Deep Modular Multimodal Fusion on Multiple Sensors for Volcano Activity Recognition

Hiep V. Le[1](✉), Tsuyoshi Murata[1], and Masato Iguchi[2]

[1] Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan
hiep@net.c.titech.ac.jp, murata@c.titech.ac.jp
[2] Disaster Prevention Research Institute, Kyoto University, Kyoto, Japan
iguchi.masato.8m@kyoto-u.ac.jp

Abstract. Nowadays, with the development of sensor techniques and the growth in a number of volcanic monitoring systems, more and more data about volcanic sensor signals are gathered. This results in a need for mining these data to study the mechanism of the volcanic eruption. This paper focuses on Volcano Activity Recognition (VAR) where the inputs are multiple sensor data obtained from the volcanic monitoring system in the form of time series. And the output of this research is the volcano status which is either *explosive* or *not explosive*. It is hard even for experts to extract handcrafted features from these time series. To solve this problem, we propose a deep neural network architecture called VolNet which adapts Convolutional Neural Network for each time series to extract non-handcrafted feature representation which is considered powerful to discriminate between classes. By taking advantages of VolNet as a building block, we propose a simple but effective fusion model called Deep Modular Multimodal Fusion (DMMF) which adapts data grouping as the guidance to design the architecture of fusion model. Different from conventional multimodal fusion where the features are concatenated all at once at the fusion step, DMMF fuses relevant modalities in different modules separately in a hierarchical fashion. We conducted extensive experiments to demonstrate the effectiveness of VolNet and DMMF on the volcanic sensor datasets obtained from Sakurajima volcano, which are the biggest volcanic sensor datasets in Japan. The experiments showed that DMMF outperformed the current state-of-the-art fusion model with the increase of F-score up to 1.9% on average.

Keywords: Multimodal fusion · Volcano Activity Recognition
Time series · Convolutional Neural Network

1 Introduction

Volcanic eruption causes severe damage to human and society, hence it is one of the main concerns of many people in the world, especially to volcano experts.

U. Brefeld et al. (Eds.): ECML PKDD 2018, LNAI 11053, pp. 602–617, 2019.
https://doi.org/10.1007/978-3-030-10997-4_37

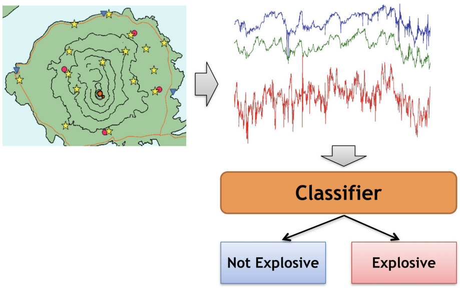

Fig. 1. The overview of Volcano Activity Recognition (VAR). The volcanic monitoring system (left) includes sensors to obtain the signals in the form of time series (right). The goal is to build an explosive eruption classifier. The input is raw time series, and the output in this research is the status of the volcano that is either *explosive* or *not explosive*.

A popular method for volcano research is to analyze sensor signals obtained from the volcanic monitoring system. Multiple sensors are deployed in this system and each sensor is responsible for measuring a specific type of data. Examples of volcanic sensor data are ground deformation, ground surface vibration, and gas emission. These data are represented in the form of time series whose values are numeric and are recorded periodically in real time. As there are correlations between these time series data and volcanic eruption, the data are very valuable for the mining of volcano activities [1]. Because the data is gathered continuously in real time, the amount of data is increasing in size. This opens an opportunity for both volcano and machine learning researchers to mine the data in large scale. Our main focus in this paper is Volcano Activity Recognition (VAR). VAR is the task of classifying time series sensor data into multiple categories of volcano activity. Figure 1 shows the overall structure of VAR. In this paper, we classify the two most important statues of a volcano: *explosive* and *not explosive*. If the classification is successful, this research can give an insight to the mechanism of the eruption. In the context of this paper, the eruption means explosive eruption.

