Ruapehu Eruption Forecasting: Classifying Anthropogenic and Weather Noise

Q. Gao and T. Waldin

Final Year Projects, 2021
Dept. of Civil and Natural Resources Engineering
University of Canterbury
Project supervisor(s): D. Dempsey

Keywords: *Time-series, classification, eruption, forecasting.*

ABSTRACT

Mount Ruapehu is an active volcano in the North Island of New Zealand. Hydrothermal eruptions and phreatomagmatic explosions can happen rapidly and without warning, with destructive and potentially fatal consequences. It has been shown that a structured machine learning approach can detect eruption precursors in real-time seismic data from Whakaari, another volcano in New Zealand. However, unlike Whakaari, there are ski fields on Ruapehu, and it also experiences a greater frequency and intensity of storms. Both phenomena generate noise that can contaminate seismic signals, obscuring eruption precursors and confusing automated algorithms. We adapted a machine learning model to recognize storm and chairlift noise in seismic signals. The noise classifying models we trained could identify unseen chairlift and storm events with up to 95% certainty. Even the worst classification results still had a certainty of approximately 85%, suggesting the models were operating reasonably well. When pre-eruption unrest is present in the data, neither model recognised it as a chairlift or storm event, suggesting that the model will not falsely classify the unrest periods that are utilised by eruption forecasters. The next steps for this project could be to combine the two models into one surface noise classification model, test the model on real-time data and if successful, implement the model in an automated eruption forecaster at Ruapehu.

1. INTRODUCTION

Ruapehu is an active volcano located in Tongariro National Park in the North Island of New Zealand. The mountain and surrounding national park are utilised by people for tourism and recreation. Whakapapa and Turoa ski fields have capacity for up to 5500 skiers and snowboarders each (Stuff, 2019). A third, smaller ski field, Tukino, also operates on the mountain. Figure 1 shows a map of the mountain with all relevant locations and features labelled, including the GeoNet seismic station denoted FWVZ used in this study.

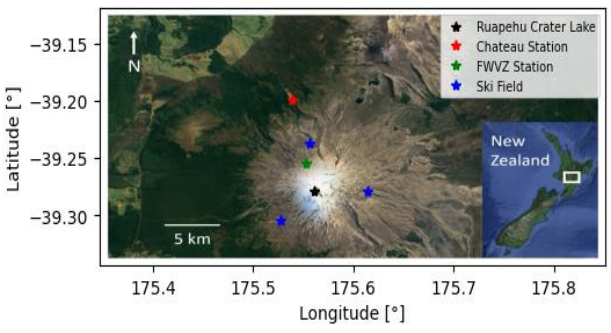


Figure 1. A map of Ruapehu with relevant locations and features labelled (Google Earth, 2021).

There have been several eruptions at Ruapehu during the past few decades, notably in 1995 - 1996, 2006 and 2007 (Waikato Regional Council, 2019). Past eruptions have spread ash, rocks, and water throughout and beyond the summit area and caused lahars. Lahars are flowing slurries of mud and debris that can be extremely destructive (Kilgour et al, 2010).

The 1995-1996 eruptions spread ash as far away as Auckland and Wellington, closing airports (Te Ara, 2006). The closure of Whakapapa, Turoa, and Tukino ski fields caused an estimated $100 million loss in revenue (Te Ara, 2006). The 2006 eruption was speculated to be entirely underwater and did not cause significant damage (A. Mordret et al, 2010). The most recent eruption was on the 25th of September 2007. The eruption produced lahars in two valleys, one in the Whakapapa ski field where several people were injured (Wunderman, 2007).

The effective communication of eruption forecasting results is a critical aspect of an early warning system (Potter et al, 2014). In New Zealand, a volcanic alert level (VAL) system is used, where VAL 0 represents no unrest, VAL 1-2 represent states of unrest, and VAL 3-5 represent eruptive events of varying severity. At the time of writing (October 2021), Whakaari is at VAL 2,

with moderate to heightened unrest, and Ruapehu is at VAL 1, with minor unrest (GeoNet, 2021). Ruapehu was last raised to VAL 2 in December 2020, which prompted an exclusion zone of a 2 km radius from the centre of the crater lake (DOC, 2020). VAL does well to classify the current state of the volcano, however, it does not provide a forecast of when or how likely the volcano is to erupt.

