this paper and classification of the eruption timeline outlined in Section 4.3.1.1. 76

Figure 4.4: Daily SO_2 emission rate values for the measurements made at Augustine between 2002 and 2008 and values interpolated using GPR (top panel) and CLVF (bottom panel). 83

Figure 4.5: Results from novelty detection classification applied to the features derived from Augustine data. 91

Figure 4.6: Results from the multi-class results from the random forest classification applied to the seismic, geodetic and gas features derived from Augustine data. 92

Figure 4.7: Two-dimensional visualisation of the seismic, gas flux and geodetic features derived from Augustine. 96

Figure 5.1: Representative examples of waveforms of each class from the datasets used in this study. 113

Figure 5.2: (a) Illustration of the SeismicNet neural network architecture (Canario *et al*., 2020a). 114

Figure 5.3: Illustration of the active learning experiment presented here. 115

Figure 5.4: (a) Learning curves for the active model (blue) and random model (orange) for the Nevado del Ruiz dataset. 118

Figure 5.5: (a) Learning curves for the active model (blue) and random model (orange) for the Llaima dataset. 119

Figure 5.6: Visualisation of the final model features using a t-Distributed Stochastic Nearest-Neighbour Embedding projection..... 122

Figure 6.1: Illustration of a potential framework for integrating information from qualitative records such as activity reports into a machine learning approach. 134

Figure 6.2: Results of a Latent Dirichlet Allocation model applied to over 20 years of GVP weekly activity reports from Etna and Karymsky volcanoes..... 134

Figure 6.3: Timeline of eruptive activity at Soufrière Hills Volcano, Montserrat from 1995 – 2016. 137

List of Tables

Table 1.1: Summary of modelling approaches for identifying or modelling eruption onset with categories after Whitehead and Bebbington (2021). 5

Table 1.2: Summary of current approaches for identifying or modelling eruption end..... 10

Table 1.3: Summary of machine learning (ML) approaches in volcanology..... 14

Table 2.1: Summary of network parameters for the seismic data used in this thesis..... 24

Table 3.1: Summary of results from all machine learning models applied to the Nevado del Ruiz and Telica datasets..... 52

Table 4.1: Summary of observable changes and underlying physical processes which are linked to observations of seismicity, gas and geodetic measurements. 72

Table 5.1: Testing performance (AUC and accuracy) for the models presented in this study. 118

Table 6.1: Summary of the future work ideas presented in this thesis, including the necessary data and model types to execute these projects..... 139

Chapter 1 - Introduction

Volcanic eruptions can have devastating social and economic consequences. To mitigate these consequences, early identification of impending eruption is crucial to take actions such as evacuating nearby populations or anticipating diversions to infrastructure and transport networks. Despite this, accurate and timely forecasts of volcanic eruptions are not yet routinely possible (Sparks, 2003; Poland and Anderson, 2020). Understanding the critical thresholds which govern the onset, size and duration of volcanic eruptions has been identified as one of the grand challenges in volcanology (NASEM, 2017). As well as understanding the beginning and progression of volcanic eruptions, this challenge inherently involves understanding their end, which at present do not have a systematic definition (Phillipson *et al*., 2013).

Figure 1.1 is an illustration of the generic eruption cycle after Pyle *et al*. (2013). Volcanic eruptions can be preceded and followed by elevated activity, termed unrest (Siebert *et al*., 2015). The duration and nature of volcanic unrest can be variable (e.g., Stix, 2018). For example, the 2009 eruption of Okmok volcano, Alaska, displayed few signs of geophysical unrest before eruption with only 5 hours of precursory seismic activity (Johnson *et al*., 2010). For other volcanic systems, such as Shishaldin volcano, Alaska and Turrialba volcano, Costa Rica, unrest can extend for significantly longer periods than the duration of observable eruption, and the unrest may show long-term patterns of elevated seismicity or degassing on timescales of years to decades (Roman *et al*., 2019).

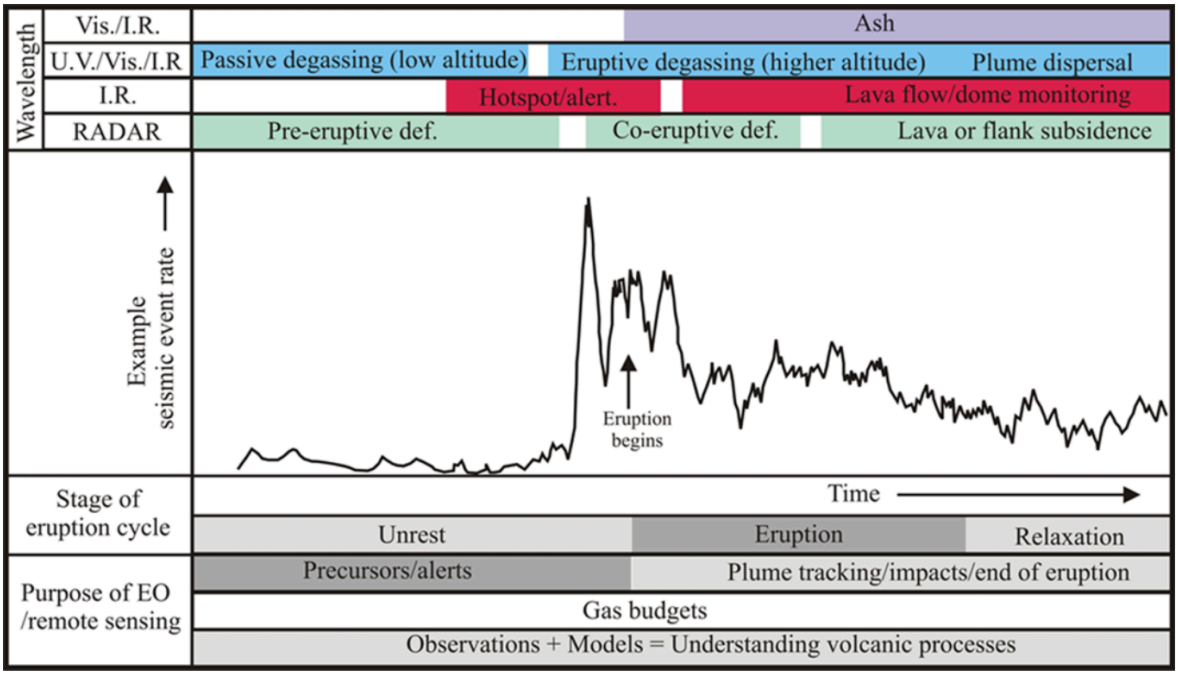


Figure 1.1: Illustration of a generic eruptive cycle and remote sensing techniques which can be used to monitor volcanic activity. Complementary ground-based measurements for monitoring volcanic activity over the course of an eruptive cycle include seismometers, geodetic, and ground-based gas sensors. The seismic event rate corresponds to the 1980 eruption of Mt St Helens, USA. From Pyle et al. (2013).

Monitoring the condition of systems using machine learning is well-developed in other fields, such as healthcare or jet engine monitoring (e.g., Clifton and Tarassenko, 2006). Volcanic and healthcare systems share several conceptual similarities: both can be thought of as "high-integrity" systems in which the modes of failure are not well-known or well-characterised (Clifton *et al*., 2014). Continuous measurements made in both clinical and volcanological contexts involve using instruments based on the surface of the system, for example, wearable sensors or seismometers as respective examples. This surface-based instrumentation is subject to noise from external factors, such as human motion affecting wearable sensor data or wind affecting seismic data quality. Moreover, measurements in both settings are often made at a variety of timescales: in a volcanological context, measurements range from continuous records of ground motion via seismometers to discrete satellite observations or observatory reports which occur on a weekly – monthly basis. These conceptual similarities motivate new approaches for identifying critical changes in volcanic activity, based on established approaches in the healthcare engineering for identifying patient deterioration (e.g., Clifton *et al*., 2014; Shamout *et al*., 2019). This thesis aims to explore knowledge and techniques that are

well-developed in healthcare and apply them to better understand the timing of transitions in volcanic state.

The following sections provide some background to the current state-of-the-art in terms of identifying eruption onset, eruption end, and the application of machine learning approaches in volcanology.

1.1) Eruption onset

Volcanic eruptions are often, but not always, preceded by recognisable precursory activity. Piecing together geophysical and geochemical observations to ascertain whether the origin of unrest is magmatic or non-magmatic (Pritchard *et al*., 2019), or will lead to eventual eruption (Moran *et al*., 2011), are fundamental challenges in volcanic monitoring. Forecasting eruptions and their nature is an additional challenge, as forecasts typically involve a statement or inference on the likely geographic location and size/style of any potential activity (Poland and Anderson, 2020), as well as the likelihood of its occurrence.

Table 1.1 details six types of methods for identifying eruption onset in the short-medium term as categorised by Whitehead and Bebbington (2021). These approaches can involve volcano-specific expert interpretation of observational data, such as the detection of ash-rich eruptions from infrasound alerts at Etna (Ripepe *et al*., 2018). Event trees are a method of computing probability of outcomes of volcanic activity from the probabilities of individual events within an eruptive timeline, which can be formulated based on local eruptive history and Bayesian statistics (e.g., Marzocchi *et al*., 2004; Wright *et al*., 2019). Bayesian Belief Networks are another way to graphically represent the states of a volcanic system and the probabilities associated with transitions between them (e.g., Aspinall and Woo, 2014). The Failure Forecast Method (FFM) is a power-law approximation, based on the underlying assumption that accelerating rates of observables (such as seismic event rate or gas emission) directly correspond to underlying material failure in the subsurface (Voight, 1988; Cornelius and Voight, 1995; Kilburn, 2003). This approach has been applied to retrospectively forecast eruption times for a variety of volcanic systems, including lava dome growth and explosive

eruptions (e.g., Smith and Kilburn, 2010; Lavallée *et al*., 2008). However, the FFM is limited in its ability to successfully forecast eruptions in real-time. Boué *et al*. (2015) applied a probabilistic implementation of the FFM retrospectively to several explosive episodes at Colima volcano, Mexico and were only able to successfully apply the method where there were simple patterns of accelerating seismicity. Process and source models are a general term for models which involve the identification of patterns in observational data and assume a source mechanism for these patterns. An example of this type of model is the recognition of pre-eruptive patterns in seismicity at dormant stratovolcanoes after White and McCausland (2019). Recently, machine learning has been trialled as a retrospective method for identifying eruption onset: Dempsey *et al*., (2020) designed a machine learning approach to issuing short-term alerts of increased eruption probability at Whakaari, New Zealand, using a random forest approach applied to tremor data.

Table 1.1: Summary of modelling approaches for identifying or modelling eruption onset with categories after Whitehead and Bebbington (2021).

Type of model	Method	Data type	Example application(s)	Comments
Expert interpretation	Volcano-specific interpretation of available data, e.g., noting increases in seismic event rate at Alaskan volcanoes; infrasound array-based detection of ash-rich explosions at Etna	Many – depends on volcanic system	Cameron *et al*., 2018 Pesicek *et al*., 2018	Can be on a regional or volcano-specific scale
Event trees	Expert elicitation; historical data	All available – represents an integrated framework	Marzocchi *et al*., 2004 Marzocchi *et al*., 2008 Wright *et al*., 2019	Can also be formulated for long-term forecasting and hazard assessment
Belief networks	Expert elicitation (for constructing the network); observational data; computational approaches (for estimating probabilities)	All available – represents an integrated framework	Aspinall and Woo, 2014	Can also be formulated for long-term forecasting and hazard assessment
Failure forecasting	Failure Forecast Method (FFM)	Mostly seismic data (e.g., event rates, RSAM); gas observations	Voight, 1988 Cornelius and Voight, 1995 Kilburn *et al*., 2003 Lavallée *et al*., 2008 Smith and Kilburn, 2010 Boué *et al*., 2015	Loosely physical, based on the theory of damage accumulation in materials; applied retrospectively
Process/source models	Combination of observation with an assumption of the source process explaining the source	Depends on process/source, but commonly seismic data	e.g., White and McCausland, 2019 White and McCausland 2016	
Machine learning	Random forest model	Seismic tremor data	Dempsey *et al*., 2020	Applied retrospectively

1.2) Eruption end

There is no systematic definition for the end of a volcanic eruption (Phillipson *et al*., 2013). Existing definitions for the end of eruption come under either generic definitions such as the 90-day rule, where an eruption is considered to have ended if there are no visible signs of eruption after 90 days (Simkin and Siebert, 1994), or volcano-specific definitions of eruption end. Examples of volcano-specific definitions include Piton de la Fournaise Volcano, La Réunion, where eruptive episodes are defined based on an increase or decrease in seismic tremor amplitudes (Battaglia and Aki, 2003), and Soufrière Hills Volcano, Montserrat, where the end of eruption is defined by the MVO Scientific Advisory Committee via three thresholds on gas (less than 50 t/day of SO_2 output), geodetic (less than 5 mm/year displacement across all GPS sites) and seismic activity (no low-frequency earthquake swarms) (SAC, 2004).

The GVP database incorporates end-dates for periods of eruption, deformation, and gas data in its records. However, Siebert *et al.* (2011) found that of 10,415 eruptions in the Volcanoes of the World (VOTW) database at the time of writing, 59% of the entries did not have an associated end-date. This lack of data was attributed to the gradual nature of eruption end, which made assigning a discrete date difficult. Phillipson *et al.* (2013) suggest that end-dates in the GVP database have a temporal uncertainty on the order of days. Other databases do not attempt to populate the field of eruption end-dates. The Large Magnitude Explosive Volcanic Eruption (LaMEVE) database contains an unpopulated end-date category due to uncertainties in securing the date of eruption end from non-historical events (Crossweller *et al.*, 2012).

Table 1.2 summarises the current approaches for identifying or modelling the end of volcanic eruptions. These modelling approaches can be split into the broad categories of general physically-motivated models, eruption-specific models and probabilistic models.

Wadge (1981) formulated a simple model of eruption rate over time for effusive eruptions. Figure 1.2 illustrates the asymmetric distribution of effusion rate; whereby effusion rate increases to a maximum, followed by logarithmic decay in effusion rate. The evolution of effusion rate ($Q_{f,t}$) throughout the eruption may be modelled in terms of the maximum effusion rate (Q_{max}) and an exponential decay constant (λ):

$$\boldsymbol{Q_{f,t} = Q_{max} e^{-\lambda t}} \quad (1)$$

In this model the end of the eruption is physically represented by a minimum effusion rate below which eruption cannot continue. Bonny and Wright (2017) used MODIS (MODerate Resolution Imaging Spectroradiometer) to determine Q_{max} and l for 104 eruptions over the past 15 years to test the applicability of the Wadge model and found that approximately 1/3 of eruptions conformed to this pattern. However, a wide range of decay constants was observed and there was not a representative average decay constant even between eruptions at the same volcano, thereby limiting this technique as a real-time forecasting tool.

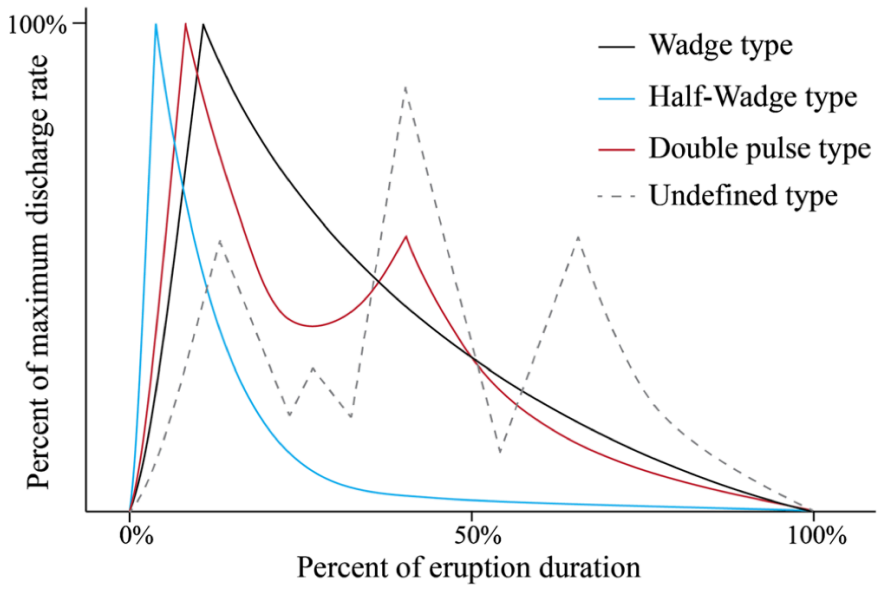


Figure 1.2: Schematic illustration of the asymmetric distribution of lava discharge rate time-series proposed by Wadge (1981). For comparison are schematic illustrations of other forms of discharge rate identified by Bonny and Wright (2017), including half-Wadge (blue), double pulse (red) and undefined (dashed). From Bonny and Wright (2017).

Pinel *et al*. (2010) modelled cyclic volcanic activity and edifice growth using a simple physical model in which replenishment and eruption depend upon edifice size and dimensions. This model assumes that the magma reservoir remains hydraulically connected to a deep magma replenishment source, embedded within an elastically-deforming crust. Edifice growth leads to greater pressure exerted on the magmatic system hence lower magma supply rate as the eruption progresses. The model encompasses a definition of eruption end: the point at which the eruptive fissure closes due to confining stress, at the point where the reservoir pressure equals the hoop stress (the normal stress in the direction of the chamber azimuth). However, other physical mechanisms which can contribute to eruption end such as degassing and crystallisation are not resolved within the model.

Eruption-specific models can yield insights into eruption end where elements of the underlying structure of the volcanic system are known. Hooper *et al*. (2015) forecasted the end-date of 2014-2015 Bárðarbunga eruption to within an error of 9 days by modelling the caldera as piston collapse, combined with GPS ground deformation measurements. This model produces exponential decay of effusion rate similar to that of Wadge (1981), but the overall driver of the effusion change is caldera collapse. A piston-like model has potential application to future caldera-forming eruptions. However, the extent to which silicic systems would conform to the underlying assumption of a slow collapse where material exits laterally is not clear. Where a volcanic system is well-observed, the drivers of eruption end can be qualitatively determined. The Mt St Helens dome-forming eruption of 2004 - 2008 was monitored with seismic data, GPS, tilt, remote cameras, and gas emissions, from which Dzurisin *et al*. (2015) formulated a conceptual model of eruption end. In this model, eruption end was primarily due to decreasing overpressure attributed to overburden from dome growth and an emptying reservoir where there was little/no reservoir recharge during the eruption. They also suggest that friction between semi-solid magma and conduit walls could have generated friction which acted against the progression of eruption.

Anderson and Segall (2013) used a physics-based forward model in combination with a Monte Carlo Markov Chain inversion of GPS and extrusion rate to estimate eruption parameters of the Mt St

Helens eruption of 2004 - 2008. They find that the model underfits GPS observations close to the end of eruption and does not capture the relatively rapid post-eruptive reinflation after eruption end, thereby indicating that more complex processes governing chamber influx were occurring than those currently included within the model. Segall (2013) describes the deterministic forecasting of the final eruptive volume using this physics-based model, and retrospectively suggests that a model incorporating no eruptive recharge fitted after approximately 2 years of eruptive data has good agreement with the final eruptive volume. However, key physical processes which could contribute to eruption cessation, such as crystallisation or cooling, are not parameterised within the model.

Probabilistic models can be used to better understand eruption end without assumption of any physical parameters. Gunn *et al*. (2014) produced a probabilistic model of eruption duration at Etna based on survivor function analysis using historical records. It is possible with this model to forecast the probability of an eruption exceeding a given duration based on the time since a previous eruption. However, although this model can be applied to other systems, there are limitations as there are not many volcanoes which have a dense enough record of eruption for this type of analysis. Moreover, the model is based on recorded eruption start and end dates, which may in themselves be uncertain in the absence of well-defined criteria.

Table 1.2: Summary of current approaches for identifying or modelling eruption end.

Type of model	Model name	Data types	Parameters which control eruption end	References	Comments
Physically-motivated models	Wadge model	Remotely-sensed thermal data	Release of magma chamber overpressure	Wadge, 1981 Bonny and Wright, 2017	
	Open hydraulic model	N/A	Reservoir pressure decrease (due to magma withdrawal)	Pinel *et al*., 2010	Simple physical model (not validated with observational data)
Eruption-specific physical models	Forecasting the cessation of laterally propagating dyke during the 2014-15 Bárðarbunga eruption	Geodetic measurements (GPS and InSAR)	Lateral outflow of material; piston collapse of caldera	Hooper *et al*., 2015	
	End of Mt St Helens 2004 - 2008 eruption	Seismic data; tilt measurements	Decreased overpressure due to dome growth; friction in the conduit; no magmatic recharge	Dzurisin *et al.,* 2015	
		GPS; erupted volumes	Lack of recharge	Anderson and Segall (2013) Segall (2013)	
Probabilistic models	Eruption duration forecast	Duration of previous eruptions at Etna	N/A	Gunn *et al*., 2014	

1.3) Machine learning

Machine learning is the process by which computers learn patterns in 'seen' data, which can give insights on unseen data. Supervised machine learning classification is when models are trained on labelled data which is representative of the classes being identified, and then tested on unseen data withheld during training (Bishop, 2006). Multi-class classification tasks involve selecting training data which is representative of each class for model training. One-class classification or novelty detection is the process of classifying testing data which differ from the data observed during training (Pimentel *et al*., 2014). Novelty detection problems typically involve a training dataset in which there are many more examples of the "normal" class than the "abnormal" class.

Machine learning models make inferences based on a representation of the data, which can be referred to as "features". There are two main approaches towards deriving these features (e.g., Majtner *et al*., 2016; Nanni *et al*., 2017):

(i) Hand-crafted features are those which are designed by the user, which can represent fundamental physical characteristics of the data which are expected to vary according to the classification task at hand (e.g., for classifying volcano-seismicity, the frequencies contained within the waveform, or the peak amplitude of the waveform might be two attributes expected to change with event type).

(ii) Learned features are representations of the data which are created by passing the original data (whether that be, for example, a raw waveform or a spectrogram representation of a volcano-seismic event) through a series of hidden layers which perform operations (e.g., convoluting and downsampling the original data).

Both approaches to deriving features have advantages and disadvantages. Hand-crafted features can be beneficial as they do not require storing the raw data in memory which, for continuous measurements such as a seismic waveform, can have thousands of individual samples (where a sample represents each reading) for one example. Moreover, using hand-crafted features can enhance

the interpretability of the final models as inferences on the relative importance of features can be directly interpreted in terms of physical characteristics of the input data (Lin *et al*., 2020). In the case of the seismic analysis in Chapters 3 and 4, features were selected which share fundamental similarities with those used in current monitoring approaches, such as seismic event rates or waveform frequency information. However, hand-crafted features come at the expense of requiring the user to perform the necessary pre-processing steps on the data, which may not be trivial. For example, multiplet analysis is computationally expensive (Rodgers *et al*., 2015b). Moreover, due to the selection of features for a specific task, one set of hand-crafted features may not scale well to alternative classification tasks (e.g., Chan *et al*., 2015).

Learned features are calculated directly from the input observations as opposed to being manually designed for the classification task, which can mean that they are transferable to other tasks (Nanni *et al*., 2017). Additionally, learned features are thought to represent more abstract elements of the original data, which are more robust to the variability within each class (Chan *et al*., 2015). However, learned features require large amounts of training data to generate representative features (e.g., Le, 2013) which, for supervised models, can require significant human cost (in time or resources) to annotate accurately. Moreover, these large datasets require considerable computational power to train models.

Table 1.3 is a summary of existing machine learning approaches applied so far in volcanology. A considerable number of machine learning approaches have been developed within the field of volcano-seismic analysis. These include the detection, classification, and characterisation of phases of volcano-seismic events. Early volcano-seismic models were primarily specific to seismic events from one volcanic system; however, recent contributions have produced models which can be extended to multiple volcanic systems (Bueno *et al*., 2020; Cortés *et al*., 2021). Models to classify volcanic state as eruptive or non-eruptive have been successfully applied to continuous monitoring data such as seismic (Chapter 3; Ren *et al*., 2020; Dempsey *et al*., 2020), geodetic and gas times series (Chapter 4). However, machine learning approaches applied to remote sensing measurements including InSAR-

derived ground deformation, OMI-derived SO_2 emission data or video camera footage have had success in the automatic detection of volcanic unrest. Machine learning approaches have also been applied to geochemical data such as major element, trace element or isotopic compositions, some of which can be accessed through open-access petrological databases. These studies have classified the source of volcanic rocks, whether that be the broad tectonic setting (Petrelli and Perugini, 2016; Ueki *et al*., 2018) or the specific source volcano within a regional setting (Bolton *et al*., 2020). Petrological data have also been used as input to machine learning-derived thermobarometers to estimate magma storage depths prior to eruption (Petrelli *et al*., 2020).

The ability of machine learning approaches to learn patterns in the data is based on the accuracy of labels within the training dataset. Consideration of how to label data is necessary for volcanic applications. When classifying volcanic state as eruptive or non-eruptive, obtaining "true" labels for activity on a given day can be ambiguous in the absence of volcano-specific definitions of when a volcano is erupting. Ren *et al*. (2020) classify activity at Piton de la Fournaise Volcano, La Réunion, where the eruptive activity is strongly associated with the occurrence of volcanic tremor (Battaglia and Aki, 2003; Battaglia *et al*., 2005) and therefore ground-truth labels are well-defined. For volcanoes which lack these specific definitions, the start- and end-dates contained within the Smithsonian Global Volcanism Programme (GVP) catalogue can be used for comparison with the output of the machine learning classification against. However, these labels can be uncertain, with end-dates suggested to have uncertainty on the order of days or more (Phillipson *et al*., 2013). Hence, other metrics to evaluate the classifications are the occurrence of visual eruption (such as ash emission) or the consistency of the classification. These metrics are explored in Chapter 3. For other classification tasks, labels may be imprecise depending on how complex the task of labelling the data is (e.g., Bouveyron and Girard, 2009). In the field of automatic volcano-seismic event classification, Cortés *et al*. (2021) suggest that the agreement between analysts labelling the same data is approximately 80%, though they do not provide a quantitative basis for this estimate.

Table 1.3: Summary of machine learning (ML) approaches in volcanology.

Field	Aim	Methods	Data types	References
Volcano-seismic analysis	Detection of volcano-seismic events	Hidden Markov Models (HMMs)	Seismic waveforms	Gutiérrez *et al*., 2009 Bhatti *et al*., 2016 Lara-Cueva *et al*., 2016 Bueno *et al*., 2020
	Classification of volcano-seismic events	Support Vector Machine (SVM); Random forest; Convolutional Neural Networks (CNNs); Bayesian Neural Networks (BNNs)	Seismic waveforms; Spectrograms;	Masotti *et al*., 2006 Langer *et al*., 2006 Apolloni, 2009 Malfante *et al*., 2018 Bueno *et al*., 2019 Bueno *et al*., 2020 Canario *et al*., 2020a Cortés *et al*., 2021 Lara *et al*., 2021
	Identification of seismic phases	CNNs	Seismic waveforms	Lapins *et al*., 2021
Classifying volcanic activity	Identifying eruptive transitions	Random forest classification, SVM, logistic regression, GPC, k-means clustering	Seismic time series; geodetic and gas time series; Infrasound	Ren *et al*., 2020 Dempsey *et al*., 2020 Manley *et al.,* 2020 Witsil and Johnson, 2020a
Volcanic monitoring	Identifying unrest via satellite-derived ground deformation	CNNs; clustering methods	InSAR	Anantrasirichai *et al*., 2018; Gaddes *et al*., 2019
	Identifying unrest via satellite-derived SO_2 emissions	Logistic regression	Ozone Monitoring Instrument daily SO_2 emissions	Flower *et al*., 2016
	Identifying unrest via video camera observations	Artificial Neural Network (ANN)	Continuous video data	Witsil and Johnson, 2020b
Thermobarometry	Estimating magma storage depths and pre-eruptive temperatures	ML-based stochastic gradient boosting; extremely randomised trees; random forest; k-Nearest Neighbour; decision tree regression	Major element compositions	Petrelli *et al*., 2020
Classifying a source from geochemical data	Associating rock compositions with tectonic setting	SVM; random forest; sparse multinomial regression	Major element compositions; trace element compositions; isotopic compositions	Petrelli and Perugini, 2016 Ueki *et al*., 2018
	Associating tephra compositions with Alaskan source volcano on a regional scale	Random forest; SVM; Artificial Neural Networks	Major element glass compositions	Bolton *et al*., 2020

1.4) Thesis structure

This thesis is structured as follows:

Chapter 2:

This chapter is a discussion of the data used in this thesis. I describe the process of data collection and processing of the data used in Chapters 3 - 5, and the extraction of representative features used to classify volcanic activity in Chapters 3 and 4.

Chapter 3:

This chapter is an investigation into the end of volcanic eruptions, exploring the utility of machine learning classification to identify eruptive transitions. Four multi-class methods (Support Vector Machine, Logistic Regression, Random Forest, and Gaussian Process Classification) were applied to features derived from seismic data from Nevado del Ruiz and Telica volcanoes. The resulting classifications are compared to independent constraints on volcanic activity, such as the occurrence of visual eruption (e.g., ash emission or explosions).

Chapter 4:

The approaches applied in Chapter 3 are extended to include seismic, ground deformation and gas measurements from the 2005 – 2006 unrest and eruption of Augustine Volcano, Alaska. Both novelty detection, in which only non-eruptive days are used to train the model, and multi-class classification are applied to features derived from these datasets. Two contrasting approaches are applied to interpolate missing values in SO_2 emission rate time series. I compare the results of the novelty detection models which are trained on all types of data to models which are only formulated on each type of data in isolation. Additionally, I compare the separability of features which are derived from a seismic catalogue with more stringent criteria for inclusion to a seismic catalogue containing the full set of events.

Chapter 5:

Considerable prior work has been done to develop machine learning approaches for the classification of volcano-seismic events. Previous approaches have required a large, labelled training dataset to successfully train a model. Active learning is an approach wherein a model learns quicker by self-selecting the training data from which it learns. This paradigm has great potential utility in the field of volcano monitoring, in which there are large burdens on the time of analysts. An active learning approach is applied to two datasets for which the analyst-derived labels are withheld, using an existing volcano-seismic event classifier, and the performance of this classifier using actively-sampled data and randomly-sampled data is evaluated.

Chapter 6:

This chapter contains a summary of the work presented in this thesis. Future extensions of the approaches contained within this thesis are discussed, given the availability of datasets.

1.5) Related references

Chapter 3:

A very similar version of this work was published as:

Manley, G.F., Pyle, D.M., Mather, T.A., Rodgers, M., Clifton, D.A., Stokell, B.G., Thompson, G., Londoño, J.M. and Roman, D.C., 2020. Understanding the timing of eruption end using a machine learning approach to classification of seismic time series. *Journal of Volcanology and Geothermal Research*, *401*, p.106917.

GM lead this work and undertook the vast majority of the analysis and writing. MR performed the frequency analysis and multiplet analysis on both datasets, and the detection on the Telica dataset. JML provided access to the Nevado del Ruiz catalogue. Preliminary research on the methods used in this chapter were performed by BS whilst on a NERC-funded Research Experience Placement. GM then expanded and refined the methods, in addition to contextualising the results and producing all the published figures.

Chapter 4:

A very similar version of this work was published as:

Manley, G.F., Mather, T.A., Pyle, D.M., Clifton, D.A., Rodgers, M., Thompson, G. and Roman, D.C., 2021. Machine learning approaches to identifying changes in eruptive state using multi-parameter datasets from the 2006 eruption of Augustine Volcano, Alaska. *Journal of Geophysical Research: Solid Earth*.

GM lead this work and undertook the vast majority of the analysis and writing. Specifically: GM performed the detection and feature extraction of the Augustine seismic data and developed the CLVF and GPR interpolation and code to run the models with guidance from TM, DP, DC, and MR. GT provided access to the Augustine waveform data and AVO catalogue.

Chapter 5:

The work in Chapter 5 has been accepted at Frontiers in Earth Science.

Manley, G.F., Mather, T.A., Pyle, D.M., Clifton, D.A., Rodgers, M., Thompson, G. and Londoño, J.M., 2022. A deep active learning approach to the automatic classification of volcano-seismic events. *Frontiers in Earth Sciences.*

GM lead this work and undertook the vast majority of the analysis and writing. JML provided access to the Nevado del Ruiz catalogue. The Llaima catalogue was accessed from Canario *et al*. (2020b). GM performed the analysis in this chapter with guidance from TM, DP, DC and MR.

Chapter 2 - Data processing

This chapter is an introduction to the acquisition of the seismic, gas and geodetic data used in this thesis. For each data type, I discuss the data collection and processing, and discuss how representative hand-crafted features can be extracted to build machine learning models (explored in Chapters 3 and 4).

2.1) Seismic data

2.1.1) Data collection

Seismic data are collected from seismometers, which are sensors designed to respond to ground vibrations. The seismic data used within this thesis are collected from three-component seismometers, designed to detect simultaneous ground motion in the up-down, north-south, and east-west directions. Broadband seismometers are designed to record ground motion over a wide range of frequencies (typically in the region of 0.01 - > 25 Hz, though individual instrument ranges will vary). Short-period seismometers can detect a smaller range of frequencies (in the region of 1 - > 25 Hz), and are typically used for monitoring local ground motion. Seismometers are a form of inertial sensor, which comprise a 'proof mass' suspended within a frame (Ackerley *et al*., 2014). Motion is sensed by detecting the differential motion between the mass and the frame using a transducer, which outputs the displacement as a voltage. The seismometer output is coupled to a digitiser, which converts the voltage output of the seismometer to a count reading. Typically, data from the seismic network is transmitted to a remote station via telemetry (Calderoni *et al*., 1997).

2.1.2) Continuous data processing

Figure 2.1 and Table 2.1 illustrate the seismic networks and configurations, type of seismometers, sampling rate and data formats for the datasets used throughout this thesis. Waveform data are obtained in either SAC or SEED formats for all datasets apart from the Llaima dataset, where the waveforms time series were obtained from Canario *et al*. (2020b) as NumPy files with frequency filters already applied. The Nevado del Ruiz dataset is obtained in a catalogue format with each event

file stored separately in SAC format. For all datasets, noise-reduction is applied by filtering each waveform to a given frequency range by using a band pass filter with 2 poles, as illustrated in Figure 2.2. Filtering out low frequencies (e.g., < 0.5 Hz) removes low frequency noise or microseism which can be caused by natural processes including water motion. Filtering out high frequencies removes high frequency noise which can be due to wind motion (e.g., > 20 Hz). Additionally, linear trends are removed from each waveform for all datasets.

For the Telica and Augustine datasets, the continuous seismic waveform data are processed to detect events contained within it. This detection is performed using a Short-Term Average / Long-Term Average (STA/LTA) ratio triggering method (e.g., Trnkoczy, 2009). The STA/LTA trigger works by comparing the average energy within a short-term window to the average energy contained within a preceding long-term window. If the ratio is found to be larger than a given threshold value, then a trigger is recorded for that station. The STA/LTA algorithms are run on the vertical component of the seismic data twice: once using a low-frequency filter and another using a high-frequency filter. The triggers of the STA/LTA detection are processed to exclude detections which are unlikely to have a volcanic origin. The timing of detections on all stations within the multi-station seismic network is compared within a network-specific tolerance (typically in the region of seconds), and detections which do not occur at three stations are discarded. Regional seismic events which have a non-volcanic origin but are large enough to trigger a detection are excluded from the datasets by one of two methods, depending on the available software. For the Telica dataset, the approximate origin locations for the detected events are computed using Antelope Environmental Data Collection software and events which do not have an origin within a pre-specified grid within the region of the volcano are excluded (e.g., Rodgers *et al*., 2015a). For the Augustine dataset, the times of regional events of a given size in the time spanned by the dataset, and within the regional area, are obtained. Then, the detections which correspond to these times are manually inspected and discarded. In Chapter 3 for the Telica dataset, the continuous seismic data are also processed to compute the Real-time Seismic Amplitude Measurement (RSAM). RSAM is a measure of the average signal amplitude in a user-specified time window (e.g., per minute or hourly) (Endo and Murray, 1991).

2.1.3) Waveform processing

A seismic catalogue is created by extracting the waveforms corresponding to the detection times from the continuous data at a single seismic station. The seismic monitoring infrastructure available at volcanoes can be variable (Moran *et al*., 2008), and the number of seismic stations can be increased following the onset of unrest. For example, the seismic network at Agung, Indonesia was expanded from two stations to eight stations when unrest peaked in September 2017 (Sahara *et al*., 2021). Therefore, a consideration for developing methods which could be expanded to a monitoring context is to establish whether these methods yield successful results using minimal quantities of data, i.e., from a single seismic station. The station is selected based on the proximity to the volcano and completeness of the seismic record for the station during the study period. It should be noted that, though the waveforms in the catalogue are derived from continuous data from a single station, multi-station information has been used to exclude detections with a non-volcanic origin as described in Section 2.1.2. Where only one station is available for the detection of events, consideration of how non-volcanic events which cannot be excluded using multi-station information might affect the quality of the seismic catalogue would be necessary. Once a catalogue of detections has been obtained, the waveforms are processed to extract fundamental spectral and amplitude characteristics from each waveform. These attributes could be better-suited to distinguishing between non-eruptive and eruptive activity than the raw waveform time-series data.

The frequency content of waveforms can vary with source process; thus, frequency characteristics of each waveform are a first-order method to distinguish between types of volcanic earthquake. The dominant frequencies of a waveform are calculated by taking the Fourier spectrum of the waveform and extracting the top N-frequencies with greatest amplitude within this spectrum. More than one frequency can be chosen, rather than the top frequency, because the top dominant frequency may not adequately characterise volcano-seismic events with a bimodal frequency distribution (Buurman and West, 2010). Dominant frequency has previously been used to distinguish between types of volcanic events (e.g., Latter, 1981). The band ratio or frequency index is an additional frequency metric after

Buurman and West (2010), which is calculated via the ratio of the mean spectral amplitude in a high-frequency band (A_{high}) and the mean spectral amplitude in a low-frequency band (A_{low}):

$$\boldsymbol{Band\ ratio} = \mathbf{log_{10}}\left(\frac{\boldsymbol{mean(A_{high})}}{\boldsymbol{mean(A_{low})}}\right) \quad (2)$$

A negative band ratio indicates that the waveform has primarily low-frequency energy, whereas positive band ratio indicates that the energy is primarily in high-frequencies. The high- and low-frequency ranges depend on characteristic frequencies of waveforms at a given volcanic system. For Augustine and Nevado del Ruiz, the low- and high-frequency bands were set as 1 – 2 Hz and 10 – 20 Hz respectively, after Buurman and West (2010). For Telica, the low- and high-frequency bands were set as 1 – 6 Hz and 6 – 11 Hz respectively, and the band ratio was taken as the base 2 log rather than base 10 log (Rodgers *et al*., 2015a). Peak amplitudes of each waveform are computed from the peak-peak amplitude of each event, and are expected to vary with the size of the generating event, e.g., the size of an explosion (Murphy and Lahoud, 1969). Additionally, the spectral width of a waveform can vary with event type, with LPs generating signals with a relatively narrow spectral peak compared to other events (e.g., Clarke *et al*., 2021). To approximate the spectral width of the waveform, the frequency range which spans either side of the Fourier spectrum peak to within half of the peak amplitude is calculated.

Multiplets are families of repeating earthquakes which share similar event characteristics (Rodgers *et al*., 2015b). The multiplet analysis on the Telica and Nevado del Ruiz data in Chapter 3 was computed using the program Peakmatch (Rodgers *et al*., 2015b). Within Peakmatch, repeating events are identified from a cross-correlation of waveform-pairs, which is computed by calculating the sliding dot product of two waveform time-series vectors. Multiplets or waveform families are then identified by matching waveforms to a "master" event which defines the waveform family.

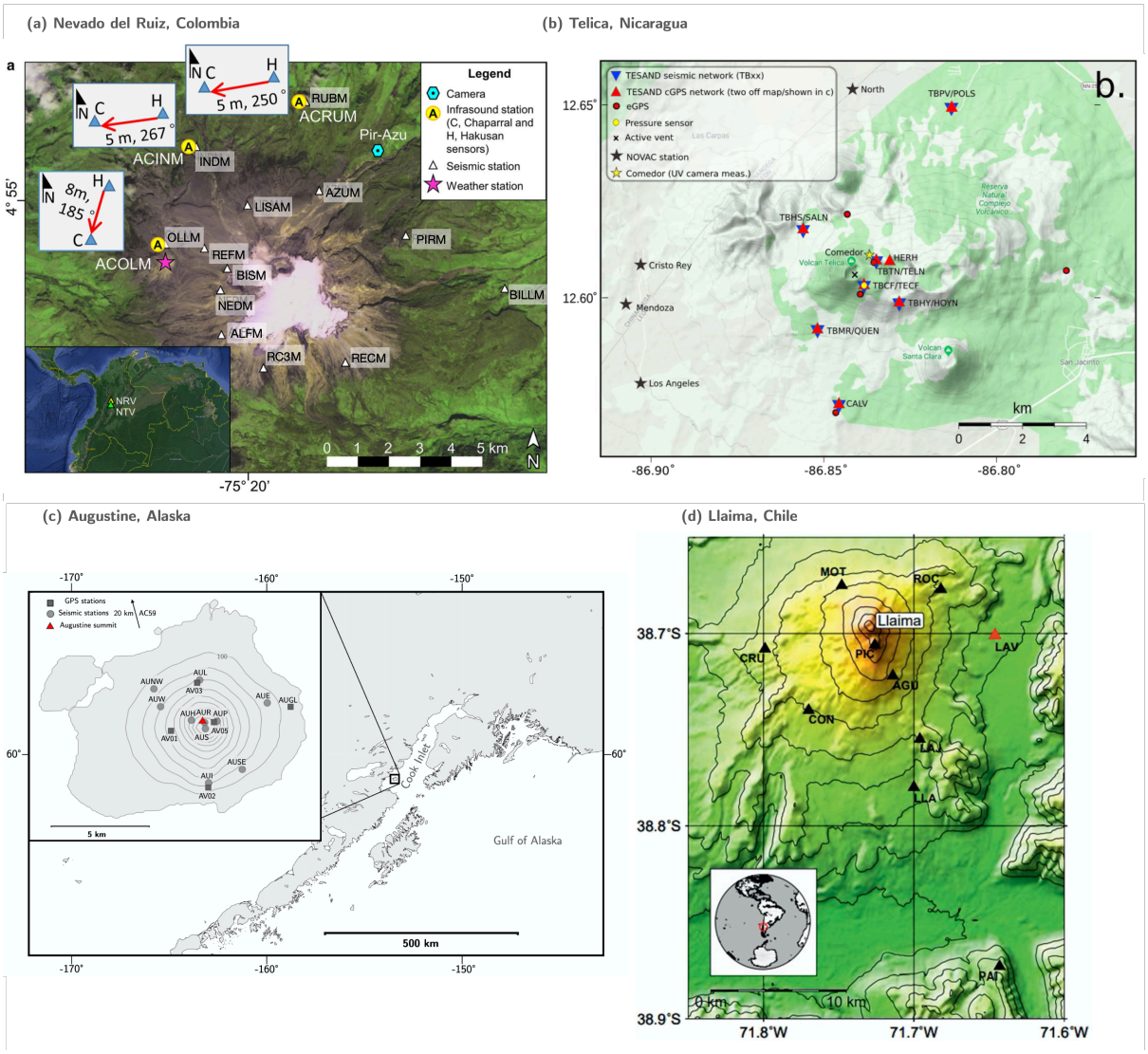


Figure 2.1: Map of seismic network for the seismic data used in this thesis. (a) Nevado del Ruiz seismic station map, modified from Castaño *et al.* (2020). (b) Telica seismic station map, from Roman *et al.* (2019). (c) Augustine seismic station map, from Chapter 4 of this thesis. (d) Llaima seismic station map, from Canario *et al.* (2020b)

Table 2.1: Summary of network parameters for the seismic data used in this thesis.