VAR can be solved using raw sensor signal, but the features extracted from this raw data could have more potential in terms of class discrimination than the raw data itself [2]. However, handcrafted feature extraction is time-consuming and is hard to decide even for volcano experts. Recently, deep learning with many layers of nonlinear processing for automatic feature extraction has proven to be effective in many applications such as image classification, speech recognition, and human activity recognition [5]. In this research, we propose a deep neural network architecture called VolNet which adapts Convolutional Neural Network (CNN), a particular type of deep learning model, for VAR on a single sensor. VolNet with its deep architecture is able to learn a hierarchy of abstract features automatically, and then use these features for the classification task. Based on the VolNet, we extend the model to multimodal fusion which takes into account all related sensors for this task. This is important because the eruption is controlled by many different factors which are measured by different sensors. In our context of multimodal fusion, one type of data obtained from a sensor is called as a modality. Recent multiple sensor fusion models fuse the features of all modalities at once and then feed them to the classifier [6–8]. In this paper, we call this "one-time fusion". This way of fusion ignores the properties of each modality and treats all modalities as the same. However, we consider this is not a good approach in the problems related to interdisciplinary study like VAR where the properties of data are different and important to design the solutions. Our assumption is some modalities are more likely to be correlated than others, and hence better to be fused together before they will be fused with other modalities. Based on that idea, we propose a simple but effective fusion model called Deep Modular Multimodal Fusion (DMMF) which uses VolNet as building block. DMMF is able to fuse relevant modalities in each module separately in a hierarchical fashion.

We have conducted extensive experiments for VAR on real world datasets obtained from Sakurajima volcanic monitoring system. Sakurajima is one of the most active volcanoes in Japan. In this paper, we propose two models for VAR: VolNet on a single sensor and DMMF on multiple sensors. First, we compared the performance of VolNet with conventional time series classification on a single sensor. Second, we compared DMMF with the best results obtained from the first experiment on single sensor and the one-time fusion model. The result shows that our proposed VolNet and DMMF outperformed all other state-of-the-art models.

To the best of our knowledge, this work is the first attempt to employ deep neural network for the study of VAR. Our deep model learns various patterns and classifies volcanic eruption accurately. The following are the contributions of this paper:

- Propose an accurate VolNet architecture for VAR on a single sensor.
- Propose a simple but effective fusion model called Deep Modular Multimodal Fusion (DMMF) which fuses the modalities in different modules in a hierarchical fashion.
- Outperform volcano experts on the task of VAR.
- Conduct extensive experiments in real volcanic sensor datasets.

The rest of the paper is organized as follows: We briefly introduce the dataset used in our experiment in Sect. 2. Next, we explain our approaches for VAR in Sect. 3. In the Sect. 4, we show detailed experiments on proposed method and baseline models. Related work will be summarized in Sect. 5. And finally, we conclude the paper in Sect. 6.

2 Datasets

We use volcanic sensor data obtained from Osumi Office of River and National Highway, Kyushu Regional Development Bureau, MLIT[1]. The data is about

[1] http://www.mlit.go.jp/en/index.html.

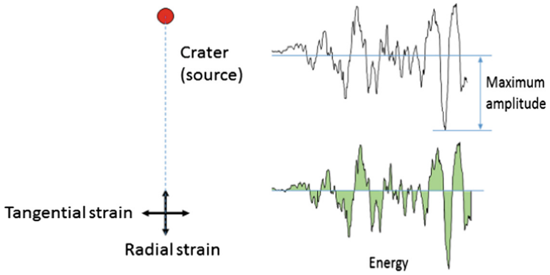

Fig. 2. The explanation of strain data (left) and seismic data (right). Strain data is measured by Strainmeters. These instruments measure linear strain by detecting horizontal contraction or extension in a length. They are installed in the direction of the crater (radial component) and perpendicular to the radial direction (tangential component). Seismic data is measured by Seismometer. This measures ground surface vibration as the velocity of a particle. Square sum of the velocity is proportional to seismic energy to evaluate the intensity of long-term tremor. Maximum amplitude velocity in the seismic records is treated as the instantaneous intensity of the event.

the explosive eruptions of Sakurajima volcano[2] which is one of the most active volcanoes in Japan. There are many explosive eruptions occurring in this volcano every week, so the data is good for VAR. This data includes four types of sensor data arranged into two groups. The first group is the seismic data including "seismic energy" and "maximum amplitude". These data are related to the ground surface vibration. The second group is the strain data including "tangential strain" and "radial strain". These two strains measure the ground deformation horizontally and vertically respectively. The data in each group are correlated with each other as they measure one type of data but in different ways. The details of data measurement and the instruments are shown in Fig. 2.

The data for all sensors are numeric values and recorded in every minute. The total data includes eight years from 2009 to 2016, and this is the biggest dataset about the volcanic monitor in Japan.

3 Proposed Methods

3.1 Problem Definition

VAR takes D sensors as the input and each sensor is a time series of length N. Formally, the input is a matrix of size $D \times N$ with the element x_i^d is the i^{th} element of the time series obtained from sensor d, where $1 \leq i \leq N$ and $1 \leq d \leq D$. In case of VAR for single sensor, D is equal to 1 and the input is a time series of length N. In this paper, the output of VAR is the class of the input, which is either *explosive* or *not explosive*. The task is a two-class classification. In VAR, the majority of the input are not explosive, but the explosive cases attract

[2] https://en.wikipedia.org/wiki/Sakurajima.