The material Failure Forecast Method (FFM) is a common technique that can provide a forecast time for an eruption. (Chardot et al, 2015). However, this method provides a prediction of when an eruption is most likely to occur instead of a forecast of how likely an eruption is to occur.

It has been shown that a structured machine learning approach can detect eruption precursors in real-time seismic data from Whakaari, another volcano in New Zealand (Dempsey et al, 2020). Similar seismic data can be obtained for Ruapehu from GeoNet's FWVZ station located at the base of the Far West T-Bar. An example of these signals leading up to the 2006 eruption is shown in Figure 2.

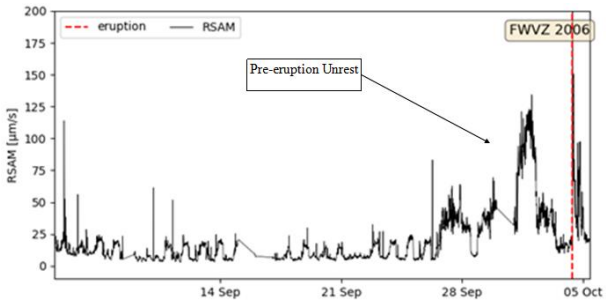


Figure 2. A plot showing the Real-time Seismic Amplitude Measures (RSAM) leading up to an eruption. The pre-eruption unrest is labelled.

The approach at Whakaari involved a machine learning model, based on time-series feature engineering. The model was trained to identify time-series features in the data, such as means, standard deviation, standard error and slope of a linear regressor, and Fourier coefficients. The model is designed to record and analyse these features, so it can then automatically identify similar trends of features and forecast in real-time data. (Dempsey et al, 2020).

Unlike Whakaari, there are three ski fields on Ruapehu. Ruapehu also experiences a greater frequency and intensity of storms than Whakaari. The vibrations caused by chairlift operation and the increased noise due to adverse weather can contaminate seismic signals. This contamination can obscure precursors and confuse automated algorithms. This project aims to adapt and train the machine learning model developed by Dempsey et al (2020) to identify and classify noise caused by weather and chairlift operation on Ruapehu.

2. METHODOLOGY

2.1. Data

Four sets of data were obtained from GeoNet's FWVZ station, each containing a set of values averaged over 10-minute intervals. These values exist for every 10-minute period from January 2006 to July 2021. Real-time Seismic Amplitude Measures, or RSAM, measures the relative intensity of vibrations in a particular frequency band, and has units of μms^{-1}. Separate data sets containing high frequency (HF) and medium frequency (MF) bands are available with the same units. The latter measures are useful as surface noise tends to have higher frequency seismic vibrations. The fourth set of data was the Displacement Seismic Amplitude Ratio or DSAR. DSAR was computed by integrating the MF and HF signals and computing the average absolute signal over 10-minute windows (Dempsey et al, 2020). The ratio of the two quantities was taken as the DSAR and is therefore unitless. Each set of data had a corresponding data set with outlying peaks filtered out, as these usually represented earthquakes which are independent of weather or anthropogenic noise. These filtered sets of data were the sets used in this study.

Vibrations that were generated by the operation of chairlifts were identified as a significant increase in HF that was consistent for approximately 8 hours (Figure 3). As MF does not increase by as much, there is a reduction of DSAR (the ratio of MF to HF). Periods of chairlift operation, or chairlift 'events' in the data only occur during the ski season between June and October, and this makes them easy to identify.