Dataset	Type of seismic stations within network	Number of seismometers in network	Seismic station used	Instrument type	Sampling rate (Hz)	Frequency range filtered (Hz)	Data format	Obtained in catalogue format?
Nevado del Ruiz	Mixed (short-period, broadband	10	BISZ	Short-period	100	[0.5,20]	SAC	Yes
			TOLZ	Short-period	100	[0.5,20]	SAC	Yes
Telica	Broadband	6	TBTN	Broadband (Guralp 6TD/6 T)	100	[0.5,20]	SEED	No
Augustine	Mixed (short-period, broadband)	11	AUI	Short-period	100	[0.5, 18]	SEED	No
Llaima	Mixed (short-period, broadband)	9	LAV	Broadband	100	[1, 10]	Time series	Yes

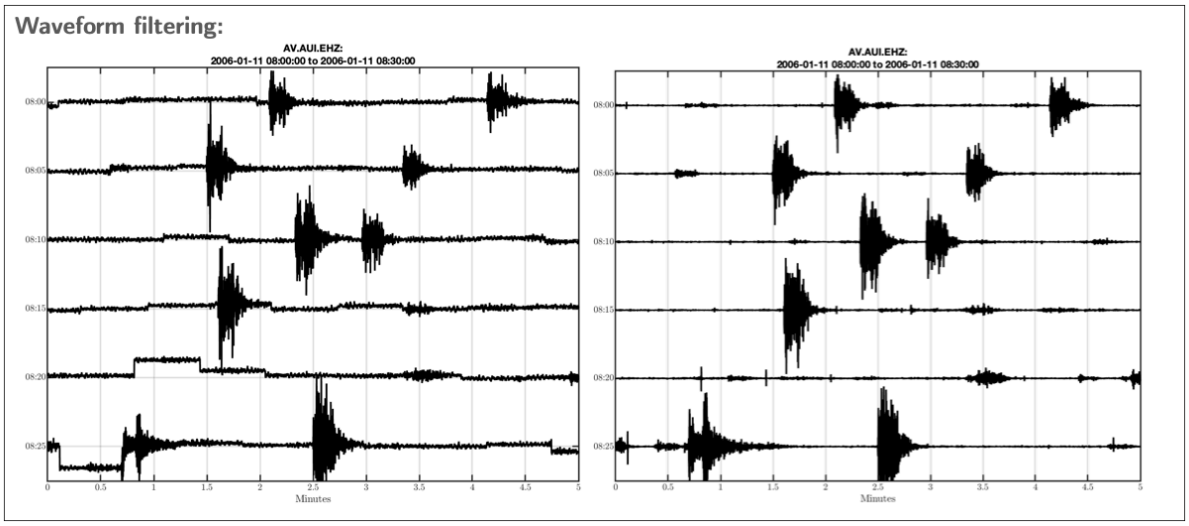


Figure 2.2: Illustration of the noise reduction process for the seismic data. The left-hand panel is a 30-minute sample of continuous seismic waveform data and the right hand panel is the same 30-minute sample with a band pass filter from 1 Hz – 18 Hz applied and linear trends removed.

2.1.4) Feature extraction

The waveform characteristics described in Section 2.1.3 are computed for each waveform. These individual waveform characteristics are then transformed into representative daily values by taking summary statistics over a daily time period. These summary statistics have been widely-used for

representing physiological time-series, and provide advantages such as being simple to compute and more robust to noise than the raw time series data (e.g., Guo *et al*., 2020). The summary statistics used to calculate features in Chapters 3 and 4 are the mean, median, variance, minimum, maximum, 10^{th} percentile, 90^{th} percentile and change in mean from the previous day.

For characteristics which are not based on individual waveform characteristics, different summary statistics are used to calculate a representative daily value. Event rate, such as increasing rate of VT seismicity, is considered to be an important precursor to eruption for some volcanoes including basaltic systems (e.g., Cameron *et al*., 2018). The event rate per day and the change in event rate from the previous day are calculated as features. For Telica, where classifications are automatically generated based on the frequency content (Rodgers *et al*., 2015a), the event rates for High-Frequency and Low-Frequency events are also included. For RSAM data, mean and variance of the per-hour readings and the change in RSAM mean from the previous day is also computed.

2.2) Gas data

2.2.1) Data collection

Here I discuss the data collection process for the Augustine gas time series data used in Chapter 4 of this thesis, which was described by McGee *et al*. (2010). The gas measurements recorded at Augustine prior to, during, and after the 2005 – 2006 unrest and eruption were measured on airborne traverses of the area using aircraft-mounted instruments. The SO_2 emissions were measured using a correlation spectrometer (COSPEC) combined with a GPS receiver to measure the latitude, longitude, and altitude of each measurement. The COSPEC instrument works by comparing the ultraviolet absorption spectrum of SO_2 in the measurement region (i.e., the volcanic plume) against a reference absorption spectrum (Moffat and Millan, 1971). Each measurement of SO_2 emission rate at Augustine was computed from between 1 and 7 airborne traverses. McGee *et al.* (2010) transform the traverse measurements of the COSPEC SO_2 signal into a daily emission rate to determine the area under the signal curve (units km * ppm * m). This signal is converted to a daily emission rate (E) (in t / d) via

the following calculation, where S_w represents the average wind speed, W_p represents the plume width, A represents the resultant average of the plume traverse, and 0.00023 is the ratio of the COSPEC calibration cells used during the airborne traverse:

$$E_{COSPEC}^{SO_2} = \mathbf{0.00023} * (S_w * W_p * A) \quad (3)$$

The SO_2 emission rates used in this thesis are taken from McGee *et al*. (2010) in a time series format in units of tonnes per day.

2.2.2) Data processing

An increase in measured SO_2 at the surface can reflect the ascent of new magma into the shallow volcanic system (e.g., Pritchard *et al*., 2019). This increase can reflect either the process of magmatic degassing, as SO_2 emissions are favoured by low-pressure conditions (relative to other species such as CO_2), or the boiling and subsequent removal of a shallow hydrothermal system which scrubbed and therefore suppressed SO_2 emissions at the surface. Hence, the daily SO_2 emission rate is the quantity from which gas features are extracted. However, the SO_2 emission rate measurements at Augustine are made at irregular intervals. Therefore, the individual daily SO_2 emission rates after McGee *et al*. (2010) must be interpolated to estimate a measurement on the intervening days.

Two approaches are taken towards interpolating values for the irregularly-sampled SO_2 emission rate time series. The first approach is the Carry Last Value Forward (CLVF) method, where the daily value of SO_2 emission rate is equal to the last measured value of SO_2 emission rate. This approach has previously been used to interpolate measurements in health time series data for input into machine learning models (Shamout *et al*., 2019). This method uses no future information to interpolate the daily values and is computationally inexpensive to compute. Moreover, daily values are based upon the measured values and, unlike a higher-order interpolation method, no functional form is assumed for the nature of transitions in SO_2 emission rate (e.g., a linear interpolation would assume that SO_2

emission rates are linearly increasing, and use a future point to interpolate the intervening values). However, CLVF can permit a large increase in the daily interpolated value of SO_2 emission rate, especially where there is a large interval between measurements.

The second approach to interpolating daily values of SO_2 emission rate is Gaussian Process Regression (GPR). This method is non-parametric, whereby no functional form for the model is specified, and probabilistic, therefore producing an estimate of the distribution of the SO_2 emission rate on a given day. A GPR model is formulated by fitting a latent function to the data which can be specified by a mean function $m(\boldsymbol{x})$ and covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$ (Rasmussen and Williams, 2006):

$$\boldsymbol{f}(\boldsymbol{x}) \sim \boldsymbol{GP}(\boldsymbol{m}(\boldsymbol{x}), \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}')) \quad (4)$$

The mean function is set as the mean of the training dataset. The covariance function maps the similarity of pairs of input data $\boldsymbol{x}$ and output data $\boldsymbol{x}'$. The variations in SO_2 emission rate measured at Augustine spanned three orders of magnitude, therefore a matérn function which permits transitions of this nature was chosen for the GPR models (Rasmussen and Williams, 2006), with smoothing parameter 5/2 chosen as this formulation has been successfully used for applying GPR to data with limited samples (Zhang *et al*., 2018). For this covariance function $k(x, x')$, σ_l represents the characteristic length-scale of the covariance function, σ_f represents the signal standard deviation, and $r = \sqrt{(x - x')^T (x - x')}$ represents the Euclidean distance between x and x'.

$$\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\sigma_f^2} \left(\mathbf{1} + \frac{\sqrt{5}\boldsymbol{r}}{\boldsymbol{\sigma_l}} + \frac{\sqrt{5}\boldsymbol{r}^2}{3\boldsymbol{\sigma_l^2}}\right) \mathbf{exp}\left(-\frac{\sqrt{5}\boldsymbol{r}}{\boldsymbol{\sigma_l}}\right) \quad (5)$$

The GPR model is fitted using the in-built function fitrgp in MATLAB, which uses a quasi-Newton solver to estimate the hyperparameters of the covariance function (such as the length-scale) and the noise variance of the model by maximising the marginal log likelihood of the model (Mathworks,

2021). After a GPR model is fitted to the measured SO_2 emission rate values, the model is evaluated at every day in the study period.

2.2.3) Feature extraction

From the interpolated SO_2 emission rate time series, two features are derived: the daily value for SO_2 emission rate (in the case of GPR, the mean value of daily emission rate) and the change in daily value for SO_2 emission rate from the previous day.

2.3) Geodetic data

2.3.1) Data collection

Here I discuss the data collection process for the geodetic time series data used in Chapter 4 of this thesis, which was described by Pauk *et al*. (2010). The geodetic data from Augustine Volcano were collected from the Plate Boundary Observatory network GPS stations at Augustine. Figure 2.2c illustrates the configuration of the GPS network at Augustine. GPS stations collect precise position information at a given interval (e.g., 15 seconds), which are then processed to create a time series of displacement relative to a given reference frame. The GPS stations on Augustine Island were situated in areas of competent bedrock, where short drilled-braced monuments were installed. Each station was equipped with a Trimble NetRS geodetic grade GPS receiver, and a radome covered Trimble dual frequency (L1/L2) choke ring antenna.

For the geodetic data at Augustine, the Nevada Geodetic Laboratory (NGL) daily solutions for stations AV01, AV02 and AC59 were used (Blewitt *et al*., 2018). The NGL daily solutions are calculated from the GPS data by using a point positioning approach (Zumberge *et al*., 1997) with GIPSY/OASIS software and NASA Jet Propulsion Laboratory orbit and clock products for the GPS constellation. The daily solutions used in this thesis are linked to the IGS14 model in the ITRF14 reference frame (Rebischung and Schmid, 2016; Altamimi *et al*., 2016). We obtain these daily solutions from the NGL website in a time series format corresponding to metres of displacement.

2.3.2) Data processing

To represent the deformation signal at Augustine, station components and combinations are chosen such that a vertical component and a horizontal component of deformation are represented. The daily vertical position of station AV01 is used as an approximation for the vertical component of deformation, as the deformation during the 2005 - 2006 eruptive period exceeds the typical seasonal signal at this station. For the horizontal component of deformation, the daily change of the N-S components of the station pair AV02-AC59 baseline vector were used. For the geometry of these two station the baseline variation is a proxy for N-S extension or shortening across the volcano (e.g., Dixon *et al.*, 1991; Cervelli *et al.*, 2010; Saballos *et al.*, 2014).

2.3.3) Feature extraction

The features used are the individual position values for the vertical (AV01 vertical component) and horizontal (change in N-S components of AV02-AC59) component of deformation described above, and the change in this individual position value from the previous day.

Chapter 3 - Understanding the timing of eruption end using a machine learning approach to classification of seismic time series

3.1) Abstract

The timing and processes that govern the end of volcanic eruptions are not yet fully understood, and there currently exists no systematic definition for the end of a volcanic eruption. Currently, end of eruption is established either by generic criteria (typically 90 days after the end of visual signals of eruption) or criteria specific to a given volcano. We explore the application of supervised machine learning classification methods: Support Vector Machine, Logistic Regression, Random Forest, and Gaussian Process Classifiers and define a decisiveness index D to evaluate the consistency of the classifications obtained by these models. We apply these methods to seismic time series from two volcanoes chosen because they display contrasting styles of eruption: Telica (Nicaragua) and Nevado del Ruiz (Colombia). We find that, for both volcanic systems, the end-date we obtain by classification of seismic data is 2 - 4 months later than end-dates defined by the last occurrence of visual eruption (such as ash emission). This finding is in agreement with previous, general definitions of eruption end and is consistent across models. Our classifications have a higher correspondence of eruptive activity with visual activity than with database records of eruption start and end. We analyse the relative importance of the different features of seismic activity used in our models (e.g., peak event amplitude, daily event counts) and find little consistency between the two volcanic systems in terms of the most important features which determine whether activity is eruptive or non-eruptive. These initial results look promising, and our approach may offer a robust tool to help determine when an eruption has ended in the absence of visual confirmation.

3.2) Introduction

The processes which govern large-scale change in volcanic systems are not yet fully understood. Volcanic systems are dominated by complex and non-linear processes. This complexity has implications for both understanding and forecasting the onset of volcanic activity (e.g., Sparks, 2003), and for managing transitions in behaviour during prolonged eruptions (e.g., Sparks and Aspinall, 2004; Hicks and Few, 2017; Barclay *et al*., 2019). While much attention has been focussed on forecasting the timing of eruption onset, and the timing of alerts, warnings and calls for evacuation in the run-up to eruption, or as the eruption escalates (e.g., Marzocchi and Woo, 2007; Winson *et al*., 2014; Cameron *et al*., 2018), less attention has been focussed on the ends of volcanic eruptions (Bonny and Wright, 2017).

There is no widely-accepted systematic definition for the end of an eruptive period (e.g., Phillipson *et al*., 2013) and, as a result, end-dates are often poorly reported. Although some global compilations of volcanism contain a field for eruption end-date, the end-dates are often not recorded. The Smithsonian Global Volcanism Program (GVP) highlight that, of the 10,415 eruptions in the Volcanoes of the World (VOTW) database at the time of writing, there were no available termination dates for 59% of the entries (Siebert *et al.,* 2011). This lack of data was attributed to the gradual nature of eruption endings, which made assigning a discrete date difficult. Phillipson *et al.* (2013) suggest that end-dates in the GVP database have a temporal uncertainty on the order of days, but inspection of the slow decline in observable activity at some systems (such as Mont Pelée after 1905; or Soufrière Hills Volcano Montserrat since 2010; Lacroix, 1908; Wadge *et al*., 2014) suggests that the uncertainty could be much larger in some cases - not least in the absence of a definition of what constitutes the 'end of eruption'. Even in well-monitored cases eruption end dates are hard to determine. For example, even though the activity at Mt St Helens from 2004 – 2008 was closely monitored with networks of instruments (including seismic, tilt and gas measurements), determining the end of the eruption was impeded by poor weather conditions throughout the month of December 2007. It could

not be conclusively determined that small-scale extrusion was finished until visual observations were made in January 2008 (Dzurisin *et al.,* 2015).

The lack of a specific definition for eruption end has implications for volcanic hazard. Tilling (2009) cites the mistaken identification of decreased volcanic activity as the end of eruption as one of the primary reasons for major loss of life during the 1982 El Chichón eruption. De la Cruz-Reyna *et al*. (2017) identify the definition of eruption end as a particular difficulty during sustained periods of activity, such as at Popocatépetl in Mexico. Popocatépetl has been in continuous eruption since 1994, exhibiting both effusive and explosive activity, but with lulls in activity which have led to uncertainty over whether or not they represent the end. Similar stop-start activity has characterised the long-lived and ongoing dome-forming eruptions of Santaguito, Guatemala (1922 to present), and Soufrière Hills Volcano, Montserrat (1995 to present). Better understanding of the timing of eruption end could have implications for the allocation of resources and management of populations living adjacent to volcanoes during both acute and sustained volcanic eruptions.

Obtaining an operational definition for the end of an eruption relies on piecing together various measurements of volcanic activity to determine when a break in volcanic activity represents the end of the eruptive period. Definitions for eruption end in a monitoring context come under two main categories:

(i) Generic rules on the end of eruptions: For example, Simkin and Siebert (1994) used a generic 90-day (or 3-month) rule for the end of eruption, i.e., that if a volcano displays no visible signs of eruption for 90 days, then the eruption can be considered over.

(ii) Definitions that are volcano-specific. For example, eruptions at Piton de la Fournaise, Réunion, are defined solely on increases or decreases in seismic tremor amplitude (Battaglia and Aki, 2003).

Volcanic systems undergo periods of activity and repose, on varying timescales (e.g., Barmin *et al.*, 2002; Lamb *et al.*, 2014). Identifying the critical thresholds which govern when large-scale changes in

volcanic behaviour occur is acknowledged as one of the fundamental research questions associated with understanding the beginning, evolution, and termination of volcanic activity (NASEM, 2017). Development of models for these critical thresholds is necessary to understand the processes which drive large-scale change in volcanic behaviour, but this in turn requires knowing when these changes occur in the timeline of an eruption.

To understand the timing of large-scale change in the system, it is important to define the various states of volcanic behaviour. Siebert *et al.* (2015) define eruption according to the observation of the following: explosive ejection of either fragmented new magma or older solid material, and/or the effusion of liquid lava. Activity outside eruption is defined as unrest, and a volcano in no state of eruption or unrest is said to be in repose (unless extinct). Phillipson *et al*. (2013) define two further categories of unrest: pre-eruptive and non-eruptive unrest, which are based on the presence of observable parameters such as deformation or seismicity. Neither of these categories can be assigned in real time, it being necessary to wait either until the crisis has subsided or an eruption has begun to determine whether the unrest was pre-eruptive or not.

Volcanic state has been previously characterised in models of seismic evolution over the eruptive period. McNutt (1996) developed the generic earthquake swarm model, which describes the evolution of seismic activity over an eruptive cycle. In this model, it is suggested that the rate and type of seismicity observed over an eruption can be generalised, and that the physical processes governing the type of seismicity at each eruptive stage may be inferred. Carniel (2014) describes how time series which have undergone a process of data reduction can be used to identify and infer the timing of transitions between different states of volcanism.

Machine learning is the process by which computers learn without being explicitly programmed. In fields such as healthcare, jet engine monitoring or economics, the use of machine learning methods for both data analysis and real-time monitoring is already established. Volcanic systems have conceptual parallels with these systems: they can be described as a “high-integrity” system (Clifton *et*

al., 2014) in which observation of failure (i.e., eruption) is rare in comparison with stable behaviour, and the number of failure modes are not known or not well characterised. The use of machine learning techniques in volcanology is an emerging field. Pattern recognition techniques have previously been applied to volcano-seismic data, with a particular focus on detection and classification of seismic event types from raw waveform data (e.g., Langer *et al*., 2006; Curilem *et al*., 2009; Apolloni, 2009; Bicego *et al.*, 2013; Maggi *et al*., 2017; Malfante *et al*., 2018). Machine learning has also been applied to satellite data, to detect signs of unrest in large numbers of acquisitions: Anantrasirichai *et al*. (2018) used deep learning to detect ground deformation in Sentinel-1 data, and Flower *et al*. (2016) used logistic regression to detect volcanic eruptions in global daily observations of SO_2 measured using the Ozone Mapping Instrument. Ren *et al*. (2020) used multi-station seismic tremor measurements to classify behaviour at Piton de la Fournaise volcano and identify fundamental frequencies of the tremor.

In this paper we use classification machine learning models (see Section 3.3.1 for a full description) (i) to classify eruptive and non-eruptive patterns in volcanic time series data, and (ii) to observe how these patterns differ from inferences based on visual observation or conventional monitoring techniques. Similar classification techniques have been previously successfully applied in a healthcare context to classify patient state (e.g., Clifton *et al*., 2014) and therefore have potential for characterising volcanic state. Our approach is distinct from previous work in volcanology (discussed above) as we classify overall volcanic state as eruptive or non-eruptive, as opposed to aiming to detect distinct change in one observable. We present a proof-of-concept study in classifying seismic time series for two volcanoes selected to cover a range of eruptive styles.

3.3) Methods

3.3.1) Machine learning methods used

Figure 3.1 illustrates the four multi-class methods used for the analysis in this paper. The methods used are all supervised methods, in which we select days from time series data (e.g., seismic event

counts) for training that include both eruptive and non-eruptive examples to train the model. During training, some training points are held back and used to test the trained model to increase the model accuracy (a process known as model validation). Once a model has been trained, it is then tested using days from the time series which were not presented during the training period.

For this preliminary work, we use machine learning models where features are calculated and chosen as inputs to the model, as opposed to other methods such as deep learning wherein features are calculated and chosen within the model. Choosing features as model input is preferred so that we can use features derived from the seismic data that are similar to those used in current monitoring practices. These features, such as event rate or peak signal frequencies, have had widespread success in a monitoring context as a basis for distinguishing between states of eruption (Carniel, 2014). The established use of these features in a monitoring context means that results such as the relative importance of a given feature in the models is directly applicable to current observatory practices.

Each machine learning model we apply is distinct from the others in its method of determining the boundary between non-eruptive and eruptive data. We choose to apply multiple methods which have the same training period, to observe whether the classification of eruptive and non-eruptive behaviour is consistent with several distinct methods. Each method determines the boundary between classes in a different way, and thus the methods have their own advantages and disadvantages. Support Vector Machine models (SVMs, Section 3.3.1.1) are good at handling non-linear relationships between data. Logistic regression (LR, Section 3.3.1.2) is a straightforward model, and thus rarely overfits data, while offering a fully interpretable approach whereby its parameters are informative of the relative contribution to the classification of the input variables. Random forest models (RFs, Section 3.3.1.3) are generally associated with high classification accuracy and involve taking an ensemble of individual decision trees (where the latter are weak classifiers). Gaussian Process models (GPs, Section 3.3.1.4) directly capture the uncertainty associated with the prediction and offer a principled approach to dealing with artefact in time-series data.

3.3.1.1) Support vector machine (SVMs)

SVMs (Figure 3.1a) involve finding the hyperplane between two classes of data which maximises the margin of the classification, where the margin is defined as the perpendicular distance between the decision boundary and the closest data points (Bishop, 2006). SVMs use the "kernel trick" to transform the data to a higher dimensional space, in which potential non-linearities in the original data can be separated (which would not be possible for logistic regression and other generalised linear models, for example). The choice of kernel depends on the properties of the dataset, such as non-linearity of the data. SVMs have been widely used in the field of seismic detection (e.g., Ruano *et al*., 2013), and they are well suited to general models even with few training examples (Mountrakis *et al*., 2011). We use the LibSVM libraries (Chang and Lin, 2011) to formulate models using both Radial Basis Function (RBF) and even-order polynomial kernels. The values of the hyperparameters for these models are selected using standard 5-fold cross-validation (Hastie *et al*. 2001).

3.3.1.2) Logistic regression (LR)

LR models (Figure 3.1b) are a form of generalised linear model: this means that the classification linearly depends on the features, where each feature has a coefficient in the linear model. Logistic regression models the posterior probability of a given day being eruptive as a continuous (sigmoid, or S-shaped) function of a linear expression of the features. The probabilistic output of these models means that for each day of results, we can infer the certainty of a classification on that day. A full discussion of logistic regression is included in McCullagh and Nelder (1989) and Hastie *et al*. (2001). In the Earth Sciences, LR models have previously been applied in estimation of landslide hazard (Pradhan and Lee, 2010).

3.3.1.3) Random forest (RF)

RF classification (Figure 3.1c) involves the averaging of an ensemble of decision trees: each decision tree comprises a series of operations that consecutively compare available features in the input data to randomly-selected thresholds on those features (Hastie *et al*., 2001). Many possible decision trees are combined in random forest classification: the result of each tree contributes a vote towards the final

classification. The hierarchical nature of decision trees means that these methods can be used to determine the relative importance of the features, where features that appear towards the top of the decision tree contribute more to the classification and are therefore associated with greater importance (Breiman *et al*., 1984). Random forest models have been applied extensively in remote sensing, due to the high accuracy of classification obtained and their ability to identify important variables (discussed more in Section 3.4.4; and Belgiu and Drăguţ, 2016).

3.3.1.4) Gaussian process classification (GPs)

Gaussian processes (GPs) (Figure 3.1d) fit stochastic models to obtain a probability that a given data point is in a given class (Bishop, 2006). The classification that results from a Gaussian process classification is therefore associated with a given uncertainty. Like SVMs, GP classification involves a choice of kernel function to train the model. We use an Automatic Relevance Determination (ARD) kernel for training models, which allows the importance of each feature input into the model to be evaluated (Williams and Rasmussen, 2006).

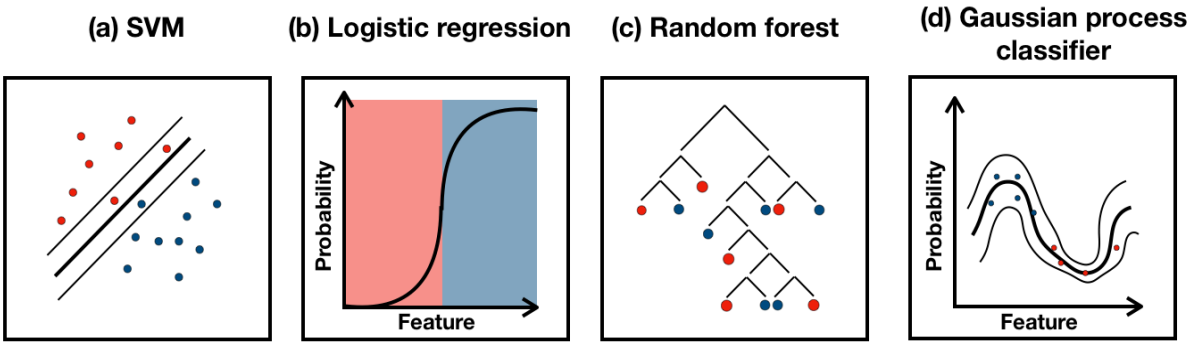


Figure 3.1: Visualisation of how boundaries are defined within the four machine learning methods used in this paper. (a) Support Vector Machine (SVM) (b) Logistic regression (c) Random forest decision tree (d) Gaussian process classifier with uncertainty bounds marked. Red and blue dots denote two classes of data, which in this study represent non-eruptive and eruptive data. The y-axis in (b) and (d) represents the probability that a given day belongs to the red class or blue class.

3.3.2) Model assumptions

The models used in this paper require the fundamental assumption that the input variables are Independent and Identically Distributed (IID). Assuming data which are IID implies that each day of features is independent from other days of features, and that the features on each day are drawn from

the same underlying statistical distribution (Cover and Thomas, 2006). While data are rarely perfectly IID in practice, the assumption typically holds to the degree that models involving such approaches yield satisfactory results. To determine whether or not this hypothesis is appropriate for our data, and therefore whether or not our models can perform well using previously-unseen data, we train and test (using data held-out during the training process) separate models for each volcanic system.

Though we are considering a physical system, which may not provide perfectly IID data, we hypothesise that this assumption is valid to a first order approximation. Other machine learning models exist which can take into account non-IID behaviour; however, these methods are beyond the scope of this paper.

3.3.3) Nevado del Ruiz and Telica volcanoes

We analyse single-station seismic datasets from two volcanic systems: Nevado del Ruiz Volcano, Colombia and Telica Volcano, Nicaragua. We choose to apply models to the datasets from these two volcanic systems for our test study as they represent contrasting styles of activity. Telica displays near-continuous levels of seismic activity, whereas Nevado del Ruiz represents more punctuated volcanic activity. Therefore, successful classification of these differing styles is a useful proof-of-concept that these methods have the potential to be extended to a range of volcanic situations.

Nevado del Ruiz is a stratovolcano in Colombia which primarily erupts products of andesite-basaltic andesite composition (Cuellar-Rodriguez and Ramirez-Lopez, 1987; Londoño, 2016). The dataset is from 21st March 2007 to 25th February 2015 and covers two eruptive periods, both recorded in GVP. The first phase has a start-date of 22nd February 2012 and end-date of 12th July 2013, and the second, much shorter, phase has a start-date of 15th December 2014 and end-date of 7th January 2015. An increase in seismic activity began in September 2010 and ash emissions from Nevado del Ruiz were observed from early 2012 onwards (Global Volcanism Program, 2017). The 2012 ash emissions of Nevado del Ruiz were the first emission of ash since the VEI 3 eruption of 1985, which led to the

lahar inundation of Armero and > 25,000 fatalities (Lowe *et al.,* 1986; Naranjo *et al.,* 1986). The end-date of the first phase is a day later than the last recorded ash emission and coincides with the last advisory of the Washington Volcanic Ash Advisory Centre (VAAC, 2013). It is unclear how the dates for the second phase are chosen, as ash emission was observed both in November 2014 and later in January 2015.

Telica volcano is a persistently restless volcano in Nicaragua which undergoes small (VEI 1 - 2) eruptions every few years (Geirsson *et al.,* 2014; Rodgers *et al.,* 2013; Rodgers *et al.,* 2015a). Persistently restless volcanic systems are characterised by high and variable rates of seismicity and degassing, with frequent explosive activity (Rodgers *et al.,* 2015a; Geirsson *et al.,* 2014; Roman *et al.,* 2019; Stix, 2007). The Telica data used for this analysis were obtained from the TESAND network, for the period 1st April 2010 to 18th March 2013. This data period contains one VEI 2 eruption, recorded in GVP with a start-date of 7th March 2011 and an end-date of 14th June 2011. The end-date is 3 days after the last ash-and-gas explosion sequence (comprising 17 explosions) was observed on 14th June 2011 (Geirsson *et al.,* 2014).

3.3.4) Classification scheme

Figure 3.2 illustrates the process of training and testing the multi-class methods introduced in Section 3.2.1. We use supervised machine learning algorithms: these models are trained on labelled data and subsequently tested on unseen data. Data are labelled in "eruptive" and "non-eruptive" classes according to the eruption dates recorded by the GVP database. Non-eruptive data comprises all the data which does not fall under the dates recorded in GVP for the eruption. We select training periods to represent non-eruptive and eruptive data as input to the models. These training periods are selected such that they do not overlap with the GVP start and end dates, because we want to independently constrain the timing of transitions between eruptive and non-eruptive activity. The eruptive and non-eruptive training periods are chosen to represent times in which the presence or absence of visual eruption was confirmed from activity reports, archived by the Servicio Geológico Colombiano (SGC) and Instituto Nicaragüense de Estudios Territoriales (INETER).

For Nevado del Ruiz, the non-eruptive period we select to train the model is from 15th June 2009 to 30th September 2011. This non-eruptive period is approximately 4 months before the weekly reports mark the first possible ash emission, clear deformation signal and increase in SO_2 emission (Global Volcanism Program, 2012a). The eruptive training periods selected are 23rd March 2012 – 26th February 2013 and 9th April 2013 – 25th April 2013. These periods are selected because they coincide with ash emissions confirmed by Manizales observatory (Appendix B, B.1; SGC). We use 938 days of the Nevado del Ruiz daily time series for training the models and 1364 days of the time series for subsequently testing the models.

For Telica, we choose non-eruptive training periods from 29th June 2010 – 8th January 2011 (before eruption) and 1st September 2012 – 20th November 2012 (a year after the eruption). The second half of the training period is selected as it is over a year after the end of eruption. The eruptive training period is from 28th March 2011 – 1st June 2011, starting after ash emission had been confirmed by visual observation (Global Volcanism Program, 2011; Geirsson *et al.,* 2014). We use 333 days of the Telica daily time series for training the models and 730 days of the time series for subsequently testing the models.

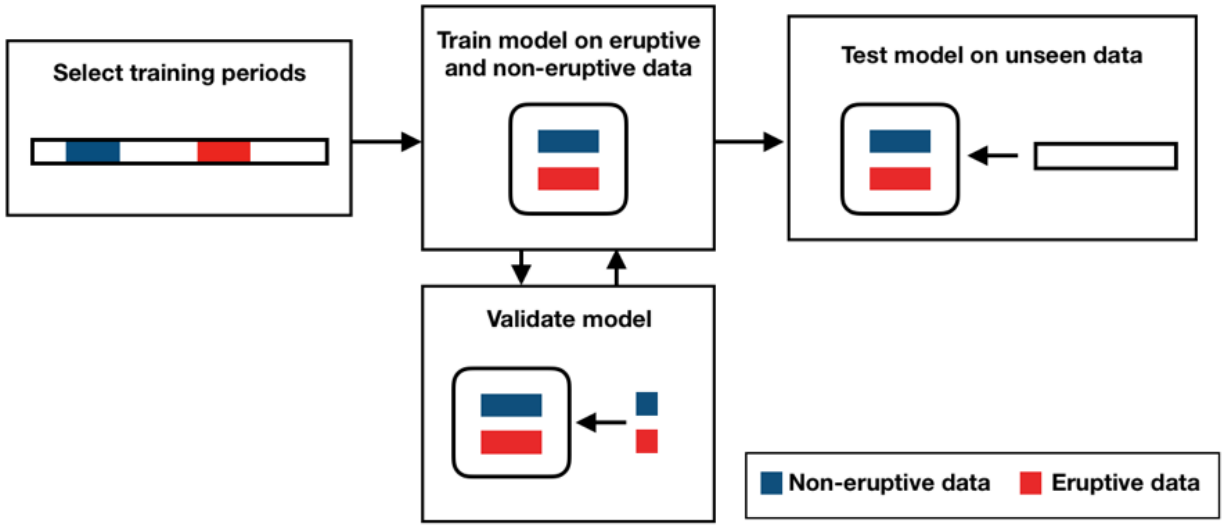


Figure 3.2: The framework of training and testing supervised multi-class classification models. Training periods which include non-eruptive and eruptive data are selected. The model is trained on a subset of these training data, and concurrently validated using the rest of the training data. After a model has been trained, the model can be run again with testing data. Blue represents data labelled as non-eruptive and red represents data labelled as eruptive.

3.3.5) Feature extraction

Feature extraction is the process of selecting variables which will be used as inputs into machine learning models. Figure 3.3 describes the process of feature extraction from raw seismic data. The inputs to machine learning models are time series derived from raw seismic data. To produce these derived time series, several categories of features are selected from the seismic data (see Appendix A; A.1 and A.2). For all features apart from event rates, the data are taken from a single seismic station. Telica data are from the TBTN station of the TESAND network, located approximately 1 km east of the active vent (Roman, 2009). Nevado del Ruiz data are from the BISZ station, located approximately 2 km west of Arenas crater (Global Volcanism Program, 2012a). For Telica, there are 45 features (A.2) and for Nevado del Ruiz (A.1) there are 36 features in total. There are a greater number of features for Telica due to the inclusion of features from individual event classifications and RSAM data.

Total event rates per day from network detections are used for both volcanoes. For the Telica data, two additional features derive from the automatic spectral classification of Low Frequency (LF) and

High Frequency (HF) as defined by Rodgers *et al*. (2015a). Band ratio is defined as the base 2 log of the ratio of high-frequency to low-frequency energy (Rodgers *et al*., 2015a; c.f. Buurman and West, 2010). The distinction between high- and low-frequency bands is dependent on the typical frequencies of the volcanic system: for Telica, low-frequency activity is defined as 1 - 6 Hz, and high-frequency activity is defined as 6 - 11 Hz (Rodgers *et al*., 2015a). Dominant frequencies are obtained by recording the 5 peak frequencies from each event during the day. Peak amplitude is calculated from the maximum peak-peak amplitude of each event. Waveform standard deviation is obtained by calculating the width of the largest peak of the spectra for each event during the day. RSAM measurements are calculated hourly during the day for the Telica dataset. Multiplet information is the number of waveform families active on a given day, obtained by waveform cross-correlation using Peakmatch (Rodgers *et al*., 2015b).

From the categories of observations described above, features are calculated on a per-day basis by taking the mean, median, variance, minimum, maximum, 10^{th} percentile, 90^{th} percentile and change in mean from the previous day. For multiplets and event rates, only the per-day value and change in value from the previous day is calculated. The RSAM features are mean and variance of the per-hour readings and change in mean from the previous day. For a full list of features see Appendix A.

For days in the time series with zero events, the whole day is omitted from the time series as no features can be extracted for this day. The gaps in the dataset could be filled using a method such as imputation in which missing data are replaced by a substitute, such as the mean of the whole dataset (Schafer and Graham, 2002). However, given that the days which have no associated data represent a small proportion of the dataset, we choose to leave these gaps within the time series.

Models can be limited by large quantities of features. High-dimensional systems (those with many features) are not ideal to work with: as the number of dimensions of data increases, the number of training examples required to train a consistent model grows exponentially (Bishop, 2006). This phenomenon is known as the "curse of dimensionality" (Bellman, 1961). For generalised linear

models such as logistic regression, high-dimensional systems are especially poor to work with (Johnstone and Titterington, 2009). For this reason, we apply regularisation to the logistic regression model, to reduce the number of dimensions as input to the model. We use a technique known as the Least Absolute Shrinkage and Selection Operator (LASSO) to reduce the number of dimensions as input to the logistic regression models. The LASSO acts to penalise large coefficients in linear models, so that a smaller subset of the full set of features is chosen to model on for each dataset. A full discussion of the LASSO formulation is included in Hastie *et al*. (2001).

We use features derived from single-station seismic data, hence do not include derived event parameters such as location or depth. Seismic data are the only type of data which is input to the model. We do not use other observables, such as gas or deformation data. The end-date obtained by classification therefore corresponds to the end of seismicity associated with the eruption, and therefore represents the seismic end-date. Seismicity can often continue longer than the end of visible eruption, as it reflects the processes occurring at depth within the volcanic system. Seismic data are one of the most ubiquitous monitoring datasets collected at volcanoes and understanding the path to cessation of processes at depth is crucial in terms of understanding the end of eruptions, supporting the value of focussing only on seismic data for this preliminary study.

The features which we use, except for those associated with RSAM, are derived from detected seismic events. The advantage of using these discrete features is that the method could be easily extended to seismic catalog data from other systems with waveforms attached. However, depending on the seismic characteristics of a given volcano, the inclusion of more features derived from continuous data (such as dominant tremor frequency) may be necessary.

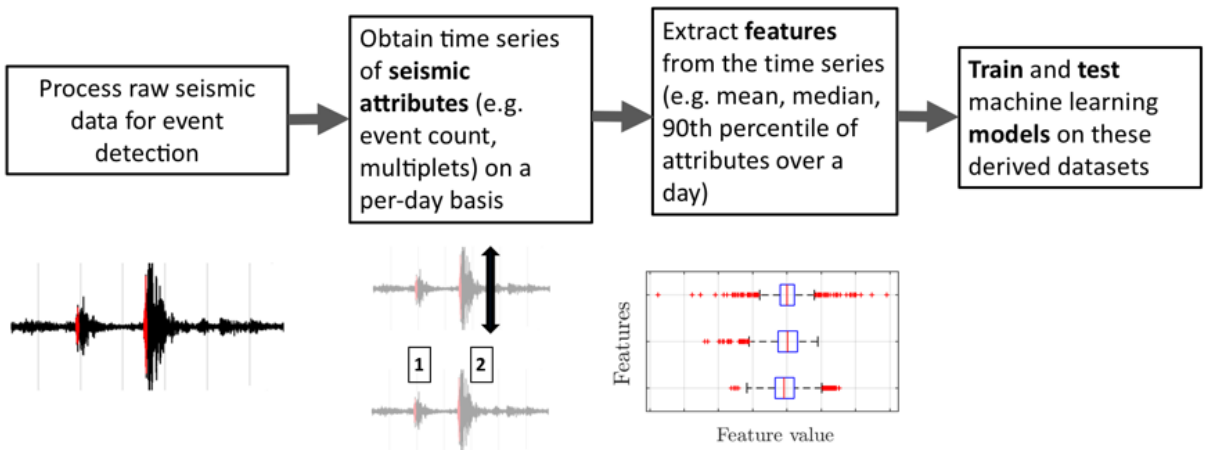


Figure 3.3: Diagram of the framework for extracting features from raw seismic waveform data. Events are detected from raw waveform data. From each event waveform we extract features including the peak amplitudes and band ratio. We then calculate features including the mean and variance from all the waveforms in each day. The resulting time series are used as input for the machine learning model.

3.3.6) Eruption classification

Each day in the time series is classified independently from the other days. To pick out the large-scale changes in classification, we define a rolling threshold filter criterion for a day to be classified as eruptive, which is based on a moving average of the classifications. For a day to be classified as eruptive, the day itself and the seven days preceding that day must also be classified as eruptive. By applying this filter to the model output, classification of observations as eruptive is more conservative than if the results are left unfiltered. The choice to require 7 days of eruptive classification is made as the timescale of a week corresponds to typical timescales on which observations of volcanic activity are communicated to the public where the eruption circumstances are ongoing or chronic, for example in weekly reports of activity.

3.3.7) Quality assessment: Decisiveness index

Our aim when classifying volcanic state is not necessarily to achieve maximum accuracy relative to GVP labels, due to the issue with GVP definition of volcanic state (discussed in Section 3.2). We therefore define an alternative index to model accuracy for evaluating our models. This index is a measure of how consistent or decisive the model classification is over the whole dataset, expressed as a percentage of the total number of days containing data. As we are looking to classify overall

patterns of eruptive or non-eruptive activity, the decisiveness index favours classifications with less noise.

To define the decisiveness index D, we take the number of transitions between classes in our models (Nt_M; where a transition can be from non-eruptive to eruptive or eruptive to non-eruptive) and subtract the expected number of transitions corresponding to the number of eruptive periods in the dataset (Nt_E; where one eruptive period would have two transitions, at the beginning and end of the eruption) then normalise by the number of days contained within the data period (N_d). An index of 0 means that the number of transitions in the model are exactly equal to the number of assumed transitions. A larger index means that there is more inconsistency (i.e., more indecision) in the final model classification.

$$\boldsymbol{D} = \frac{\boldsymbol{Nt_M - Nt_e}}{\boldsymbol{N_d}} * \boldsymbol{100} \quad (6)$$

The decisiveness index is reported as a percentage. The worst classification which one might produce would alternate between classes every day, therefore contain a transition for each day. As the number of days is three orders of magnitude than the number of transitions in the datasets presented here, the number of transitions would be equal to the total number of days and the decisiveness index will approach 100 %.

The definition of the decisiveness index is a method to evaluate which models make the most consistent classifications of eruptive state. This index makes no assumptions about when the transitions occur within the dataset. Therefore, the index is used in conjunction with comparison to visual observation of volcanic state to evaluate the success of the models presented within this paper.

3.4) Results

We independently trained 4 different classification models for each volcanic system, with each type of model trained and tested on each volcano separately. The analysis could be extended by training a model on several different seismic datasets, which would be a general classification model. However, a general model would require datasets from a greater variety of volcanic settings to ensure that the non-eruptive and eruptive distributions were well-characterised by the machine learning models.

3.4.1) Nevado del Ruiz

The results from each machine learning method are summarised in Table 3.1. For all 4 machine learning methods, the end-date of the first phase of eruption at Nevado del Ruiz obtained by classification of the data is later than the end-date contained in the GVP database. Figure 3.4 illustrates the results from the SVM classification of Nevado del Ruiz data in a time series plot for the whole data period. Several observations can be made which are consistent features of all of the models summarised in Table 3.1, though we only plot the SVM model (Figure 3.4) as it has the highest model accuracy of 82.6 % (Table 3.1):

1. There is a sustained classification of non-eruptive activity before the beginning of eruptive activity, with only 1 pre-eruptive day erroneously classified as eruptive in 2007 for the SVM model.

2. The eruption end-date is 4 – 5 months later than the end-date recorded in GVP. Though the eruption end-date was recorded as the 12th July 2013, just the previous day, active ash emission was observed on the 11th July 2013, with further reports of gas and steam emission until November 2013 (Global Volcanism Program, 2017).

3. The second phase of eruption was classified as longer-lived than the GVP start- and end-dates would suggest. Though recorded in GVP as commencing in December 2014 and finishing in

January 2015, the SVM classification of eruptive behaviour lasted from July 2014 to February 2015.

Table 3.1 summarises the results for the decisiveness index when applied to Nevado del Ruiz. It can be seen that the Gaussian process classifier has the best result for the decisiveness index with 3.13, followed by SVM with 3.65. Logistic regression and random forest classifications have a poorer score for the decisiveness index of 4.78 and 4.17 respectively. The range of D for the Nevado del Ruiz models is 1.65.

Figure 3.4 also contains information for the days in the Nevado del Ruiz dataset on which there were insufficient data to calculate features. Where there are many data gaps in the sequence, for example, during non-eruptive activity in 2007 or during the eruption in mid-2012, the classification is the same on either side of and during the data gap. From this observation we can determine that the decision to leave data gaps and not to fill them with a method such as imputation is justified (Section 3.3.5).