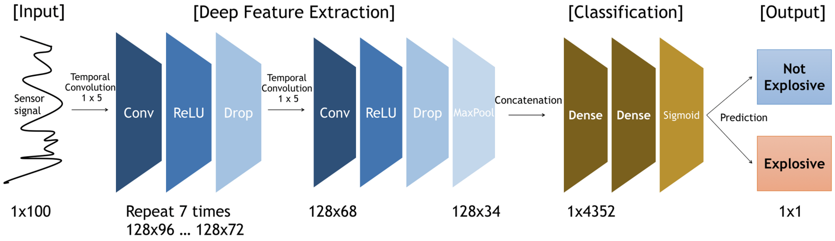

Fig. 3. The architecture of VolNet on a single sensor for VAR. The layers includes Conv (Convolution), ReLU, Drop (Dropout), MaxPool (Max Pooling non-overlapping pooling region with length of 2), Dense (fully-connected), and Sigmoid. The temporal convolution has kernel size of 5. The numbers before and after "x" on the bottom refer to the number of feature maps and the size of a feature map, respectively.

more attention than not explosive cases. Therefore, the goal of this research is not only to optimize the misclassification rate in general, but also maximize the precision and recall of explosive class.

3.2 Challenges

There are two main challenges in this task. The first is extreme class imbalance: Although Sakurajima is an active volcano, it is not explosive for most of the time. This poses a big challenge to the classification task as most classifiers tend to favor the majority class. The second challenge is sensor fusion. Because the eruption is complicated and controlled by many factors, multiple different time series sensor data should be used in order to improve the performance of the classifier. The fusion model should handle multiple data effectively.

3.3 Proposed VolNet for Single Sensor

CNN with its deep architecture is able to learn a hierarchy of abstract features automatically. Given a time series, CNN extracts features from the input using convolution operation with a kernel. Since the convolution operation will spread in different regions of the same sensor, it is feasible for CNN to detect the salient patterns within the input, no matter where they are. Because the nature of time series data is temporal, we use one-dimensional (1D) kernel for each time series independently. The feature map extracted by 1D convolution is obtained as:

$$m_i^{l+1}(x) = \sigma\left(\sum_{f=1}^{F^l}\left[\sum_{p=0}^{P^l-1} K_{if}^l(p) m_f^l(x+p)\right] + b_i^{l+1}\right), \tag{1}$$

where m_i^{l+1} is the feature map i in layer $l+1$, F^l is the number of feature maps in layer l, P^l is the length of the kernel in layer l, K_{if}^l is the convolution kernel of the feature map f in the layer l, and b_i^{l+1} is the bias vector in layer $l+1$. Pooling layer is also used in our model to increase the robustness of features to the small variations. In general, given a feature map, the pooling layer is obtained as:

$$m_i^{l+1}(x) = f_{1 \leq n \leq N^l}\left(m_i^l(x+n)\right), \tag{2}$$

where f is the pooling function, $1 \leq n \leq N^l$ is the range of value function f applies for, and N^l is the length of pooling region.

In this part, we will construct the architecture of VolNet especially for VAR. The overall structure is shown in Fig. 3. There are four main blocks in the architecture. The first block is "Input" and the last one is "Output". The network takes a time series of raw volcanic sensor signal as input and outputs the status of the volcano which is either *explosive* or *not explosive.* The second block called "Deep Feature Extraction" to automatically extract deep features from the time series input. This block includes the following eight small blocks in order: (1) a convolution layer, (2) a rectified linear unit (ReLU) layer that is the activation function mapping the output value using the function $relu(x) = max(0, x)$, and (3) a dropout layer that is a regularization technique [3] where randomly selected neurons are ignored during training, hence reduce over-fitting. We employ a max pooling at the end of this block to decrease the dimension of the feature maps. All the convolution layer has kernel size of 5 and 128 feature maps. Dropout has the probability of 0.5 and we only use dropout for three small blocks (the first, fourth and last small blocks). The reason is more dropout layers can lead to much randomness which is not good in our case. Max pooling has non-overlapping pooling region with length of 2. These hyper parameters are proved to be effective in our task through experiments and chosen using a validation set. The third block called "Classification" is a fully-connected network taking the learned features in previous layer and output the class using sigmoid function $S(x) = \frac{1}{1+e^{-x}}$. We also use dropout layer in this block to reduce over-fitting. One remark in designing the architecture of VolNet especially for VAR is that there are no normalization layers. The experiments showed that adding batch normalization [4] did not improve but worsen the performance.