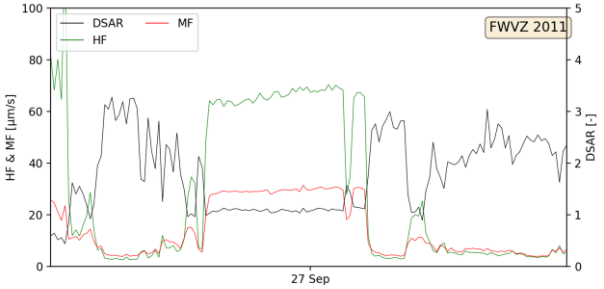


Figure 3. HF, MF, and DSAR data during a period of chairlift operation during September 2011.

Storm events do not generate the same distinctive shape as the vibrations from chairlifts, and they can occur at any time of year. To identify storms, we compared rainfall alongside the seismic data. Like the chairlift noise, an increase in HF was recorded, but it tended to increase gradually to a peak and then slowly decrease back to a normal value. MF also increased in this pattern but not as significantly, once again suppressing the DSAR signal. The peak in the HF and MF tended to match the peak in the rainfall as shown in Figure 4.

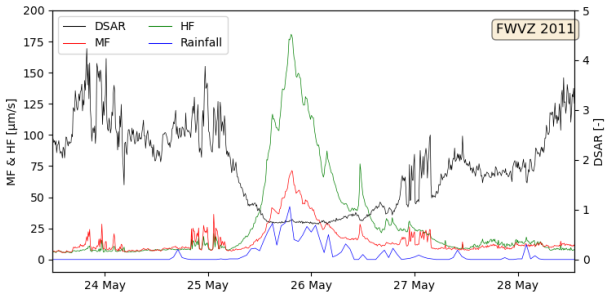


Figure 4. A plot of HF, MF, DSAR, and rainfall for a storm event in May 2011.

The seismic data shown in Figures 3 and 4 represent clear chairlift and storm events. A human can identify these, but for these classifications to be useful for a monitoring system they will need to be automated. This will require the training of classification models.

2.2. Modelling

The goal of modelling is to identify anthropogenic and weather noise automatically. The ski fields' chairlift noise can be identified by their regular occurrences during business hours, and the corroborating trends in other datasets can identify the noise of storms. Two separate modifications of the time-series pattern recognition algorithm model developed by Dempsey et al (2020) were trained to classify these events.

First, 20 storm and 20 chairlift periods were manually identified between 2011 and 2020. Of these, 15 were used to train classification models, denoted the 'in-sample' events. The excluded 5 events were considered 'out-of-sample' events and were used for validation of the models. A flowchart demonstrating the process of training and testing the models is presented in Figure 5.

The models are trained using time-series features, such as moving means or maximum values calculated within a window that slides over the data. The overall model is called a Random Forest, made up of many small models called Decision Trees. Each Decision Tree evaluates different combinations of time-series features. Both the Random Forest and Decision Trees are a class of model, called classifiers: their output assesses whether a particular combination of features is likely to be associated with a storm/chairlift event, or not. Increasing the number of classifiers increases the accuracy of the modelling but also increases the computation time. The chairlift classification model used 180 Decision Trees and the storm classification model used 100 Decision Trees.