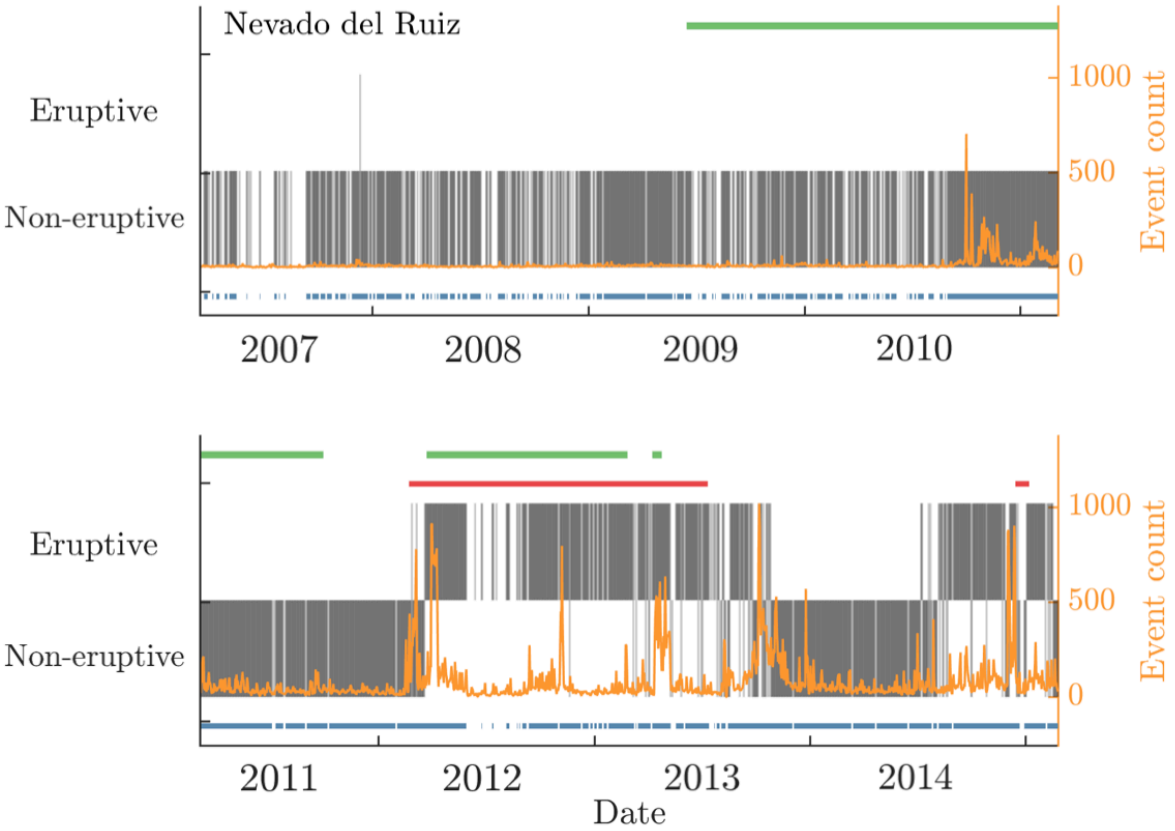


Figure 3.4: Results from SVM classification on Nevado del Ruiz data over the study period. This study period runs from 21st March 2007 to 6th March 2011 (top panel) and from 6th March 2011 to 25th February 2015 (bottom panel). Results of the classification are denoted by the grey rectangles where rectangles in the top half of each panel denote a classification of eruptive and rectangles in the bottom half of the panel denote a classification of non-eruptive. As each classification is made independently, consecutive days of the same classification together – such as non-eruptive classification in the top panel – illustrate the decisiveness of the classifier. Above the classification, the green horizontal line at the top of each panel denotes the timing of the training period and the red horizontal line second from the top of each panel denotes the timing of the eruption as recorded by GVP. The right axis and orange line within the plot denote daily event count for all events. Below the classification, the blue horizontal line denotes the days for which we could derive features (listed in AP 1.1). In the top panel, gaps in data were primarily due to low event count, whereas in the bottom panel during the eruption there is a gap which corresponds to instrument failure.

All of the models apart from random forest yield greater than 70% accuracy (where the model result is compared to the GVP label of whether a day is eruptive or non-eruptive) after filtering. The greatest accuracy of 82.6 % is achieved by using the SVM model. Logistic regression models have the second highest accuracy (79.6 %). High model accuracy is therefore consistent over multiple types of classification, including both non-linear and linear models. Random forest and Gaussian process classifiers have a similar accuracy.

Figure 3.5 displays the Receiver Operating Characteristic (ROC) curves from all classification models applied to the Nevado del Ruiz dataset. ROC curves plot the false positive rate against the true positive rate for a binary classifier as the threshold of classification is changed. A better classifier will

have a higher true positive rate at low false positive rate, as more points will be correctly classified. Better classifiers will also have a greater Area Under the Curve (AUC), which can be seen where the curve is higher than the diagonal line through the origin of the graph. From Figure 3.5 it can be seen that logistic regression and random forest have a very similar structure, with an AUC of 0.89. The SVM has a slightly better performance at low false positive rate, but overall has a lower AUC of 0.85. The Gaussian process classifier has a relatively poor performance relative to the SVM, logistic regression and random forest, despite having a similar model accuracy to the random forest models.

Though good accuracy is achieved by the models, it should be highlighted that this represents a comparison of model output for each day compared to the GVP label of whether a day is eruptive or non-eruptive. We anticipate that the GVP labels are not entirely reliable due to uncertainty in the GVP definition of end of eruption (Phillipson *et al*., 2013). This error in labels leads to a number of false positives after the GVP end-date, where the day has been classified as eruptive by the models though the GVP label is non-eruptive.

Figure 3.6 is a summary of the feature importance results for logistic regression, random forest, and Gaussian process classification, the three methods for which this analysis is available. From the results we can observe that there is not much consistency between the three methods as to the most important features within the dataset, either in the group of features (e.g., Event rate, dominant frequencies) or the type of features within a group (e.g. Rate, Δ, mean; for a full list of features see Appendix A). For the Gaussian Process classifier, the 10^{th} percentile of peak amplitude per day is the feature that has the greatest importance.

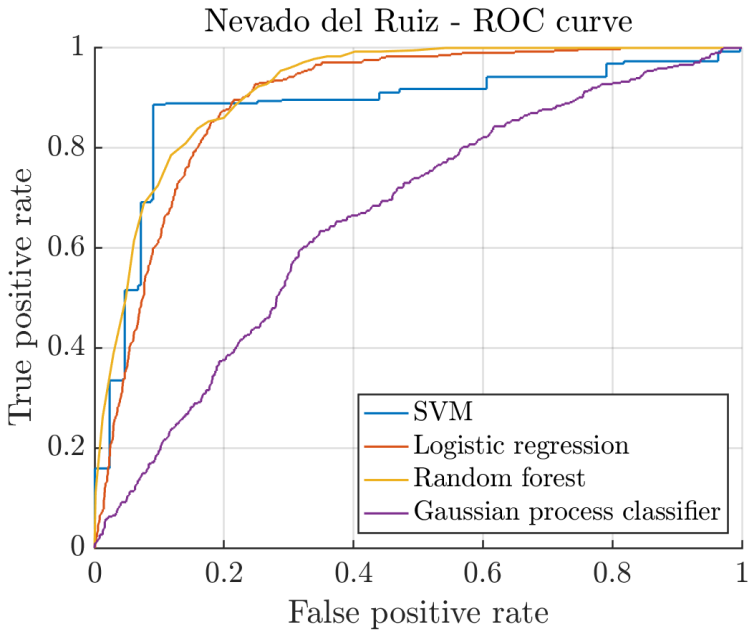


Figure 3.5: Receiver Operating Characteristic (ROC) curve for all the methods applied to the Nevado del Ruiz data. The ROC curve plots the true positive rate, false positive rate and AUC value. SVM, logistic regression and random forest have similar performance, compared to Gaussian process classification.

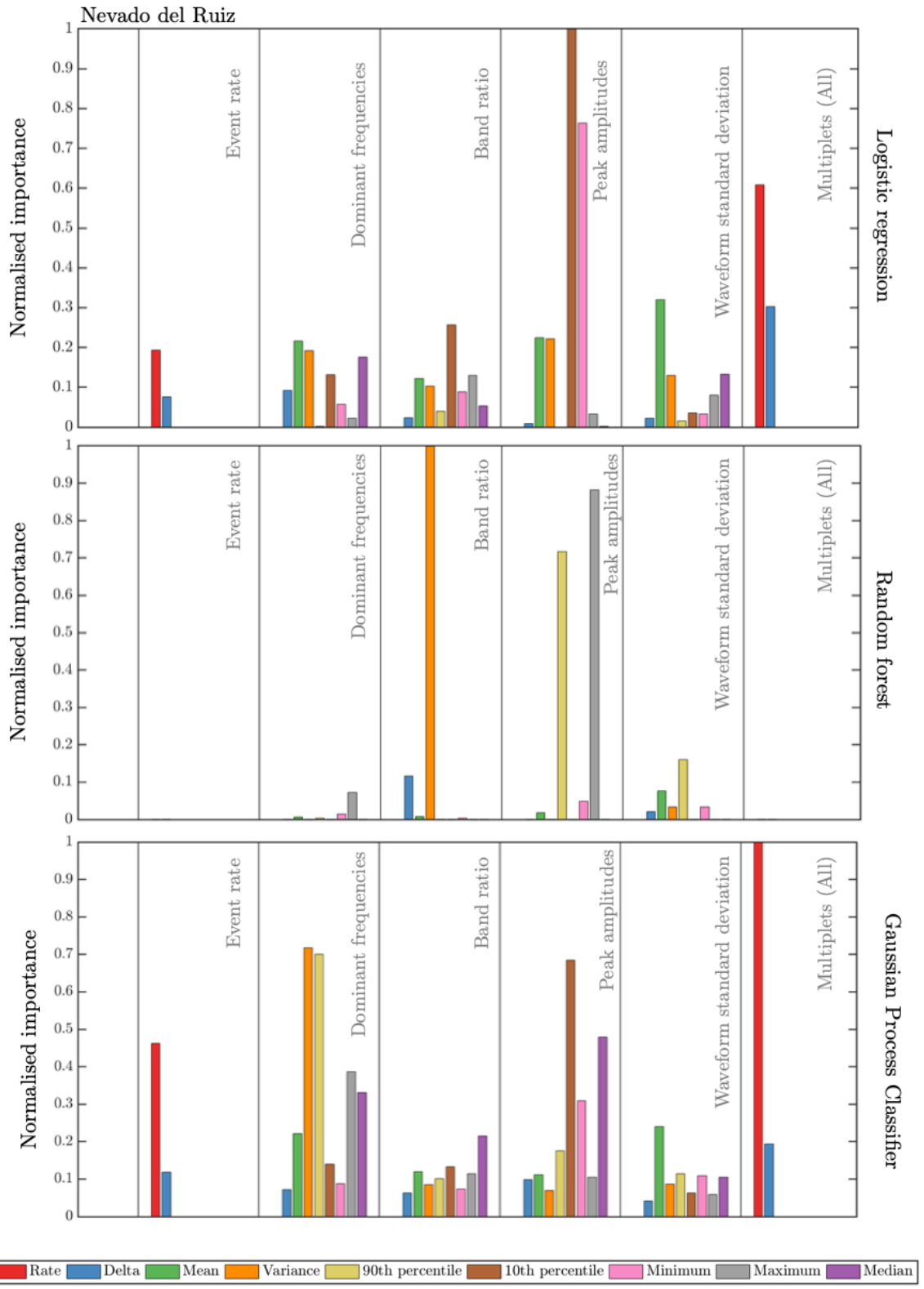


Figure 3.6: Results from feature importance analysis methods on Nevado del Ruiz data. Feature importance is derived from logistic regression (top), random forest (middle) and Gaussian process classification (bottom). The y-axis denotes absolute importance which varies depending on the models, normalised by the maximum value for the model. There is a much greater range in importance for the Gaussian process classification than for the random forest model. Vertical lines separate the categories of features. For a full list of input features see Appendix A.

Table 3.1: Summary of results from all machine learning models applied to the Nevado del Ruiz and Telica datasets.

Volcano	Method	Start-date for first phase (GVP)	End-date for first phase (GVP)	Start-date for first phase (model)	End-date for first phase (model)	Decisiveness Index (%)*	Model accuracy (unfiltered)	Model accuracy (filtered)
Nevado del Ruiz	SVM	22nd February 2012	12th July 2013	February 2012	November 2013	3.65	76.2 %	82.6 %
	Logistic regression			February 2012	December 2013	4.78	75.4 %	79.6 %
	Random forest			February 2012	November 2013	4.17	62.8 %	66.9 %
	Gaussian process classifier (GPC)			February 2012	November 2013	3.13	67.6 %	72.7 %
Telica	SVM	7th March 2011	14th June 2011	March 2011	August 2011	3.01	83.6 %	87.9 %
	Logistic regression			March 2011	October 2011	3.76	65.1 %	71.6 %
	Random forest			March 2011	August 2011	3.39	62.6 %	69.5 %
	Gaussian process classifier (GPC)			March 2011	August 2011	1.69	86.2 %	90.5 %

*Decisiveness Index is defined in Section 3.3.7. Lower D scores are better, and D is comparable across different datasets.

3.4.2) Telica

Table 3.1 presents the summary of results of models run on Telica Volcano. As with the Nevado del Ruiz data, end-dates obtained by classification of Telica data for successful models (unsuccessful modelling is summarised below) are all approximately 2 months later than the end-date contained within the GVP database.

Gaussian process classification had the best value of decisiveness for data both for models applied to Telica data, and over both Nevado del Ruiz and Telica classifications overall, with a value of 1.69. Gaussian process classification also had the highest accuracy (87.9 %) of all the models applied to Telica. SVM had the second-best value of decisiveness index of all the models (3.01), followed by random forest (3.39) and logistic regression (3.76) models. The range of D for models on the Telica dataset is 2.07.

Figure 3.7 summarises the classification from the Gaussian process classifier on the data from Telica, the model which had the highest accuracy and lowest decisiveness index of any model applied in this study. The attributes of the classification which we identify consistently over all models for the Telica time series are as follows:

1. For all the machine learning approaches, the end-date inferred from the models was later than the end-date recorded by GVP. The end-date for the eruptive phase at Telica was 14th June 2011, whereas the end-dates obtained by successful classifications were all in August 2011.

2. After the end-date of the eruption inferred by the models, there are 2-3 short periods of elevated event count in November 2011 and March 2012 which correspond to classifications of eruptive activity.

Figure 3.8 summarises the ROC curves for all the methods. Although Gaussian process classification has the best values for decisiveness and accuracy, this model has the lowest Area Under the Curve (AUC) at 0.64. SVM and logistic regression classifiers have similar AUC values at 0.94 and 0.91 respectively. Overall, the AUC values for Telica are less smooth than for Nevado del Ruiz, a consequence of the smaller dataset that we have for Telica.

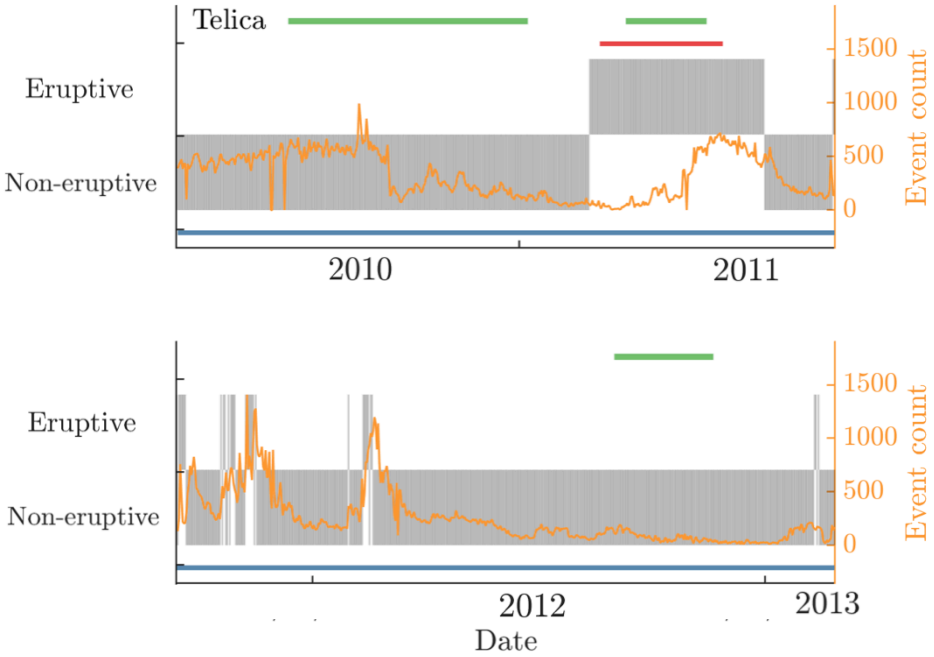


Figure 3.7: Results from GPC classification on Telica data over the study period. The study period runs from 1st April 2010 to 6th October 2011 (top panel) and from 7th October 2011 to 18th March 2012 (bottom panel). Results of the classification are denoted by the grey rectangles where rectangles in the top half of each panel denote a classification of eruptive and rectangles in the bottom half of the panel denote a classification of non-eruptive. As each classification is made independently, consecutive days of the same classification together – such as non-eruptive classification in the top panel – illustrate the decisiveness of the classifier. Above the classification, the green horizontal line at the top of each panel denotes the timing of the training period and the red horizontal line second from the top of each panel denotes the timing of the eruption as recorded by GVP. The right axis and yellow line within the plot denote daily event count for all events. Below the classification, the blue horizontal line denotes the days for which we could derive features. There are no significant periods of data shortage throughout the Telica data period.

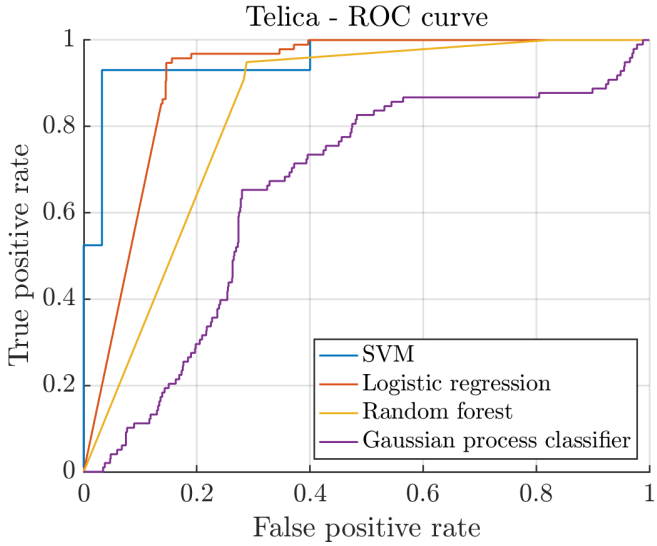


Figure 3.8: Receiver Operating Characteristic (ROC) curve for all the methods applied to the Telica data. The ROC curve plots the true positive rate, false positive rate and Area Under Curve (AUC) value. SVM, logistic regression and random forest have similar performance, compared to Gaussian process classification.

Figure 3.9 is a summary of the feature importance results for the methods applied to Telica data. Event rate features do not have high values for variable importance. This finding can be confirmed by observing the classification in Figure 3.7: the overall event rate spans a similar range (from 0 – 500 events per day) during classification of both eruptive and non-eruptive activity by our models. As seen for the models applied to Nevado del Ruiz, there is not much consistency in the individual features which are associated with a higher importance.

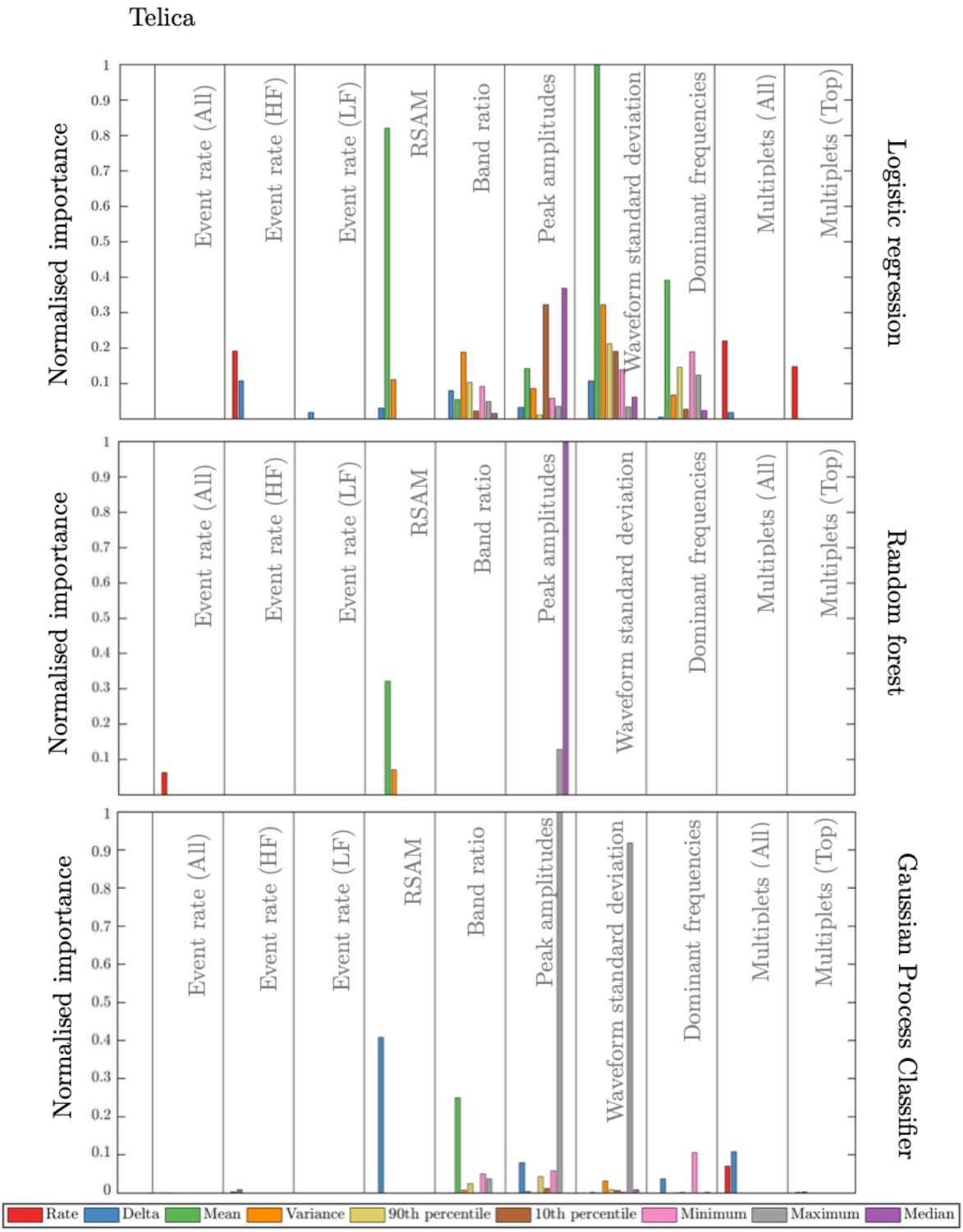


Figure 3.9: Results from feature importance analysis methods on Telica data. Feature importance is derived from logistic regression (top), random forest (middle) and Gaussian process classification (bottom). The y-axis denotes absolute importance which varies depending on the models, normalised by the maximum value for the model. Vertical lines separate the categories of features.

3.4.3) Training models with data from the end of volcanic eruption

The Nevado del Ruiz models are trained using two training periods: a period before the beginning of the first eruption and a period during the first phase of the eruption (Figure 3.4). In these models we do not train over any transitions between eruptive and non-eruptive behaviour. We now extend the modelling to train over two extra periods using SVM:

(i) Over the GVP start-date, with training period 23rd June 2009 – 23rd December 2011 (non-eruptive) and 2nd February 2012 – 24th March 2013 (GVP start of eruption).

(ii) Over the GVP end-date, with training period 23rd June 2009 – 23rd September 2011 (non-eruptive) and 29th December 2012 – 26th September 2013 (GVP eruption end).

We would expect the models to successfully classify non-eruptive and eruptive behaviour if the GVP dates represent reliable labels of the transitions in the dataset.

The results from model (i) are very similar to those presented for a model trained over no transitions in behaviour (Figure 3.4). However, for model (ii), we obtain a very poor classification: eruptive behaviour is not classified until July 2012 despite visual evidence of eruption from February 2012 onwards (Global Volcanism Program, 2012a). Moreover, no activity during the second phase of eruption is classified as eruptive. We conclude from this result that the eruption end-date recorded in GVP does not provide a reliable label for the transition between eruptive and non-eruptive behaviour.

3.5) Discussion

3.5.1) Classification compared to visual observations

The results presented in Section 3.4 suggest that in both phases of eruption at Nevado del Ruiz the classification of eruptive activity is more prolonged than the GVP start- and end-dates would suggest. A possible reason for the discrepancy between model classification and GVP eruption duration is that GVP classifications are based on visual observation of volcanic activity, whereas we are running models on the seismicity, with 7-day rolling window filtering. Seismicity can indicate processes occurring at depths of several kilometres within the volcanic system (Moran *et al.,* 2011), which it is reasonable to expect would continue after visual signals of volcanic eruption had ended. The classification of eruption until November 2013 (Table 3.1) could represent continued or declining seismogenic processes at depth during the declining phase of the eruption. In this respect, the seismic end-dates presented here are hypothesised to represent the most generous bound on the end-date of the eruption.

In Figure 3.10 we compare the event rate and model classification to the alert level recorded at Nevado del Ruiz for the duration of the data period, as event rate is a commonly-used parameter for investigating volcanic state. The majority of alert level changes were concentrated in the period leading up to eruption, and the first seven months of eruptive activity. There are no alert level changes following 5th September 2012, on which date the alert level was downgraded from II (Orange) to III (Yellow) (Global Volcanism Programme, 2012a). The alert level changes are therefore too coarse to provide insights into the processes occurring at the end of the eruption.

Figure 3.10 also summarises the event rate and recorded ash emissions according to weekly reports and confirmed visual reports of ash emission (Appendix B, B.1; SGC). The classification of the second phase of eruptive activity precedes the ash emission during July 2014, and continues until further ash emission during November 2014. Our model results show a high correlation of eruptive classification with ash emission during the second phase of volcanic eruption recorded from weekly

reports from the Manizales observatory and observatory records (Londoño and Galvis, 2018), having trained our model on the seismic signals associated with ash emissions during the first phase of volcanic activity.

The good agreement between classification and observation can also be noted with a comparison to event rate: though the GVP end-date of 11th July 2013 coincides with a consistent low event rate for Nevado del Ruiz, our model continues to classify behaviour as eruptive for 4 further months, which spans a spike in event rate to 1000 events per day in September - October 2011, and culminated in ash emission at the end of November (Global Volcanism Program, 2017).

In the Telica results, we observed deviations between our model classification and that of the GVP in terms of end-dates from August – October 2011, in addition to two periods of elevated event rate from October – November 2011 and February – March 2012. These periods are united by records of "jet-turbine" sounds from the crater initially reported by nearby communities (Global Volcanism Program, 2012b), and by crater incandescence and gas emission in September 2011 and February 2012. Overall, comparison to visual observations is more difficult for Telica data as visual reports of activity are released monthly rather than weekly in Colombia, and there is no system of alert level classifications.

For both volcanoes, models without the rolling threshold filter applied yielded classification of eruptive activity before the GVP eruption beginning date. However, these classifications were not sufficiently consistent to remain as eruptive after applying the 7-day rolling filter. Further work is required to see how these results could be analysed for the case of pre-emptive classification of eruption to evaluate whether those classifications of eruptive activity are truly eruptive or represent a false positive result.

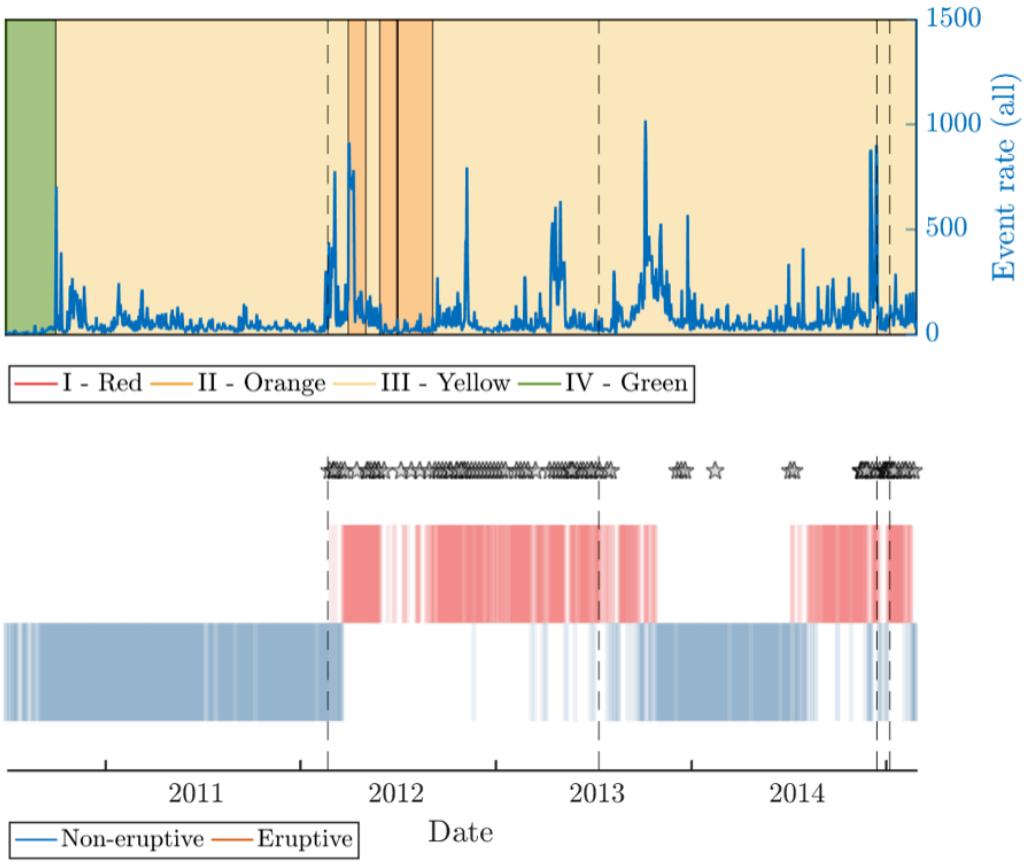


Figure 3.10: Comparison of the event rate and alert level (top panel) and SVM classification and recorded ash emissions (bottom panel) at Nevado del Ruiz between the dates of 28th June 2010 to 25th February 2015. Vertical dashed lines in each plot indicate the GVP start- and end-dates of the two phases of eruption during the data period. Stars indicate confirmed ash emissions. Rectangles in the bottom panel represent the classification from SVM (as in Figure 3.4) where blue rectangles are non-eruptive classification and red rectangles are eruptive classification. We choose to plot the SVM as it was the best-performing classifier for Nevado del Ruiz. The alert level was consistently green from the beginning of the data period to the first alert level change on 30th September 2010.

3.5.2) Possible application of methods and future work

In this study we have shown that retrospective classification of volcanic activity can yield timing of change with a greater correspondence to heightened activity and ash emission than end-dates denoted by visual activity judged to be the last of an eruptive phase (Figure 3.10). The coarseness of alert level changes and the success of the classification model discussed in Section 3.5.1 presents the possibility to use these machine learning classification methods to identify potential start- and end-dates for seismically monitored volcanoes. Estimates for seismic end of eruption obtained by classification could be combined with other indicators of activity, such as deformation, thermal or gas data, to make a judgement on whether the eruption has ended.

In remote locations, or where conditions are unfavourable for making visual observations, classification of seismic data could yield a consistent method for determining transitions in eruptive state in the absence of other evidence. The work presented here is an example of how to distinguish between eruptive and non-eruptive seismic activity using information from one seismic station alone. For example, though no official visual observations of Telica volcano coincided with our eruptive classifications in late 2011 or early 2012, nearby communities reported jet-like sounds which coincided with these periods. Further applications for this method could include monitoring volcanoes in remote locations where regular visual observations of the volcano are not practical.

The end-dates yielded by the successful methods are between 2 – 4 months later than the end-dates judged to be the last visual indication of eruption (Section 3.4). This finding is in agreement with the generic 90-day rule discussed in Section 3.2. The validation of this months-long timescale of eruption cessation, consistent across volcanoes of differing eruption style, provides new insights on the physical processes which govern changes in seismicity at the end of volcanic eruption. These processes could include magma withdrawal or relaxation, or rheological changes in the magma (which could in turn be due to, for example, increased crystallinity or decreased gas content).

This study is a proof-of-concept of the classification of time series for the detection of large-scale changes in eruptive systems, and further work would be required to make classifications on a real-time basis. In addition, to apply these techniques to a greater number of volcanoes requires representative seismic data during eruptive and non-eruptive periods to train new models. These models also do not give any indication of the type or severity of potential eruptive activity when the classification is eruptive. Further work, incorporating a more diverse range of time-series observations and datasets into the modelling, is needed to investigate the sensitivity of the modelled end-dates determination to the nature and variety of datasets used.

The models presented in this study could be extended by defining 3 classes for model training and classification. Here, the non-eruptive class could be split into two classes: one which represents a

background class and one which represents a precursory class to eruption. However, to make this extension it would be necessary to have an independent data stream to reliably distinguish between the background and precursory states, such as a gas time series hence this analysis is not presented here.

3.5.3) Failure of logistic regression

For both of the datasets, logistic regression performed poorly relative to the other methods applied to the datasets, with end-dates 1-3 months after the other classification models which were all in agreement. As logistic regression involves linear modelling to define the classification, failure to characterise the overall volcanic state indicates that the underlying relationships are non-linear, even when features are removed from the dataset through regularisation. The processes governing volcanic eruption have been previously described as "nonlinear and stochastic" (Sparks, 2003), which could account for the failure of the logistic regression approach here.

3.5.4) Feature importance: feature ranking

The feature importance results from the 2 methods which yield full feature importance results are summarised in Figure 3.11; here the top 10 features ranked as most important in determining the transition between eruptive and non-eruptive for the Nevado del Ruiz and Telica datasets are plotted. As there may be orders of magnitude between the importance score for these two methods, it is better to value the very top features. The groups of features with high ranks for both methods are dominant frequencies, band ratio and waveform standard deviation (Nevado del Ruiz), and peak amplitude and dominant frequencies (Telica). Dominant frequencies rank in the top 5 features for all methods and for both volcanoes.

Though daily event rate is widely used as a parameter for determining changes in volcanic activity, the daily (total) event rate only appears in the top 10 features at Nevado del Ruiz. Increases in VT seismicity has been demonstrated as a common precursor to volcanic eruption at closed volcanic systems, particularly at previously dormant volcanoes (Cameron *et al.* 2018). However, studies at

basaltic systems including Kilauea Volcano, Hawaii (Chastin and Main, 2003) and Piton de la Fournaise Volcano, Réunion Island (Collombet *et al*., 2003) have found that these precursory increases in VT seismicity are often either not present, or do not always lead to an eruption. Increases in VT seismicity ending in no eruption have also been documented as "failed eruptions" (Moran *et al*., 2011), representing a challenge in determining whether a phase of unrest will lead to eruption. From the results presented here, we cannot conclusively identify a category of features which distinguishes between eruptive and non-eruptive behaviour for both volcanoes.

3.5.5) Verification of model assumptions

From the successful classification of non-eruptive and eruptive activity presented in Section 3.4, we conclude that it is at least approximately correct to make the assumption that the data are Independent and Identically Distributed (IID) for the relatively small VEI and short-lived (< 3 – 5 year) eruptions considered in this study. If similar methods are applied on different timescales, this assumption may not be valid. If data were binned on a shorter timescale than daily observations, it may not be appropriate to make this assumption as individual events in certain cases can be quasi-periodic, i.e., not independent from each other (Ignatieva *et al*., 2018). Following a catastrophic eruption, the assumption that each day is drawn from an identical distribution may not hold. Seismicity has been shown to reflect processes within the conduit (e.g., Jousset *et al*., 2003), which in turn can be eroded by several processes during eruption including volcanic tremor or wall collapse (Macedonio *et al*., 1994). Observations of precursory seismic activity at Kelud Volcano, Indonesia preceding the 2007 and 2014 eruptions found significant differences in seismic characteristics before both eruptions, which is consistent with the contrasting eruption dynamics of the two events (Hidayati *et al*., 2018). Transition periods between eruptive and non-eruptive data may last on timescales from hours to weeks (Carniel *et al*., 2003; Ripepe *et al*., 2002), and behaviour during these transitions may represent a different mode of the volcanic system (Connor *et al*., 2003; Rodgers *et al*., 2016). Though the successful models presented here indicate that there are no significant transition periods within the data periods included in this study, for volcanoes with transition periods on longer timescales (such as days – weeks) the transition period may need to be defined as a separate, third class of activity.

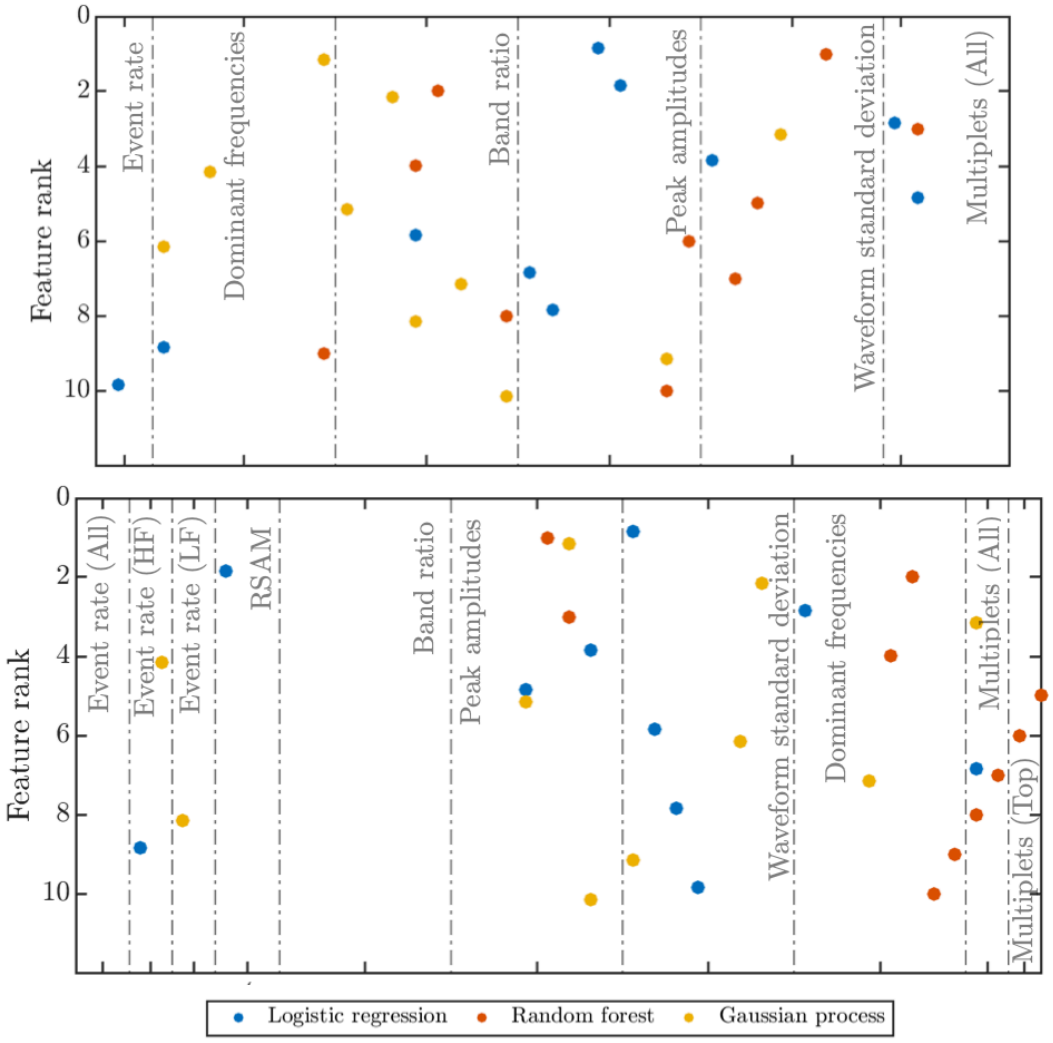


Figure 3.11: Plot of the top 10 ranked features from each feature importance method on Nevado del Ruiz (top) and Telica (bottom). Vertical dashed lines represent the distinction between groups of features (labelled on the x-axis).

3.6) Conclusions

Machine learning methods can successfully classify overall patterns of eruptive and non-eruptive behaviour in seismic time series. This study is the first to apply machine learning techniques applied to single-station seismic data to classify overall volcanic state as eruptive or non-eruptive. We define a decisiveness index D to evaluate classification of eruptive state based on the consistency of classification, which is comparable across datasets. Our models have a high agreement in terms of eruptive classification with visual indicators of eruption, such as ash emissions. The date of the eruption end is found to be consistently later than the date recorded in GVP, by approximately 60 – 120 days. This finding agrees with previous, non-physical definitions of end of volcanic eruption, such as the 90-day rule for determining the timing of eruption end (Simkin and Siebert, 1994). Classification of eruptive and non-eruptive data could be applied to seismic time series to determine when end of eruption occurred, in the absence of conclusive visual observations. Support Vector Machine and Gaussian Process Classifiers were the most successful classification models applied to Nevado del Ruiz and Telica respectively. Logistic regression, a linear classifier, had lower classification accuracy and decisiveness for both datasets, which could be due to non-linearity in the data. Feature importance methods identified little consistency between the most important seismic features used as model inputs. Work on a larger number and variety of datasets is necessary to determine whether these most important features are consistent between volcanoes, or between volcanoes with similar eruption styles or tectonic settings.

Chapter 4 - Machine learning approaches to identifying changes in eruptive state using multi-parameter datasets from the 2006 eruption of Augustine Volcano, Alaska

4.1) Abstract

Understanding the timing of critical changes in volcanic systems, such as the beginning and end of eruptive behaviour, is a key goal of volcanic monitoring. Traditional approaches to forecasting these changes have used models motivated by the underlying physics of eruption onset, which assume that geophysical precursors will consistently display similar patterns prior to transition in volcanic state. We present a machine learning classification approach for detecting significant changes in patterns of volcanic activity, potentially signalling transitions during the onset or end of volcanic activity, which does not require a model of the physical processes underlying critical changes. We apply novelty detection, where models are trained only on data prior to eruption, to the precursory unrest at Augustine Volcano, Alaska in 2005. This approach looks promising for geophysically-monitored volcanic systems which have been in repose for some time, as no eruptive data are required for model training. We compare novelty detection results with multi-class classification, where models are trained on examples of both non-eruptive and eruptive data. We contextualise the results of these classification models using constraints from petrological, satellite and visual observations from the 2006 eruption of Augustine Volcano. The transition from non-eruptive to eruptive behaviour we identify in mid-November 2005 is in agreement with previous estimates of the initiation of dike intrusion prior to the 2006 eruption. We find that models which include multiple types of data (seismic, deformation and gas emissions) can better distinguish between non-eruptive and eruptive data than models formulated on single data types.

4.2) Introduction

Volcanic eruptions are characterised by large-scale transitions in behaviour: from repose to unrest; from unrest to eruption; and from eruption to quiescence. Identifying the transitions occurring at the beginning or end of volcanic activity is a fundamental goal of volcanology (Sparks, 2003; NASEM, 2017). The beginnings of volcanic eruptions are often marked by changes in geophysically-observable indicators of unrest, including seismic, gas and ground deformation measurements; however, it is often difficult to interpret these signals and integrate multi-parameter observations to determine whether this unrest will lead to an eruption (Pritchard et al., 2019; Moran et al., 2011; Segall, 2013). Interpreting the physical meaning of changes in geophysical observations can be especially difficult in real-time. Figure 4.1 and Table 4.1 outline known physical processes that can generate various signals of unrest in seismic, gas and deformation data, and which operate across a range of depths within the volcanic system.

Geophysical parameters may not always show consistent changes leading up to eruption. For example, increasing volcano tectonic (VT) seismicity is typically considered an important precursor to eruption (Cameron *et al*., 2018) at basaltic systems including Kīlauea, Hawai‘i and Piton de la Fournaise, Réunion Island. However, these precursory increases are not always present, and even if present do not consistently culminate in an eruptive episode (Chastin and Main, 2003; Collombet *et al*., 2003). An analysis by Pesichek *et al*., (2018) of seismically-monitored eruptions spanning a variety of eruptive styles across Alaska found that only 30% were associated with a statistically-significant increase in precursory event rate. At persistently restless volcanoes, such as Telica Volcano, Nicaragua, volcanic unrest is often prolonged and seismic event rates decrease prior to eruption due to weak sealing within the hydrothermal system (Geirsson *et* al., 2014; Rodgers *et al*., 2015; Roman *et al.*, 2016; Roman *et al*., 2019). Compilations of global satellite observations of deformation illustrate that, although the majority of studied volcanoes which erupted were associated with deformation, there remain several eruptions for which there was no observable associated deformation prior to eruption, including Shishaldin and Cleveland volcanoes (Ebmeier *et al*., 2013;

Biggs *et al*., 2014; Lu and Dzurisin, 2014). In the case of Cleveland volcano, this lack of observable deformation in both long-term and short-term Interferometric Synthetic Aperture Radar (InSAR) images is attributed to the "open-conduit" nature of the shallow magma storage system (Lu and Dzurisin, 2014). Gas emission rates including SO_2 emissions generally increase prior to eruption (e.g., Sparks, 2003). However, in some instances, such as the March 1993 eruption of Galeras Volcano, Colombia, a decreasing SO_2 emission rate was observed prior to eruption which was attributed to retention of gases due to sealing processes within the volcanic carapace (Fischer *et al*., 1994).

Previous approaches to identifying critical changes preceding eruptions have focussed on incorporating physical understanding of potential drivers of eruption. For example, the Failure Forecast Method (FFM) is based on the assumption that an accelerating rate of a given observable geophysical parameter, such as the rate of seismicity, corresponds to the accumulation of subsurface damage (e.g., Voight, 1988; Voight and Cornelius, 1991; Kilburn, 2003). Eruptions may then be forecast by fitting a power-law relationship to the rate of occurrence of the observable parameter. However, physically-motivated models rely on the assumption that the mechanics underlying the model are consistent, which may not always be the case. Application of a probabilistic formulation of the FFM to multiple run-ups to explosions in the same eruptive episode at Volcan de Colima, Mexico yielded variable results for run-ups in which the pattern of unrest was more complicated than a single acceleration of seismic rate (Boué *et al*., 2015). Other physically-motivated models may be highly successful but are rather specific to one eruption or class of eruptions and limited in their applicability to a wide range of volcanic systems. For example, Anderson and Segall (2013) presented a physics-based extrusion model combined with a Monte-Carlo inversion approach, and demonstrated retrospectively that the final extruded volume of the Mt St Helens 2004 – 2008 dome-forming eruption could be successfully forecasted from up to two years prior to the end of the eruption (see also Segall, 2013).

Identifying eruption onset and end using machine learning classification approaches is a growing field. Dempsey *et al*. (2020) presented a machine learning approach to issuing short-term alerts from

real-time seismic data at Whakaari, New Zealand, using supervised random forest classification on features derived from seismic tremor data and retrospectively applied the model to five episodes of eruption. Ren *et al*. (2020) applied both supervised and unsupervised classification to seismic data at Piton de la Fournaise, La Réunion and identified consistent tremor frequencies between eruptive episodes. Manley *et al*. (2020) identified transitions based on supervised classification of data from Telica and Nevado del Ruiz volcanoes, and showed that machine learning classification of features derived from seismic data alone held the potential to identify the end of eruptive sequences. Additionally, the timing of transition between eruptive and non-eruptive seismic activity at the end of eruption was found to be 2 – 4 months after the last visual activity observed at the volcano.

While seismic data have historically been the most common observable used to monitor volcanoes, complementary information is contained within other parameters, including gas and deformation data. Following on from our previous study (Manley *et al*., 2020) which was limited to seismic data, here we use multiple types of instrumental data, including seismicity, gas emission rate and GPS-derived deformation to classify behaviour prior to and during a major eruption at Augustine volcano, Alaska, using a machine learning approach. We hypothesise that the range of physical phenomena captured from an ensemble of observables is much wider than would have been identified using a single data stream, and that the physical processes driving these phenomena will have operated across a range of depths within the volcanic system.

Augustine Volcano, Alaska is a stratovolcano located in the Cook Inlet region (Figure 4.2), approximately 280 km SW of Anchorage, and is one of the most historically active volcanoes in Alaska (Power *et al*., 2010), with major eruptions in 1976 and 1986. The most recent eruption of Augustine began in January 2006, and was documented in detail (as outlined in Section 4.3.1) (e.g., Power *et al*., 2010 and references therein). Coombs *et al*. (2010) described the chronology of deposits from the eruption, inferred from observations made during overflights, and satellite remote sensing. Timing of the large ash explosions which defined the explosive phase of the eruption were constrained using infrasound (Petersen *et al*., 2006), and these observations which extended over the

course of the eruption provide an independent constraint on the timing of transitions in activity at Augustine, and of the physical interpretations of our modelling results detailed below. The 2006 eruption of Augustine represents an ideal case study as the eruption was large, and the availability of comprehensive geophysical and observational datasets constraining the activity provides context for the results of classification modelling.

In order to discuss the transitions between the state of volcanic activity at Augustine, we first define the states of behaviour which we adopt here, namely 'repose', 'unrest' and 'eruption'. We use a definition of eruption after Siebert *et al*. (2015): eruption is characterised by the explosive ejection of fragmented new magma or older previously solidified magma, and/or the effusion of liquid lava. Unrest is defined as any non-eruptive activity above background which occurs outside the eruption, and the term repose is used for a volcano which is neither erupting or displaying above-background signs of unrest. For the Augustine system specifically, the 2006 activity can also be divided into phases distinguished by their visual products erupted during these periods (Power *et al*., 2006), and according to the Alaska Volcano Observatory system of rating volcanic activity known as the Level of Concern Color Code system. These classification schemes are described in detail in Section 4.3.1.

In this paper we compare results from classifying volcanic state using a novelty detection or one-class approach, which does not incorporate information from future days of the time series, to a multi-class approach in which models are trained on examples of both non-eruptive and eruptive data. These models are formulated on features derived from seismic, gas and geodetic data. We present a novel approach for handling missing and irregularly-sampled values in SO_2 emission rate time series. We compare the performance of novelty detection and multi-class classification approaches using SO_2 emission rates estimated using either no future data points or with a probabilistic, retrospective approach formed using the full set of data points. We also compare models formulated with all datasets to models formulated using only seismic, gas or geodetic features.

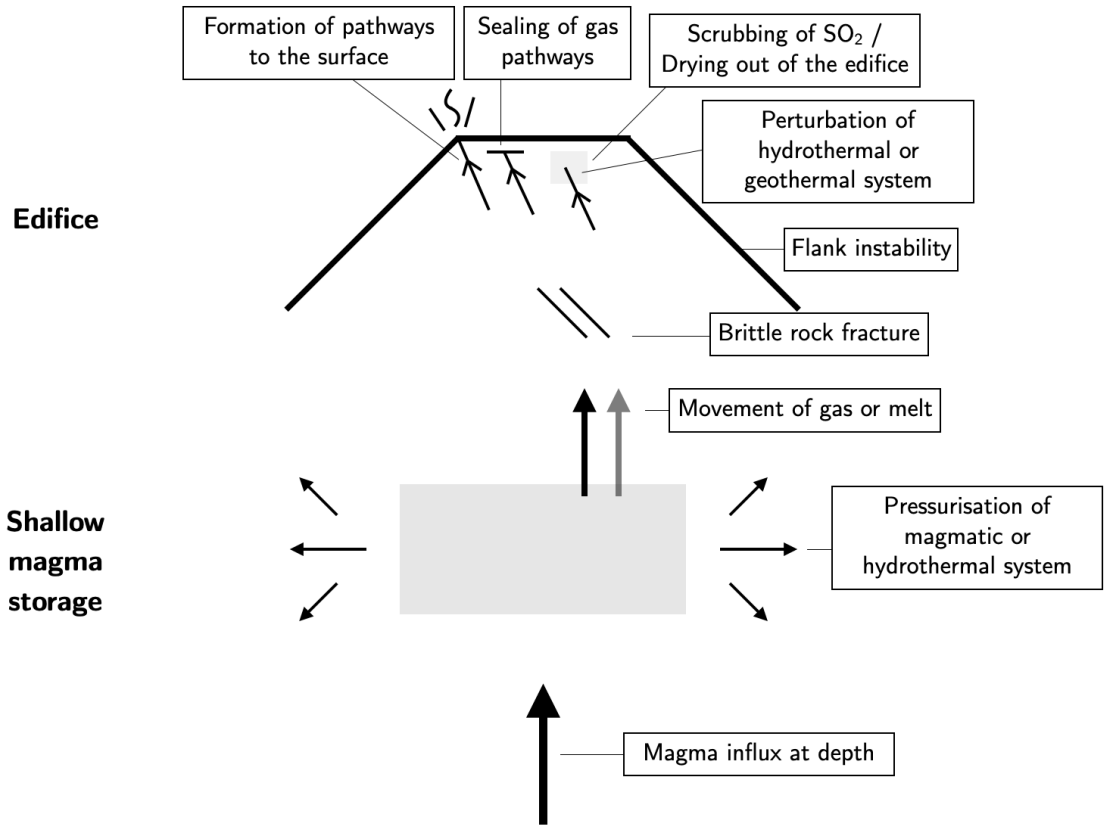


Figure 4.1: Illustration of the range and relative depths of physical processes associated with observation of seismicity, gas and geodetic measurements. Arrows are a conceptual representation of movement of material. Grey boxes are a generalised representation of magmatic or hydrothermal activity. Annotations denote the corresponding physical processes outlined in Table 4.1.

Table 4.1: Summary of observable changes and underlying physical processes which are linked to observations of seismicity, gas and geodetic measurements. All of these physical processes are conceptually illustrated in Figure 4.1.

Type	Observed	Associated physical processes
Seismic	Swarms of low frequency (LF) earthquakes Volcanic tremor Accelerating seismic rates (e.g., Cameron *et al*., 2018) or increase in Real-time Seismic Amplitude Measurement (RSAM) Declines in seismicity (if close to eruption) due to sealing (Endo *et al*., 1996)	Pressurisation of magmatic or hydrothermal system Brittle rock fracture (e.g., Chouet and Matoza, 2013) Movement of gas or melt Formation of pathways to surface
Gas	Increased SO_2 flux approaching eruption Decreased SO_2 flux due to sealing	Magma influx at depth modifying gas composition ratios (e.g., Aiuppa *et al*., 2007) Formation of pathways to surface (e.g., Sparks, 2003; Werner *et al*., 2011) SO_2 scrubbing, or drying out of edifice (Symonds *et al*., 2001)
Geodetic	Deformation which exceeds the baseline recorded deformation Changes in sign of deformation (e.g. deflation or inflation) if very close to eruption	Intrusion of new magma at depth Magma migration Perturbation of hydrothermal or geothermal system Flank instability Brittle rock fracture

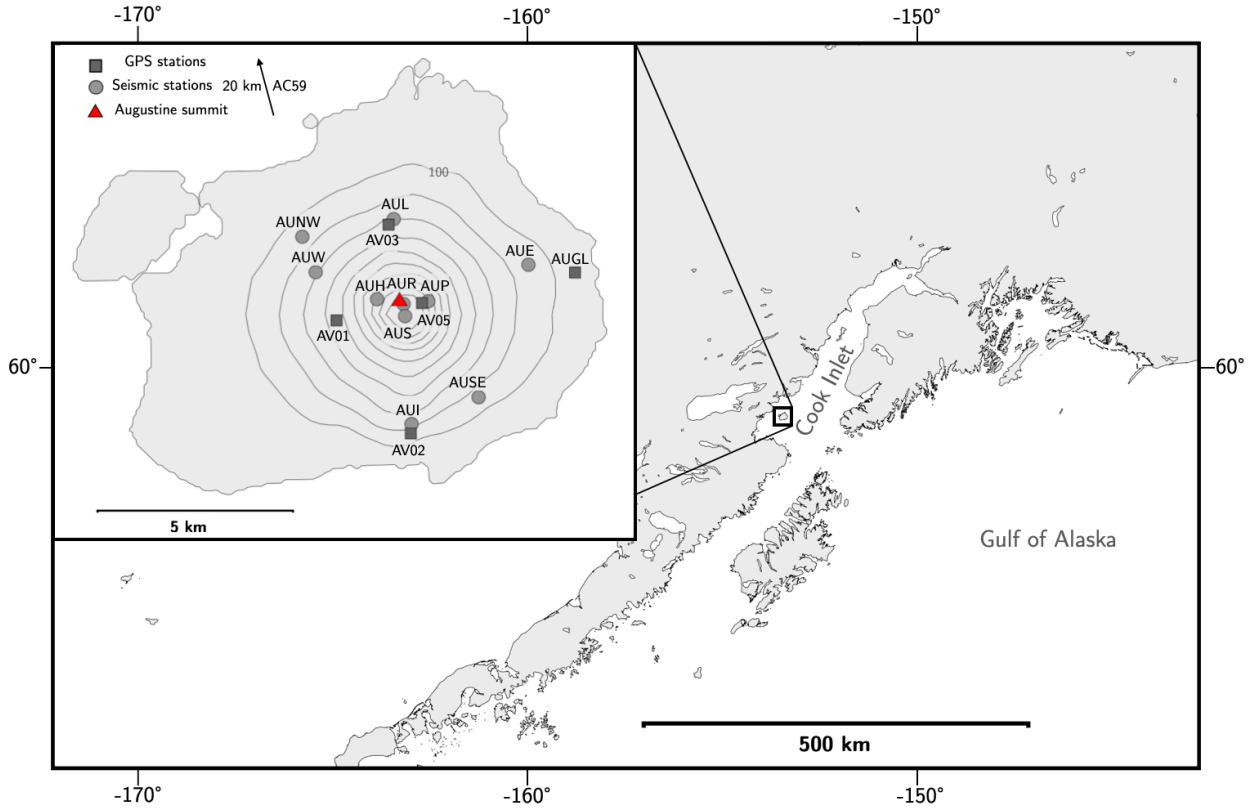


Figure 4.2: Seismic and GPS station map for Augustine Island, inset in a map of the Cook Inlet, Alaska. The location is indicated by latitude and longitude on the axes, with contours at 100m intervals. Squares denote the location of GPS stations, circles denote the location of seismic stations and the triangle represents the location of Augustine Volcano summit.

4.3) Methods

4.3.1) Augustine Volcano

4.3.1.1) Eruption of 2006

The most recent explosive eruption of Augustine began in January 2006 and is recorded in the Smithsonian Global Volcanism Programme (GVP) database as a VEI 3 event. The GVP start-date for activity is 9th December 2005 and the end-date is 27th April 2006, based on reports of changes in seismicity, though the exact days are marked with uncertainty in the GVP record (Global Volcanism Programme, 2006). Figure 4.3 illustrates the timeline of the visible eruption during the 2006 period of activity. The 2006 eruption of Augustine began with several small, phreatic explosions on 2nd, 10th, 12th and 15th December (Power and Lalla, 2010), which were not thought to contain juvenile material. The main phase of the eruption began on the 11th January 2006, with two large explosions generating substantial ash-bearing plumes. January 11th was also the day of the eruption with the highest seismic event rate. Explosive activity continued until 28th January, after which the eruption style mainly comprised low-level ash emission, and small pyroclastic flows (Vallance *et al*., 2010).

There are different ways to classify the stages of the event. At the time of its explosive eruption in 2006, Augustine volcano had a system of Concern Color codes which had 4 states: green, yellow, orange and red (Brantley, 1990). The beginning of the 2006 eruption was successfully forecast by the Alaska Volcano Observatory, with the Concern Color code being raised from Green to Yellow 3 days preceding the first phreatic explosions on 2nd December 2005 (Figure 4.3, Cameron *et al*., 2018). In the Concern Color Code system in place at the time, the Yellow level implies that an eruption is possible within the next few weeks. Activity associated with the 2006 eruption itself was split into 4 phases by Power *et al*. (2010), and activity before the eruption is known as the precursory phase, whereas activity following the designated eruption end-date is known as the post-eruptive phase. The eruption itself comprises an explosive phase, continuous phase, hiatus and effusive phase. The classification schemes associated with the Augustine eruption of 2006 are summarised in Figure 4.3.

4.3.1.2) Physical constraints on the magmatic system at Augustine

There are two primary lithologies associated with the 2006 eruption (Larsen *et al*., 2010): High-Silica Andesite pumice (HSA) and Low-Silica Andesite Scoria (LSAS). The volcano summit was located at 1.2 km above sea level (Power and Lalla, 2010). The primary magma storage was located at 3 – 5 km depth below sea level, as constrained by post-and syn-eruptive earthquake locations (Power and Lalla, 2010). Melt inclusion analyses of HSA (Webster *et al*., 2010) and lack of significant SO_2 emissions before December 2005 indicated storage at > 4 km depth below the edifice (Coombs *et al*., 2010). There was a primarily aseismic conduit linking the magma storage to the fracture zone; with few earthquakes located at these depths during the 1970s (Power and Lalla, 2010). The majority of earthquakes during the precursory swarm were located within a network of fractures from 100s of m above sea level – 1 km below sea level (Power and Lalla, 2010; DeShon *et al*., 2010; Figure 4.3).

Cervelli *et al*., (2010) suggest that there was no active magma column below the edifice prior to the 2006 eruption, inferred due to low fumarolic temperatures at Augustine. However, Jacobs and McNutt (2010) report increases in the monthly temperature average from January 2005, measured at the location of the AUS seismic station, which are attributed to a volcanic origin. The conduit which fed the 2006 eruptive episode has been suggested to have formed in the preceding weeks – months to the eruption (Cervelli *et al*., 2006; Cervelli *et al*., 2010).

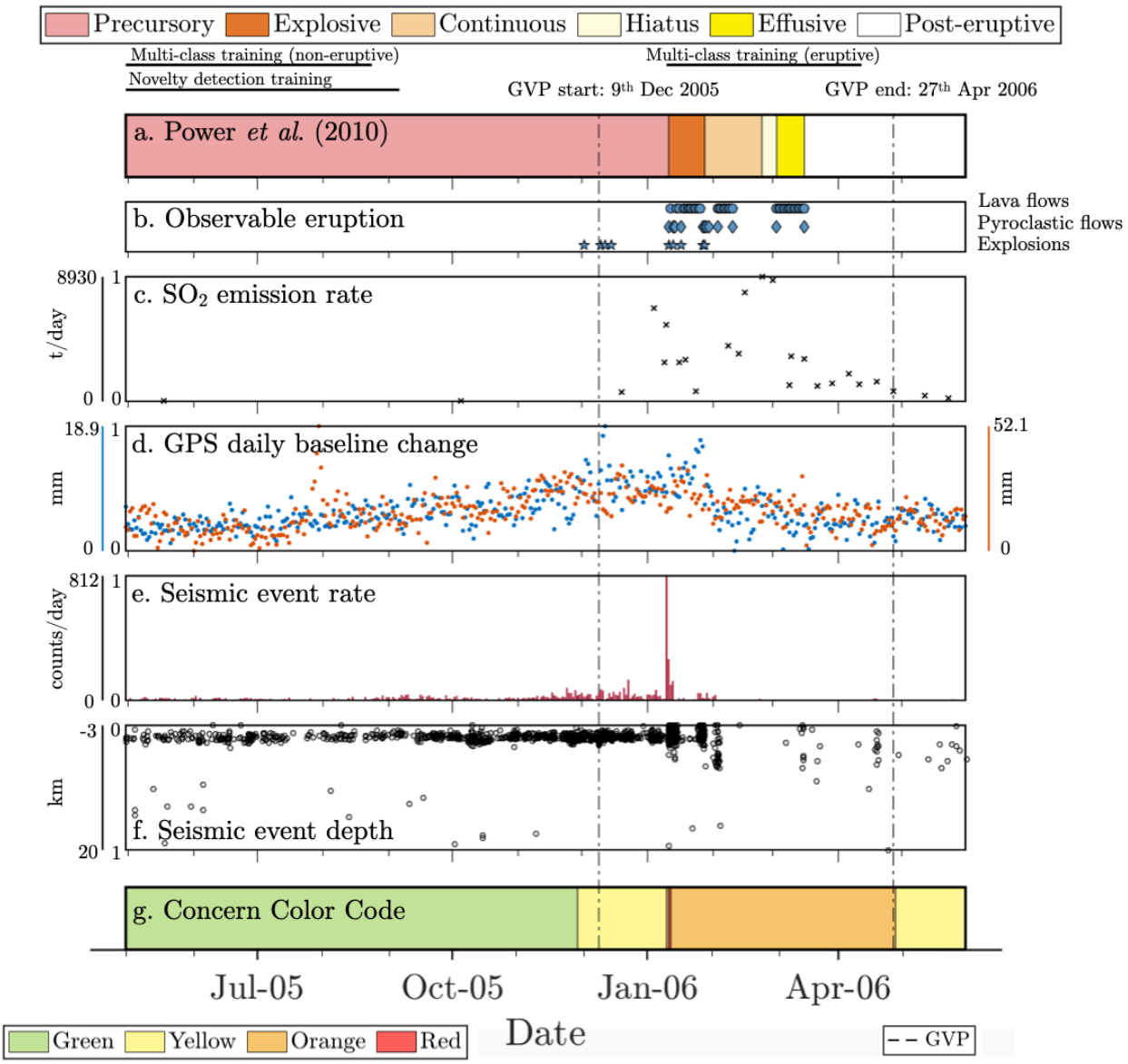


Figure 4.3: Timeline of the data period from April 2005 – May 2006 used in this paper and classification of the eruption timeline outlined in Section 4.3.1.1. Dashed vertical lines denote the GVP beginning and end dates for the eruptive period. Black horizontal lines above the top bar represent the non-eruptive and eruptive training periods for the multi-class model (upper) and non-eruptive training period for the novelty detection model (lower). The top bar (a) represents the classification of activity presented in Power *et al.* (2010) and the bottom bar (g) represents the Concern Color Code status of Augustine throughout the data period. In the observable eruption bar (b), blue symbols denote days on which visual eruption was confirmed: stars represent days in which explosions were observed, diamonds represent days in which pyroclastic flows were observed and circles represent days in which there was evidence of lava dome growth and/or effusion (compiled from Coombs et al, 2010). In the GPS panel (d), blue dots represent the baseline length changes of the N-components of AV02-AC59 and orange dots represent the AV01 vertical component, which are the station and component combinations used to derive geodetic features. Measurements of daily SO_2 emission rate (c), GPS baseline length changes (d), seismic event rate (which is one seismic feature) (e) and seismic event depth (f) are normalised between 0 and 1 and the true bounds of these quantities are contained within the second y-axes for each plot. Seismic event depths are not used to derive features.

4.3.1.3) Data processing

Before applying the machine learning methods (Section 4.3.5), we first extract a number of representative features from the seismic, gas and geodetic measurements. We derive a set of 40 features comprising 34 seismic features, 2 gas features and 4 geodetic features, which are outlined in Appendix C; C.3 – C.5. This full set of features is used to train and test the models described in Section 4.4.

4.3.2) Seismic time series

4.3.2.1) Catalogue curation

We create a catalogue by performing a full network detection on the seismic waveform data from Augustine volcano between 30th April 2005 and 31th May 2006. We use a Short-Term Averaging/Long-Term Averaging (STA/LTA) detection, with detection parameters from DeRoin and McNutt (2012). We perform this detection on the full network of 10 seismic stations (Figure 4.2), using the GISMO toolbox for MATLAB (Thompson and Reyes, 2018). The detection parameters for identifying a trigger were an STA window of 1s and an LTA window of 8s, with an STA/LTA threshold of 2 required for event triggering. The STA/LTA algorithm is used in "frozen" mode, meaning that the LTA window is fixed once an event is triggered. The detection was run twice over the data, once using a low-frequency filter (0.5 – 10 Hz) and once using a high-frequency filter (3 – 18 Hz). For a detection to be included as an event we checked that it was detected on 3 different stations within 10 seconds of each other. To exclude nearby regional events, we checked the resulting detection times against the USGS earthquake catalogue (USGS, 2021). We considered all events occurring in the data period of at least 2.5 magnitude within 500 km of Augustine Island and manually checked events in our detection which coincided with a travel time between 30 – 300s of the USGS catalogue, excluding approximately 50 events.

The event rates from the network detection are in agreement with event rates quoted in Buurman and West (2010). A comparison of the two event rates highlights good agreement on days with no events

and for days with the highest event rates during the explosive phase of the eruption (11th – 13th January 2006). For days with medium event rates, our catalogue yields a daily event count approximately 1 order of magnitude greater than the AVO catalogue. Event rates in the explosive phase onwards were affected by the outages of summit seismic stations, resulting in fewer detected events meeting the criteria for inclusion in the catalogue.

Alaska Volcano Observatory (AVO) published an analyst-reviewed catalogue comprising 2,367 seismic events associated with the 2006 eruption (Dixon and Stihler, 2009). For an event to be included in the AVO catalogue, it must have greater than 3 P phases and greater than 2 S phases in addition to locating an event hypocentre to within 15 km of the volcano and with standard hypocentral errors. Due to these criteria for inclusion, the AVO catalogue represents a subset of the total events at Augustine. Buurman and West (2010) find that 39% of the full number of events they detect at Augustine are included in the AVO catalogue, and that the criteria for an event to be included mean that small and emergent seismic events are underrepresented in the catalogue.

4.3.2.2) Seismic feature extraction

Using event times from both the AVO catalogue and our catalogue, we derive a set of seismic features from the event waveforms at station AUI, a seismic station which was in place before the eruption and was not disabled during the eruption (listed in Appendix C; C.1 and C.2). We use the number of seismic events detected per day, and any days with zero seismic events are omitted from the time series. Though the gaps in the time series could be filled with a substitute value, e.g., the mean value of the feature (Schafer and Graham, 2002), we suggest that no value should be substituted as no events were detected therefore it would not follow to interpolate a value for the event characteristics of non-existent events. Dominant frequency information is derived by recording the 5 peak frequencies from each event in a given day. The ratio of low to high frequency energy in a waveform can be expressed via the band ratio (Buurman and West, 2010; Rodgers *et al*., 2015a) and is calculated by taking the base 2 log of the ratio of energy within two windows of low frequency (1 – 2 Hz) and high frequency (10 – 20 Hz). These windows have previously been applied to data from

Augustine (Buurman and West, 2010). We calculate peak amplitude based on the maximum amplitude of the event normalised by its mean amplitude. We compute the fast Fourier transform of each waveform and calculate the spectral width of the largest peak of each event.

After calculating features for each event on a given day, we then derive the set of seismic features for each day by taking the mean, median, variance, minimum, maximum, 10^{th} percentile, 90^{th} percentile and change in mean from the previous day for all categories apart from event rate, for which the per-day value and change in value from previous day are used.

We derive two sets of features from the seismic data, and compare the two sets of features in Section 4.5.2.1:

(i) Features from seismic events from our catalogue

(ii) Features from seismic events from the AVO Augustine catalogue

4.3.3) Gas time series

SO_2 emission rate measurements for Augustine are available from July 2002 to July 2008. Measurements of SO_2 column abundance were made via airborne traverses, using COSPEC (McGee *et al*., 2010). The SO_2 emission rates which we use for this analysis were processed by McGee *et al*. (2010), who processed the raw COSPEC traverses by calculating the area under the SO_2 traverse time series and divided by the approximate width. This average signal was then multiplied by the plume width and average wind speed to determine an SO_2 emission rate in units of t/day. The time series of SO_2 emission rate is irregular, reflecting the challenges of making these airborne measurements. The median interval between measurements is 8 days, and the mean interval between measurements is 50 days. There are larger intervals between measurements during the pre-eruptive and post-eruptive phases than the eruptive phases as the measurement frequency increased when seismic unrest began in 2005.

To use the gas time series data as input for our models alongside the daily time series from geodetic and seismic datasets, we use interpolation to estimate values of SO_2 emission rate between measurements made at Augustine. Figure 4.4 illustrates the two methods of interpolation used:

(i) Gaussian Process Regression (GPR): a probabilistic approach in which a stochastic model is applied to provide estimates of intervening measurements with confidence intervals, in which no functional form is assumed.

(ii) Carry Last Value Forward (CLVF): until a new measurement of SO_2 emission rate is made, every day has the same value of emission flux.

To interpolate the gas measurements using the GPR, we fit the model to the measurements of SO_2 emission rate. The values are first logged as we do not want to fit a model which can permit negative values for SO_2 emission rate, and because the noise in the dataset is multiplicative i.e., the

uncertainties in emission rate increase as the magnitude of the emission rate increases. Before the eruption, there are values of zero in the dataset which would be undefined in log space, so we choose a value an order of magnitude smaller than the lowest SO_2 emission rate as a "detection limit" of 10^0 t/day. We use a matérn covariance function (Williams and Rasmussen, 2006), which has previously been used to capture processes which are not smoothly varying in meteorological datasets (Guttorp and Gneiting, 2006).

These methods of interpolation each have advantages and disadvantages. CLVF is a conservative estimate of the daily emission rate, using no future emission rate measurements to interpolate the daily values. However, CLVF does not provide an estimate of uncertainty on the emission rate measurements, and also permits large jumps in the daily value of SO_2 emission rate. GPR does not assume an underlying functional form for the SO_2 emission rate time series and also yields an uncertainty bound on the interpolated daily values. The upper uncertainty bound broadly follows the form of the mean, apart from either side of the unrest and eruption (2004 - 2005 and 2007) where the uncertainty is higher due to an increase or decrease of measured emission rate respectively. This upper bound is not used as a feature in the models. GPR requires the use of all of the data points to estimate values for the intervening days and thus is a retrospective method. The total SO_2 flux estimated using CLVF interpolation is 4.3×10^5 t and the mean total SO_2 flux estimated using GPR is 4.5×10^5 t with 95% confidence interval bounds of 1.3×10^5 t and 2.0×10^6 t. Though neither method is preferable to a complete time series of measured daily values of SO_2 emission rate, the methods used for interpolating intervening daily values presented here represent two contrasting approaches to the problem of missing data in gas time series, which might find applications to estimating appropriate time-averaged SO_2 fluxes as well as for detecting changes in eruptive state as explored here.

Though measurements of CO_2 and H_2S were attempted at the same time as SO_2, there are under half as many values obtained for the emission rates of these species compared to SO_2 measurements

(McGee *et al*., 2010). No value of CO_2 or H_2S emission rate is available for the continuous or hiatus phase during the eruption, and there are significant gaps in the precursory and post-eruptive phases of the eruption. Consistent measurements of SO_2 emission rate of 0 t/day during the precursory and post-eruptive phases could be due to the process of SO_2 scrubbing by groundwater (Symonds *et al*., 2001). As the precursory phase of the eruption occurred during the Alaskan winter, the presence of snowpack on Augustine could also have contributed to scrubbing and suppressed the quantity of SO_2 measured at the surface. We derive features only from the SO_2 emission rate, and take the data as reported with no modification to account for potential scrubbing processes. The features extracted from the full SO_2 emission rate time series are the daily estimated value and the change in daily value from the previous day.

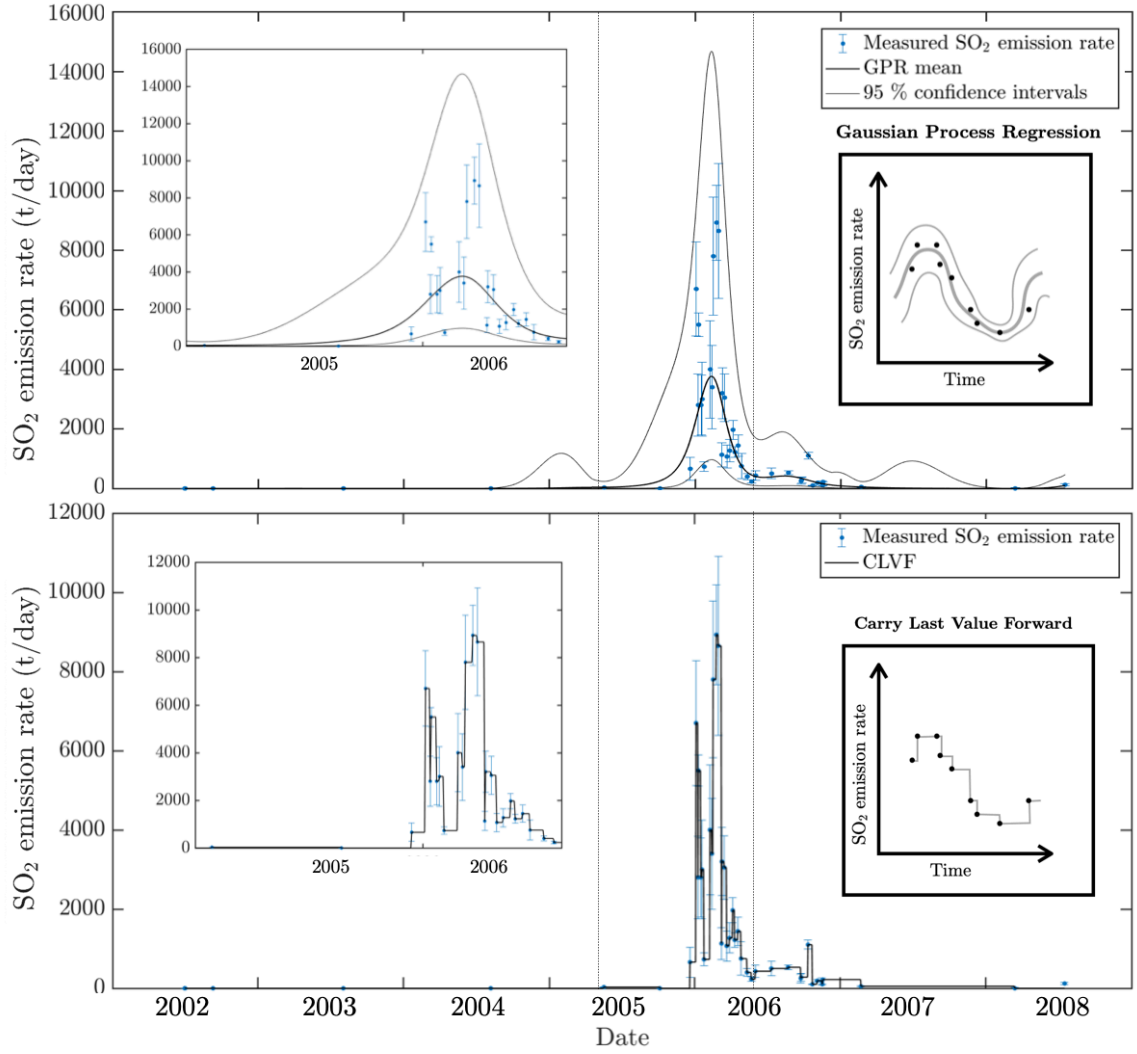


Figure 4.4: Daily SO_2 emission rate values for the measurements made at Augustine between 2002 and 2008 and values interpolated using GPR (top panel) and CLVF (bottom panel). Blue points represent the values processed and reported in McGee et al. (2010) with bars denoting two standard deviations and vertical lines indicating the SO_2 emission rates during the study period. Inset figures within the panels represent an enlarged version of the SO_2 emission rates during the study period (left) and a conceptual illustration of each method (right).

4.3.4) Geodetic measurements

A GPS network comprising 6 stations was in place before the 2006 eruption of Augustine (Figure 4.2; Cervelli *et al*., 2010). This network began to detect precursory deformation around July 2005, and the timing of the first dike intrusions is inferred from deformation measurements as occurring in November 2005 (Cervelli *et al*., 2006). Data availability from stations was variable during the eruption due to the destruction of stations by volcanic deposits during the later stages of eruption. We use the daily GPS solutions provided by Nevada Geodetic Laboratory (Blewitt *et al*., 2018) in the International GNSS Service 14 reference frame (Rebischung and Schmid, 2016). We choose stations AV01, AV02 and AC59, which had a complete record of measurements before, during and after the eruption. For station AV01 we extract the daily vertical position. The difference between the maximum and minimum value of AV01 vertical component during the study period is 52.1mm, a variation which exceeds the typical seasonal signal observed at AV01. For each day we also calculate the variation with respect to the previous day. For station pair AV02-AC59 we compute the daily change of the N-S component of the baseline vector with respect to a fixed chosen day. For the geometry of these two stations the baseline variation is a proxy for N-S extension or shortening across the volcano (e.g., Dixon, 1991; Cervelli *et al.*, 2010; Saballos *et al.*, 2014). The magnitude of change for the AV02-AC59 N-S component of the baseline is 18.9mm. For each day we also calculate the variation with respect to the previous day. The features we use are daily vertical position, daily change in the N-S component of the baseline, and the difference with respect to the previous day of both the vertical position and the N-S component of the baseline.

4.3.5) Machine learning methods

Classification methods aim to categorise input observations into a finite number of categories which are representative of different types of activity of a physical system (Bishop, 2006). The methods we use are supervised, meaning that the model is trained on a selected set of training examples with assigned labels and then subsequently tested on unseen data. During the model training process, some training points (known as validation data) are held back and used to test the model in order to maximise the training accuracy. The validation data are distinct from the testing data which is

subsequently used to evaluate the effectiveness of the models in classifying eruptive state. The models presented in Section 4.4 are trained on the seismic, gas and geodetic features described in Sections 4.3.2 – 4.3.4. Classification methods used here can be split into one-class and multi-class classification, which are described in further detail below.

4.3.5.1) Novelty detection classification

Novelty detection (also known as one-class classification) methods involve training the model on a single class of observation, and then testing subsequent observations against this model to identify patterns which are anomalous or novel, thus are often referred to as novelty detection (Pimentel *et al*., 2014). In this approach, typically the training or "normal" class is better sampled than the abnormal or "novel" data with the objective of identifying the onset of the "novel" data. A novelty detection approach is advantageous for systems in which there may be several possible "abnormal" modes, and in which those abnormal modes are not well characterised. Existing applications of novelty detection include medical diagnostics (e.g., Quinn and Williams, 2007) and jet engine condition monitoring (Clifton *et al*., 2007). In the geosciences, novelty detection has been previously applied to remote sensing for the purposes of classifying the occurrence of a single type of land use (Li *et al*., 2010; Munoz-Mari *et al*., 2010).

Use of novelty detection in volcanology is novel. Volcanic systems are conceptually well-suited to a novelty detection approach, as there is a broad spectrum of volcanic styles (thus failure types) associated with eruption. Moreover, larger eruptions are associated with longer repose times (Pyle, 1998; Passarelli and Brodsky, 2012) and thus previous eruptions may not have been fully instrumented (and therefore well-characterised). We use a nearest-neighbour-based approach to novelty detection, using the NDtool novelty detection toolbox (Clifton, 2007; Pimentel *et al*., 2014). For a nearest neighbour approach, the Euclidean distance between a test sample and its nearest neighbour within the training data is computed. The novelty score is then calculated by comparing this distance to the maximum distance between that nearest neighbour within the training data and its nearest neighbour in the entire training set. Hence, a larger novelty score is achieved where the test

sample lies far outside the region of the training data as the distance between the testing sample and its nearest neighbour will greatly exceed the distance between that nearest neighbour and any point in the training data. The novelty threshold is defined as the maximum novelty score calculated within the training dataset. Novelty detection models require a model training period which is sufficiently long to capture the distribution of the normal class, and we select the minimum possible training length for successful model training in order to test the model over the months preceding the eruption. We select a training period for the novelty detection model which begins on 30th April 2005, the first day of the data period, and ends on 6th September 2005. We hold back 1/5 of the training points to use as validation data for the model. This training period represents approximately 30% of the total number of days spanned by the dataset.

4.3.5.2) Multi-class classification

Multi-class classification methods involve training models on two or more classes of observation, and then testing on unseen data. In this study we apply binary classification, with two classes: "non-eruptive" data and "eruptive" data, and we select representative training examples from both classes. For the non-eruptive training data, we select days at the beginning of the data period in order to train as far away from the hypothesised timing of the start of eruption, from 30th April 2005 – 24th August 2005. For the eruptive training period, although the GVP start-date was a month earlier, visible eruption was limited to a short period of steam explosions and therefore we choose a training period from the main explosive phase of the eruption onwards, from 10th January 2006 – 12th April 2006. This training period represents approximately 40% of the total number of days spanned by the dataset. We apply a random forest method, in which an ensemble of decision trees is averaged to produce a classification, wherein the decision represents the conversion from model inputs to final classification. Individual decision trees represent a series of operations which compare the features to randomly selected thresholds (Hastie *et al*., 2001). The aggregation of decision trees means that the final models have good generalisation. We use standard 5-fold cross validation in order to test the model performance during the training process (Hastie *et al*., 2001). Once a final aggregated tree is formulated, the model is then tested on the unseen data. This method was successfully applied in

previous work classifying volcanic state from features derived from seismic data (Manley *et al*., 2020).

4.3.6) Assumptions

The classification approaches applied here involve the underlying assumption that the variables used as input are Independent and Identically Distributed (IID). This assumption implies that features on a given day are mutually independent from features from other days, and features are drawn from the same underlying distribution between days (Cover and Thomas, 2006). Data are rarely perfectly IID in practice; however, we hypothesise that this approximation holds to a first-order and observe whether this hypothesis is supported by the resulting models distinguishing between previously-unseen eruptive and non-eruptive data. This assumption was shown to be acceptable to a first order in previous work involving these classification methods (Manley *et al*., 2020) and we therefore retain this assumption for the seismic features. In this paper, we also assume that the individual gas and geodetic changes will be approximately IID for similar reasons to the seismic data: the generative process which produces measured geodetic and gas emission rates is assumed to not have memory of values on previous days. This assumption is less constrained for the interpolated GPR and CLVF values of SO_2 emission rate as the values interpolated for the missing days depend on the nearest measured values. Nonetheless, interpolation of the missing gas data is preferable to excluding the days with no measurements altogether. Gill *et al*., (2007) establish a similar methodology to fill missing values in hydrological datasets prior to applying machine learning algorithms, and find significantly better performance for models with interpolated values compared to models formulated with missing values omitted from the dataset.

4.4) Results

We present results from one-class and multi-class classification models trained and tested independently on time series of features derived from the Augustine seismic, geodetic and gas data. We compare classification results using two end-member approaches for interpolation of the SO_2 emission rate: Carry Last Value Forward (CLVF) and Gaussian Process Regression (GPR), introduced in Section 4.3.3.

4.4.1) Novelty detection results

Figure 4.5 illustrates the results from novelty detection methods using features derived from seismic, geodetic and both gas data interpolation methods, which are also summarised in Figure 4.6c and Appendix C (C.6). These models were both formulated using a nearest-neighbour approach to novelty detection. Each day of the model is classified independently, therefore there is no influence of previous or future novelty scores on the current day. There are several observations which can be made from the novelty detection results (Figure 4.5):

(i) Both models begin to consistently classify days as eruptive prior to the phreatic explosions in December 2005. The novelty scores begin to exceed the novelty threshold in mid-November 2005 for both gas interpolation methods (GPR and CLVF). This classification of abnormal behaviour precedes the change in Concern Color code at Augustine (Cameron *et al*., 2018).

(ii) The overall shape of the novelty score is comparable to the method of SO_2 emission rate interpolation, thus the CLVF novelty scores appear step based and GPR novelty scores are smoothly varying.

(iii) There is agreement between the models on the end-date of eruption – GPR novelty scores and CLVF novelty scores do not return below novelty threshold within the data period end-date of May 2006.

(iv) Precursory classification of abnormal behaviour in the CLVF model occurs whilst SO_2 emission rates remain at 0 t/day, indicating that changes in the geodetic and seismic features are primarily responsible for the novelty scores.

We compare the novelty detection model results to the GVP labels and find 84% testing agreement for the CLVF-based model compared to GVP labels and 58% agreement for GPR-based model compared to GVP labels. The misclassifications with respect to the GVP labels are entirely false positives, i.e., more days are classified as eruptive compared to the dates recorded in GVP. The model agreements reported here are computed relative to the labels recorded in GVP, where the beginning of the eruption is recorded as 9th December 2005 and the end-date is 27th April 2006. Labels from GVP are defined by the presence or absence of visual evidence of volcanic activity. However, as discussed in Section 4.2 and by Manley *et al*., (2020), a range of subsurface processes are sampled by seismic, gas and geodetic observations which are not necessarily accompanied by surface activity and therefore reflected in the GVP labels.

4.4.2) Multi-class results

The results from random forest classification are illustrated in Figure 4.6 and outlined in Appendix C (C.6). As in the one-class models, each day of the model is classified independently thus the classification on any given day is not influenced by the classification on another day. The following key observations can be made from the two random forest approaches:

(i) The model formulated using seismic, geodetic and GPR-interpolated gas features (Figure 4.6a) begins to classify days as eruptive in mid-November 2005, whereas the model based on seismic, geodetic and CLVF-interpolated gas features (Figure 4.6b) begins to classify days as eruptive in mid-December 2005.

(ii) There is some agreement between the models on the end-date of eruption – the CLVF-based model returns to non-eruptive classification in early-mid May 2006, whereas the GPR-based model end-date is slightly earlier, in early May.