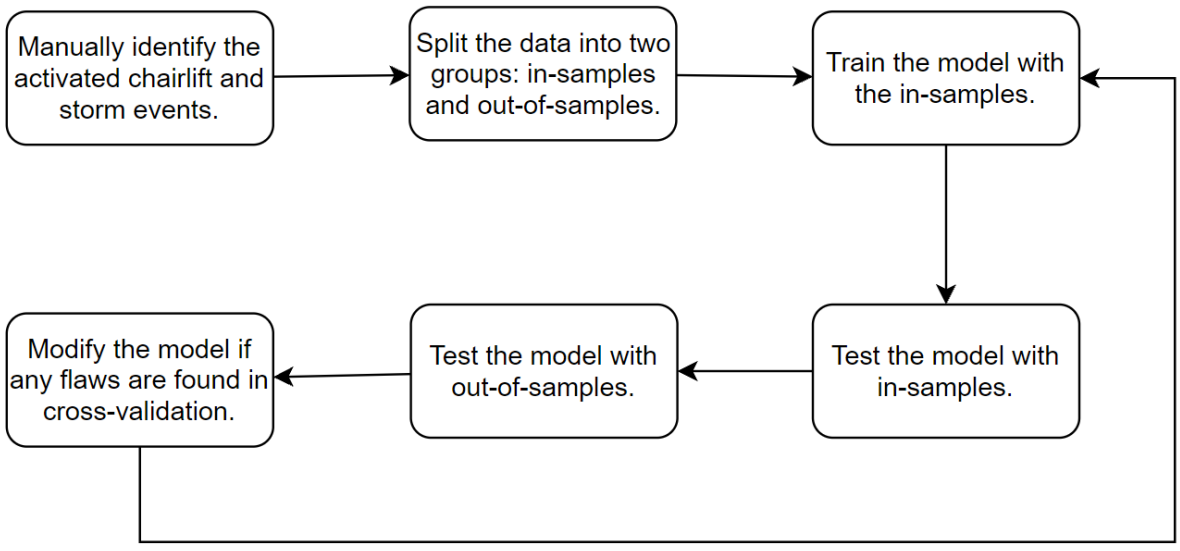


Figure 5. A flowchart demonstrating the process of training and testing the models.

The models were calibrated using the in-sample events, the results of which is shown in Figure 6.

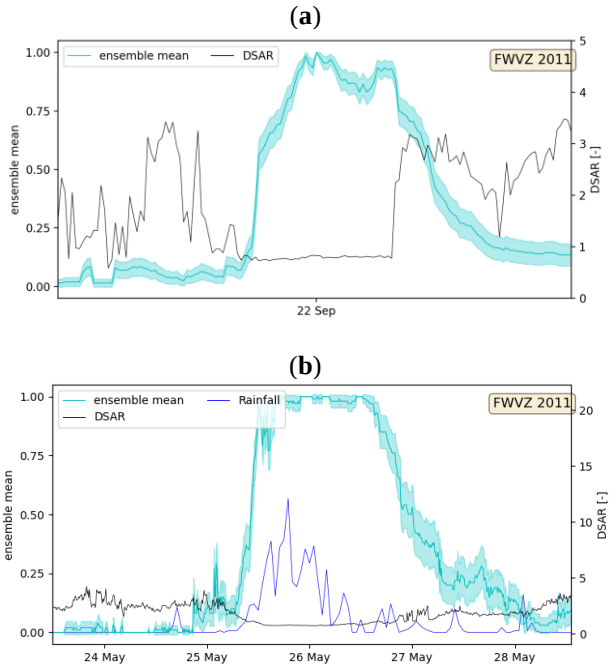


Figure 6. (**a**). The results of the classification model for an in-sample chairlift event. (**b**). The results of the classification model for an in-sample storm event. The blue line (with error envelope) is the output of the classification model. Values closer to 1 indicate the model is interpreting the data (black line) as a chairlift or storm event.

The models output by an ensemble mean, which is a number between 0 and 1 plotted on the left axes of Figure 6. The ensemble mean is a measure of certainty of the classifying model, with a 1 meaning an event (chairlift operation, storm) is definitely occurring and 0 meaning an event is definitely not occurring. Since the classification models were trained using the time-series features of in-sample events, we expect them to produce good results attempting to identify them. The ensemble means for the events shown in Figure 6 are between 0.85 and 1.00. This demonstrates that the model is well calibrated to its training data.

The ensemble mean can also be used with a threshold value. While the model is unlikely to be certain about the occurrence of an event, a suitable threshold can classify an event each time it is surpassed. For example, a threshold of 0.9 would cause the chairlift event in Figure 6a to be identified approximately 3 hours after it started, whereas a threshold of 1.0 would miss it completely. However, identifying a suitable threshold is beyond the scope of this project.

2.3. Validation

The models were tested on out-of-sample chairlift and storm events for validation purposes. A high accuracy in identifying anthropogenic and weather noise in an independent data set such as the out-of-sample events, would suggest the model can generalize to unseen data and is thus useful in operation. It is also important to see how the models perform leading up to eruptions to determine if they will falsely identify pre-eruption unrest as chairlift or storm noise.