The agreement of the random forest models with GVP labels is higher for both interpolation approaches than the one-class approach. Overall model agreement for the model with CLVF-interpolated gas features is 86% and model agreement for the model with GPR-interpolated gas features is 82%. The multi-class model with CLVF-interpolated gas features begins to classify days as

eruptive later than all other novelty detection or multi-class models. However, the classification of eruptive activity still begins approximately a month before the main explosive phase of the eruption, in which the first juvenile material associated with the eruption was ejected.

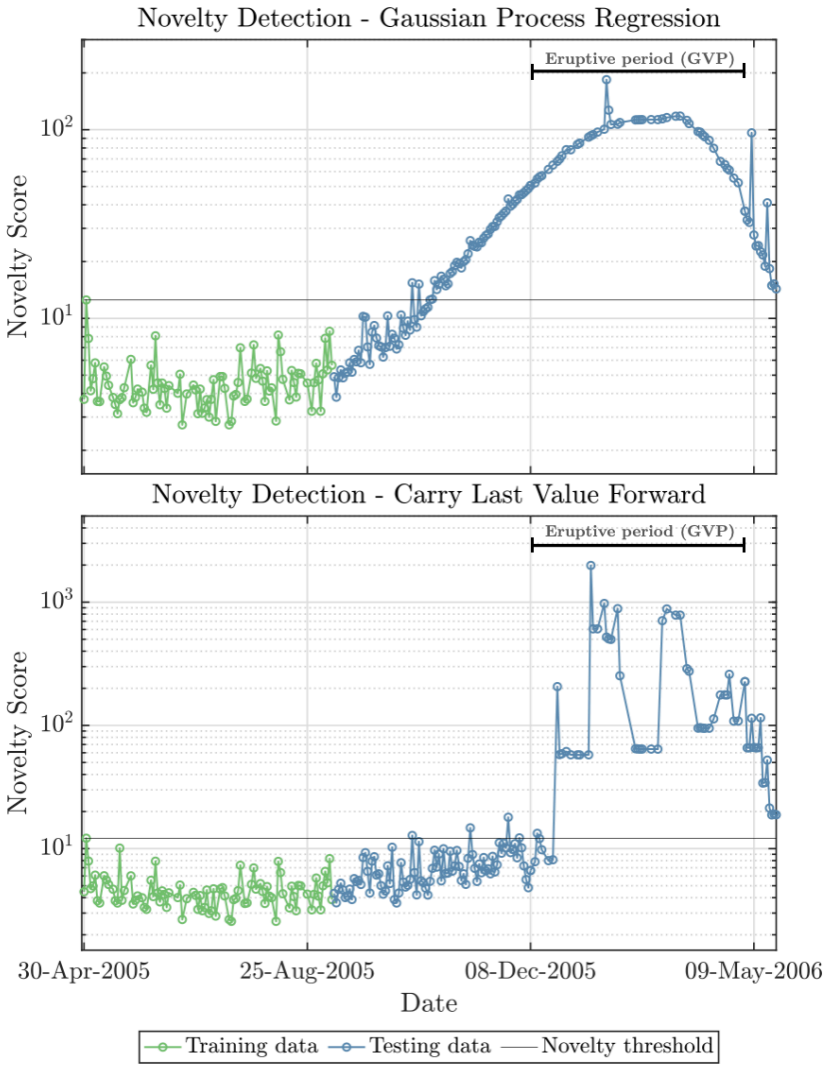


Figure 4.5: Results from novelty detection classification applied to the features derived from Augustine data. Green line denotes the training data and blue line denotes the testing data. The grey horizontal line represents the novelty threshold determined during model training. The black bar at the top of each panel represents the dates on which eruption was occurring according to the GVP database. Novelty score is dimensionless, therefore the y-axis has no associated units but note the different y-axis scales between the top and the bottom panels.

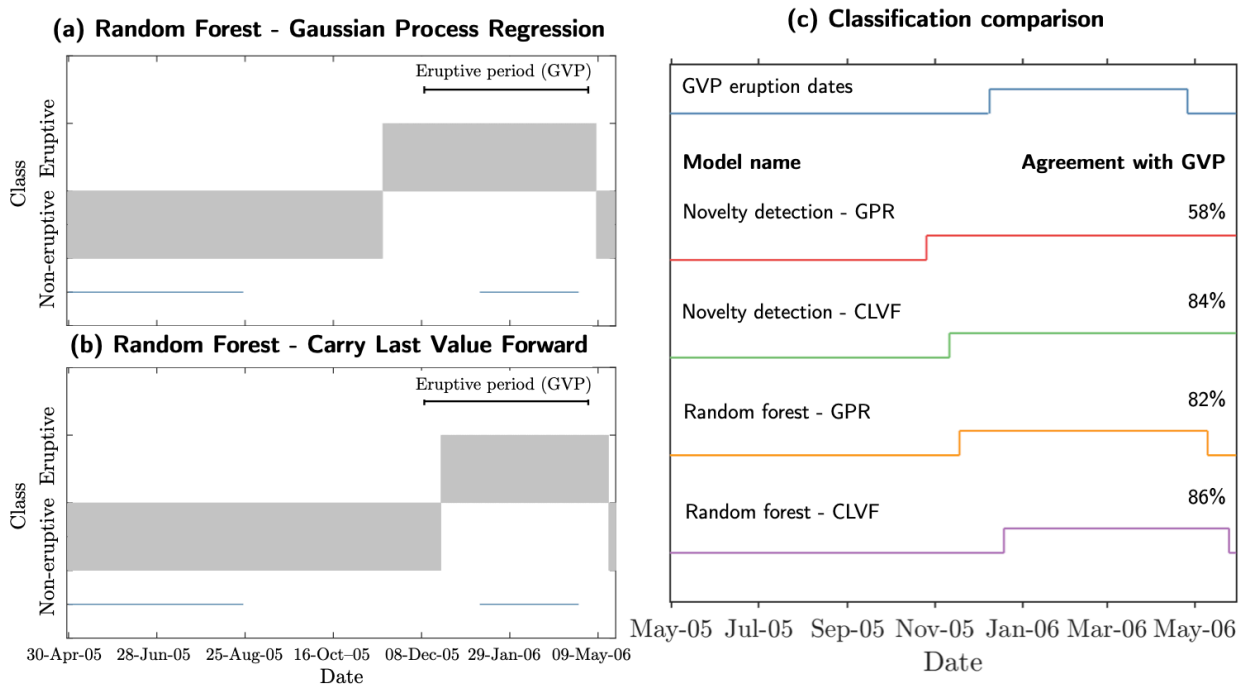


Figure 4.6: Results from the multi-class results from the random forest classification applied to the seismic, geodetic and gas features derived from Augustine data. These results are from either (a) Gaussian Process Regression and (b) Carry Last Value Forward interpolation of gas time series data. The grey squares in panels (a) and (b) represent the classification output of the model corresponding to either non-eruptive or eruptive. The horizontal blue line beneath the classification results represents the training period of the model. The black bar at the top of each panel represents the dates on which eruption was occurring according to the GVP database. (c) Model comparison of the classification results from novelty detection and random forest methods with GVP dates. Each line represents the eruptive state (non-eruptive or eruptive) in GVP or as classified by the model, with the eruptive state beginning when a model begins to classify data as eruptive and ending when a model begins to classify data as non-eruptive. For the novelty detection models, the novelty score does not come below the threshold before the end of the study period. Model agreement with GVP is calculated by dividing the number of classified days which match the GVP classifications by the total number of classified days, expressed as a percentage. Please refer to Appendix C (C.6) for more details.

4.5) Discussion

4.5.1) Physical interpretation of model results

Both CLVF-based and GPR-based novelty detection models and GPR-based multi-class models begin to classify days as eruptive during the precursory phase of the Augustine eruption, around November 2005. For the GPR-based novelty detection model, the novelty score does not return below the novelty threshold for the remainder of the study period whereas the CLVF-based novelty scores have isolated peaks above the novelty threshold. Previous qualitative models of activity during the precursory phase of eruption are consistent with transition at this time. On 17th November 2005, there was a marked offset in the GPS-derived offsets between AV04 and AV05, which Cervelli *et al*. (2010) attribute to a crack opening at a depth near sea level, and visual observations of surface cracks at the summit. A day later on 18th November, the rate of edifice inflation also increased (Cervelli *et al*., 2010).

Previous modelling of ascent duration based on 40 – 100 micron reaction rims in amphiboles from the LSA magma is consistent with a ~20 – 57 day duration of ascent (Larsen *et al*., 2010). This petrologically-derived constraint is consistent with dike initiation around 17th-18th November, as the shortest ascent duration would be associated with an eruption date around 9th December 2005 (consistent with phreatic explosions) and the longest ascent duration would be associated with an eruption date around 12th January 2006 (consistent with beginning of the explosive phase). Therefore, the transition which we identify using machine learning approaches in the precursory phase could be associated with physical changes associated with dike initiation.

Earthquakes in the early precursory phase of activity at Augustine, prior to dike initiation, are attributed to the accumulation of gas below the conduit (Larsen *et al*., 2010). The depth of earthquakes which would be generated by this mechanism is consistent with the model of Huppert and Sparks (2016), who model the pressure distribution associated with the rise of compressible vapour through a porous conduit. In this model, the gas pressure scales with the square root of depth,

whereas the lithostatic pressure scales linearly with depth. Therefore, a critical overpressure is thought to occur approximately 2/3 of the distance between the source depth (i.e., magma storage zone) and the surface (i.e., the edifice) when the gas pressure exceeds the lithostatic pressure and causes rock fracture, thereby nucleating VT events. The observed depths of precursory VTs during the unrest at Augustine are consistent with this model, as the events are concentrated in a fracture zone 3 – 4 km above the primary magma storage depth. A similar process has been observed at Soufrière Hills Volcano, Montserrat, where VT swarms occurred in an overpressure zone ahead of the ascending magma (Rodgers *et al*., 2016). The eruptive transition we identify from our classification modelling could correspond to the source mechanism of precursory earthquakes changing as dike intrusion began. This hypothesis is supported by Syracuse *et al*. (2011) who relocated seismic events at Augustine with an improved velocity model, and found that from November 2005 onwards seismic event locations began to align in a NW-SE trend consistent with intrusive volumes modelled by Cervelli *et al*. (2010) and the regional stress field (Buurman *et al*., 2014).

The transport of magma to the surface via dike intrusion is also associated with changes in gas measurements and ground deformation (e.g., Sparks *et al*., 2012). Increased emission of SO_2 can represent shallow magmatic ascent of gas-rich magma, as the degassing of SO_2 occurs at lower pressures than CO_2 (Moran *et al*., 2011). Where gas scrubbing processes have suppressed the surface emission of SO_2 prior to eruption, increases in SO_2 emission rate can represent the boiling and subsequent removal of a hydrothermal system to allow previously scrubbed SO_2 emissions to reach the surface (Pritchard *et al*., 2019). Magmatic ascent is often associated with deformation which, in the case of the Augustine eruption, was recorded as an increase in edifice inflation rate (Cervelli *et al*., 2010).

4.5.2) Importance of feature groups

4.5.2.1) Comparison of seismic catalogues

In this study, we derive features from seismic, gas and geodetic data. Figure 4.7 is a 2-dimensional visualisation of the full set of seismic, gas flux and geodetic features (with 3-dimensional representation in Appendix C; C.7). This visualisation is performed using the t-Distributed Stochastic Nearest Neighbour Embedding (t-SNE), which is a probabilistic machine learning algorithm designed to project high-dimensional data—the set of 40 features—into lower dimensional space such as 2- or 3-dimensions (Van der Maaten and Hinton, 2008). This process is performed by calculating a joint probability for points in the high-dimensional space using the conditional probability that a given datapoint would pick another as its neighbour, given a probability distribution described by a Gaussian centred at that point. The low-dimensional joint probability is calculated using the Student t-distribution with one degree of freedom. The final mapping is calculated by minimising the Kullback-Leibler divergence between these two joint probabilities, such that points which are close in high-dimensional space remain close in the two-dimensional representation. We apply the t-SNE to four sets of features: the two sets of seismic and geodetic features used for classification in Section 4.4 (with GPR or CLVF derived daily values for gas), and two sets of features derived from the AVO seismic catalogue (with the same geodetic and GPR or CLVF derived daily values for gas flux).

Comparison of the t-SNE projections for the set of seismic, gas and geodetic features containing seismic features derived from our seismic event catalogue shows that the non-eruptive and eruptive days are clustered in separate areas of the projected space, with little overlap between the two clusters. The transition between non-eruptive and eruptive data is smoother for GPR-based interpolation than CLVF-based interpolation, which reflects the continuous nature of GPR-based interpolation. Novelty detection classification methods rely on the ability to fully characterise the training class of data, so that outliers can be reliably identified. The distinction between clusters in the projection of the Augustine dataset illustrates that there are systematic differences between the eruptive and non-eruptive days of the dataset which can be exploited by a one-class approach. Comparison of the

projections of features from our catalogue and AVO catalogue-derived features indicates that the differences in eruptive and non-eruptive data are less distinct for the AVO catalogue, with no clear separation of data into clusters on this basis. The increased separability of features from our catalogue suggests that the smaller, more emergent events not included in the AVO analyst-reviewed catalogue contain meaningful information which contributes to distinguishing between eruptive and non-eruptive data.

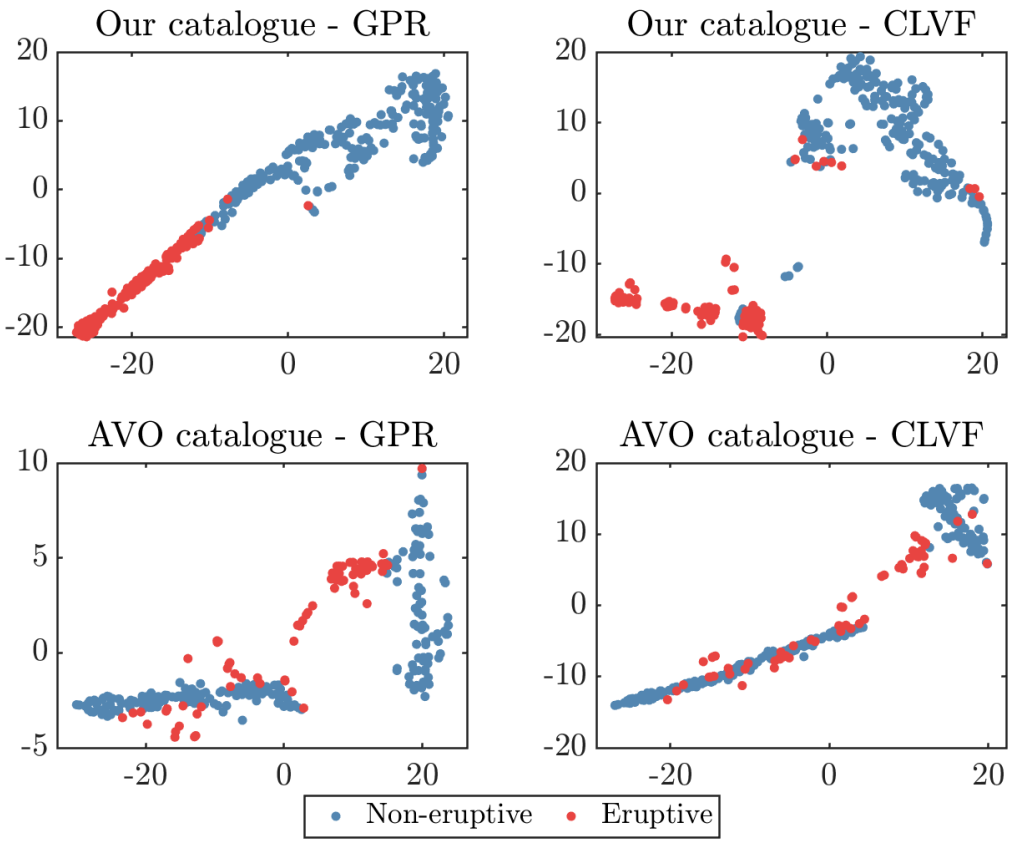


Figure 4.7: Two-dimensional visualisation of the seismic, gas flux and geodetic features derived from Augustine. This visualisation is performed using the t-distributed stochastic nearest neighbour embedding (t-SNE; van der Maaten and Hinton, 2008). Blue points represent non-eruptive days and red points represent eruptive days according to the GVP database. The top two panels, which include seismic features from our seismic catalogue, display a greater separation between non-eruptive and eruptive days than the bottom two panels, which include seismic features from the AVO analyst-reviewed catalogue. A three-dimensional visualisation of these datasets using t-SNE is plotted in Appendix C (C.7). The x and y axes represent non-linearly scaled distances projected into a two-dimensional space and thus have no units, and are not comparable across axes.

4.5.2.2) Comparison of multi-parameter and single-parameter models

For comparison against the full sets of features used for models in Section 4.4, we trained and tested three additional novelty detection models using features derived from only the seismic, gas or geodetic data with the same training dates (Appendix C; C.8). The models formulated using seismic data and geodetic data alone have a similar trend to the model using the full set of features in the precursory months, with novelty scores from both models exceeding the novelty threshold in November 2005. However, both seismic and geodetic models fail to consistently classify days from the main eruptive sequence above the novelty threshold. In the case of the seismic data, we propose that this model failure arises due to outages in summit seismic stations during the explosive phase of the eruption, which means that smaller events are underrepresented in our catalogue during the later stages of the eruption. For the geodetic data, the inflation signal was concentrated in the precursory phase and, once the main phase of the eruption commenced, the daily displacement decreased until similar to pre-inflation values. Models formulated using gas features with a GPR interpolation classified behaviour as eruptive as soon as model training ended, whereas models formulated using gas with a CLVF interpolation fail to identify an appropriate novelty threshold as the interpolated SO_2 emission rates (hence novelty scores) remained very low until significant emissions were measured on 20th December 2005. The GPR failure likely results from the smoothly varying nature of the interpolated time series where there is no stable normal behaviour to characterise during model training, whereas the CLVF failure is due to the low surface emission of SO_2 during the precursory phase.

The results of this single data modelling indicate that, in the case of the 2006 Augustine eruption, use of multiple types of data for classification is able to discern between non-eruptive and eruptive data better than single data-streams alone for novelty detection models. When all of the features are included in the model, the precursory increasing trend in novelty score remains similar to the individual seismic and geodetic models and eruptive days are better identified due to the inclusion of the SO_2 emission rate features. Moreover, the single data analysis indicates that the inclusion of multiple datasets can compensate for data quality issues due to instrument failures during a large

eruption. Applying multi-class models to single data-streams was not attempted due to the issues identified above pertaining to isolating suitable eruptive and non-eruptive training periods for each of the data-streams separately.

4.5.3) Applications of the method

4.5.3.1) Forecasting eruptions

Both one-class and multi-class methods, explored in this paper, have potential utility in forecasting. Poland and Anderson (2020) note that machine learning approaches represent an as-yet underutilised resource in eruption forecasting, which is in part due to limited quantities of available training data. The choice to apply one type of machine learning model over another involves consideration of the past volcanic activity and/or monitoring infrastructure available at a given volcanic system, as these factors affect the data available for training models. Here, we discuss possible applications of novelty detection versus multi-class approaches to the detection of eruptive transitions in different volcanic scenarios.

Novelty detection approaches – those in which models are trained only on non-eruptive data – would be applicable to monitored volcanoes which have been in repose for some time (e.g., 10s years), as these systems would not have eruptive data available for training but a large amount of data from repose to characterise the non-eruptive class. This class of volcanoes would include many of those that have experienced major eruptions in recent decades (Pyle and Barclay, 2020). Other potential applications include volcanoes which have displayed several types of activity in the past or have highly variable manifestations of eruptive activity. Here, a novelty detection approach would not assume a consistent mode of eruption. Additionally, volcanoes where the monitoring infrastructure has significantly changed since the previous eruption are a good candidate for novelty detection, as the eruptive data from the previous eruption would not necessarily characterise the eruptive data on the updated monitoring network.

Multi-class approaches – those in which models are trained on both non-eruptive and eruptive data – could be more useful at volcanoes which display consistent signals associated with eruption. Here, eruptive training data available from past eruptions would well-characterise the eruptive behaviour of the system in order to identify future eruptive signals. For example, Piton de la Fournaise volcano has a consistent tremor signal that is correlated with eruption (Battaglia and Aki, 2003) and most eruptive episodes are preceded by precursory VT swarms (De Barros *et al*., 2013), which has made it a good candidate for previous multi-class studies (Ren *et al*., 2020). Volcanoes with short repose time between eruptions could be better-suited to a multi-class approach, as there would be insufficient time to characterise the non-eruptive class with a one-class approach.

For the purposes of forecasting eruptions, consideration of which methods are used to interpolate values for missing data points are necessary. CLVF-based interpolation involves no future information to fill in missing values, therefore is well-suited to a forecasting context. GPR-based interpolation uses the full set of data points to construct the model from which the daily values for the training period are taken and therefore is a purely retrospective analysis in this study. However, Colopy *et al*. (2018) demonstrate that Gaussian Processes can be successfully applied in healthcare for forecasting patient vital signs beyond the model training periods. Future work could take Gaussian Processes into the domain of real-time volcanic forecasting.

4.5.3.2) Elucidating timings of transition in volcanic state

Future applications of classification could extend the two-class approach here to investigate intra-eruptive transitions in activity. Prolonged eruptions may undergo intra-eruptive transitions; for example, dome-forming eruptions often switch between effusive and explosive behaviour (e.g., Edmonds and Herd, 2007; Preece *et al.*, 2016). Characterising and understanding the nature and timing of changes in eruptive intensity and style, and anticipating future transitions in behaviour, is critical for managing hazards during eruptive crises (e.g., Fearnley and Beaven, 2018; Barclay *et al.*, 2019). For systems in which the transitions between activity are sufficiently long to generate enough

training data, one-class or multi-class methods trained during eruption could help to distinguish intra-eruptive transitions.

Consideration of the methods for interpolating gaps in irregularly sampled time series is necessary. As discussed in Section 4.4, we observe different end-dates for the 2006 eruption depending on the SO_2 interpolation methods used. We find that models formulated based on GPR interpolation of the SO_2 time series yield an earlier estimate of eruption end-date than models formulated based on CLVF-based interpolation. This observation is explained due to the 'carry forward' nature of CLVF: at the end of the eruption where the SO_2 emission rate is consistently decreasing between observations, CLVF represents an overestimate of the daily emission rate and therefore days are classified as eruptive for a longer period. Consideration of interpolation methods becomes more important as the sampling frequency of the time series decreases, as more points need to be estimated for sparser time series.

Application of the approach presented in this paper is not limited to the data streams used here. Additional ground-based data streams from which features could be derived include infrasound or thermal measurements. Satellite-derived observations of ground displacement (e.g., InSAR) are being made in increasing frequency, and could be incorporated into a machine learning framework. Indeed, there is a growing body of literature examining the automatic detection of unrest in InSAR time series (Anantrasirichai *et al*., 2018; Biggs *et al*., 2019; Gaddes *et al*., 2019; Valade *et al*., 2019).

4.6) Conclusions

Identifying transitions in multi-parameter time series is a challenge in volcanology. We apply novelty detection (one-class) and multi-class machine learning methods to features derived from seismic, gas and geodetic data from the 2006 eruption of Augustine Volcano, Alaska. We present the first study applying a one-class or novelty detection approach to classify volcanic state, which does not require training examples characterising eruptive behaviour. We suggest a framework for processing multi-parameter data to produce daily representative features, including two distinct approaches towards interpolating missing values in SO_2 emission rate time series. These approaches include Carry Last Value Forward (CLVF) interpolation in which the daily value remains constant since the previous measurement of emission rate, and Gaussian Process Regression (GPR) which is a probabilistic approach using the full available data to estimate intervening values with uncertainty bounds. We compare the results of models formulated using seismic, geodetic and both CLVF and GPR estimates of daily gas emission rates. Both novelty detection and multi-class random forest models identify a key change-point from non-eruptive to eruptive state in mid-November. This change-point is in agreement with previous conceptual models at Augustine which suggest the initiation of dike intrusion from the shallow magmatic storage at this time. We find that novelty detection models formulated using multiple types of data are able to distinguish between non-eruptive and eruptive data better than models formulated on single data-streams alone. We suggest possible applications of one-class novelty detection and multi-class modelling and discuss considerations in applying these approaches in a monitoring or retrospective context. Novelty detection methods show promising application for monitoring volcanoes where previous eruptions have not been observed, or where the monitoring infrastructure has significantly changed since previous eruptions.

Chapter 5 - A deep active learning approach to the automatic classification of volcano-seismic events

5.1) Abstract

Volcano-seismic event classification represents a fundamental component of volcanic monitoring. Recent advances in techniques for the automatic classification of volcano-seismic events using supervised deep learning models achieve high accuracy. However, these deep learning models require a large, labelled training dataset to successfully train a generalisable model. We develop an approach to volcano-seismic event classification making use of active learning, where a machine learning model actively selects the training data which it learns from. We apply a diversity-based active learning approach, which works by selecting new training points which are most dissimilar from points already in the model according to a distance-based calculation applied to the model features. We combine the active learning with an existing volcano-seismic event classifier and apply the model to data from two volcanoes: Nevado del Ruiz, Colombia and Llaima, Chile. We find that models with data selected using an active learning approach achieve better testing accuracy and AUC (Area Under the Receiver Operating Characteristic Curve) than models with data selected using random sampling. Additionally, active learning decreases the labelling burden for the Nevado del Ruiz dataset but offers no increase in performance for the Llaima dataset. We visualise the features from the two datasets and suggest that active learning can reduce the quantity of labelled data required for noisy and less separable datasets. This study represents the first evaluation of an active learning approach in volcano-seismology.

5.2) Introduction

Understanding the evolution of seismic activity prior to and during eruptions is critical for understanding transitions in volcanic state (e.g., Power *et al*., 1994). Variations in the type of volcano seismicity can reflect the underlying source processes associated with magmatic or hydrothermal transport, and stress changes (Chouet and Matoza, 2013; McNutt and Roman 2015). Timely characterisation of the type of volcano seismicity (e.g., event types outlined in Table 5.1) is therefore imperative for assessing evolving volcanic hazards, a task typically performed by analysts in volcano observatories. Hence, the automatic detection and classification of volcano-seismic event types would be valuable for reducing the workload of analysts during periods of heightened volcanic activity (Scarpetta *et al*., 2005).

Machine learning is a powerful tool in the processing and study of volcano-seismic events and there have been advances in their automatic detection and classification using supervised machine learning approaches (e.g., Langer *et al*., 2006; Apolloni, 2009; Malfante *et al*., 2018). Recent contributions to the literature have explored the use of convolutional neural networks (CNNs) for the purposes of volcano-seismic event classification (e.g., Canario *et al*., 2020a; Lara *et al*., 2021). Rather than acting on a set of user-designed features, CNNs operate by learning a series of filters to extract useful features from the data and output a classification. Neural networks are formulated from an input layer, a number of hidden layers which transform the data, and an output layer from which predictions are made. The resulting set of features is proposed to be more robust of intra-class variability than a set of manually chosen or hand-crafted features (Chan *et al*., 2015).

Supervised machine learning approaches require labelled data for training and larger amounts of data are required for deep learning models to acquire an accurate representation of the input data (Krizhevsky *et al.,* 2012; Hinton *et al.*, 2012). Hence, the labelling time required to curate a data volume large enough to establish an automatic classification system can be significant depending on the difficulty of the labelling task. Active learning is the process by which a machine learning model

chooses the optimum data for training to achieve higher accuracy using a smaller amount of training data (Settles, 2009). Labels for these chosen training points are then assigned by an expert. Active learning is useful for contexts in which obtaining labels for unlabelled samples is difficult or time-consuming. For this reason, active learning has previously been applied in healthcare engineering with the motivation to develop tools for automated classification of health-related data, whilst minimising the time burden placed upon healthcare professionals to label large quantities of examples for training (e.g., Smailagic *et al*., 2018; Kiyasseh *et al*., 2020). This motivation is shared in the volcano monitoring setting, where the manual classification of events requires expert knowledge. During periods of heightened volcanic activity, the time demands upon observatory staff will increase, thus minimising the time spent on manual classification for establishing an automatic classifier is advantageous (Scarpetta *et al*., 2005).

We use an approach to active learning which has been primarily applied to choose an optimum training dataset from a large pool of unlabelled data. This problem is important for volcano-seismology, where campaign-based monitoring may mean that a glut of data is received at once. For volcanoes which have had a long period of repose prior to eruption, the seismic network may be sparse or non-existent (Moran *et al*., 2008). During times of volcanic crisis, new seismic stations may be installed within a seismic network, thus there may be no backlog of previous data to train an automatic classifier upon. Alternatively, digitisation of legacy data may also generate a large volume of unlabelled data for which there is a time-burden for digitising and verifying the associated labels (Thompson *et al*., 2020).

Previous machine learning techniques in volcano seismology have used transfer learning to make the most of limited labelled training data. Transfer learning is the process of using knowledge (e.g., model structure or weights) from a previous classification task to obtain a model for a new classification task in which limited or no training data is available (Pan and Yang, 2009). Transfer learning schemes are generally more successful where the size of the training dataset is small in comparison to the previous classification task. Lapins *et al*. (2021) apply transfer learning techniques

for detecting phase arrivals at Nabro Volcano, Eritrea and show that a model with feature extraction layers based on an established seismic phase picking model can be used to label seismic phases where only a limited training set is available. Bueno *et al.* (2020) incorporate transfer learning and uncertainty-based active learning to create a program for volcano-seismic analysis and classification via a model pre-trained on events from both Mount St Helens and Bezymianny volcanoes. However, no analysis of the potential for active learning to reduce the size of the training set in the field of volcano-seismic event classification has yet been done.

In this paper, we present an active learning approach (after Sener and Savarese, 2017) to select optimal training data. We then combine this active learning approach with an existing volcano-seismic event classification network SeismicNet (Canario *et al*., 2020a). We apply this method to two datasets of detected volcano-seismic events which are fully labelled with analyst-reviewed classifications (these event types are outlined in Table 5.1):

(i) 5614 detected seismic events from the 2012 unrest and eruption of Nevado del Ruiz Volcano, Colombia. These events are labelled as VT, LP and Hybrid events by analysts from the Servico Geológico Colombiano (SGC).

(ii) 3592 detected seismic events from Llaima Volcano, Chile over the 2010 - 2016 non-eruptive period (Canario *et al*., 2020b). These events are labelled as VT, LP, Tremor and Tectonic events by analysts from the Observatorio Volcanólogico de los Andes del Sur (OVDAS). Canario *et al*., (2020a) used this dataset to train and test neural network architectures for automatic event classification.

We evaluate the success of the active learning approach in selecting the training dataset against an equivalent model trained with randomly selected training points. We withhold the labels from the training data until points are selected, where we add the points and associated labels to the training dataset. After training the model, we test the model performance on unseen data.

5.3) Methods

5.3.1) Data

For this experiment we use two seismic datasets which are annotated with analyst-reviewed labels. These labels are withheld from the model during the active learning experiment (outlined in Section 5.3.3) until a point is selected to be included in the training dataset (through either active or random sampling), and the process of labelling the waveforms is simulated by providing the true labels to the model when they are selected.

Table 5.1 summarises the frequency characteristics and potential source processes of the event types which are contained in the two datasets used for this study and representative examples of these event types are plotted in Figure 5.1. Volcano-tectonic (VT) events are characterised by impulsive onsets and dominant frequencies above 5 Hz (Lahr *et al.,* 1994). They are thought to be brittle failure events that can be triggered by mechanisms such as stress change from magma intrusion (Roman and Cashman, 2006), or an increase in pore fluid pressure due to volatile transport (Huppert and Sparks, 2016). Long Period (LP) events are characterised by an emergent onset followed by a set of decaying harmonic oscillations, and are typically concentrated within the frequency range of 0.5 – 5 Hz (Lahr *et al.,* 1994). The source mechanism of LP events can be attributed to resonance of a fluid-filled crack or conduit (Chouet 1996; Neuberg 2000). These signals are also termed Low Frequency (LF) events and can reflect a broader set of potential source mechanisms including slow-rupture earthquakes in unconsolidated edifice (Bean *et al.,* 2014), magma-hydrothermal interaction (Jolly *et al.,* 2017) and magmatic degassing (Rodgers *et al.,* 2015a). Throughout this manuscript we use the term LP as this is the term assigned within the labelling schemes of the datasets used here. Hybrid events (HB) contain a mixture of both VT and LP signals, typically with a high frequency impulsive onset similar to a VT event and a decay which is prolonged similar to an LP event (Lahr et al., 1994). Potential source mechanisms for hybrid events include a brittle-failure triggering event which induces conduit resonance (e.g., Neuberg *et al.,* 2006). Harrington and Brodsky (2007) suggest that hybrids can be attributed to brittle failure events with low rupture velocity which have undergone complex path

effects. Rodgers *et al.,* (2016) identify families of repeating events that transition from hybrids into LPs, suggesting similar source processes with subtle changes over time in the triggering process. Tremor events (TR) are a continuous low-frequency signal, usually associated with fluid migration. Suggested source mechanisms for tremor include sustained triggering of LP events in magma conduits (Powell and Neuberg 2003), restricted or choked flow of fluids in the sub-surface (Chouet 1996; Hellweg 2000; Lesage *et al.,* 2006) or magma degassing or fissure eruptions at the surface (Nadeau *et al.,* 2011; Battaglia and Aki 2003), In the Llaima dataset, the tectonic event type (TC) includes events with a non-volcanic origin, such as regional or local earthquakes (Canario *et al.,* 2020b). Source mechanisms of tectonic events are indistinguishable from VTs, i.e., brittle shear failure due to stress changes, but these events can be distinguished from VTs as being located outside of the volcanic network. Cataloguing tectonic events is important for volcano monitoring as interactions between volcanoes and regional tectonic activity have been documented in many places (Marzocchi *et al.,* 1993; Linde and Sacks, 1998; LaFemina *et al.,* 2004), and distal VTs may in some cases be considered tectonic events and excluded from volcano-seismic catalogues (Roman *et al.,* 2008).

For this study, we trial the methods on two datasets from Nevado del Ruiz, Colombia and Llaima Volcano, Chile. We choose these datasets for trialling the active learning approach for two reasons. Firstly, the datasets each contain a different number of classes: the Llaima dataset contains 4 event types (VT, LP, tremor and tectonic events) and the Nevado del Ruiz dataset contains 3 event types (VT, LP and hybrid events). Hence, we can test whether the active learning approach is more beneficial where a greater number of event types are being classified. Secondly, the datasets are from different types of volcanic activity: the Llaima dataset is from 2010 – 2016 which is a non-eruptive period, and the Nevado del Ruiz dataset is from 2012, which covers precursory unrest and eruption from 22nd February 2012 onwards. This distinction is important for trialling an active learning approach on seismic data from both non-eruptive activity and heightened unrest as the spectral contents of seismic events can vary during unrest and eruption (Rodgers *et al*., 2016). For example, Buurman and West (2010) observe an increase in events dominated by low frequency energy prior to

explosions during the 2006 eruption of Augustine Volcano. This distinction allows us to test whether the active learning approach performs better on events with greater variability.

5.3.1.1) Nevado del Ruiz dataset

We use data from the 2012 eruptive period of Nevado del Ruiz Volcano, Colombia, an ice-covered stratovolcano located within the Andes (Londoño, 2016). The 2012 eruption of Nevado del Ruiz represented the first emission of ash since the 1985 eruption, which caused over 25,000 fatalities when lahars inundated surrounding communities, including the city of Armero (Lowe *et al.,* 1986; Naranjo *et al.,* 1986). The data are from the vertical component of the seismic station TOLZ. The dataset includes 5614 waveforms of three types: 2363 VT events, 2867 LP events and 384 hybrid events. These events have been classified by analysts at the Servicio Geológico Colombiano (SGC). The waveforms have been filtered using a band pass filter to include frequencies between 0.5 and 20 Hz. Each input waveform is 64 s long, and shorter waveforms are padded with 0 values to reach this length. Seismic data from between 2005 – 2006 at Nevado del Ruiz have previously been used to train and test an automatic classifier using Hidden Markov Models, which achieved a classification accuracy of approximately 88% (Cárdenas-Peña *et al.,* 2013). This dataset was significantly smaller than the dataset used here and involved an extensive data pre-processing stage to extract frequency-based features and select the features which would provide the greatest information over time. This approach contrasts with our method, in which very little pre-processing is required and no prior knowledge about waveform properties from the whole study period is used to optimise the approach. Long-term data from Nevado del Ruiz from 2007 – 2014 were automatically classified using a random forest classifier (Rodgers *et al*., 2015b). This study used the full spectrogram as input vector and using ~ 7% of analyst-labelled events for training achieved a classification accuracy of approximately 77%. Hand-crafted features derived from this catalogue of detected events have previously been used to classify eruptive and non-eruptive activity during the 2012 eruption of Nevado del Ruiz (Manley *et al*., 2020).

5.3.1.2) Llaima dataset

We use the dataset presented by Canario *et al*. (2020b) which encompasses 4 event types (VT, LP, Tremor and Tectonic) from 2010 – 2016 at Llaima Volcano, Chile. The data are from the vertical component of the seismic station LAV, with labels from Observatorio Volcanológico de los Andes del Sur (OVDAS). The total dataset size is 3592 waveforms, which comprises 304 VT events, 1310 LP events, 490 tremor events and 1488 tectonic events (e.g., regional events). Therefore, the Llaima dataset is approximately 36% smaller than the Nevado del Ruiz dataset. As described in Canario *et al*. (2020b) the waveforms have been filtered using a band-pass filter to include frequencies between 1 and 10 Hz and subsequently normalised by the maximum value. Each input waveform is 64 s long, and shorter waveforms are padded with 0 values to reach this length. This dataset was previously used for studying the performance of a range of neural network architectures by Canario *et al*. (2020a) and achieved validation accuracies of over 90% using the SeismicNet, though the models were not tested on unseen data in their study.

5.3.2) Machine learning classification

We combine the active learning selection of training data with an existing volcano-seismic event classifier. We use an existing CNN-based volcano-seismic event classifier, known as SeismicNet, proposed by Canario *et al*. (2020a), illustrated in Figure 5.2a. The SeismicNet classifier is based on a similar architecture designed for classifying waveforms from audio, known as SoundNet (Aytar *et al.,* 2016). The architecture takes the raw waveforms as input rather than derived information such as a spectrogram representation or wavelet transform. We choose this classifier as it achieved comparable results to other CNN-based approaches (e.g., Curilem *et al*., 2018; Lara *et al*., 2021) but required significantly fewer pre-processing steps, which is concordant with the aim of this study to develop a method for the prompt labelling of a pool of unlabelled events.

This classifier is a Convolutional Neural Network (CNN), a neural network which contains a number of layers which execute convolutions on the input data and downsample (or downsize) the signal.

These layers act as the "feature extraction" process for the CNN. Convolutional layers apply convolution over the data with a given size (known as the filter size). These convolutions are used in combination with pooling layers, which reduce the size of the data, and dropout layers which are randomly set to 0 with a given probability, and are used to increase the generalisability of a model and prevent overfitting to the training data (Srivastava et al., 2014). Finally, fully connected layers are used to compute the final classification output of the model. We use the categorical cross entropy loss function and the stochastic gradient descent optimiser to fit the model. The models were implemented in Keras v2.4.0 with a Tensorflow backend. The network architecture is outlined in Figure 5.2a, with the parameters shown for each layer. Before running the experiments, we tested the model architecture, varying parameters such as dropout and hyperparameters such as the model batch size and number of epochs for training, and chose the parameters which achieved the greatest validation accuracy in training. We choose a model batch size of 16 and number of training epochs of 150 for the Nevado del Ruiz dataset and 200 for the Llaima dataset.

5.3.3) Active learning

Diversity-based active learning works on the principle that a model will achieve better performance if more variation across the training set is sampled by the model during the training period. Because the extent of the variability in the training set has been sampled during the model training period, in theory the model will perform well on unseen testing data under the assumption that the training data captures the full diversity of the dataset. In contrast to diversity-based active learning, uncertainty-based active learning schemes involve choosing examples to be labelled based on the uncertainty of the model in the given classification, thereby prioritising examples which lie close to the boundary between classes (Jain and Kapoor, 2009). However, for deep learning classifiers, the probability output of the model is not representative of the model confidence (Gal and Ghahramani, 2016) and therefore the classifier must be adapted to a Bayesian framework to adequately quantify the uncertainty of the classification. We therefore select a diversity-based method of active learning such that the method can be combined with previous classifiers.

We apply a core-set selection approach to active learning after Sener and Savarese (2017), known as KCenterGreedy, illustrated in Figure 5.2b. The KCenterGreedy method works by selecting new data points which are the most dissimilar from any data point within the current training set. Deep learning models are formulated on batches of data, making labelling single data points logistically challenging, and meaning that obtaining a single label is not likely to significantly affect the model performance. The KCenterGreedy algorithm is implicitly designed to work on selecting batches of data. We choose this algorithm as it is extensible to any CNN-based classifier for volcano-seismic events.

Figure 5.3 illustrates the process of the active learning experiment. The model is trained with an initial randomly-chosen batch of labelled waveforms. The full set of unlabelled data is then passed through the feature extraction layers of the network to generate the features for the algorithm. Within the feature space, the minimum pairwise Euclidean distance (a measure of feature similarity) between all datapoints to a point within the training set is then calculated. This calculation results in 0 distance for points already labelled within the training set. A new point which is the maximum distance away from any labelled point is selected and added to the set of labelled points. The minimum distances are then re-calculated with this point included (as unlabelled points located near the newly-chosen point may now be closer to a labelled point). This process is repeated until the user-selected batch size is reached and all the newly chosen points are labelled. The user-selected batch size is the quantity of data selected in one active learning cycle as illustrated in Figure 5.3. Next, the model is retrained on all the data chosen so far and the AUC (Area Under the Curve, a measure of the ability of a classifier to distinguish between classes – see below) is re-evaluated. Once all the training data have been selected, the model is then tested on unseen data.

We split the full dataset into training, validation and testing datasets with a split of 2/3, 1/6 and 1/6 of the dataset respectively. For the models in this paper, the initial randomly-selected batch of waveforms is chosen to be approximately 50% of the total training size. We choose the active learning user-selected batch size as 216 datapoints. Experienced volcano-seismic analysts can classify

in the region of hundreds of volcano-seismic events in a day (e.g., Thompson *et al*., 2020). Therefore, we pick a figure for selecting batches in the lower region of this estimate as a reasonable quantity to be labelled at once before the models can be re-trained with the new data included.

We compare the active learning experiment to an experiment where the next batch is randomly-selected. The performance metric we use to compare the performance of the models is the Area Under the Receiver Operating Characteristic Curve (AUC-ROC, referred to as AUC). The ROC curve is a plot of the False Positive Rate (FPR) vs the True Positive Rate (TPR) of a model at various threshold values. The AUC is computed by integrating under this curve, whereas model accuracy is computed by setting a FPR threshold and computing the percentage of examples which were correctly classified. A higher AUC corresponds to a higher TPR at lower FPR and therefore better model performance. For comparing models, the AUC is a preferable metric to other metrics such as the training accuracy as the AUC is independent of a threshold which, in the case of accuracy, can vary between models.

Table 5.1: Description of the volcano-seismic event types contained in the datasets used for this study.

Event type	Typical frequency range	Selected potential source mechanisms	Dataset with these events included	References for potential source mechanisms
VT	5 - 15 Hz	Brittle rock failure	Nevado del Ruiz, Llaima	Roman and Cashman 2006 Roman and Power, 2011 Lahr *et al*., 1994
LP	0.5 – 5 Hz	Resonance of fluid-filled crack or conduit; magma-hydrothermal interactions; magmatic degassing; brittle melt fracture	Nevado del Ruiz, Llaima	Chouet, 1996 Chouet *et al.,* 1994 Lahr *et al.,* 1994 Neuberg *et al*., 2000 Thomas and Neuberg 2012 Bean *et al.,* 2014

Hybrid	High-frequency onset (> 5 Hz) with low-frequency decay (0.5 – 5 Hz)	Brittle triggering event followed by conduit resonance	Nevado del Ruiz	Lahr *et al.,* 1994 Neuberg *et al*., 2000 Neuberg *et al.*, 2006 Harrington and Brodsky, 2007 Rodgers *et al.,* 2016
Tremor	1 – 12 Hz	Extended magma movement, explosions	Llaima	Julian, 1994 Baptie *et al.,* 2002 Neuberg *et al.* 2000 Hellweg, 2000 Jellinek and Bercovici, 2011
Tectonic	Over 5 Hz	Non-volcanic brittle failure events e.g., local or regional seismic events	Llaima	Marzocchi *et al.,* 1993 LaFemina *et al.,* 2004 Canario *et al*., 2020b

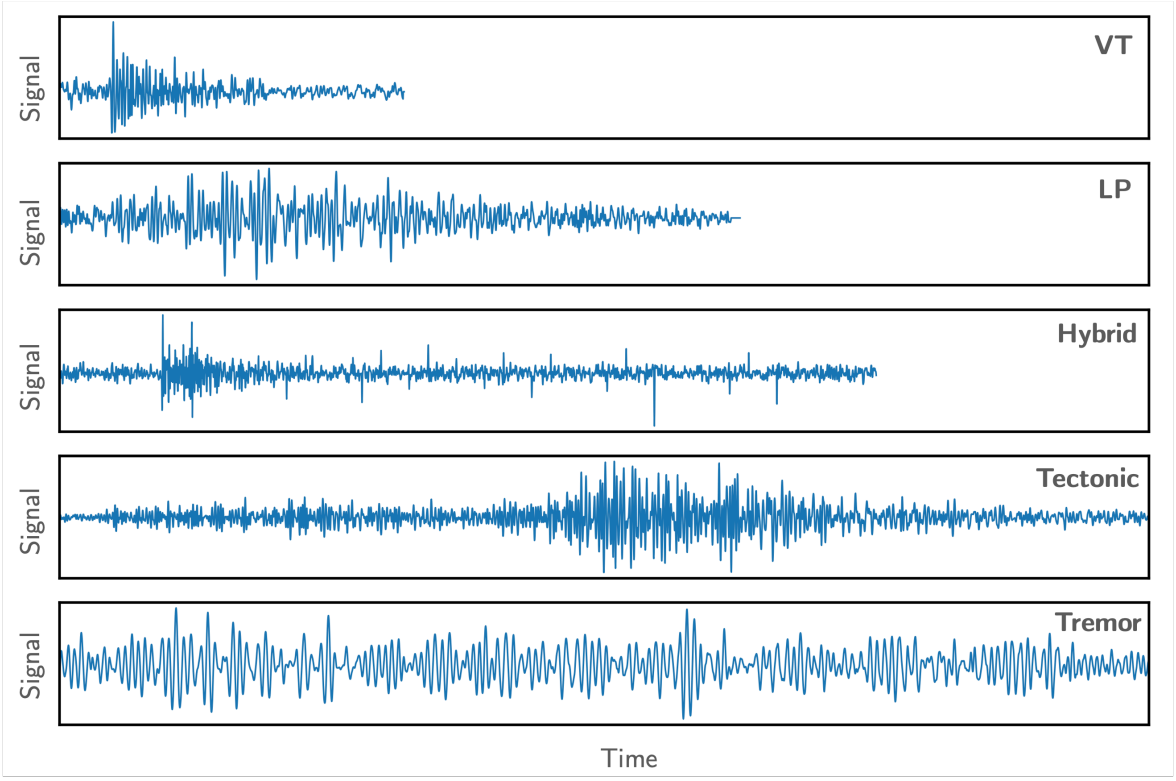


Figure 5.1: Representative examples of waveforms of each class from the datasets used in this study, where VT represents a volcano-tectonic event and LP represents a long-period event. The VT, LP, tectonic and tremor examples are from the Llaima dataset, and the hybrid example is from the Nevado del Ruiz dataset. Each waveform has an amplitude normalised to between 0 and 1 and the range of the x-axis in each panel is 64s (which corresponds to the length of the input waveform to the model).

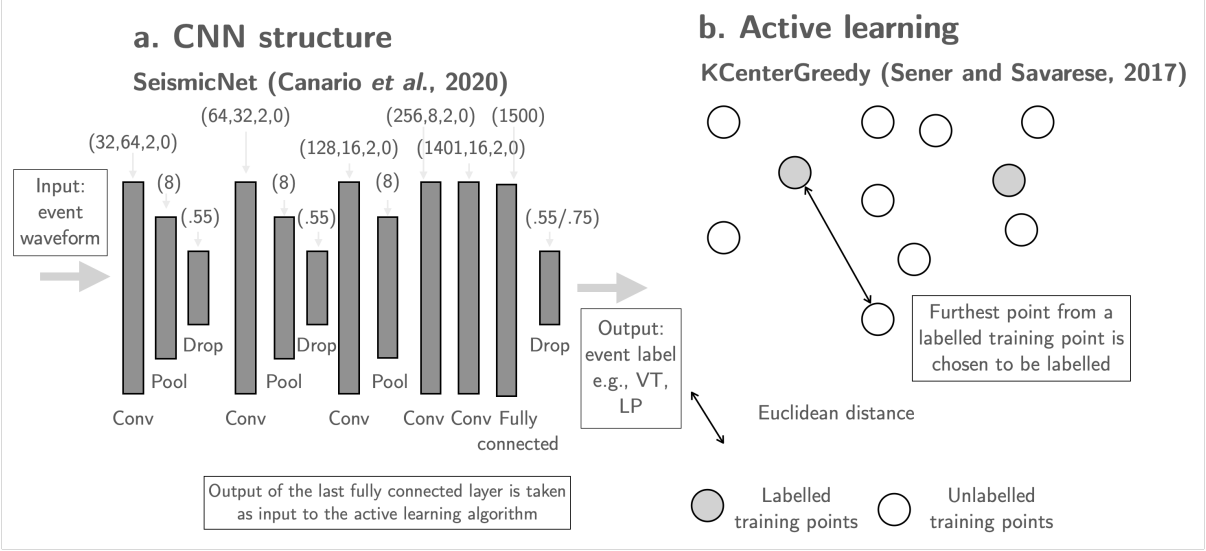


Figure 5.2: (a) Illustration of the SeismicNet neural network architecture (Canario *et al.*, 2020a). For the convolutional layers 'Conv', figures in brackets refer the number of filters, kernel size, strides and padding respectively. For the pooling layers 'Pool', the figure in brackets is the pool size. For the dropout layers 'Drop', the figure in brackets is the dropout rate, where the two rates in the final dropout layers represents the rate for the Nevado del Ruiz dataset (0.55) and Llaima dataset (0.75), which is the only model architecture parameter which was changed between the two datasets. For the fully connected layer, the figure in brackets represents the size of the final dropout layer (1500). (b) Conceptual illustration of the active learning and deep learning components of the model. Conceptual illustration of the KCenterGreedy algorithm (Sener and Savarese, 2017), where grey points represent points within the training dataset and white points represent unlabelled points. Arrows represent the minimum pair-wise Euclidean distance in the feature space (n-space with dimensions equal to the number of neurons in the final fully-connected layer i.e., 1500 in our study) of the furthest point from any currently within the training dataset

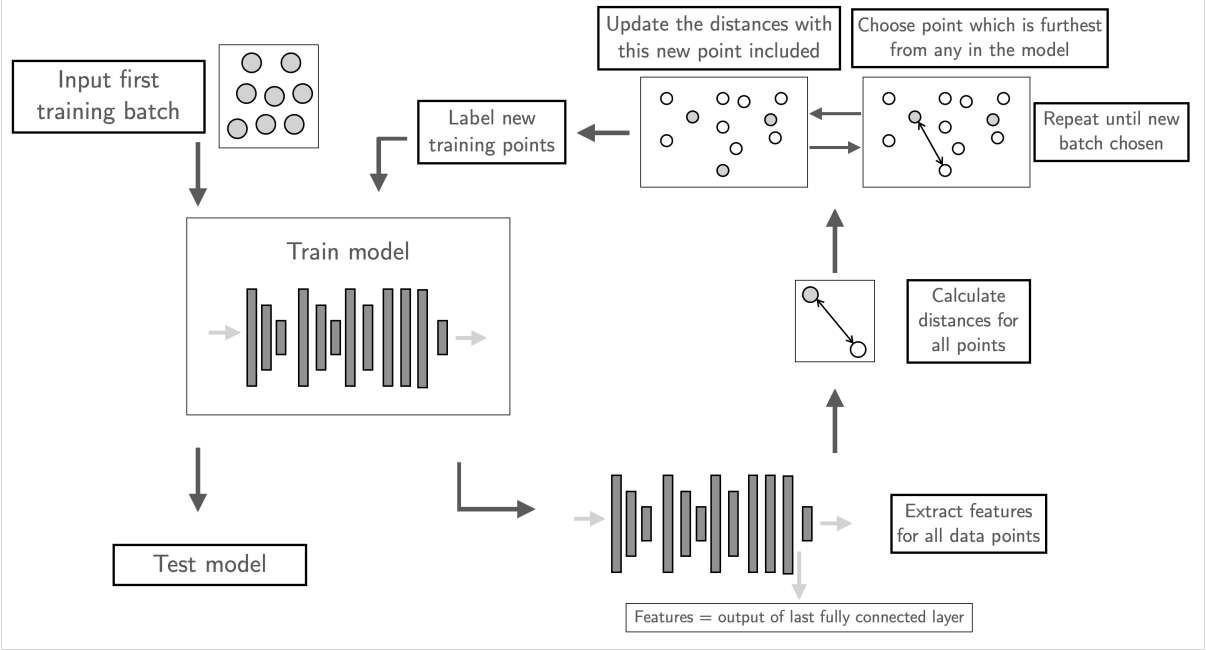


Figure 5.3: Illustration of the active learning experiment presented here. Grey points represent points within the training dataset and white points represent unlabelled points. Double-headed arrows represent the pair-wise Euclidean distance of points within the dataset (see Figure 5.1b).

5.4) Results

To assess the results of the active learning frameworks described in Section 5.3, we compare the model performance via the AUC against the proportion of the dataset which is being used for model training, where greater AUC corresponds to better model performance. As illustrated in Figure 5.3, the model is trained with a subset of labelled data points and new training points are added to the labelled set either through active or random selection. Hence, the proportion of labelled points increases over the course of the experiment. We compare the active learning selection of training data against a random selection of training data. For a successful active learning approach, the models will achieve higher performance at a lower proportion of training data than the randomly-selected case.

5.4.1) Nevado del Ruiz dataset

Figure 5.4 illustrates the active learning and randomly-selected learning curves for models trained on the Nevado del Ruiz dataset. Table 5.2 indicates the testing AUC and accuracy of the model. The active learning model achieves greater training AUC at a smaller proportion of the dataset used for training with an AUC comparable to that achieved at the final training step obtained with just approximately 60% of the training datapoints labelled. The randomly-sampled model achieves greater AUC than the actively-sampled model once approximately 85% of the training data are labelled. However, the active-sampling testing AUC of 0.8848 is greater than the random-sampling testing AUC of 0.8626 when tested on the same set of unseen datapoints. The final testing accuracy of the active learning model, at 70.73%, is higher than the randomly-sampled model at 68.27%. These testing accuracies are comparable to classification accuracies achieved on noisy seismic data (e.g., Cortés *et al*., 2021) but lower than those achieved for the less noisy Llaima dataset.

Figure 5.4 contains the confusion matrix of both models. For the actively-sampled model, the incorrect classifications are primarily the erroneous classification of VT events as LP events. For the randomly-sampled model the opposite occurs, where LP events are primarily misclassified as VT events. Both models classify the majority of hybrid events as LP events (though hybrids represent a

small proportion of the dataset), with the random classifier failing to identify any hybrid events. Previous automatic classifiers applied to Nevado del Ruiz data had a high misclassification rate of hybrid events, and it was suggested that discriminating between hybrid and LP events is particularly difficult from observations made at a single recording station (Cárdenas-Peña *et al*., 2013). Additionally, the a-priori classification of hybrid events may be subject to human analyst bias (Langer *et al*., 2006).

5.4.2) Llaima dataset

Figure 5.5 illustrates the active learning and randomly-selected learning curves for models trained on the Llaima dataset. Table 5.2 indicates the testing results of the models. The testing AUC of the active learning model, at 0.9898, is better than the testing AUC of the randomly-sampled model, at 0.9332. The testing accuracy is similar to the training and validation accuracy reported by Canario *et al*. (2020a), at 93.32% for the active learning model and 94.82% for the randomly-sampled model. The active learning approach achieves similar training AUC until approximately 60% of the dataset is labelled, after which the randomly-selected model achieves greater AUC. The models approach a similar training AUC towards the end of model training.

Figure 5.5 includes the confusion matrix of the active learning compared to the randomly-sampled case. The correct classifications are similar between both models. For the active learning model, misclassifications are primarily tectonic events misclassified as VT events, and tremor events misclassified as tectonic events. For the randomly-sampled case, misclassifications of LP or tectonic events as VT are the main source of wrong classifications.

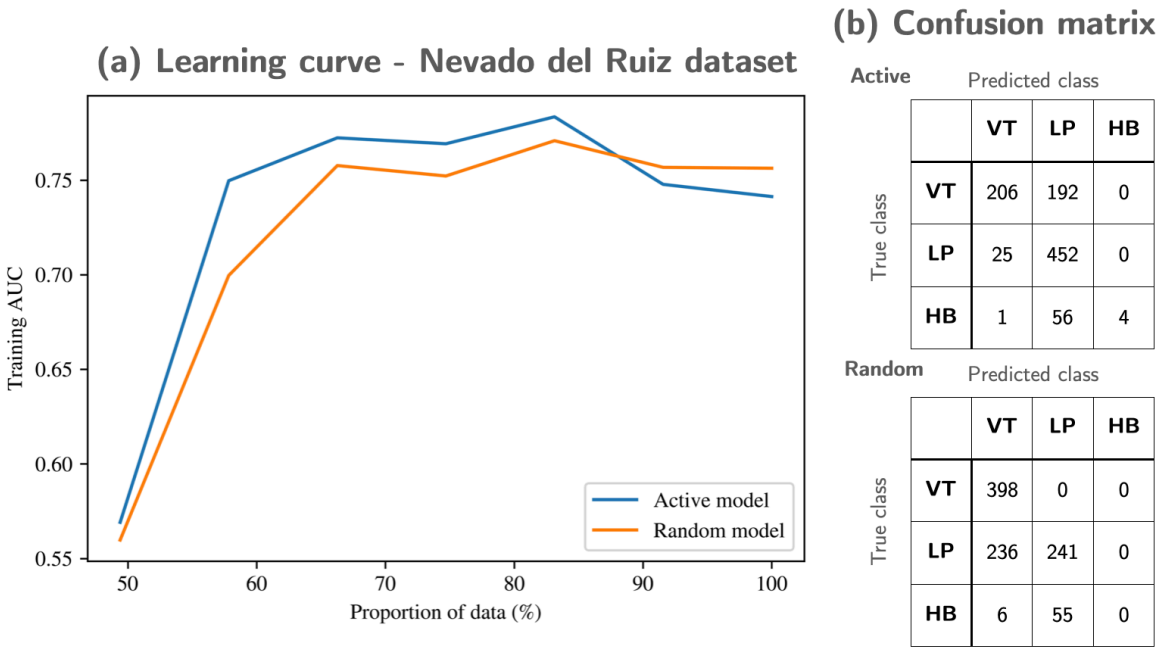


Figure 5.4: (a) Learning curves for the active model (blue) and random model (orange) for the Nevado del Ruiz dataset. The proportion of data is the amount of labelled data which the model is trained on, and the training AUC is the final AUC for the model training. (b) Confusion matrices for the active model (top) and the random model (bottom). Numbers indicate the number of events which were classified in each category, such that true positive or correct classifications are on the diagonal of the matrix and misclassifications appear in the off-diagonal elements.

Table 5.1: Testing performance (AUC and accuracy) for the models presented in this study.

Dataset	Model type	Testing AUC	Testing accuracy
Nevado del Ruiz	Actively-sampled	0.8848	70.73%
	Randomly-sampled	0.8626	68.27%
Llaima	Actively-sampled	0.9898	93.32%
	Randomly-sampled	0.9332	94.82%

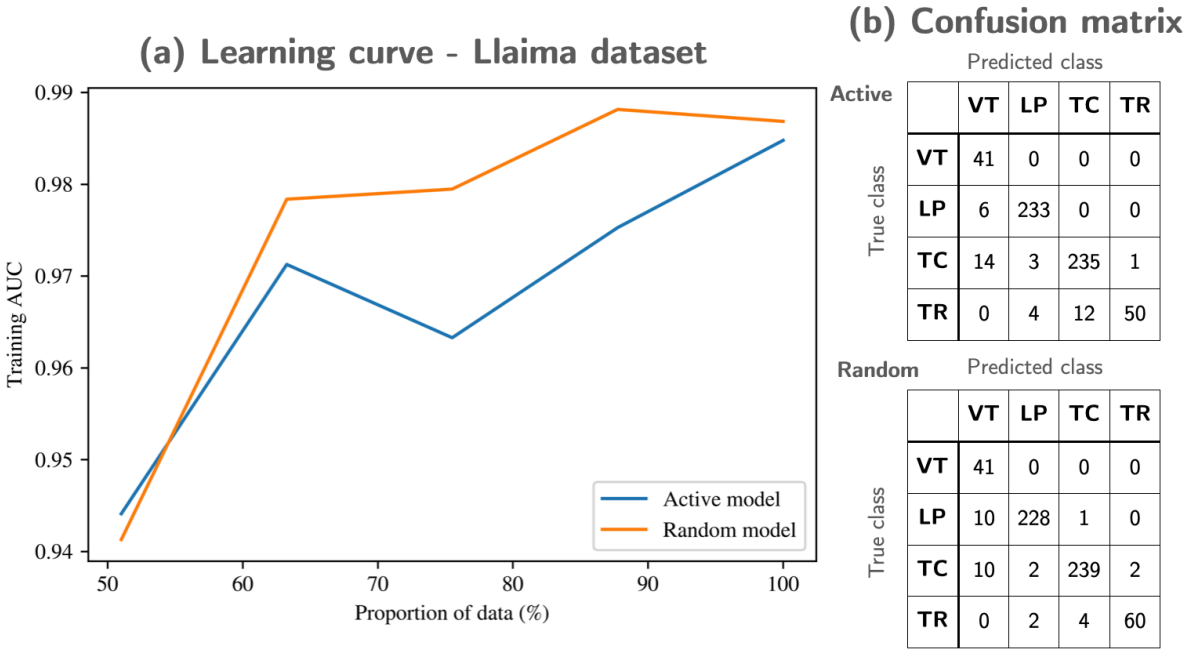


Figure 5.5: (a) Learning curves for the active model (blue) and random model (orange) for the Llaima dataset. The proportion of data is the amount of labelled data which the model is trained on, and the training AUC is the final AUC for the model training. (b) Confusion matrices for the active model (top) and the random model (bottom). Numbers indicate the number of events which were classified in each category, such that true positive or correct classifications are on the diagonal of the matrix and misclassifications appear in the off-diagonal elements.

5.5) Discussion

5.5.1) Model comparison

As discussed in Section 5.4, the models trained with actively-sampled data achieved better training AUC than the models trained with randomly-sampled data for the Nevado del Ruiz dataset. The Nevado del Ruiz active learning model achieved similar training AUCs at 55% of the data labelled to 100% of the data labelled, indicating the potential of an active learning approach to reduce the amount of labelling required for training a model. However, the improvement for using the active sampling was significantly smaller for the Llaima dataset. To contextualise these results, we can investigate the structure of the two datasets. Figure 5.6 is a 2-dimensional projection of the features from the final model using t-distributed Stochastic Nearest Neighbour Embedding (t-SNE), with individual features coloured by the order in which they are selected into the model. The t-SNE is a probabilistic machine learning algorithm which is designed to preserve the high-dimensional distance when the data ***are*** projected into two dimensions, such that points which are near in high-dimensional space remain near in the resulting two-dimensional representation (Van der Maaten and Hinton, 2008). This projection is achieved by minimising the Kullback-Leibler divergence of a high-dimensional joint probability and a low-dimensional joint probability that a point would pick another as its neighbour.

Figure 5.6a and 5.6d indicate some separability of the features based on the true label for the Nevado del Ruiz dataset, though the LP events show some overlap with both VT and hybrid events. There is excellent separability of the features based on the true label for the Llaima dataset. For the Nevado del Ruiz data, the hybrid events plot in a similar space to the LP events when projected which could explain the misclassification of hybrids as LP events which was discussed in Section 5.3.1. Figure 5.6c and 5.6f represent the point selection for the random models and we observe that the colours are evenly distributed across the space, which is consistent with the selection method as points are randomly sampled.

Figure 5.6b and 5.6e represent the point selection for the active learning models. For the Nevado del Ruiz active learning model (Figure 5.6b), a greater concentration of dark blue points (indicating earlier selection into the training set) is observed in the primarily-hybrid and primarily-VT clusters. This observation indicates that a greater proportion of characteristic events of those classes are included at an earlier stage in the experiment, which could explain the better performance of the active learning model with less training data. For the Llaima active learning model (Figure 6e), the dark blue points (indicating earlier selection of the point into the training set) are primarily in tremor events and at the edges of the LP and tectonic clusters, with more red points concentrated in the middle of the LP and tremor clusters. This observation could explain the relatively poor performance of the active sampling on the Llaima dataset early in the active learning experiment, as the active sampler is primarily picking outliers of the event types and tectonic events, which encompass a range of non-volcanic source processes. These observations indicate that a diversity-based active learning approach could be more beneficial for datasets where the event types are less separable, such as the Nevado del Ruiz dataset, as the active selection more efficiently samples the characteristic events in each class than random sampling.

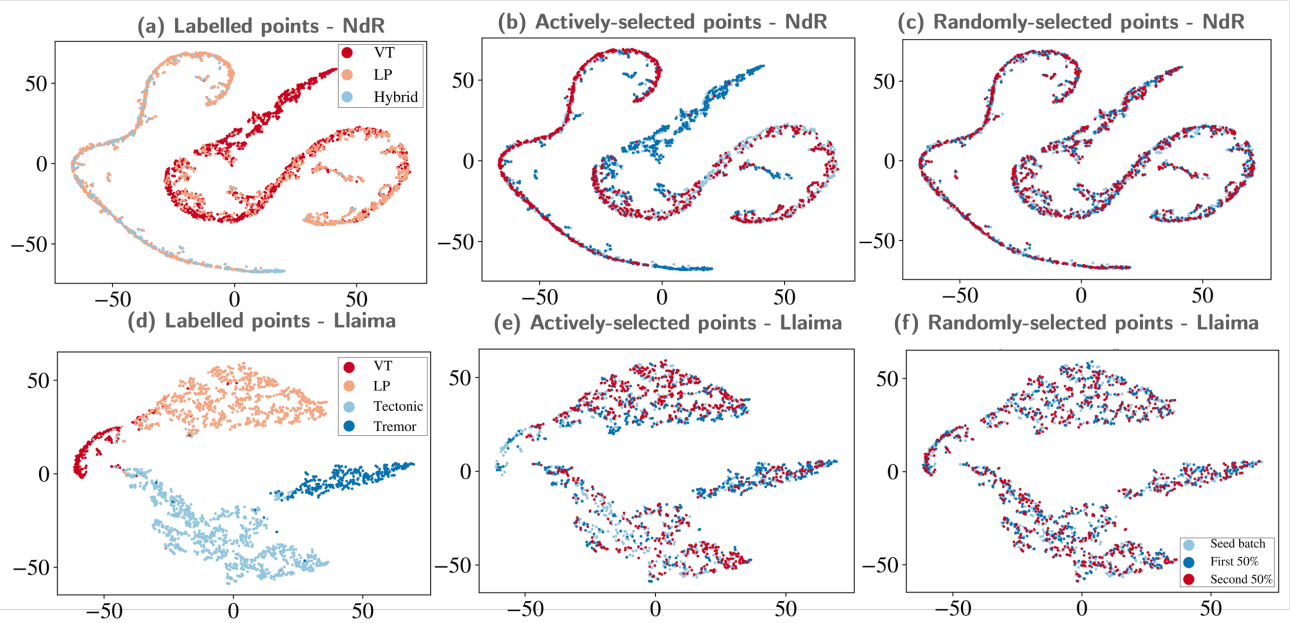


Figure 5.6: Visualisation of the final model features using a t-Distributed Stochastic Nearest-Neighbour Embedding projection (Van der Maaten and Hinton, 2008) from the actively-sampled model for the Nevado del Ruiz (NdR) dataset (a - c) and Llaima dataset (d – f) with the individual point colours in each panel: (a) Points coloured according to the Nevado del Ruiz labels as VT (red), LP (pink), hybrid (light blue). (b) Active-sampling model for Nevado del Ruiz with points coloured according to when they were selected to be added to the model, with the seed batch (initial, randomly selected waveforms) coloured in light blue, the first 50% of points selected into the model coloured in dark blue and the final 50% of points selected into the model coloured in red (according to legend in panel f). (c) Random-sampling model to Nevado del Ruiz with points coloured in the same scheme as (b). (d) Points coloured according to the Llaima labels as VT (red), LP (pink), tremor (light blue), tectonic (dark blue). (e) Active-sampling model for Llaima with points coloured in the same scheme as (b). (f) Random-sampling model for Llaima with points coloured in the same scheme as (b). The x and y axes represent non-linearly scaled distances projected into a 2-dimensional space and thus have no units, and are not comparable across axes.

5.5.2) Applications for active learning

The active learning approach, such as that presented here, has practical applications for volcano monitoring and volcano seismology research. Seismic monitoring is one of the most common and earliest-employed tools to be deployed as part of a monitoring network (e.g., Tilling, 1989), and large seismic catalogues are created for most well-monitored volcanoes (Newhall *et al*., 2017). Digitisation and curation of legacy seismic data can yield a large quantity of previously-unlabelled or analysed data, or inconsistently labelled data due to changes in monitoring network, observatory staff and/or procedures (Thompson *et al*., 2020). An active learning approach could be used in these contexts to accelerate the process of event classification to provide retrospective insights on the characteristics of seismicity prior to eruptions. Alternatively, active learning can be utilised at awakening volcanoes

where limited (or no) prior seismic catalogue is available, e.g., the re-awakening of Chaitén Volcano, Chile in 2008 (Lara, 2009). In such cases, an active learning approach running in real-time would minimise the burden on observatory analysts and speed up event classification tasks, which are both crucial aspects of volcano crisis management.

The active learning framework presented here has the advantage of not being intrinsically linked to any specific classification scheme. Other methods, such as automated classifiers trained on existing waveforms, might rely on a consistent set of event types. However, depending on the volcanic system being studied, the set of geologically-relevant event types may vary. For example, glacially-derived seismic events, also known as 'icequakes', are not always included in classification schemes and can mimic other volcanic event types (Thelen *et al*., 2013). Though icequakes are not always included in analyst-reviewed catalogues, identifying and monitoring the occurrence of these events is key for hazard assessment as debris can be mobilised by water from glaciers (Carrasco-Nunez et al., 1993). Rockfall events can be linked to changes in volcanic activity. DeRoin and McNutt (2012) analyse the occurrence of rockfall events at Augustine Volcano, USA and found a greater occurrence of rockfall events preceding the 2005 eruption, which they attribute to higher ground instability due to increased steaming at the summit. Timely identification of diverse events such as icequakes or rockfalls is important so that they may either be removed from the seismic record or analysed for their relevance to volcanic activity. Unlike using an existing classifier trained on data from other volcanic systems, an active learning approach to classification can be easily expanded to accommodate more event types based on the geological setting and unique properties of the volcanic environment.

5.5.3) Limitations

The active learning strategies presented here assume that the labels provided by the human expert (also termed an "oracle" in the field of active learning) are correct. This assumption may not be perfect. Seismic waveforms can be modified depending on external effects which may include path effects, such as the soil/bedrock characteristics, or the station geometries. Indeed, Cortés *et al*. (2021) suggest that the agreement between seismic analysts is approximately 80%. Incorrect classifications

are a source of noise, and Linville *et al*. (2019) show that for a tectonic seismic catalogue 70% of the machine learning classifier error may be attributed to analyst error during event labelling. A full analysis of the effect of oracle noise in our data is beyond the scope of this study and, for the analysis presented here, we do not distinguish error due to analyst mislabelling and error due to misclassification by the machine learning model.

This study has involved the use of single-station data for classifying seismic events. In practice for many volcanoes, a seismic network comprising multiple stations is used for monitoring volcano-seismic activity. The analyst-reviewed labels used as the ground-truth labels may therefore be made with reference to multi-station information. As discussed above, path and site effects can modify waveforms and therefore inclusion of waveforms from multiple stations for event classification could mitigate this effect. The approach presented here could be extended to include waveforms from multiple stations as channels of input into the CNN. This multi-channel input approach has previously been applied to use a 3-channel waveform as input for seismic phase arrival detection (Lapins *et al*., 2021). Using channels from a variety of station locations in the network (for example, a combination of summit and distal stations), could mitigate against the location-specific events which include, but are not limited to, the attenuation of high-frequency seismic energy within the waveform (e.g., Clarke *et al*., 2021).

Additionally, we have not considered any change in event classes during this study, using events which are from a limited time range. Cárdenas-Peña *et al*. (2013) used Hidden Markov Models to classify seismic activity at Nevado del Ruiz data and select features by choosing the set of features which contained the most time-varying information. We do not make any adjustments to our model for the time-dependent variation of seismicity. We suggest that if models are trained over a longer period than presented here (e.g., 10s of years), examples are selected from the full extent of the period covered by the training dataset to account for any variability over time.

5.6) Conclusions

We present an application of active learning to volcano-seismic event classification. We apply a method of active learning designed to select an optimum training dataset from a large pool of unlabelled training data. We combine this active learning with a CNN-based volcano-seismic event classifier and apply the model to two volcano-seismic datasets from Nevado del Ruiz, Colombia and Llaima, Chile. For both datasets, the active learning models achieve greater model performance when tested on unseen data than models trained using randomly-selected data. The active learning approach to data selection achieved greater training AUC at a smaller fraction of data for the Nevado del Ruiz dataset; however, for the Llaima dataset the gain from using an active learning approach was significantly smaller. To contextualise these results, we compare the separability of features for both dataset and suggest that the diversity-based active learning approach presented here can provide particular utility in accelerating model performance for datasets where the event types are less separable, as at Nevado del Ruiz. We discuss the potential applications of an active learning approach, which include the analysis of large amounts of unlabelled data from legacy datasets, rapid event classification at awakening volcanoes with no prior catalogue, or where volcano-specific event types (such as rockfall events) are widespread.

Chapter 6 - Conclusions

6.1) Thesis summary of science chapters

Chapter 3:

- I applied four multi-class classification methods to features derived from seismic data at two volcanoes to classify daily volcanic activity as eruptive and non-eruptive.
- These volcanoes: Nevado del Ruiz, Colombia, and Telica Volcano, Nicaragua, have contrasting eruptive styles therefore represent a good choice for a proof-of-concept study for these methods.
- I defined a new metric, the decisiveness index, for evaluating the success of machine learning models where a key quality of the resulting classification is the consistency of the classification.
- Support Vector Machine and Gaussian Process Classification approaches yielded the best decisiveness index scores and testing accuracies.
- Logistic regression did not perform as well as other classification methods, which could be due to underlying non-linearities in the data.
- The model classification of 'eruptive' or 'non-eruptive' had greater agreement with visual observations of eruption (e.g., ash emission) than the GVP-inferred labels of eruptive or non-eruptive activity.
- I demonstrated that features derived from single-station seismic data could be used to classify volcanic state.

Chapter 4:

- I extended the approaches developed in Chapter 3 on seismic datasets, to apply machine learning approaches to features derived from seismic, geodetic and gas data.
- I applied one-class classification or novelty detection, in which models are only trained on non-eruptive data, and multi-class classification, in which models are trained on both non-eruptive and eruptive data.

- Novelty detection approaches could be applied at volcanic systems where continuous geophysical measurements for previous eruptions have not been made, or where the monitoring infrastructure has significantly changed since previous eruptions.
- Non-dimensional visualisations of the full set of features with either seismic features from my detection, or seismic features from the Alaska Volcano Observatory (AVO) catalogue, indicate that smaller and more emergent events not contained in the AVO catalogue contribute towards distinguishing between non-eruptive and eruptive activity.
- I trained novelty detection models on each set of features individually (seismic, geodetic or gas features) and found that models formulated on all data types could better-distinguish between eruptive and non-eruptive activity than any of the single types of data alone. For gas-only models, the non-eruptive data was not well-characterised, whereas seismic- and geodetic-only models failed to adequately identify eruptive data.

Chapter 5:

- I combined an active learning approach with an existing volcano-seismic event classifier, to actively select the training data which the model learns from.
- This active learning approach does not require modifications to existing network architectures, therefore can be applied to other classification methods.
- I applied this model to two datasets of analyst labelled waveforms from Nevado del Ruiz Volcano, Colombia and Llaima Volcano, Chile.
- I compared the performance of the model trained with actively-sampled data with the performance of a model trained with randomly-selected data.
- For both datasets, the active learning model achieved better testing performance (measured via the AUC).
- For the Nevado del Ruiz dataset, the model with actively-sampled data achieved better training performance than the model with randomly-selected data. For the Llaima dataset, the gain from using an active learning approach was significantly smaller.

- From these results, I suggest that diversity-based active learning can reduce the size of the training set for noisy datasets where the events are less separable.

6.2) Implications for monitoring

The work presented in this thesis has demonstrated the potential use of machine learning classification for identifying large-scale transitions in eruptive activity. The retrospective approach to classifying activity explored in this thesis provides a data-driven constraint on the end of volcanic eruptions (Chapter 3), and on pre-eruptive changes in seismicity, deformation, and gas (Chapter 4). This constraint can be used to better identify and understand the timescales of pre- and post-eruptive processes.

Machine learning approaches to classifying activity, such as those presented in Chapters 3 and 4, could have promising utility in the forecasting of volcanic eruptions. Adapting the retrospective approach taken throughout this thesis to a forecasting context would be necessary, as forecasting requires a probabilistic approach (Marzocchi and Woo, 2007; Poland and Anderson, 2020; Whitehead and Bebbington, 2021). Probabilistic outputs are a feature of the methods explored in Chapter 3 such as Gaussian Process Regression and, for deterministic methods such as the Support Vector Machine (SVM), outputs can be converted to probabilities via methods such as the Platt scaling (Platt, 1999). For novelty detection approaches such as those explored in Chapter 4, Clifton *et al*. (2014) formulated a real-time novelty detection approach to monitoring patient deterioration using a probabilistic form of a one-class SVM, which has the potential to be adapted for volcanological purposes.

Novelty detection methods involve the identification of "abnormal" volcanic activity and could provide the basis of an alarm system for monitoring changes in volcanic activity. However, the goal of eruption forecasting involves a statement of the probability, size, style, and location of activity. The novelty detection methodology in this thesis represents a step towards this ultimate goal and could be further developed by application to a greater number of eruptive episodes. To adapt the novelty

detection methods in Chapter 4 for real-time volcanic monitoring, several steps would be required and would need to be integrated with previous monitoring infrastructure. A feature extraction system would need to automatically extract daily features from the volcano-seismic events recorded on each day. The waveform characteristics from which features are derived for input to the novelty detection models in Chapter 4 can be calculated using open-source software such as ObsPy (Beyreuther *et al*., 2010). A novelty score could then be computed from a pre-trained model. For real-time analysis, novelty detection-based condition monitoring has previously been designed for the monitoring of rural handpumps in remote locations (Greeff *et al*., 2018). This method involves processing data from accelerometers which are mounted on the instrument, and computing novelty scores using either an 8-bit microprocessor on the handpump or a cloud-based classifier. In the context of hand-pump monitoring, novelty detection models represent the foundation of an alarm system for "abnormal" behaviour of the device so that maintenance on the pump can be performed.

Supervised machine learning approaches require large amounts of labelled data for training (Chan *et al*., 2015). Throughout this thesis I have explored the consequences of using reduced amounts of data, as the seismic monitoring infrastructure available at volcanoes is variable (e.g., Moran *et al*., 2008). The modelling undertaken in Chapters 3 and 4 used seismic features derived from single-station data hence did not include parameters which can only be estimated from a seismic network with many stations such as event location or depth – though did use multiple stations to perform the detection of events. Additionally, most features were derived from detected events rather than continuous data, meaning that the approach could easily be extended to seismic catalogue data (with attached waveforms) from other volcanoes. I found that the use of single-station seismic data was sufficient to classify the activity as eruptive or non-eruptive. In Chapter 4, I derived two sets of seismic features: one from my seismic detection applied to the Augustine data, and one from the AVO analyst-reviewed catalogue. Comparing the separability of these features via non-dimensional visualisations indicated greater separability of eruptive and non-eruptive data when the smaller, more emergent events not included in the AVO catalogue were present. This finding has implications for applying

these methods to other volcano-seismic event catalogues, as those catalogues with stringent criteria on the inclusion of events may perform worse than catalogues which include events of varying type.

For the multi-parameter study explored in Chapter 4, I demonstrated that novelty detection models which were trained on multiple datasets (seismic, geodetic and gas) better-distinguished between eruptive and non-eruptive data than models trained on these datasets alone. However, the large quantity of dedicated ground-based instrumentation installed during the 2005-6 eruption of Augustine is not widely available for many volcanoes (e.g., Kumagai *et al*., 2007; Moran *et al*., 2008). With this in mind, remote instrumentation (such as satellite-based observations or regional infrasound networks) has provided multidisciplinary measurements for eruptions at sparsely instrumented systems such as Raikoke and Ulawun (McKee *et al*., 2021a; McKee *et al*., 2021b) and could be integrated into future machine learning approaches.

In the deep active learning study explored in Chapter 5, I demonstrated that an active learning approach could be used to reduce the amount of labelled training data required to build a volcano-seismic event classifier for datasets with less-distinct event types. Active learning could be expanded to other volcanic datasets for which labelling is time consuming. For example, computer vision machine learning approaches have been used to classify volcanic activity from analysis of the frames of video data from Villarrica, Chile (Witsil and Johnson, 2020b). Manual analysis of video frames can be time consuming, thus an active learning approach would be appealing to build a training dataset to apply this methodology over long timescales. Additionally, the classification of volcanic structures within satellite data could be a promising application for an active learning approach, as active learning approaches are established for the purposes of classifying land use in satellite imagery (Tuia *et al*., 2009).

6.3) Future perspectives

There are several potential avenues for expanding the work in this thesis. The methods and approaches explored in Chapter 3 and 4 could only be applied to a selection of datasets due to the limited availability of long-term datasets, and future work could apply these approaches to a wider variety of datasets from more volcanic systems. Here, I suggest some additional approaches for the analysis of volcanological time series data which could be used to address research gaps, which are summarised in Table 6.1.

6.3.1) Uniting quantitative and qualitative information in models of volcanic state

The amount of data collected by volcano observatories is ever-increasing, and this data represents both discrete and often qualitative observations e.g., weekly reports of visibly observed volcanic activity and continuous observations e.g., seismic or ground deformation. The qualitative monitoring of volcanoes has a rich and long-lived history and there are many volcanoes for which only historic eruptive observations have been made (Pyle and Barclay, 2020). In addition, there is a wealth of machine learning approaches for analysing the contents of free-text (e.g., Soderland, 1999).

Electronic Health Records (EHR) systems are used in a healthcare context to contain the broad array of medical data associated with a patient. The data contained within these records are varied in format including free-text clinical notes, diagnostic codes, and sensor measurements. The use of deep learning for analysis of EHRs has been recently established in the field of clinical machine learning (Shickel *et al*., 2017). These methods leverage the ability of deep learning models to handle high-dimensional datasets and account for complex and non-linear interactions between variables inferred from the input data. This application has involved the development of automatic clinical text analysis, to automatically extract key information from often complex clinical notes. The range of observational data routinely collected by volcano observatories for monitoring purposes has fundamental similarities to EHR data. Hence, a clear area for future study is to extract complementary

features from qualitative sources such as observatory reports and to investigate the effect of including these features within a machine learning framework.

Figure 6.1 demonstrates a potential methodology for leveraging data from qualitative sources to yield new insights on volcanic activity. This process would require the collection of written accounts of volcanic activity, which could include weekly activity bulletins or reports of volcanic activity. These sources must be processed to extract the raw text contained within them which describes the volcanic activity in question. For sources such as letters or historic reports, this process may require the digitisation of the contents before the text can be extracted. The free text must then be processed to extract the keywords which are representative of the volcanic activity. This processing can involve removing filler or "stop" words (such as "and" or "if") which do not add significant meaning to the sentence. Additionally, suffixes and prefixes of keywords can be removed to transform them into a word root which encompasses words of the same meaning without plurals or tenses, e.g., the word emissions and emission would both be transformed into the word root "emiss". The processed text can then be used as input to a model.

A promising approach is the method of topic modelling via Latent Dirichlet Allocation or LDA (Blei *et al*., 2003), which is an unsupervised approach to representing the latent structure in a collection of text-based records. This representation is formulated by assigning the probability that a given keyword belongs to a number of user-specified "topics". The results of this topic modelling could be investigated in isolation, e.g., to map the occurrence of different topics over time. This approach could be used for the automatic generation of timelines of volcanic activity to illustrate the timing of large-scale transitions in volcanic activity. Additionally, the topic modelling results could be combined with continuous features (such as those explored in Chapters 3 and 4) to perform supervised classification. Time series features derived from an LDA approach have been used in a healthcare context to augment physiological signals, and increased classification performance for predicting patient mortality in Intensive Care Unit (ICU) patients when compared to models formulated on physiological signals alone (Ghassemi *et al*., 2015).

Figure 6.2 illustrates the results of a pilot study of the LDA method applied to GVP weekly activity reports from Etna and Karymsky volcanoes. These volcanoes were chosen for study as they have had multiple eruptions during the period that weekly eruption activity reports were archived at GVP, thus have a relatively complete record over a 20-year period from late-2000 to May 2021. An LDA model with 3 topics was applied to both records of written weekly reports and both the keywords contained within each topic (Figure 6.2a and Figure 6.2c) and the occurrence of keywords from these topics over time is plotted (Figure 6.2b and Figure 6.2d). For Etna, the occurrence of Topic 1, where keywords were more associated with effusive activity (e.g., "flow"), is greater around 2004 – 2005, when the eruptive activity was dominated by lava flows (Global Volcanism Program, 2005). Additionally, the occurrence of Topic 3, which included words associated with explosive activity (e.g., "explos"), was greater during the paroxysmal episodes of activity during the emergence of the New Southeast crater in 2010 (Global Volcanism Program, 2013).

For Karymsky, the record of weekly activity reports in GVP is more complete, but visual observations are limited by the remote location of the volcano. Topic 3 keywords (Figure 6.1c) have the greatest occurrences during 2000 – 2006, where eruptive activity was dominated by ash-and-gas explosions and elevated seismicity (e.g., Global Volcanism Program, 2003). The occurrence of Topic 1, which encompasses keywords associated with plume ejection (e.g., plume, ash, km), is greatest during the period 2006 - 2013 where large plumes were erupted, extending up to 200 km from the volcano (Global Volcanism Program, 2009; Flower and Kahn, 2018). Topic 2 encompasses keywords associated with the remote observation of activity such as thermal anomalies via satellite imagery or ash emission via the observations of pilots (e.g., "satellit", "aviat"). This topic also reflects the reasonably stable activity during the period 2013 - 2021, where the alert level was consistently at orange ("remain", "orang").

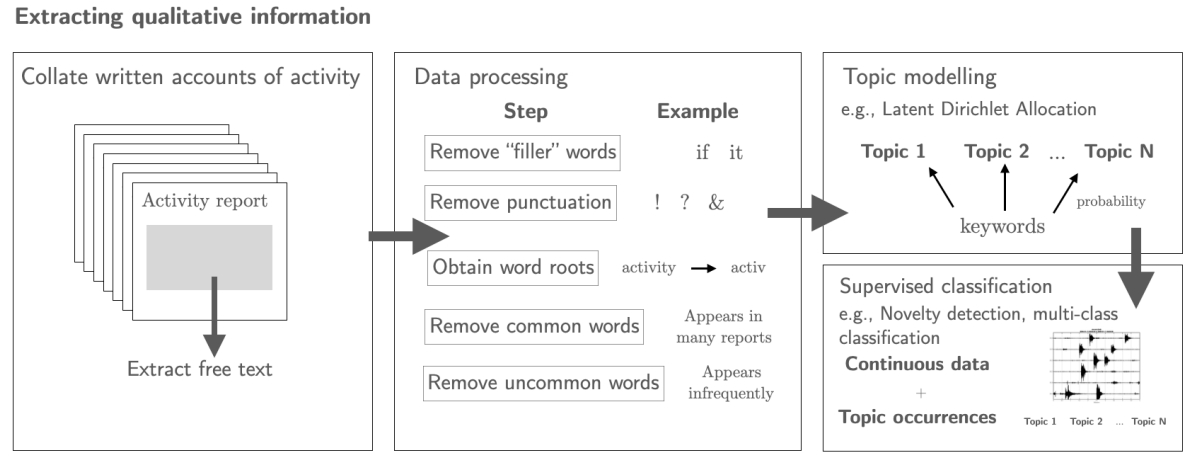


Figure 6.1: Illustration of a potential framework for integrating information from qualitative records such as activity reports into a machine learning approach.