3. RESULTS AND DISCUSSION

3.1. Model Output and Analysis

The models have been developed using the time-series features of 15 storm events and 15 chairlift events. The results of that analysis have been used to predict that a given out-of-sample event is a storm or chairlift event. The results of the models' performance on out-of-sample events are demonstrated in Figures 7 and 8.

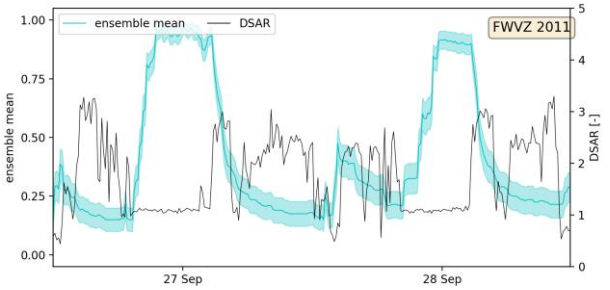


Figure 7. Performance of the chairlift classification model on an unseen chairlift event.

Figure 7 shows the performance of the model when given data from two out-of-sample chairlift events on consecutive days. The chairlift events can be visually identified by the DSAR suppression in the distinctive 8-hour pattern. The ensemble mean increases to above 0.85 during the period of the chairlift events and drops to 0.25 when the chairlifts are not operating. This would suggest the model is successful in this case. The other 5 out-of-samples chairlift events were classified with ensemble means ranging between approximately 0.7 to 0.95, with all ensemble means reaching at least 0.85.

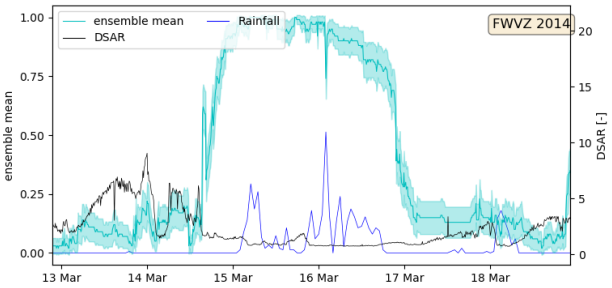


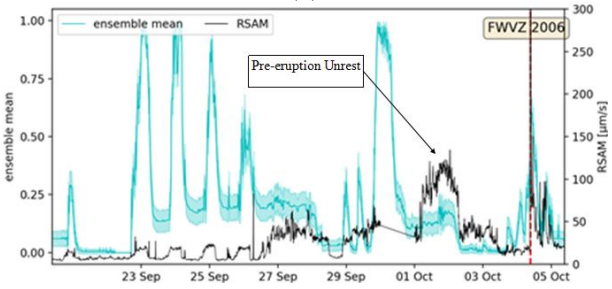
Figure 8. The results of the storm classification model on an unseen storm event.

Figure 8 shows the performance of the model classifying an out-of-sample storm event. The storm event is clear due to the suppression of the DSAR and the increased rainfall. During this period the ensemble mean increases to values between approximately 0.8 and 0.95 for the duration of the event.

The ensemble mean stays below 0.25 when the storm event is not occurring. This would suggest the model is successful at classifying the storm. The other 4 out-of-sample storm events were classified with ensemble means ranging from approximately 0.75 to 0.95, with all ensemble means reaching at least 0.85.

As demonstrated, the models appear to be successful for classifying chairlift and storm noise. However, it is also important that the models do not misclassify eruption precursors as well. Figures 9 and 10 show the models' performance during a period of two weeks leading up to eruptions.

(**a**)



(**b**)

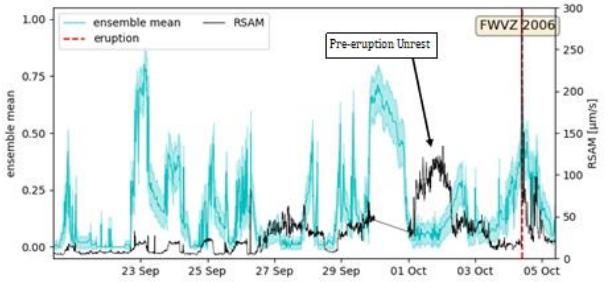
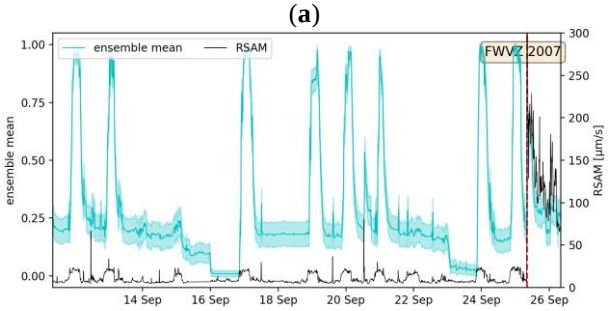


Figure 9. (**a**). Performance of the chairlift model on the October 2006 Eruption. (**b**). Performance of the storm model on the October 2006 eruption.

The October 2006 eruption displayed clear pre-eruption unrest in the RSAM signals in the weeks prior to the eruption. The important feature of these plots is that neither model identifies the unrest as a storm or chairlift event in this case which is successful. Successful classification of chairlift operation can be seen from the 23rd to the 26th of September.

(**a**)



(**b**)

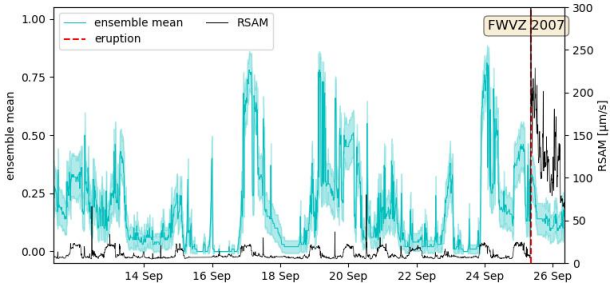


Figure 10. (**a**). Performance of the chairlift model on the September 2007 eruption. (**b**). Performance of the storm model on the September 2007 eruption.

The September 2011 eruption did not have any clear unrest in the RSAM signal leading up to the eruption. All high ensemble means appear to be correctly classifying chairlift and storm noise, so the model should not misclassify any pre-eruption unrest. A potential limitation of the model would be classification of storms or periods of chairlift operation that occur simultaneously to pre-eruption unrest. This is not a situation that could be easily tested as there is a limited amount of pre-eruption data available and it is difficult to predict how the models will respond. An automated system could identify chairlift or storm noise is contaminating the signal and remove signals that are related to eruption precursors at the same time, creating potential for confusion.

As different models were used for chairlift and storm noise, it was important to see how the models performed on the events that they were not trained to identify. The results of the chairlift model on a storm event and vice versa are demonstrated in the Figures 11 and 12.

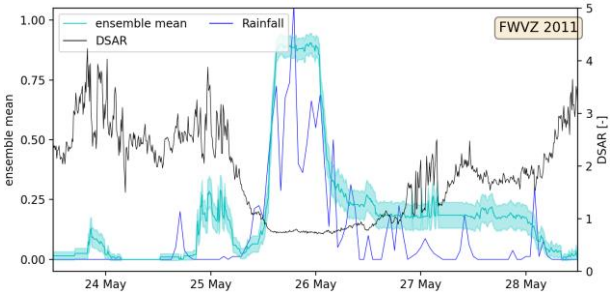


Figure 11. The results of the chairlift classification model on an unseen storm event.

Figure 11 shows the chairlift classification model on a storm event which can be identified by the rainfall increase and the DSAR suppression. This model is not performing as expected as it identifies a storm event as a chairlift event with an ensemble mean of approximately 0.85. While this highlights a limitation of the model, it does not necessarily invalidate the approach, because multiple classification models could operate in parallel.