(a) Top 5 words for each topic - Etna

Rank	Topic 1	Topic 2	Topic 3
1	lava	crater	crater
2	flow	emiss	lava
3	vent	bocca	explos
4	km	nuova	cone
5	strombolian	strombolian	strombolian

(b) Top 5 words for each topic - Karymsky

Rank	Topic 1	Topic 2	Topic 3
1	km	aviat	seismic
2	plume	report	ash
3	ash	satellit	concern
4	ft	orang	ga
5	l	remain	level

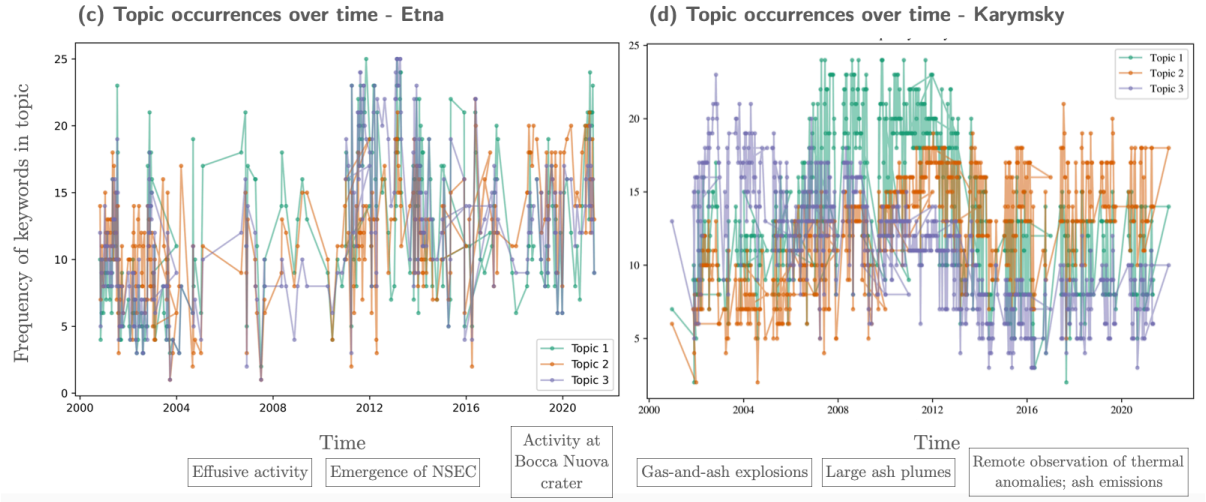


Figure 6.2: Results of a Latent Dirichlet Allocation model applied to over 20 years of GVP weekly activity reports from Etna and Karymsky volcanoes. (a) and (b) represent the top 5 words appearing in each topic ranked by their weight for inclusion into the topic. (c) and (d) represent the number of words from the top 25 words in each topic appearing in each report over time. Below these panels, broad descriptions of the activity over time which are discussed in the text are annotated in boxes, where NSEC is short for the New Southeast Crater. For both panels, Topic 1 is plotted in green, Topic 2 is in orange and Topic 3 is in purple. This unsupervised modelling is a method for automatically representing long-term time-series of volcanic activity, or could be combined with continuous data for the supervised classification of volcanic activity.

6.3.2) Repose on long timescales

The proof-of-concept study on eruption cessation presented in Chapter 3 covered three eruptive periods (two from Nevado del Ruiz, and one from Telica). An as-yet unaddressed question within the literature is whether the end of a volcanic eruption represents a return to a pre-eruptive baseline, or whether post-eruptive activity is distinct from pre-eruptive activity, and whether this is consistent across volcanoes of various eruptive styles.

Future work could address this question by observing a greater range of eruptive transitions and comparing the similarity of non-eruptive data prior to eruptions with non-eruptive data from after eruptions. A trial study to test this hypothesis could be to extract features from volcanic data from periods of extrusion and quiescence from the same long-lived eruption. An ideal candidate for a study of this nature would be the long-lived eruption of Soufrière Hills Volcano, Montserrat, due to the long record of eruption and pause at the volcano, which is illustrated in Figure 6.3 after Rodgers *et al*. (2016). A key research question for Soufrière Hills Volcano is whether the seismicity during the current pause in eruption is distinct from previous pauses in eruption.

Additionally, there is a dense record of SO_2 measurements made at Soufrière Hills Volcano (e.g., Christopher *et al*., 2010; Figure 6.3) and inclusion of gas-derived features to this analysis could yield additional insights into the nature of repose on long timescales. Edmonds (2008) interpreted the high emission rates of SO_2 during pauses in eruption at Soufrière Hills Volcano as the result of efficient segregation of gas from the melt, in combination with high bulk permeability, which allowed high quantities of SO_2 to reach the surface whilst magma remained stagnant at depth. Comparison of seismic- and gas-derived features (and a combination of data types) could be a mechanism to investigate the nature of this proposed decoupling of the gas-magma system over long timescales and over multiple episodes of pause in eruption.

To execute this study, unsupervised machine learning methods could be used. Unsupervised machine learning is the process of training a machine learning model with no labels (e.g., Bishop, 2006). Clustering algorithms are a type of unsupervised method which are implemented with the goal of grouping similar examples within the dataset and include methods such as k-means clustering or Gaussian mixture models. Supervised methods could also be used, such as the multi-class classification methods used in Chapters 3 and 4. These supervised methods could be used by training a model on data from a selected number of pauses and validating the model on the remaining pauses in eruption. Then, the classification of the validation data (and uncertainty of this classification, if using a probabilistic method such as Gaussian Process Regression) could be used to make inferences on the statistical similarity of data from pauses in eruption.

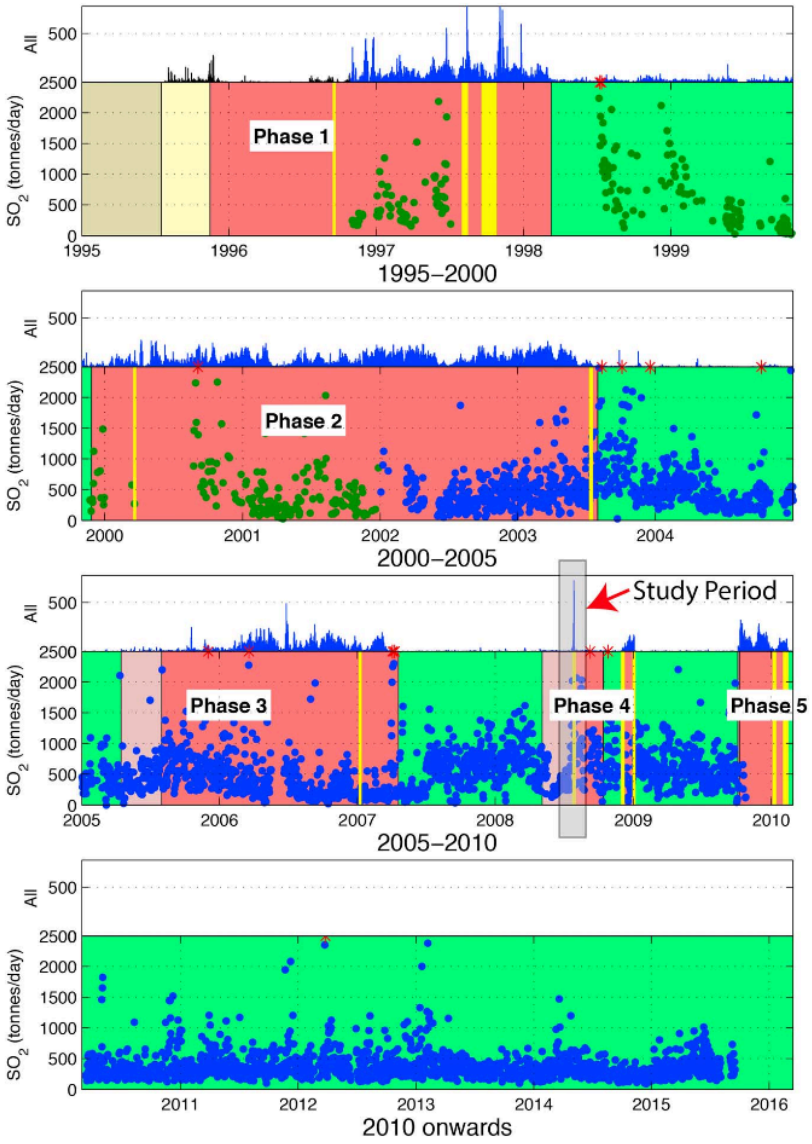


Figure 6.3: Timeline of eruptive activity at Soufrière Hills Volcano, Montserrat from 1995 – 2016. Periods of extrusion (Phases 1 – 5) are in pink, pauses in eruption are in green, and major explosions are in yellow. Total counts of volcanic earthquakes are plotted as vertical bars, and daily SO_2 emission rate measurements are plotted as dots. From Rodgers *et al.* (2016).

6.3.3) Understanding the physical processes governing eruption end

The work detailed in Chapter 3 explored the timescales governing the end of volcanic eruptions. One of the key findings of this work was that the seismic end of eruption identified using machine learning approaches agreed with the timescale of eruption end from generic definitions or the "90-day rule", whereby the end of eruption is defined based on a lack of visual eruption for 90 days. This finding provides a data-driven constraint on the timescale of eruption cessation processes, which was consistent across these two volcanoes with contrasting eruptive styles.

Future work could take this finding further by investigating the physical processes which underpin this timescale. Individual studies on end of eruption (discussed in Chapter 1) have indicated that the processes governing the end of volcanic eruptions could include (but may not be limited to) decreased magma supply, where the reservoir is not recharged during the eruption. For eruptions where the edifice size increases during the eruption, such as dome-growth, the decreased overpressure due to eruption at the surface is suggested to be a primarily driver of eruption end (e.g., Wadge, 1981). Additionally, friction between semi-solid magma and the conduit wall and crystallisation could act against the continuation of eruption. An investigation of the physical processes governing eruption end for the inversion of data from specific eruptions could build on previous work in the realm of physics-based forward models after Anderson and Segall (2013) and Segall (2013) by adding parameterisations for processes such as crystallisation and wall friction. Alternatively, a modelling approach could build on recent advances in the modelling of multiphase reactive transport (e.g., Keller and Suckale, 2019), to formulate a forward model of conduit processes and validate this model alongside observations.

6.3.4) Intra-eruptive transitions

The models explored in Chapters 3 and 4 of this thesis have classified volcanic data as eruptive or non-eruptive. A natural extension of this approach is to consider a greater number of classes of volcanic activity. Volcanic activity can be prolonged, and can undergo transitions in eruptive style during the course of a major and/or prolonged eruption. For example, transitions between explosive and effusive behaviour are often observed in dome-forming eruptions (e.g., Edmonds and Herd, 2007; Preece *et al*., 2017). Previous work in this field has explored the use of semi-Markov chain models to model the probability of transition between 8 major states of eruptive activity based on the eruption and quiescence durations from a database of multi-phase eruptions (Bebbington and Jenkins, 2019).

Future advances in the field of volcanic state characterisation could take this modelling into the realm of continuous monitoring data including one-class methods, for detecting changes in volcanic style which are unknown or unobserved in nature, or multi-class methods, for recognising previously-

observed or repetitive changes in volcanic activity. A study of this nature would require a long time series of data, which covered multiple phases of intra-eruptive activity with sufficient length to train a one-class or multi-class model. For a one-class model, potential datasets could include eruptions which have an extended dome-forming phase prior to explosive eruption, such as the 2020 - 2021 eruption of Soufrière St Vincent (Global Volcanism Program, 2021). For a multi-class model, repetitive changes could include those associated with dome growth and collapse during the eruption of Soufrière Hills Volcano, Montserrat (Miller *et al*., 1998). The methodologies for applying these modelling approaches could be similar to those applied in Chapters 3 and 4 of this thesis, though careful selection of training periods which characterise these distinct states of activity would be required.

Table 6.1: Summary of the future work ideas presented in this thesis, including the necessary data and model types to execute these projects.

Future work project	Aim(s)	Type of model	Type of data
Uniting quantitative and qualitative information in models of volcanic state	Visualise volcanic activity over time Characterise volcanic state	Natural Language Processing e.g., LDA Novelty detection Multi-class classification	Qualitative (e.g., weekly activity reports; letters; historical documents) Continuous monitoring data (e.g., seismic, gas, geodetic)
Repose on long timescales	Understand the extent to which pre-eruptive activity is similar to post-eruptive activity	Unsupervised classification (e.g., k-means clustering) Multi-class classification	Continuous monitoring data e.g., seismic, gas, geodetic)
Understanding the physical processes governing eruption end	Evaluate the different processes governing eruption end	Physics-based forward model of conduit processes	Continuous monitoring data e.g., seismic, gas, geodetic) for validating a forward model
Intra-eruptive transitions	Classify different types of activity which occur within a single eruption	Novelty detection Multi-class classification	Continuous monitoring data e.g., seismic, gas, geodetic) on sufficiently long timescales to characterise each type of activity

References

Ackerley, N., Beer, M., Kougioumtzoglou, I.A., Patelli, E. and Au, S.K., 2014. Principles of broadband seismometry. *Encyclopedia Earthquake Engineering*, *1941*, pp.1-35.

Aiuppa, A., Moretti, R., Federico, C., Giudice, G., Gurrieri, S., Liuzzo, M., Papale, P., Shinohara, H. and Valenza, M., 2007. Forecasting Etna eruptions by real-time observation of volcanic gas composition. *Geology*, *35*(12), pp.1115-1118.

Alaska Volcano Observatory/USGS. (1988). Alaska Volcano Observatory. International Federation of Digital Seismograph Networks. https://doi.org/10.7914/SN/AV

Altamimi, Z., Rebischung, P., Métivier, L. and Collilieux, X., 2016. ITRF2014: A new release of the International Terrestrial Reference Frame modeling nonlinear station motions. *Journal of Geophysical Research: Solid Earth*, *121*(8), pp.6109-6131.

Anantrasirichai, N., Biggs, J., Albino, F., Hill, P. and Bull, D., 2018. Application of Machine Learning to Classification of Volcanic Deformation in Routinely Generated InSAR Data. *Journal of Geophysical Research: Solid Earth*, 123(8), pp.6592-6606.

Anderson, K. and Segall, P., 2013. Bayesian inversion of data from effusive volcanic eruptions using physics-based models: Application to Mount St. Helens 2004–2008. *Journal of Geophysical Research: Solid Earth*, *118*(5), pp.2017-2037.

Apolloni, B., 2009. Support vector machines and MLP for automatic classification of seismic signals at Stromboli volcano. In *Neural Nets WIRN09: Proceedings of the 19th Italian Workshop on Neural Nets*, Vietri Sul Mare, Salerno, Italy May 28-30 2009 (Vol. 204, p. 116). IOS Press.

Aspinall, W.P. and Woo, G., 2014. Santorini unrest 2011–2012: an immediate Bayesian belief network analysis of eruption scenario probabilities for urgent decision support under uncertainty. *Journal of Applied Volcanology*, *3*(1), pp.1-12.

Aytar, Y., Vondrick, C. and Torralba, A., 2016. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, *29*, pp.892-900.

Baptie, B., Luckett, R. and Neuberg, J., 2002. Observations of low-frequency earthquakes and volcanic tremor at Soufriere Hills Volcano, Montserrat. *Geological Society, London, Memoirs*, *21*(1), pp.611-620.

Barclay, J., Few, F., Armijos, M.T., Phillips, J.C., Pyle, D.M., Hicks, A.J., Brown, S.K., and Robertson, R.E.A., 2019. Livelihoods, wellbeing and the risk to life during volcanic eruptions, *Frontiers in Earth Science*, 7: 205, doi: 10.3389/feart.2019.00205

Barmin, A., Melnik, O. and Sparks, R.S.J., 2002. Periodic behavior in lava dome eruptions. *Earth and Planetary Science Letters*, 199(1), pp.173-184.

Battaglia, J. and Aki, K., 2003. Location of seismic events and eruptive fissures on the Piton de la Fournaise volcano using seismic amplitudes. *Journal of Geophysical Research: Solid Earth*, *108*(B8).

Battaglia, J., Ferrazzini, V., Staudacher, T., Aki, K. and Cheminée, J.L., 2005. Pre-eruptive migration of earthquakes at the Piton de la Fournaise volcano (Réunion Island). *Geophysical Journal International*, *161*(2), pp.549-558.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y. and Wassermann, J. (2010), ObsPy: A Python Toolbox for Seismology, *Seismological Research Letters*, 81(3), 530-533.

Bean, C.J., De Barros, L., Lokmer, I., Métaxian, J.P., O'Brien, G. and Murphy, S., 2014. Long-period seismicity in the shallow volcanic edifice formed from slow-rupture earthquakes. *Nature geoscience*, *7*(1), pp.71-75.

Bebbington, M.S. and Jenkins, S.F., 2019. Intra-eruption forecasting. *Bulletin of Volcanology*, *81*(6), pp.1-17.

Belgiu, M. and Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, pp.24-31.

Bellman, R.E. (1961) *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, New Jersey, USA.

Bhatti, S.M., Khan, M.S., Wuth, J., Huenupan, F., Curilem, M., Franco, L. and Yoma, N.B., 2016. Automatic detection of volcano-seismic events by modeling state and event duration in hidden Markov models. *Journal of Volcanology and Geothermal Research*, *324*, pp.134-143.

Bicego, M., Acosta-Muñoz, C. and Orozco-Alzate, M., 2013. Classification of seismic volcanic signals using hidden-Markov-model-based generative embeddings. *IEEE Transactions on Geoscience and Remote Sensing*, 51(6), pp.3400-3409.

Biggs, J., Ebmeier, S.K., Aspinall, W.P., Lu, Z., Pritchard, M.E., Sparks, R.S.J. and Mather, T.A., 2014. Global link between deformation and volcanic eruption quantified by satellite imagery. *Nature communications*, *5*(1), pp.1-7.

Biggs, J., Albino, F., Dogru, F., Yip, S., Yu, C., Li, Z. and Anantrasirichai, N., 2019, December. Automated methods for detecting volcanic deformation using Sentinel-1 InSAR time series. In *AGU Fall Meeting Abstracts* (Vol. 2019, pp. G22A-01).

Bishop, C. M., 2006. *Pattern recognition and machine learning*. New York, Springer.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *The Journal of machine Learning research*, *3*, pp.993-1022.

Blewitt, G., W. C. Hammond, and C. Kreemer (2018), Harnessing the GPS data explosion for interdisciplinary science, *Eos, 99,*https://doi.org/10.1029/2018EO104623.

Bolton, M.S., Jensen, B.J., Wallace, K., Praet, N., Fortin, D., Kaufman, D. and De Batist, M., 2020. Machine learning classifiers for attributing tephra to source volcanoes: an evaluation of methods for Alaska tephras. *Journal of Quaternary Science*, *35*(1-2), pp.81-92.

Bonny, E. and Wright, R., 2017. Predicting the end of lava flow-forming eruptions from space. *Bulletin of Volcanology*, 79(7): 52, DOI 10.1007/s00445-017-1134-8

Boué, A., Lesage, P., Cortés, G., Valette, B. and Reyes-Dávila, G., 2015. Real-time eruption forecasting using the material Failure Forecast Method with a Bayesian approach. *Journal of Geophysical Research: Solid Earth*, 120(4), pp.2143-2161.

Bouveyron, C. and Girard, S., 2009. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, *42*(11), pp.2649-2658.

Brantley SR (ed) (1990) The eruption of Redoubt Volcano, Alaska, December 14, 1989–August 31, 1990. US Geological Survey Circular 1061, p. 33

Breiman L., Friedman R.A., Olshen R.A., and Stone C.G. (1984) *Classification and Regression Trees.* Pacific Grove, CA: Wadsworth.

Bueno, A., Benitez, C., De Angelis, S., Moreno, A.D. and Ibanez, J.M., 2019. Volcano-seismic transfer learning and uncertainty quantification with Bayesian neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(2), pp.892-902.

Bueno, A., Zuccarello, L., Díaz-Moreno, A., Woollam, J., Titos, M., Benítez, C., Álvarez, I., Prudencio, J. and De Angelis, S., 2020. PICOSS: Python Interface for the classification of seismic signals. *Computers & Geosciences*, *142*, p.104531.

Buurman, H., West, M., 2010. *Seismic precursors to volcanic explosions during the 2006 eruption of Augustine Volcano*. In: Power, J., Coombs, M., Freymueller, J. (Eds.), *The 2006 eruption of Augustine Volcano. U.S. Geological Survey Professional Paper 1769*, Alaska (U.S. Geological Survey Professional Paper 1769).

Buurman, H., Nye, C.J., West, M.E. and Cameron, C., 2014. Regional controls on volcano seismicity along the Aleutian arc. Geochemistry, Geophysics, Geosystems, 15(4), pp.1147-1163.

Calderoni, G., De Simoni, B., De Simoni, F.M. and Merucci, L., 1997. Real-time monitoring of seismic data using satellite telemetry. *Annals of Geophysics* 40, no. 4. August 1997.

Cameron, C.E., Prejean, S.G., Coombs, M.L., Wallace, K.L., Power, J.A. and Roman, D.C., 2018. Alaska volcano observatory alert and forecasting timeliness: 1989–2017. *Frontiers in Earth Science*, 6:86, DOI=10.3389/feart.2018.00086

Canário, J.P., Mello, R., Curilem, M., Huenupan, F. and Rios, R., 2020a. In-depth comparison of deep artificial neural network architectures on seismic events classification. *Journal of Volcanology and Geothermal Research*, *401*, p.106881.

Canário, J.P., de Mello, R.F., Curilem, M., Huenupan, F. and Rios, R.A., 2020b. Llaima volcano dataset: In-depth comparison of deep artificial neural network architectures on seismic events classification. *Data in brief*, *30*, p.105627.

Cárdenas-Peña, D., Orozco-Alzate, M. and Castellanos-Dominguez, G., 2013. Selection of time-variant features for earthquake classification at the Nevado-del-Ruiz volcano. *Computers & Geosciences*, *51*, pp.293-304.

Castaño, L.M., Ospina, C.A., Cadena, O.E., Galvis-Arenas, B., Londono, J.M., Laverde, C.A., Kaneko, T. and Ichihara, M., 2020. Continuous monitoring of the 2015–2018 Nevado del Ruiz activity, Colombia, using satellite infrared images and local infrasound records. *Earth, Planets and Space*, *72*, pp.1-18.

Carniel, R., Di Cecca, M. and Rouland, D., 2003. Ambrym, Vanuatu (July–August 2000): spectral and dynamical transitions on the hours-to-days timescale. *Journal of Volcanology and Geothermal Research*, 128(1-3), pp.1-13.

Carniel, R., 2014. Characterization of volcanic regimes and identification of significant transitions using geophysical data: a review. *Bulletin of Volcanology*. 76: 848, DOI 10.1007/s00445-014-0848-0

Carrasco-Núñez, G., Vallance, J.W. and Rose, W.I., 1993. A voluminous avalanche-induced lahar from Citlaltépetl volcano, Mexico: implications for hazard assessment. *Journal of Volcanology and Geothermal Research*, *59*(1-2), pp.35-46.

Cervelli, P.F., Fournier, T., Freymueller, J. and Power, J.A., 2006. Ground deformation associated with the precursory unrest and early phases of the January 2006 eruption of Augustine Volcano, Alaska. *Geophysical Research Letters*, 33(18).

Cervelli, P.F., Fournier, T.J., Freymueller, J.T., Power, J.A., Lisowski, Michael, and Pauk, B.A., 2010, *Geodetic constraints on magma movement and withdrawal during the 2006 eruption of Augustine Volcano*, chapter 17 of Power, J.A., Coombs, M.L., and Freymueller, J.T., eds., *The 2006 eruption of Augustine Volcano, Alaska: U.S. Geological Survey Professional Paper 1769*, p. 427–452 [https://pubs.usgs.gov/pp/1769/chapters/p1769_chapter17.pdf].

Chan, T.H., Jia, K., Gao, S., Lu, J., Zeng, Z. and Ma, Y., 2015. PCANet: A simple deep learning baseline for image classification?. *IEEE transactions on image processing*, *24*(12), pp.5017-5032.

Chang, C.C. and Lin, C.J., 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST),* 2(3), p.27.

Chastin, S.F. and Main, I.G., 2003. Statistical analysis of daily seismic event rate as a precursor to volcanic eruptions. *Geophysical Research Letters,* 30:13, DOI: 10.1029/2003GL016900

Chouet, B.A., Page, R.A., Stephens, C.D., Lahr, J.C. and Power, J.A., 1994. Precursory swarms of long-period events at Redoubt Volcano (1989–1990), Alaska: their origin and use as a forecasting tool. *Journal of Volcanology and Geothermal Research*, *62*(1-4), pp.95-135.

Chouet, B.A., 1996. Long-period volcano seismicity: its source and use in eruption forecasting. *Nature*, *380*(6572), pp.309-316.

Chouet, B.A. and Matoza, R.S., 2013. A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption. *Journal of Volcanology and Geothermal Research*, *252*, pp.108-175.

Christopher, T., Edmonds, M., Humphreys, M.C. and Herd, R.A., 2010. Volcanic gas emissions from Soufrière Hills Volcano, Montserrat 1995–2009, with implications for mafic magma supply and degassing. *Geophysical Research Letters*, *37*(19).

Clarke, J., Adam, L. and van Wijk, K., 2021. LP or VT signals? How intrinsic attenuation influences volcano seismic signatures constrained by Whakaari volcano parameters. *Journal of Volcanology and Geothermal Research*, *418*, p.107337.

Clifton, D. and Tarassenko, L., 2006. Condition monitoring of gas-turbine engines. *Transfer Report, Department of Engineering Science, University of Oxford.*

Clifton, D.A., Bannister, P.R. and Tarassenko, L., 2007. A framework for novelty detection in jet engine vibration data. In *Key engineering materials (*Vol. 347, pp. 305-310). Trans Tech Publications Ltd.

Clifton, L.A., 2007. *Multi-channel novelty detection and classifier combination*. The University of Manchester (United Kingdom).

Clifton, L., Clifton, D.A., Pimentel, M.A.F., Watkinson, P.J., and Tarassenko, L., 2014. Predictive Monitoring of Mobile Patients by Combining Clinical Observations with Data from Wearable Sensors. *IEEE Journal of Biomedical and Health Informatics* 18(3), 2014, pp. 722-730

Collombet, M., Grasso, J.R. and Ferrazzini, V., 2003. Seismicity rate before eruptions on Piton de la Fournaise volcano: Implications for eruption dynamics. *Geophysical Research Letters*, 30(21): 2099, doi:10.1029/2003GL017494

Colopy, G.W., Roberts, S.J. and Clifton, D.A., 2018. Bayesian Optimization of Personalized Models for Patient Vital-Sign Monitoring. *IEEE Journal of Biomedical and Health Informatics*, 22(2), pp.301-310.

Connor, C.B., Sparks, R.S.J., Mason, R.M., Bonadonna, C. and Young, S.R., 2003. Exploring links between physical and probabilistic models of volcanic eruptions: The Soufriere Hills Volcano, Montserrat. *Geophysical Research Letters,* 30(13).

Coombs, M.L., Bull, K.F., Vallance, J.W., Schneider, D.J., Thoms, E.E., Wessels, R.L., and McGimsey, R.G., 2010, *Timing, distribution, and volume of proximal products of the 2006 eruption of Augustine Volcano*, chapter 8 *of* Power, J.A., Coombs, M.L., and Freymueller, J.T., eds., *The 2006 eruption of Augustine Volcano, Alaska: U.S. Geological Survey Professional Paper 1769*, p. 145–185, 1 plate, scale 1:20,000, and GIS data [https://pubs.usgs.gov/pp/1769/chapters/p1769_chapter08].

Cornelius, R.R. and Voight, B., 1995. Graphical and PC-software analysis of volcano eruption precursors according to the Materials Failure Forecast Method (FFM). *Journal of Volcanology and Geothermal Research*, *64*(3-4), pp.295-320.

Cortés, G., Carniel, R., Lesage, P., Mendoza, M. and Della Lucia, I., 2021. Practical Volcano-Independent Recognition of Seismic Events: VULCAN. ears Project. *Frontiers in Earth Science*, *8*, p.702.

Cover, T.M. and Thomas, J.A., 2006. *Elements of information theory.* 2nd edition. Wiley-Interscience: NJ.

Crosweller, H.S., Arora, B., Brown, S.K., Cottrell, E., Deligne, N.I., Guerrero, N.O., Hobbs, L., Kiyosugi, K., Loughlin, S.C., Lowndes, J. and Nayembil, M., 2012. Global database on large magnitude explosive volcanic eruptions (LaMEVE). *Journal of Applied Volcanology*, *1*(1), pp.1-13.

Cuellar-Rodriguez J.V., Ramirez-Lopez C., 1987. *Descripcion de los volcanes Colombianos.* Rev CIAF, Bogota, p 189-222.

Curilem, G., Vergara, J., Fuentealba, G., Acuña, G. and Chacón, M., 2009. Classification of seismic signals at Villarrica volcano (Chile) using neural networks and genetic algorithms. *Journal of Volcanology and Geothermal Research*, 180(1), pp.1-8.

Curilem, M., de Mello, R.F., Huenupan, F., San Martin, C., Franco, L., Hernández, E. and Rios, R.A., 2018. Discriminating seismic events of the Llaima volcano (Chile) based on spectrogram cross-correlations. *Journal of Volcanology and Geothermal Research*, *367*, pp.63-78.

De Barros, L., Bean, C.J., Zecevic, M., Brenguier, F. and Peltier, A., 2013. Eruptive fracture location forecasts from high-frequency events on Piton de la Fournaise Volcano. *Geophysical Research Letters*, 40(17), pp.4599-4603.

De la Cruz-Reyna, S., Tilling, R.I. and Valdés-González, C., 2017. Challenges in responding to a sustained, continuing volcanic crisis: the case of Popocatepétl volcano, Mexico, 1994-present. In *Observing the Volcano World* (pp. 235-252). Springer, Cham.

Dempsey, D.E., Cronin, S.J., Mei, S. and Kempa-Liehr, A.W., 2020. Automatic precursor recognition and real-time forecasting of sudden explosive volcanic eruptions at Whakaari, New Zealand. *Nature communications*, *11*(1), pp.1-8.

DeRoin, N. and McNutt, S.R., 2012. Rockfalls at Augustine Volcano, Alaska: The influence of eruption precursors and seasonal factors on occurrence patterns 1997–2009. *Journal of Volcanology and Geothermal Research*, 211, pp.61-75.

DeShon, H.R., Thurber, C.H., and Power, J.A., 2010, *Earthquake waveform similarity and evolution at Augustine Volcano from 1993 to 2006*, chapter 5 of Power, J.A., Coombs, M.L., and Freymueller, J.T., eds., *The 2006 eruption of Augustine Volcano, Alaska: U.S. Geological Survey Professional Paper 1769,* p. 103–118 [https://pubs.usgs.gov/pp/1769/chapters/p1769_chapter05.pdf].

Dixon, T.H., Gonzalez, G., Lichten, S.M. and Katsigris, E., 1991. First epoch geodetic measurements with the Global Positioning System across the northern Caribbean plate boundary zone. *Journal of Geophysical Research: Solid Earth*, *96*(B2), pp.2397-2415.

Dixon, J.P., and Stihler, S.D., 2009, *Catalog of earthquake hypocenters at Alaskan Volcanoes: January 1 through December 31, 2008*: U.S. Geological Survey Data Series 467, 88 p.

Dzurisin, D., Moran, S. C., Lisowski, M., Schilling, S. P., Anderson, K. R., and Werner, C., 2015. The 2004–2008 dome-building eruption at Mount St. Helens, Washington: epilogue. *Bulletin of Volcanology.* 77:17. doi: 10.1007/s00445-015-0973-4.

Ebmeier, S.K., Biggs, J., Mather, T.A. and Amelung, F., 2013. On the lack of InSAR observations of magmatic deformation at Central American volcanoes. *Journal of Geophysical Research: Solid Earth*, 118(5), pp.2571-2585.

Edmonds, M., and Herd, R., 2007. A volcanic degassing event at the explosive-effusive transition. *Geophysical Research Letters* 34, L21310

Edmonds, M., 2008. New geochemical insights into volcanic degassing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1885), pp.4559-4579.

Endo, E.T., Murray, T.L. and Power, J.A., 1996. A comparison of preeruption real-time seismic amplitude measurements for eruptions at Mount St. Helens, Redoubt Volcano, Mount Spurr, and Mount Pinatubo. *Fire and Mud*, pp.233-247.

Fearnley, C.J. and Beaven, S., 2018. Volcano alert level systems: managing the challenges of effective volcanic crisis communication. *Bulletin of Volcanology*, 80(5), p.46.

Fischer, T.P., Morrissey, M.M., Calvache, V.M.L., Gomez, M.D., Torres, C.R., Stix, J. and Williams, S.N., 1994. Correlations between SO_2 flux and long-period seismicity at Galeras volcano. *Nature*, *368*(6467), pp.135-137.

Flower, V.J., Oommen, T. and Carn, S.A., 2016. Improving global detection of volcanic eruptions using the Ozone Monitoring Instrument (OMI). *Atmospheric Measurement Techniques*, 9(11), pp.5487-5498.

Flower, V.J. and Kahn, R.A., 2018. Karymsky volcano eruptive plume properties based on MISR multi-angle imagery and the volcanological implications. *Atmospheric chemistry and physics*, *18*(6), pp.3903-3918.

Gaddes, M.E., Hooper, A. and Bagnardi, M., 2019. Using machine learning to automatically detect volcanic unrest in a time series of interferograms. *Journal of Geophysical Research: Solid Earth*, 124(11), pp.12304-12322.

Gal, Y. and Ghahramani, Z., 2016, June. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*(pp. 1050-1059). PMLR.

Geirsson, H., Rodgers, M., LaFemina, P., Witter, M., Roman, D., Muñoz, A., Tenorio, V., Alvarez, J., Jacobo, V.C., Nilsson, D. and Galle, B., 2014. Multidisciplinary observations of the 2011 explosive eruption of Telica volcano, Nicaragua: implications for the dynamics of low-explosivity ash eruptions. *Journal of Volcanology and Geothermal Research*, 271, pp.55-69.

Ghassemi, M., Pimentel, M., Naumann, T., Brennan, T., Clifton, D., Szolovits, P. and Feng, M., 2015, February. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 29, No. 1).

Gill, M.K., Asefa, T., Kaheil, Y. and McKee, M., 2007. Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique. *Water resources research*, 43(7).

Global Volcanism Program, 2003. Report on Karymsky (Russia) (Venzke, E., ed.). *Bulletin of the Global Volcanism Network*, 28:11. Smithsonian Institution. https://doi.org/10.5479/si.GVP.BGVN200311-300130

Global Volcanism Program, 2005. Report on Etna (Italy) (Wunderman, R., ed.). *Bulletin of the Global Volcanism Network*, 30:1. Smithsonian Institution. https://doi.org/10.5479/si.GVP.BGVN200501-211060

Global Volcanism Program, 2006. Report on Augustine (United States) (Wunderman, R., ed.). *Bulletin of the Global Volcanism Network*, 31:1. Smithsonian Institution. https://doi.org/10.5479/si.GVP.BGVN200601-313010.

Global Volcanism Program, 2009. Report on Karymsky (Russia) (Wunderman, R., ed.). *Bulletin of the Global Volcanism Network*, 34:8. Smithsonian Institution. https://doi.org/10.5479/si.GVP.BGVN200908-300130

Global Volcanism Program, 2011. Report on Telica (Nicaragua). In: Sennert, S K (ed.), *Weekly Volcanic Activity Report*, 11 May-17 May 2011. Smithsonian Institution and US Geological Survey.

Global Volcanism Program, 2012a. Report on Nevado del Ruiz (Colombia). In: Sennert, S K (ed.), *Weekly Volcanic Activity Report*, 7 March-13 March 2012. Smithsonian Institution and US Geological Survey.

Global Volcanism Program, 2012b. Report on Telica (Nicaragua). In: Sennert, S K (ed.), *Weekly Volcanic Activity Report,* 12 September-18 September 2012. Smithsonian Institution and US Geological Survey.

Global Volcanism Program, 2013. Report on Etna (Italy) (Wunderman, R., ed.). *Bulletin of the Global Volcanism Network*, 38:9. Smithsonian Institution. https://doi.org/10.5479/si.GVP.BGVN201309-211060

Global Volcanism Program, 2017. Report on Nevado del Ruiz (Colombia). In: Venzke, E (ed.), *Bulletin of the Global Volcanism Network*, 42:6. Smithsonian Institution.

Global Volcanism Program, 2021. Report on Soufriere St. Vincent (Saint Vincent and the Grenadines) (Bennis, K.L., and Venzke, E., eds.). *Bulletin of the Global Volcanism Network*, 46:3. Smithsonian Institution. https://doi.org/10.5479/si.GVP.BGVN202103-360150.

Greeff, H., Manandhar, A., Thomson, P., Hope, R. and Clifton, D.A., 2018. Distributed inference condition monitoring system for rural infrastructure in the developing world. *IEEE Sensors Journal*, *19*(5), pp.1820-1828.

Gunn, L.S., Blake, S., Jones, M.C. and Rymer, H., 2014. Forecasting the duration of volcanic eruptions: an empirical probabilistic model. *Bulletin of Volcanology*, *76*(1), pp.1-18.

Guo, C., Lu, M. and Chen, J., 2020. An evaluation of time series summary statistics as features for clinical prediction tasks. *BMC medical informatics and decision making*, *20*(1), pp.1-20.

Gutiérrez, L., Ibanez, J., Cortés, G., Ramírez, J., Benítez, C., Tenorio, V. and Isaac, A., 2009, July. Volcano-seismic signal detection and classification processing using hidden Markov models. Application to San Cristóbal volcano, Nicaragua. In *2009 IEEE International Geoscience and Remote Sensing Symposium* (Vol. 4, pp. IV-522). IEEE.

Guttorp, P. and Gneiting, T., 2006. Studies in the history of probability and statistics XLIX on the Matérn correlation family. *Biometrika*, *93*(4), pp.989-995.

Harrington, R.M. and Brodsky, E.E., 2007. Volcanic hybrid earthquakes that are brittle-failure events. *Geophysical Research Letters*, *34*(6).

Hastie, T., Tibshirani, R. and Friedman, J., 2001. *The elements of statistical learning*. New York. NY: Springer.

Hellweg, M., 2000. Physical models for the source of Lascar's harmonic tremor. *Journal of Volcanology and Geothermal Research*, *101*(1-2), pp.183-198.

Hicks, A., and Few, R., 2015. Trajectories of social vulnerability during the Soufrière Hills volcanic crisis. *Journal of Applied Volcanology*, 4:10. doi: 10.1186/s13617-015-0029-7

Hidayati, S., Triastuty, H., Mulyana, I., Adi, S., Ishihara, K., Basuki, A., Kuswandarto, H., Priyanto, B. and Solikhin, A., 2018. Differences in the seismicity preceding the 2007 and 2014 eruptions of Kelud volcano, Indonesia. *Journal of Volcanology and Geothermal Research* 382, 50-67, https://doi.org/10.1016/j.jvolgeores.2018.10.017

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hooper, A.J., Gudmundsson, M.T., Bagnardi, M., Jarosch, A.H., Spaans, K., Magnússon, E., Parks, M., Dumont, S., Ofeigsson, B., Sigmundsson, F. and Hreinsdottir, S., 2015, December. Forecasting of flood basalt eruptions: lessons from Bárðarbunga. In *AGU Fall Meeting Abstracts* (Vol. 2015, pp. V24B-03).

Huppert, H.E. and Sparks, R.S., 2016. Compressible vapour flow in conduits and fractures. *Journal of Fluid Mechanics*, *802*, pp.750-759.

Ignatieva, A., Bell, A. and Worton, B. (2018). Point Process Models for Quasi-Periodic Volcanic Earthquakes. Statistics in Volcanology. 4. 10.5038/2163-338X.4.2.

Jacobs, K.M., and McNutt, S.R., 2010, *Using seismic b-values to interpret seismicity rates and physical processes during the preeruptive earthquake swarm at Augustine Volcano 2005–2006*, chapter 3 *of* Power, J.A., Coombs, M.L., and Freymueller, J.T., eds., *The 2006 eruption of Augustine Volcano, Alaska: U.S. Geological Survey Professional Paper 1769*, p. 59–83 [https://pubs.usgs.gov/pp/1769/chapters/p1769_chapter03.pdf].

Jain, P. and Kapoor, A., 2009. Active learning for large multi-class problems. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 762-769). IEEE.

Jellinek, A.M. and Bercovici, D., 2011. Seismic tremors and magma wagging during explosive volcanism. *Nature*, *470*(7335), pp.522-525.

Johnson, J.H., Prejean, S., Savage, M.K. and Townend, J., 2010. Anisotropy, repeating earthquakes, and seismicity associated with the 2008 eruption of Okmok volcano, Alaska. *Journal of Geophysical Research: Solid Earth*, *115*(B9).

Johnstone, I.M. and Titterington, D.M., 2009. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A* 367: https://doi.org/10.1098/rsta.2009.0159

Jolly, A.D., Lokmer, I., Thun, J., Salichon, J., Fry, B. and Chardot, L., 2017. Insights into fluid transport mechanisms at White Island from analysis of coupled very long-period (VLP), long-period (LP) and high-frequency (HF) earthquakes. *Journal of Volcanology and Geothermal Research*, *343*, pp.75-94.

Jousset, P., Neuberg, J. and Sturton, S., 2003. Modelling the time-dependent frequency content of low-frequency volcanic earthquakes. *Journal of Volcanology and Geothermal Research*, 128(1-3), pp.201-223.

Julian, B.R., 1994. Volcanic tremor: Nonlinear excitation by fluid flow. *Journal of Geophysical Research: Solid Earth*, *99*(B6), pp.11859-11877.

Keller, T. and Suckale, J., 2019. A continuum model of multi-phase reactive transport in igneous systems. *Geophysical Journal International*, *219*(1), pp.185-222.

Kilburn, C.R., 2003. Multiscale fracturing as a key to forecasting volcanic eruptions. *Journal of Volcanology and Geothermal Research,* 125(3-4), pp.271-289.

Kiyasseh, D., Zhu, T. and Clifton, D.A., 2020. SoCal: Selective Oracle Questioning for Consistency-based Active Learning of Cardiac Signals. *arXiv preprint arXiv:2004.09557*.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems,* 25, pp.1097-1105.

Kumagai, H., Yepes, H., Vaca, M., Caceres, V., Naga, T., Yokoe, K., Imai, T., Miyakawa, K., Yamashina, T., Arrais, S. and Vasconez, F., 2007. Enhancing volcano-monitoring capabilities in Ecuador. *Eos, Transactions American Geophysical Union*, *88*(23), pp.245-246.

La Femina, P.C., Connor, C.B., Hill, B.E., Strauch, W. and Saballos, J.A., 2004. Magma–tectonic interactions in Nicaragua: the 1999 seismic swarm and eruption of Cerro Negro volcano. *Journal of Volcanology and Geothermal Research*, *137*(1-3), pp.187-199.

Lacroix, A., 1908. *La montagne Pelée après ses éruptions, avec observations sur les éruptions du Vésuve en 79 et en 1906*. Masson et cie.

Lahr, J.C., Chouet, B.A., Stephens, C.D., Power, J.A. and Page, R.A., 1994. Earthquake classification, location, and error analysis in a volcanic environment: Implications for the magmatic system of the 1989–1990 eruptions at Redoubt Volcano, Alaska. *Journal of Volcanology and Geothermal Research*, *62*(1-4), pp.137-151.

Lamb, O.D., Varley, N.R., Mather, T.A., Pyle, D.M., Smith, P.J. and Liu, E.J., 2014. Multiple timescales of cyclical behaviour observed at two dome-forming eruptions. *Journal of Volcanology and Geothermal Research*, 284, pp.106-121.

Langer, H., Falsaperla, S., Powell, T. and Thompson, G., 2006. Automatic classification and a-posteriori analysis of seismic event identification at Soufriere Hills volcano, Montserrat. *Journal of Volcanology and Geothermal Research*, *153*(1-2), pp.1-10.

Lapins, S., Goitom, B., Kendall, J.M., Werner, M.J., Cashman, K.V. and Hammond, J.O., 2021. A little data goes a long way: automating seismic phase arrival picking at Nabro Volcano with transfer learning. *Journal of Geophysical Research: Solid Earth*, *126*(7), p.e2021JB021910.