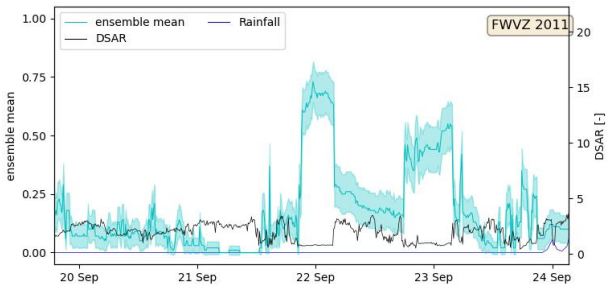


Figure 12. The results of the storm classification model on an unseen chairlift event.

Figure 12 contains two chairlift events which can be identified by the periods of DSAR suppression on the 22nd and the 23rd of September. The ensemble mean clearly increases, suggesting that both models are using several common time-series features to classifies storm and chairlift events. However, it doesn't reach a value greater than 0.75. This suggests this model is successful in this regard and shouldn't falsely identify chairlift events as storms with a suitable threshold in place.

The models have clear limitations. The chairlift classification model falsely identifying a storm with a high ensemble mean highlights this. There were also events that did not perform as well as expected, these are demonstrated in Figure 13.

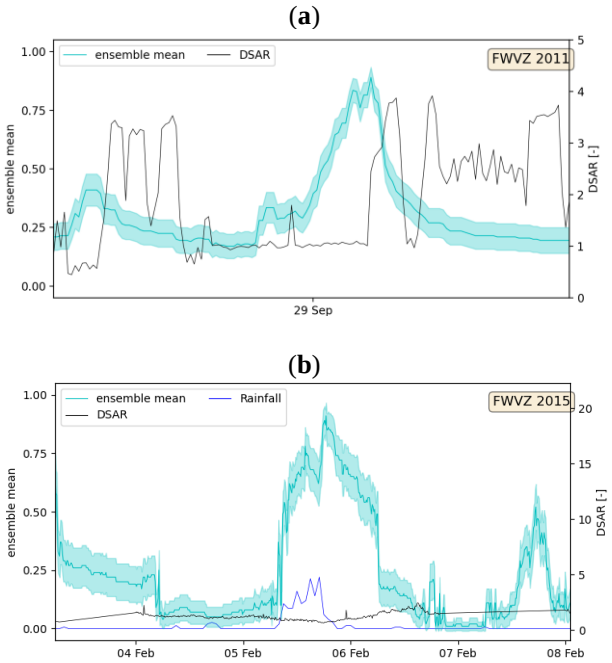


Figure 13. (**a**). Poor results of the classification model on an out-of-sample chairlift event. (**b**). Poor results of the classification model on an out-of-sample storm event.

While these examples did not perform as well as expected, the ensemble means still reach values of approximately 0.85, so would still be identified with a suitable threshold. However, the chairlift classification model performed better identifying the out-of-sample storm event shown in Figure 11, than it did at identifying the chairlift event in Figure 13a. A combined model designed to classify both storm and chairlift events simultaneously could be a next step to alleviate this limitation. This would also be a logical next step, as there is no reason an eruption forecaster would need to differentiate between noise generated by a storm or a period of chairlift operation, it would just need to be able to classify surface noise that could be contaminating seismic signals.

3.2. Analysis of Features

All but three classifiers that the chairlift model utilised used the 'index mass quantile' feature. This calculates a value of the time-series where a certain percentage of the mass lies left of the value. The top four most common features used were variations of this feature. The two most common were based on the MF time-series, and the third and fourth were based of the HF time-series. The Decision Trees would define a velocity to use as a threshold value that a certain percentage of the signals are above. An immediate significant increase in HF and MF to steady values are characteristic of chairlift noise and presumably the threshold values would less than these steady values. The classifiers that the storm model utilised did not use this feature as frequently. Instead, the most common feature was the 'linear trend slope'. This calculates a linear regression for the values of the time-series. In this case, the slope of the regression of HF was the feature used. This is unusual, as over a sliding window, the shape of the plot does not suggest that this feature would be useful. Further investigation could reveal why this feature was the most common. The feature used most frequently in both models was the 'index mass quantile'. Interestingly, features regarding DSAR were not prevalent in the models, with it only showing up in approximately 50% of the classifiers. This is interesting, as the DSAR suppression is such a clear signal visually and this is not the case for the machine learning algorithm.

4. CONCLUSIONS AND RECOMMENDATIONS

The machine learning model that we adapted to recognize storm and chairlift noise in seismic signals was able to classify these events with reasonable certainty. The noise classifying models we trained could identify unseen chairlift and storm events with up to 95% certainty. Even the worst classification results still had a certainty of approximately 85%, suggesting the models were operating reasonably well. When pre-eruption unrest is present in the data, neither model recognised it as a chairlift or storm event, suggesting that the model will not falsely classify the unrest periods that are utilised by eruption forecasters.

The next steps for this project could be to combine the two models into one surface noise classification model, test the model on real-time data and if successful,

implement the model in an automated eruption forecaster at Ruapehu.

Using one model for both classifications is logical, as there is no reason an eruption forecaster would need to differentiate between noise generated by a storm or a period of chairlift operation, it would just need to be able to classify surface noise that could be contaminating seismic signals. The models were found to share a lot of time-series features, so a combined model would likely function reasonably well. It also has the benefit of removing the potential for confusion of the chairlift model misclassifying a storm event, and vice versa.

Implementation of this model in an automated eruption forecasting system could reduce confusion from contaminated seismic signals and could substantially reduce the workload of human operators. The utility of this model could go beyond forecasting at Ruapehu and could be implemented at any active volcano that experiences severe weather or significant amounts of anthropogenic activity.

5. ACKNOWLEDGEMENTS

We would like to acknowledge D. Dempsey and A. Ardid for their significant contribution and assistance in this project. Also, to Dempsey et al (2020) for developing the machine learning algorithm that was adapted in this project. We would also like to acknowledge GeoNet for the seismic data used to train the model and NIWA for the rainfall data that helped us to identify storm events.

6. REFERENCES

Brooker, L. (2019). “Increasing visitor numbers on Mt Ruapehu proves a challenge to keep skiers and daytrippers happy.” *Stuff Ltd.*

Google LLC. (2021). Map retrieved from Google Earth October 2021.

Waikato Regional Council. (2019). “Mount Ruapehu Erupts.”

Kilgour, G., Manville, V., Della Pasqua, F., Graettinger, A., Hodgson, K. A., Jolly, G. E. (2010). “The 25 September 2007 eruption of Mount Ruapehu, New Zealand: Directed ballistics, surtseyan jets, and ice-slurry lahars.”

McSaveney, E., Steward, C., Leonard, G. (2006). “Historic volcanic activity - Ruapehu since 1945.” *Te Ara - the Encyclopedia of New Zealand.* (6)

Mordret, A., Jolly, A. D., Duputel, Z., Fournier, N. (2010). "Monitoring of phreatic eruptions using Interferometry on Retrieved Cross-Correlation Function from Ambient Seismic Noise: Results from Mt. Ruapehu, New Zealand.”

Wunderman, R. (2007). “Report on Ruapehu (New Zealand).” *Bulletin of the Global Volcanism Network*, 32(11), Smithsonian Institution.

Potter, S. H., Jolly, G. E., Neall, V. E., Johnson, D. M., Scott, B. J. (2014). “Communicating the status of volcanic activity: revising New Zealand’s volcanic alert level system.”

GeoNet. (2021). Volcanic alert levels retrieved October 2021.

Department of Conservation. (2020). “Mount Ruapehu increases to volcanic alert level 2.”

Chardot, L., Jolly, A. D., Kennedy, B. M., Fournier, N., Sherburn, S. (2015). “Using volcanic tremor for eruption forecasting at White Island volcano (Whakaari), New Zealand.”

Dempsey, D. E., Cronin, S. J., Mei, S., Kempa-Liehr, A. W. (2020). “Automatic precursor recognition and real-time forecasting of sudden explosive volcanic eruptions at Whakaari, New Zealand.”