Lara, F., Lara-Cueva, R., Larco, J.C., Carrera, E.V. and León, R., 2021. A deep learning approach for automatic recognition of seismo-volcanic events at the Cotopaxi volcano. *Journal of Volcanology and Geothermal Research*, *409*, p.107142.

Lara, L.E., 2009. The 2008 eruption of the Chaitén Volcano, Chile: a preliminary report. *Andean Geology*, *36*(1), pp.125-129.

Lara-Cueva, R.A., Benítez, D.S., Carrera, E.V., Ruiz, M. and Rojo-Álvarez, J.L., 2016. Automatic recognition of long period events from volcano tectonic earthquakes at cotopaxi volcano. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(9), pp.5247-5257.

Larsen, J.F., Nye, C.J., Coombs, M.L., Tilman, M., Izbekov, P., and Cameron, C., 2010, *Petrology and geochemistry of the 2006 eruption of Augustine Volcano*, chapter 15 of Power, J.A., Coombs, M.L., and Freymueller, J.T., eds*., The 2006 eruption of Augustine Volcano, Alaska*: *U.S. Geological Survey Professional Paper 1769*, p. 335–382 and spreadsheets [https://pubs.usgs.gov/pp/1769/chapters/p1769_chapter15].

Latter, J.H., 1981. Volcanic earthquakes, and their relationship to eruptions at Ruapehu and Ngauruhoe volcanoes. *Journal of Volcanology and Geothermal Research*, *9*(4), pp.293-309.

Lavallée, Y., Meredith, P.G., Dingwell, D.B., Hess, K.U., Wassermann, J., Cordonnier, B., Gerik, A. and Kruhl, J.H., 2008. Seismogenic lavas and explosive eruption forecasting. *Nature*, *453*(7194), pp.507-510.

Le, Q.V., 2013. Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8595-8598). IEEE.

Lesage, P., Mora, M.M., Alvarado, G.E., Pacheco, J. and Métaxian, J.P., 2006. Complex behavior and source model of the tremor at Arenal volcano, Costa Rica. *Journal of Volcanology and Geothermal Research*, *157*(1-3), pp.49-59.

Li, W., Guo, Q. and Elkan, C., 2010. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE transactions on geoscience and remote sensing*, 49(2), pp.717-725.

Lin, W., Hasenstab, K., Cunha, G.M. and Schwartzman, A., 2020. Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment. *Scientific Reports*, *10*(1), pp.1-11.

Linde, A.T. and Sacks, I.S., 1998. Triggering of volcanic eruptions. *Nature*, *395*(6705), pp.888-890.

Linville, L., Pankow, K. and Draelos, T., 2019. Deep learning models augment analyst decisions for event discrimination. *Geophysical Research Letters*, *46*(7), pp.3643-3651.

Londono, J.M., 2016. Evidence of recent deep magmatic activity at Cerro Bravo-Cerro Machín volcanic complex, central Colombia. Implications for future volcanic activity at Nevado del Ruiz,

Cerro Machín and other volcanoes. *Journal of Volcanology and Geothermal Research*, *324*, pp.156-168.

Londono, J.M. and Galvis, B. (2018). Seismic Data, Photographic Images and Physical Modeling of Volcanic Plumes as a Tool for Monitoring the Activity of Nevado del Ruiz Volcano, Colombia. *Frontiers in Earth Sciences*. 6:162. doi: 10.3389/feart.2018.00162

Lowe, D.R., Williams, S.N., Leigh, H., Connort, C.B., Gemmell, J.B. and Stoiber, R.E., 1986. Lahars initiated by the 13 November 1985 eruption of Nevado del Ruiz, Colombia. *Nature*, *324*(6092), pp.51-53.

Lu, Z. and Dzurisin, D., 2014. InSAR imaging of Aleutian volcanoes. In *InSAR imaging of Aleutian volcanoes* (pp. 87-345). Springer, Berlin, Heidelberg.

Macedonio, G., Dobran, F. and Neri, A., 1994. Erosion processes in volcanic conduits and application to the AD 79 eruption of Vesuvius. *Earth and Planetary Science Letters,* 121(1-2), pp.137-152.

Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P. and Amemoutou, A., 2017. Implementation of a multistation approach for automated event classification at Piton de la Fournaise volcano. *Seismological Research Letters*, 88(3), pp.878-891.

Majtner, T., Yildirim-Yayilgan, S. and Hardeberg, J.Y., 2016. Combining deep learning and hand-crafted features for skin lesion classification. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1-6). IEEE.

Malfante, M., Dalla Mura, M., Métaxian, J.P., Mars, J.I., Macedo, O. and Inza, A., 2018. Machine learning for volcano-seismic signals: Challenges and perspectives. *IEEE Signal Processing Magazine*, *35*(2), pp.20-30.

Manley, G.F., Pyle, D.M., Mather, T.A., Rodgers, M., Clifton, D.A., Stokell, B.G., Thompson, G., Londoño, J.M. and Roman, D.C., 2020. Understanding the timing of eruption end using a machine learning approach to classification of seismic time series. *Journal of Volcanology and Geothermal Research*, p.106917.

Marzocchi, W., Scandone, R. and Mulargia, F., 1993. The tectonic setting of Mount Vesuvius and the correlation between its eruptions and the earthquakes of the Southern Apennines. *Journal of Volcanology and Geothermal Research*, *58*(1-4), pp.27-41.

Marzocchi, W., Sandri, L., Gasparini, P., Newhall, C. and Boschi, E., 2004. Quantifying probabilities of volcanic events: the example of volcanic hazard at Mount Vesuvius. *Journal of Geophysical Research: Solid Earth*, *109*(B11).

Marzocchi, W. and Woo, G., 2007. Probabilistic eruption forecasting and the call for an evacuation. *Geophysical Research Letters*, 34: 22

Marzocchi, W., Sandri, L. and Selva, J., 2008. BET_EF: a probabilistic tool for long-and short-term eruption forecasting. *Bulletin of Volcanology*, *70*(5), pp.623-632.

Masotti, M., Falsaperla, S., Langer, H., Spampinato, S. and Campanini, R., 2006. Application of Support Vector Machine to the classification of volcanic tremor at Etna, Italy. *Geophysical research letters*, *33*(20).

Mathworks, 2021. *MATLAB fitrgp.* Mathworks, accessed 27th October 2021 <https://uk.mathworks.com/help/stats/fitrgp.html>.

McCullagh, P. and Nelder, J.A., 1989. *Generalized linear models* (Vol. 37). CRC press.

McGee, K.A., Doukas, M.P., McGimsey, R.G., Neal, C.A., and Wessels, R.L., 2010, *Emission of SO_2, CO_2, and H_2S from Augustine Volcano, 2002–2008*, chapter 26 of Power, J.A., Coombs, M.L., and Freymueller, J.T., eds*., The 2006 eruption of Augustine Volcano, Alaska*: *U.S. Geological Survey Professional Paper 1769*, p. 609–627 [https://pubs.usgs.gov/pp/1769/chapters/p1769_chapter26.pdf].

McKee, K., Smith, C.M., Reath, K., Snee, E., Maher, S., Matoza, R.S., Carn, S., Mastin, L., Anderson, K., Damby, D. and Roman, D.C., 2021a. Evaluating the state-of-the-art in remote volcanic eruption characterization Part I: Raikoke volcano, Kuril Islands. *Journal of Volcanology and Geothermal Research*, *419*, p.107354.

McKee, K., Smith, C.M., Reath, K., Snee, E., Maher, S., Matoza, R.S., Carn, S., Roman, D.C., Mastin, L., Anderson, K. and Damby, D., 2021b. Evaluating the state-of-the-art in remote volcanic eruption characterization Part II: Ulawun volcano, Papua New Guinea. *Journal of Volcanology and Geothermal Research*, *420*, p.107381.

McNutt, S.R., 1996. Seismic monitoring and eruption forecasting of volcanoes: a review of the state-of-the-art and case histories. In *Monitoring and mitigation of volcano hazards* (pp. 99-146). Springer, Berlin, Heidelberg.

McNutt, S.R. and Roman, D.C., 2015. Volcanic seismicity. In *The Encyclopedia of Volcanoes* (pp. 1011-1034). Academic Press.

Miller, A.D., Stewart, R.C., White, R.A., Luckett, R., Baptie, B.J., Aspinall, W.P., Latchman, J.L., Lynch, L.L. and Voight, B., 1998. Seismicity associated with dome growth and collapse at the Soufriere Hills Volcano, Montserrat. *Geophysical Research Letters*, *25*(18), pp.3401-3404.

Moffat, A.J. and Millan, M.M., 1971. The applications of optical correlation techniques to the remote sensing of SO2 plumes using sky light. *Atmospheric Environment (1967)*, *5*(8), pp.677-690.

Moran, S.C., Kwoun, O., Masterlark, T. and Lu, Z., 2006. On the absence of InSAR-detected volcano deformation spanning the 1995–1996 and 1999 eruptions of Shishaldin Volcano, Alaska. *Journal of Volcanology and Geothermal Research*, *150*(1-3), pp.119-131.

Moran, S.C., Freymueller, J.T., LaHusen, R.G., McGee, K.A., Poland, M.P., Power, J.A., Schmidt, D.A., Schneider, D.J., Stephens, G., Werner, C.A. and White, R.A., 2008. Instrumentation recommendations for volcano monitoring at US volcanoes under the National Volcano Early Warning System. *US Geological Survey Scientific Investigations Report*, *5114*, p.47.

Moran, S.C., Newhall, C. and Roman, D.C., 2011. Failed magmatic eruptions: late-stage cessation of magma ascent. *Bulletin of Volcanology*, 73(2), pp.115-122.

Mountrakis, G., Im, J. and Ogole, C., 2011. Support vector machines in remote sensing: A review. ISPRS *Journal of Photogrammetry and Remote Sensing*, 66(3), pp.247-259.

Mũnoz-Marí, J., Bovolo, F., Gómez-Chova, L., Bruzzone, L. and Camp-Valls, G., 2010. Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE transactions on geoscience and remote sensing*, 48(8), pp.3188-3197.

Murphy, J.R. and Lahoud, J.A., 1969. Analysis of seismic peak amplitudes from underground nuclear explosions. *Bulletin of the Seismological Society of America*, *59*(6), pp.2325-2341.

Nadeau, P.A., Palma, J.L. and Waite, G.P., 2011. Linking volcanic tremor, degassing, and eruption dynamics via SO2 imaging. *Geophysical research letters*, *38*(1).

Nanni, L., Ghidoni, S. and Brahnam, S., 2017. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, *71*, pp.158-172.

Naranjo, J.L., Sigurdsson, H., Carey, S.N. and Fritz, W., 1986. Eruption of the Nevado del Ruiz volcano, Colombia, on 13 November 1985: tephra fall and lahars. *Science*, *233*(4767), pp.961-963.

National Academies of Sciences, Engineering, and Medicine, 2017. *Volcanic eruptions and their repose, unrest, precursors, and timing*. National Academies Press.

Neuberg, J., 2000. Characteristics and causes of shallow seismicity in andesite volcanoes. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *358*(1770), pp.1533-1546.

Neuberg, J., Luckett, R., Baptie, B. and Olsen, K., 2000. Models of tremor and low-frequency earthquake swarms on Montserrat. *Journal of Volcanology and Geothermal Research*, *101*(1-2), pp.83-104.

Neuberg, J.W., Tuffen, H., Collier, L., Green, D., Powell, T. and Dingwell, D., 2006. The trigger mechanism of low-frequency earthquakes on Montserrat. *Journal of Volcanology and Geothermal Research*, *153*(1-2), pp.37-50.

Newhall, C.G., Costa, F., Ratdomopurbo, A., Venezky, D.Y., Widiwijayanti, C., Win, N.T.Z., Tan, K. and Fajiculay, E., 2017. WOVOdat–an online, growing library of worldwide volcanic unrest. *Journal of Volcanology and Geothermal Research*, *345*, pp.184-199.

Pan, S.J. and Yang, Q., 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), pp.1345-1359.

Passarelli, L. and Brodsky, E.E., 2012. The correlation between run-up and repose times of volcanic eruptions. *Geophysical Journal International*, 188(3), pp.1025-1045.

Pauk, B.A., Jackson, Michael, Feaux, Karl, Mencin, David, and Bohnenstiehl, Kyle, 2010, *The Plate Boundary Observatory permanent Global Positioning System network on Augustine Volcano before and after the 2006 eruption*, chapter 19 *of* Power, J.A., Coombs, M.L., and Freymueller, J.T., eds., *The 2006 eruption of Augustine Volcano, Alaska: U.S. Geological Survey Professional Paper 1769*, p. 467–477 [https://pubs.usgs.gov/pp/1769/chapters/p1769_chapter19.pdf].

Pesicek, J.D., Wellik, J.J., Prejean, S.G. and Ogburn, S.E., 2018. Prevalence of seismic rate anomalies preceding volcanic eruptions in Alaska. *Frontiers in Earth Science*, 6, p.100.

Petersen, T., De Angelis, S., Tytgat, G. and McNutt, S.R., 2006. Local infrasound observations of large ash explosions at Augustine Volcano, Alaska, during January 11–28, 2006. *Geophysical Research Letters*, 33(12).

Petrelli, M. and Perugini, D., 2016. Solving petrological problems through machine learning: the study case of tectonic discrimination using geochemical and isotopic data. *Contributions to Mineralogy and Petrology*, *171*(10), pp.1-15.

Petrelli, M., Caricchi, L. and Perugini, D., 2020. Machine Learning Thermo-Barometry: Application to Clinopyroxene-Bearing Magmas. *Journal of Geophysical Research: Solid Earth*, *125*(9), p.e2020JB020130.

Phillipson, G., Sobradelo, R. and Gottsmann, J., 2013. Global volcanic unrest in the 21st century: an analysis of the first decade. *Journal of Volcanology and Geothermal Research*, 264, pp.183-196.

Pimentel, M.A., Clifton, D.A., Clifton, L. and Tarassenko, L., 2014. A review of novelty detection. *Signal Processing*, 99, pp.215-249.

Pinel, V., Jaupart, C. and Albino, F., 2010. On the relationship between cycles of eruptive activity and growth of a volcanic edifice. *Journal of Volcanology and Geothermal Research*, *194*(4), pp.150-164.

Platt, J., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), pp.61-74.

Poland, M.P. and Anderson, K.R., 2020. Partly cloudy with a chance of lava flows: Forecasting volcanic eruptions in the twenty-first century. *Journal of Geophysical Research: Solid Earth*, *125*(1), p.e2018JB016974.

Powell, T.W. and Neuberg, J., 2003. Time dependent features in tremor spectra. *Journal of Volcanology and Geothermal Research*, *128*(1-3), pp.177-185.

Power, J.A., Lahr, J.C., Page, R.A., Chouet, B.A., Stephens, C.D., Harlow, D.H., Murray, T.L. and Davies, J.N., 1994. Seismic evolution of the 1989–1990 eruption sequence of Redoubt Volcano, Alaska. *Journal of Volcanology and Geothermal Research*, *62*(1-4), pp.69-94.

Power, J.A., Nye, C.J., Coombs, M.L., Wessels, R.L., Cervelli, P.F., Dehn, J., Wallace, K.L., Freymueller, J.T. and Doukas, M.P., 2006. The reawakening of Alaska's Augustine volcano. *EOS, Transactions American Geophysical Union*, 87(37), pp.373-377.

Power, J.A., and Lalla, D.J., 2010, *Seismic observations of Augustine Volcano, 1970–2007*, chapter 1 of Power, J.A., Coombs, M.L., and Freymueller, J.T., eds., *The 2006 eruption of Augustine Volcano, Alaska*: *U.S. Geological Survey Professional Paper 1769*, p. 3–40 [https://pubs.usgs.gov/pp/1769/chapters/p1769_chapter01.pdf].

Power, J.A., Coombs, M.L., and Freymueller, J.T., eds., 2010*, The 2006 eruption of Augustine Volcano, Alaska*: U.S. Geological Survey Professional Paper 1769, 667 p., 1 plate, scale 1:20,000, and data files.

Pradhan, B. and Lee, S., 2010. Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. *Environmental Modelling & Software*, 25(6), pp.747-759.

Preece, K., Gertisser, R., Barclay, J., Charbonnier, S.J., Komorowski, J.C. and Herd, R.A., 2016. Transitions between explosive and effusive phases during the cataclysmic 2010 eruption of Merapi volcano, Java, Indonesia. *Bulletin of Volcanology*, *78*(8), art. 54.

Pritchard, M.E., Mather, T.A., McNutt, S.R., Delgado, F.J. and Reath, K., 2019. Thoughts on the criteria to determine the origin of volcanic unrest as magmatic or non-magmatic. *Philosophical Transactions of the Royal Society A*, *377*(2139), art. 20180008.

Pyle, D.M., 1998. Forecasting sizes and repose times of future extreme volcanic events. *Geology*, 26(4), pp.367-370.

Pyle, D.M., Mather, T.A. and Biggs, J., 2013. Remote sensing of volcanoes and volcanic processes: integrating observation and modelling–introduction. *Geological Society, London, Special Publications*, *380*(1), pp.1-13.

Pyle, D.M. and Barclay, J., 2020. Historical records of volcanic eruptions deserve more attention. *Nature Reviews Earth and Environment* 1, 183–184. https://doi.org/10.1038/s43017-020-0044-z

Quinn, J.A. and Williams, C.K., 2007, June. Known unknowns: Novelty detection in condition monitoring. In *Iberian Conference on Pattern Recognition and Image Analysis* (pp. 1-6). Springer, Berlin, Heidelberg.

Rasmussen, C.E. and Williams, C.K., 2006. *Gaussian processes for machine learning*, vol. 1. MIT press, 39, pp.40-43.

Rebischung, P. and Schmid, R., 2016. IGS14/igs14. atx: a new framework for the IGS products. In *AGU Fall Meeting 2016*.

Ren, C.X., Peltier, A., Ferrazzini, V., Rouet-Leduc, B., Johnson, P.A. and Brenguier, F., 2020. Machine learning reveals the seismic signature of eruptive behavior at Piton de la Fournaise volcano. *Geophysical Research Letters*, *47*(3), p.e2019GL085523. (https://doi.org/10.1029/2019GL085523.)

Ripepe, M., Harris, A.J. and Carniel, R., 2002. Thermal, seismic and infrasonic evidences of variable degassing rates at Stromboli volcano. *Journal of Volcanology and Geothermal Research*, 118(3-4), pp.285-297.

Ripepe, M., Marchetti, E., Delle Donne, D., Genco, R., Innocenti, L., Lacanna, G. and Valade, S., 2018. Infrasonic early warning system for explosive eruptions. *Journal of Geophysical Research: Solid Earth*, *123*(11), pp.9570-9585.

Rodgers, M., Roman, D.C., Geirsson, H., LaFemina, P., Muñoz, A., Guzman, C. and Tenorio, V., 2013. Seismicity accompanying the 1999 eruptive episode at Telica Volcano, Nicaragua. *Journal of Volcanology and Geothermal Research*, 265, pp.39-51.

Rodgers, M., Roman, D.C., Geirsson, H., LaFemina, P., McNutt, S.R., Muñoz, A. and Tenorio, V., 2015a. Stable and unstable phases of elevated seismic activity at the persistently restless Telica Volcano, Nicaragua. *Journal of Volcanology and Geothermal Research*, 290, pp.63-74.

Rodgers, M., Rodgers, S. and Roman, D.C., 2015b. Peakmatch: A Java program for multiplet analysis of large seismic datasets. *Seismological Research Letters*, 86(4), pp.1208-1218.

Rodgers, M., Pyle, D.M. and Mather, T.A., 2015c, December. Statistical Analysis of Large Volcano Seismology Datasets to Determine Patterns of Volcanic Behaviour. In *AGU Fall Meeting Abstracts* (Vol. 2015, pp. V14B-01).

Rodgers, M., Smith, P.J., Mather, T.A. and Pyle, D.M., 2016. Quiescent-explosive transitions during dome-forming volcanic eruptions: Using seismicity to probe the volcanic processes leading to the 29 July 2008 Vulcanian explosion of Soufrière Hills Volcano, Montserrat. *Journal of Geophysical Research: Solid Earth*, *121*(12), pp.8453-8471.

Roman, D.C. and Cashman, K.V., 2006. The origin of volcano-tectonic earthquake swarms. *Geology*, *34*(6), pp.457-460.

Roman, D.C., De Angelis, S., Latchman, J.L. and White, R., 2008. Patterns of volcanotectonic seismicity and stress during the ongoing eruption of the Soufrière Hills Volcano, Montserrat (1995–2007). *Journal of Volcanology and Geothermal Research*, *173*(3-4), pp.230-244.

Roman, D.C., 2009. Telica Seismic and Deformation Network. *International Federation of Digital Seismograph Networks.* (https://doi.org/10.7914/SN/6D_2009**)**.

Roman, D.C. and Power, J.A., 2011. Mechanism of the 1996–97 non-eruptive volcano-tectonic earthquake swarm at Iliamna Volcano, Alaska. *Bulletin of Volcanology*, *73*(2), pp.143-153.

Roman, D.C., Rodgers, M., Geirsson, H., LaFemina, P.C., Tenorio, V., 2016. Assessing the likelihood and magnitude of volcanic explosions based on seismic quiescence. Earth Planet. Sci. Lett. 450, 20–28.

Roman, D. C., LaFemina, P. C., Bussard, R., Stephens, K., Wauthier, C., Higgins, M., et al., 2019. Mechanisms of unrest and eruption at persistently restless volcanoes: Insights from the 2015 eruption of Telica Volcano, Nicaragua. *Geochemistry, Geophysics, Geosystems*, 20. https://doi.org/10.1029/2019GC008450.

Ruano, A.E., Madureira, G., Barros, O., Khosravani, H.R., Ruano, M.G. and Ferreira, P.M., 2014. Seismic detection using support vector machines. *Neurocomputing*, *135*, pp.273-283.

Saballos, J.A., Conde, V., Malservisi, R., Connor, C.B., Alvarez, J. and Munoz, A., 2014. Relatively short-term correlation among deformation, degassing, and seismicity: a case study from Concepción volcano, Nicaragua. *Bulletin of Volcanology*, *76*(8), pp.1-11.

SAC, 2004. *Assessment of the Hazards and Risks associated with the Soufrière Hills Volcano, Montserrat. Part II*. Scientific Advisory Committee Technical Report.

Sahara, D.P., Rahsetyo, P.P., Nugraha, A.D., Syahbana, D.K., Widiyantoro, S., Zulfakriza, Z., Ardianto, A., Baskara, A.W., Rosalia, S., Martanto, M. and Afif, H., 2021. Use of Local Seismic Network in Analysis of Volcano-Tectonic (VT) Events Preceding the 2017 Agung Volcano Eruption (Bali, Indonesia). *Frontiers in Earth Science*, *9*.

Scarpetta, S., Giudicepietro, F., Ezin, E.C., Petrosino, S., Del Pezzo, E., Martini, M. and Marinaro, M., 2005. Automatic classification of seismic signals at Mt. Vesuvius volcano, Italy, using neural networks. *Bulletin of the Seismological Society of America*, *95*(1), pp.185-196.

Schafer, J.L. and Graham, J.W., 2002. Missing data: our view of the state of the art. *Psychological methods*, 7(2), p.147.

Segall, P., 2013. Volcano deformation and eruption forecasting. *Geological Society, London, Special Publications*, *380*(1), pp.85-106.

Sener, O. and Savarese, S., 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Settles, B. Active learning literature survey. Technical report, University of Wisconsin-Madison, Department of Computer Sciences, 2009.

Shamout, F.E., Zhu, T., Sharma, P., Watkinson, P.J. and Clifton, D.A., 2019. Deep interpretable early warning system for the detection of clinical deterioration. *IEEE journal of biomedical and health informatics*, *24*(2), pp.437-446.

Shickel, B., Tighe, P.J., Bihorac, A. and Rashidi, P., 2017. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics*, *22*(5), pp.1589-1604.

Simkin, T. and Siebert, L., 1994. *Volcanoes of the World*. Geosciences Press, Inc. Tusson.

Siebert, L., Simkin, T. and Kimberly, P., 2011. Volcanoes of the World. Univ of California Press.

Siebert, L., Cottrell, E., Venzke, E. and Andrews, B., 2015. Earth's volcanoes and their eruptions: an overview. In *The Encyclopedia of Volcanoes* (pp. 239-255). Academic Press.

Smailagic, A., Costa, P., Noh, H.Y., Walawalkar, D., Khandelwal, K., Galdran, A., Mirshekari, M., Fagert, J., Xu, S., Zhang, P. and Campilho, A., 2018, December. Medal: Accurate and robust deep active learning for medical image analysis. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)* (pp. 481-488). IEEE.

Smith, R. and Kilburn, C.R.J., 2010. Forecasting eruptions after long repose intervals from accelerating rates of rock fracture: the June 1991 eruption of Mount Pinatubo, Philippines. *Journal of Volcanology and Geothermal Research*, *191*(1-2), pp.129-136.

Soderland, S., 1999. Learning information extraction rules for semi-structured and free text. *Machine learning*, *34*(1), pp.233-272.

Sparks, R.S.J., 2003. Forecasting volcanic eruptions. *Earth and Planetary Science Letters*, 210(1-2), pp.1-15.

Sparks, R.S.J. and Aspinall, W., 2004. Volcanic activity: frontiers and challenges in forecasting, prediction and risk assessment. *Geophysical Monograph*, 150, 359-373.

Sparks, R.S.J., Biggs, J. and Neuberg, J.W., 2012. Monitoring volcanoes. *Science*, 335(6074), pp.1310-1311.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), pp.1929-1958.

Stix, J., 2007. Stability and instability of quiescently active volcanoes: The case of Masaya, Nicaragua. *Geology*, 35(6), pp.535-538.

Stix, J., 2018. Understanding fast and slow unrest at volcanoes and implications for eruption forecasting. *Frontiers in Earth Science*, *6*, p.56.

Symonds, R.B., Gerlach, T.M. and Reed, M.H., 2001. Magmatic gas scrubbing: implications for volcano monitoring. *Journal of Volcanology and Geothermal Research*, 108(1-4), pp.303-341.

Syracuse, E.M., Thurber, C.H. and Power, J.A., 2011. The Augustine magmatic system as revealed by seismic tomography and relocated earthquake hypocenters from 1994 through 2009. *Journal of Geophysical Research: Solid Earth*, *116*(B9).

Thelen, W.A., Allstadt, K., De Angelis, S., Malone, S.D., Moran, S.C. and Vidale, J., 2013. Shallow repeating seismic events under an alpine glacier at Mount Rainier, Washington, USA. *Journal of Glaciology*, *59*(214), pp.345-356.

Thomas, M.E. and Neuberg, J., 2012. What makes a volcano tick—A first explanation of deep multiple seismic sources in ascending magma. *Geology*, *40*(4), pp.351-354.

Thompson G. and Reyes, C., (2018). GISMO - a seismic data analysis toolbox for MATLAB (Version 1.20b) [software package]. Available at: http://geoscience-community-codes.github.io/GISMO/. Accessed: Jan 05, 2020. http://doi.org/10.5281/zenodo.1404723

Thompson, G., Power, J.A., Braunmiller, J., Lockhart, A.B., Lynch, L., McCausland, W., Rowe, C.A., Shea, T., White, R.A. and Breithaupt, C.I., 2020. Capturing, Preserving, and Digitizing Legacy Seismic Data from the Montserrat Volcano Observatory Analog Seismic Network, July 1995–December 2004. *Seismological Research Letters*, *91*(4), pp.2127-2140.

Tilling, R.I., 2009. El Chichón's" surprise" eruption in 1982: Lessons for reducing volcano risk. *Geofísica internacional,* 48(1), pp.3-19.

Trnkoczy, A., 2009. Understanding and parameter setting of STA/LTA trigger algorithm. In *New Manual of Seismological Observatory Practice (NMSOP)* (pp. 1-20). Deutsches GeoForschungsZentrum GFZ.

Tuia, D., Ratle, F., Pacifici, F., Kanevski, M.F. and Emery, W.J., 2009. Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(7), pp.2218-2232.

U.S. Geological Survey, 2021, Earthquake Lists, Maps, and Statistics, accessed March 19, 2021 at URL https://www.usgs.gov/natural-hazards/earthquake-hazards/lists-maps-and-statistics

Ueki, K., Hino, H. and Kuwatani, T., 2018. Geochemical discrimination and characteristics of magmatic tectonic settings: A machine-learning-based approach. *Geochemistry, Geophysics, Geosystems*, *19*(4), pp.1327-1347.

VAAC, 2013. Washington VAAC 2013 Volcano Ash Advisory Archive – Satellite Products and Services Division / Office of Satellite and Product Operations, Washington Volcanic Ash Advisory Centre, <https://www.ssd.noaa.gov/VAAC/ARCH13/archive.html>.

Valade, S., Ley, A., Massimetti, F., D'Hondt, O., Laiolo, M., Coppola, D., Loibl, D., Hellwich, O. and Walter, T.R., 2019. Towards global volcano monitoring using multisensor sentinel missions and artificial intelligence: The MOUNTS monitoring system. Remote Sensing, 11(13), p.1528.

Vallance, J.W., Bull, K.F., and Coombs, M.L., 2010, *Pyroclastic flows, lahars, and mixed avalanches generated during the 2006 eruption of Augustine Volcano*, chapter 10 of Power, J.A., Coombs, M.L., and Freymueller, J.T., eds., *The 2006 eruption of Augustine Volcano, Alaska: U.S. Geological Survey Professional Paper 1769*, p. 219–267 [https://pubs.usgs.gov/pp/1769/chapters/p1769_chapter10.pdf].

Van der Maaten, L. and Hinton, G., 2008. Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

Voight, B., 1988. A method for prediction of volcanic eruptions. *Nature,* 332(6160), pp.125-130.

Voight, B. and Cornelius, R.R., 1991. Prospects for eruption prediction in near real-time. Nature, 350(6320), pp.695-698.

Wadge, G., 1981. The variation of magma discharge during basaltic eruptions. *Journal of Volcanology and Geothermal Research*, *11*(2-4), pp.139-168.

Wadge, G., Voight, B., Sparks, R.S.J., Cole, P.D., Loughlin, S.C. and Robertson, R.E.A., 2014. An overview of the eruption of Soufriere Hills Volcano, Montserrat from 2000 to 2010. *Geological Society, London, Memoirs*, 39(1), pp.1-40.

Webster, J.D., Mandeville, C.W., Goldoff, Beth, Coombs, M.L., and Tappen, Christine, 2010, *Augustine Volcano; the influence of volatile components in magmas erupted A.D. 2006 to 2,100 years before present*, chapter 16 of Power, J.A., Coombs, M.L., and Freymueller, J.T., eds., *The 2006 eruption of Augustine Volcano, Alaska: U.S. Geological Survey Professional Paper 1769*, p. 383–423 [https://pubs.usgs.gov/pp/1769/chapters/p1769_chapter16.pdf].

Werner, C.A., Doukas, M.P. and Kelly, P.J., 2011. Gas emissions from failed and actual eruptions from Cook Inlet Volcanoes, Alaska, 1989–2006. *Bulletin of Volcanology*, *73*(2), pp.155-173.

Wessel, P., Luis, J.F., Uieda, L., Scharroo, R., Wobbe, F., Smith, W.H.F. and Tian, D., 2019. The generic mapping tools version 6. *Geochemistry, Geophysics, Geosystems*, 20(11), pp.5556-5564.

White, R. and McCausland, W., 2016. Volcano-tectonic earthquakes: A new tool for estimating intrusive volumes and forecasting eruptions. *Journal of Volcanology and Geothermal Research*, *309*, pp.139-155.

White, R.A. and McCausland, W.A., 2019. A process-based model of pre-eruption seismicity patterns and its use for eruption forecasting at dormant stratovolcanoes. *Journal of Volcanology and Geothermal Research*, *382*, pp.267-297.

Whitehead, M.G. and Bebbington, M.S., 2021. Method selection in short-term eruption forecasting. *Journal of Volcanology and Geothermal Research*, p.107386.

Williams, C.K. and Rasmussen, C.E., 2006. *Gaussian processes for machine learning* (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.

Winson, A.E., Costa, F., Newhall, C.G. and Woo, G., 2014. An analysis of the issuance of volcanic alert levels during volcanic crises. *Journal of Applied Volcanology,* 3(1), p.14.

Witsil, A.J. and Johnson, J.B., 2020a. Analyzing continuous infrasound from Stromboli volcano, Italy using unsupervised machine learning. *Computers & Geosciences*, *140*, p.104494.

Witsil, A.J. and Johnson, J.B., 2020b. Volcano video data characterized and classified using computer vision and machine learning algorithms. *Geoscience Frontiers*, *11*(5), pp.1789-1803.

Wright, H.M., Pallister, J.S., McCausland, W.A., Griswold, J.P., Andreastuti, S., Budianto, A., Primulyana, S., Gunawan, H., Battaglia, M., Diefenbach, A. and Griswold, J., 2019. Construction of probabilistic event trees for eruption forecasting at Sinabung volcano, Indonesia 2013–14. *Journal of Volcanology and Geothermal Research*, *382*, pp.233-252.

Zhang, N., Xiong, J., Zhong, J. and Leatham, K., 2018, June. Gaussian process regression method for classification for high-dimensional data with limited samples. In *2018 Eighth International Conference on Information science and technology (ICIST)* (pp. 358-363). IEEE.

Zumberge, J.F., Heflin, M.B., Jefferson, D.C., Watkins, M.M. and Webb, F.H., 1997. Precise point positioning for the efficient and robust analysis of GPS data from large networks. *Journal of Geophysical Research: Solid Earth*, *102*(B3), pp.5005-5017.

Appendices

Appendix A:

List of features used as inputs to the models discussed in Chapter 3. Δ represents change in the quantity from the previous day.

A.1) Nevado del Ruiz features

1. Event rate (all types)
2. Δ Event rate (all types)
3. Mean of dominant frequencies
4. Variance of dominant frequencies
5. 90th percentile of dominant frequencies
6. 10th percentile of dominant frequencies
7. Minimum of dominant frequencies
8. Maximum of dominant frequencies
9. Median of dominant frequencies
10. Δ of dominant frequencies
11. Mean of band ratio
12. Variance of band ratio
13. 90th percentile of band ratio
14. 10th percentile of band ratio
15. Minimum of band ratio
16. Maximum of band ratio
17. Median of band ratio
18. Δ of band ratio
19. Mean of peak amplitudes
20. Variance of peak amplitudes

21. 90^{th} percentile of peak amplitudes
22. 10^{th} percentile of peak amplitudes
23. Minimum of peak amplitudes
24. Maximum of peak amplitudes
25. Median of peak amplitudes
26. Δ of peak amplitudes
27. Mean of waveform standard deviation
28. Variance of waveform standard deviation
29. 90^{th} percentile of waveform standard deviation
30. 10^{th} percentile of waveform standard deviation
31. Minimum of waveform standard deviation
32. Maximum of waveform standard deviation
33. Median of waveform standard deviation
34. Δ of waveform standard deviation
35. Multiplet rate
36. Δ of multiplet rate

A.2) Telica features

1. Event rate (all types)
2. Δ Event rate (all types)
3. Event rate (High Frequency)
4. Δ Event rate (High Frequency)
5. Event rate (Low Frequency)
6. Δ Event rate (Low Frequency)

7. Mean of daily RSAM
8. Variance of daily RSAM
9. Δ of daily RSAM
10. Mean of band ratio
11. Variance of band ratio
12. 90^{th} percentile of band ratio
13. 10^{th} percentile of band ratio
14. Minimum of band ratio
15. Maximum of band ratio
16. Median of band ratio
17. Δ of band ratio
18. Mean of peak amplitudes
19. Variance of peak amplitudes
20. 90^{th} percentile of peak amplitudes
21. 10^{th} percentile of peak amplitudes
22. Minimum of peak amplitudes
23. Maximum of peak amplitudes
24. Median of peak amplitudes
25. Δ of peak amplitudes
26. Mean of waveform standard deviation
27. Variance of waveform standard deviation
28. 90^{th} percentile of waveform standard deviation
29. 10^{th} percentile of waveform standard deviation
30. Minimum of waveform standard deviation
31. Maximum of waveform standard deviation

32. Median of waveform standard deviation
33. Δ of waveform standard deviation
34. Mean of dominant frequencies
35. Variance of dominant frequencies
36. 90^{th} percentile of dominant frequencies
37. 10^{th} percentile of dominant frequencies
38. Minimum of dominant frequencies
39. Maximum of dominant frequencies
40. Median of dominant frequencies
41. Δ of dominant frequencies
42. Multiplet rate (top 150 families)
43. Δ multiplet rate (top 150 families)
44. Multiplet rate (all families)
45. Δ multiplet rate (all families)

Appendix B

B.1) Confirmed ash emission dates from Nevado del Ruiz

Summary of confirmed ash emission dates from NdR recorded by Londoño and Gavis, 2018 and SGC observatory reports from 2012 – 2015.

Confirmed day of emission	**Confirmed week of emission**
08/03/2012	08/05/2012
19/04/2012	15/05/2012
10/05/2012	22/05/2012
22/05/2012	29/05/2012
29/05/2012	11/09/2012
09/06/2012	18/09/2012
10/07/2012	25/09/2012
30/07/2012	02/10/2012
14/08/2012	09/10/2012
01/09/2012	12/10/2012
22/10/2012	23/10/2012
30/10/2012	30/10/2012
24/05/2013	06/11/2012
27/05/2013	07/11/2012
30/06/2013	13/11/2012
18/11/2014	20/11/2012
19/11/2014	27/11/2012
20/11/2014	04/12/2012
21/11/2014	11/12/2012
28/11/2014	17/12/2012
29/11/2014	24/12/2012
01/01/2015	31/12/2012

02/01/2015	07/01/2013
03/01/2015	15/01/2013
04/01/2015	22/01/2013
05/01/2015	12/02/2013
06/01/2015	19/02/2013
07/01/2015	26/02/2013
09/01/2015	04/03/2013
10/01/2015	19/03/2013
14/01/2015	15/04/2013
15/01/2015	23/04/2013
16/01/2015	30/04/2013
17/01/2015	07/05/2013
20/01/2015	14/05/2013
09/02/2015	21/05/2013
	28/05/2013
	04/06/2013
	11/06/2013
	17/06/2013
	24/06/2013
	02/07/2013
	09/07/2013
	16/07/2013
	30/07/2013
	06/08/2013
	24/12/2013
	18/02/2014
	08/07/2014
	15/07/2014
	25/11/2014

	02/12/2014
	09/12/2014
	16/12/2014
	23/12/2014
	30/12/2014
	27/01/2015
	03/02/2015
	10/02/2015
	17/02/2015
	24/02/2015
	03/03/2015

Appendix C

C.1) List of seismic stations used in analysis

Station
AUE
AUH
AUI
AUL
AUNW
AUP
AUR
AUS
AUSE
AUW

C.2) List of GPS stations used in analysis

Station (Component)
AV01 (vertical)
AV02 (N)
AC59 (N)

C.3) List of seismic features used as input variables.

Features with Δ indicate the variation from one day to the next.

1. Event rate (all types)
2. Δ Event rate (all types)
3. Mean of dominant frequencies
4. Variance of dominant frequencies
5. 90^{th} percentile of dominant frequencies
6. 10^{th} percentile of dominant frequencies
7. Minimum of dominant frequencies
8. Maximum of dominant frequencies
9. Median of dominant frequencies
10. Δ of dominant frequencies
11. Mean of band ratio
12. Variance of band ratio
13. 90^{th} percentile of band ratio
14. 10^{th} percentile of band ratio
15. Minimum of band ratio
16. Maximum of band ratio
17. Median of band ratio
18. Δ of band ratio
19. Mean of peak amplitudes
20. Variance of peak amplitudes
21. 90^{th} percentile of peak amplitudes
22. 10^{th} percentile of peak amplitudes
23. Minimum of peak amplitudes
24. Maximum of peak amplitudes
25. Median of peak amplitudes

26. Δ of peak amplitudes
27. Mean of waveform standard deviation
28. Variance of spectral width of the largest peak of each event.
29. 90^{th} percentile of spectral width of the largest peak of each event.
30. 10^{th} percentile of spectral width of the largest peak of each event.
31. Minimum of spectral width of the largest peak of each event.
32. Maximum of spectral width of the largest peak of each event.
33. Median of spectral width of the largest peak of each event.
34. Δ of spectral width of the largest peak of each event.

C.4) List of geodetic features used as input variables

Features with Δ indicate the variation from one day to the next.

1. Vertical position of AV01
2. Δ of vertical component of displacement at AV01
3. N-S component of the baseline vector for AV02-AC59 station pair.
4. Δ of the N-S component of the baseline vector for AV02-AC59 station pair.

C.5) List of gas features used as input variables.

Features with Δ indicate the variation from one day to the next.

1. Interpolated SO_2 value
2. Δ of interpolated SO_2 value

C.6) Summary of classification results

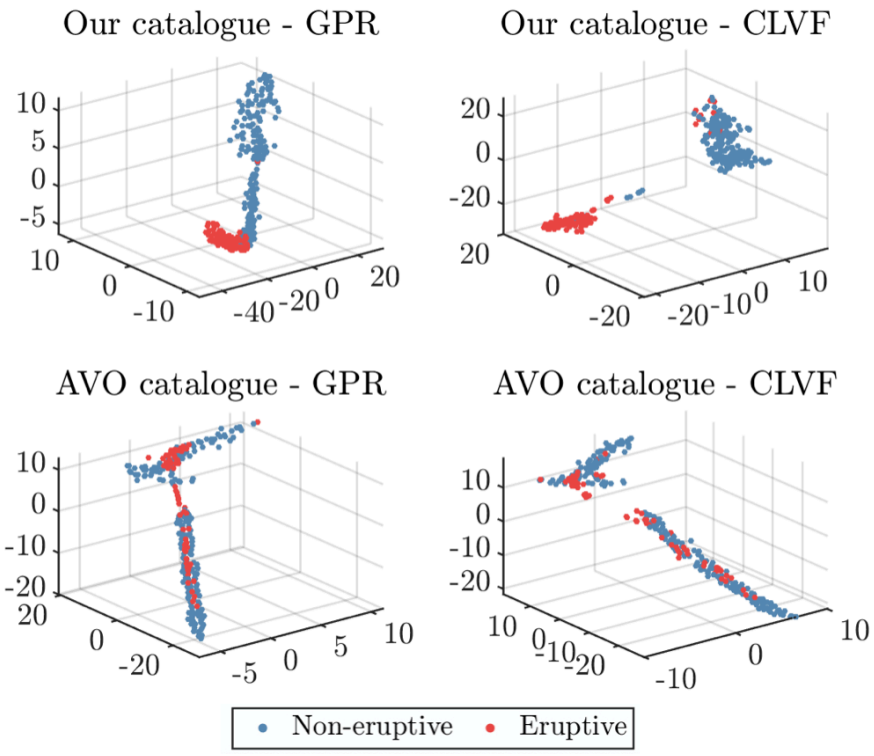
Results from both modelling approaches applied to features derived from Augustine data.

Method	**Gas interpolation method**	**Eruption start-date (GVP)**	**Eruption end-date (GVP)**	**Start-date for first phase (model)**	**End-date for first phase (model)**	**Model agreement with GVP labels[1]**
Novelty detection	GPR	9th December 2005	27th April 2006	Late-October 2005	N/A (after May 2006)	58%
	CLVF			Mid-November 2005	N/A (after May 2006)	84%
Random forest	GPR			Mid-November 2005	Early May 2006	82%
	CLVF			Early December 2005	Early-Mid May 2006	86%

[1]Model agreement is calculated by dividing the number of classified days which match the GVP classifications by the total number of classified days, expressed as a percentage.

C.7) Three-dimensional visualisation of the seismic, gas flux and geodetic features.

Three-dimensional visualisation of the seismic, gas flux and geodetic features derived from Augustine using the t-distributed stochastic nearest neighbour embedding, where blue points represent non-eruptive days and red points represent eruptive days according to the GVP database. The top two panels, which include seismic features from our seismic catalogue, display a greater separation between non-eruptive and eruptive days than the bottom two panels, which include seismic features from the AVO analyst-reviewed catalogue. The x and y axes represent non-linearly scaled distances projected into a three-dimensional space and thus have no units, and are not comparable across axes.



C.8) Single-parameter novelty detection models

Novelty detection models formulated using only seismic, geodetic, CLVF-derived gas or GPR-derived gas values respectively. Green line denotes the training data and blue line denotes the testing data. The grey horizontal line represents the novelty threshold determined during model training. The black bar at the top of each panel represents the dates on which eruption was occurring according to the GVP database. Novelty score is dimensionless, therefore the y-axis has no associated units but note the different y-axis scales between the top and the bottom panels. Y-scales are not consistent between panels.