To train VolNet, we minimize the weighted binary cross entropy loss function, and increase positive weight to deal with class imbalance problem:

$$L = \sum_{i=1}^{batch\_size} y_i \log(y'_i) \times weight + (1 - y_i) \log(1 - y'_i),$$

with y is the target and y' is the prediction. The parameter *weight* with value more than 1 is included to the loss function to penalize the cases when the target is 1 (*explosive*) but the prediction is near 0 (*not explosive*). By optimizing the loss function this way, we can force the model to favor explosive class if the ground truth is explosive. The model was trained via Stochastic Gradient Descent with the learning rate of 10^{-3}.

VolNet is designed mainly to deal with one sensor time series. In order to process multiple time series, we proposed a new fusion model built on the top of VolNet called Deep Modular Multimodal Fusion (DMMF).

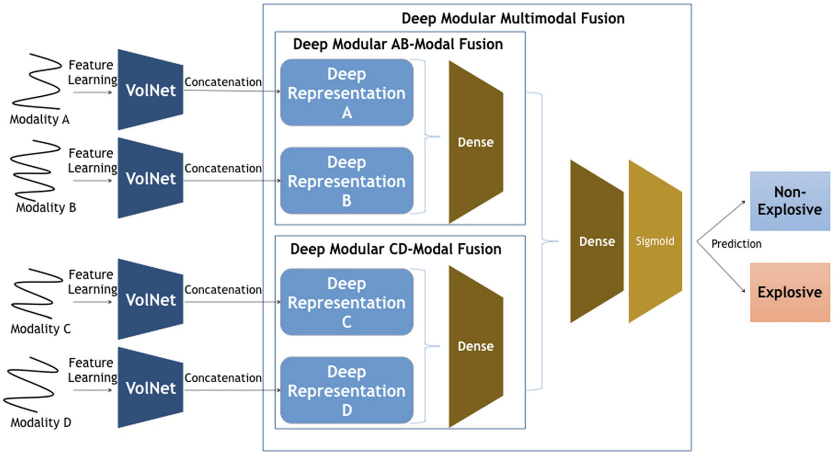

Fig. 4. The proposed Deep Modular Multimodal Fusion (DMMF) for multiple sensors based on VAR. First, we extract the deep representation features using VolNet for each modality independently. Then we group the features from the modalities which are relevant with each other into modules. In this figure, there are two modules A-B and C-D. Each module is followed by a Dense (fully-connected layer) so that they will be fused here. Then, in the next level, two modules will be fused again using one more Dense layer. Finally, we use a Sigmoid layer to compute the classification.

3.4 Proposed Deep Modular Multimodal Fusion for Multiple Sensors

While using one sensor performs well for VAR, fusing multiple sensors can potentially improve the performance. This is especially important in VAR where the eruption is affected by many factors. Also in the monitoring system, each sensor measures a different type of data and data is considered noisy due to earthquakes and typhoons. Therefore, adding more data from different sources can potentially improve the accuracy of VAR.

Recent work on multimodal fusion for sensors adapted CNN for each modality independently and then concatenate all the feature maps in one step [6], [8], which is not appropriate for VAR. In VAR, each modality is the data obtained from one sensor and some related sensors create a group of data. Our assumption is that the modalities in the same group of data are more related than other modalities and related modalities should be fused together before they will be fused with other modalities. For example, the modalities from tangential strain sensor and radial strain sensor make a group called “strain data”. They both measure the ground deformation, but in different directions which are horizontally and vertically respectively. Intuitively, the multimodal fusion model that considers fusing these two modalities first is expected to improve the performance. Based on this idea, we propose a simple but effective Deep Modular Multimodal Fusion (DMMF) which is built on the top of our proposed VolNet and is able to fuse relevant modalities in each module separately in a hierarchical fashion.

The overall architecture of DMMF is shown in Fig. 4. First, we use VolNet to extract the feature maps for each modality independently and then concatenate all feature maps to get the deep representation for each modality. Then we group the features from the modalities which are relevant with each other into modules. In each module, we concatenate the features of all modalities into one vector and add a fully-connected network so that these modalities will be fused. We then concatenate all the fused features from all modules and add one more fully-connected layer to fuse all the modules. The final layer is the sigmoid to compute the classification score based on the fused features from all modules. DMMF fuses the modalities in a hierarchical fashion, so all modalities have a chance to fused with each other. Unlike ensemble model where the final classification is made based on the classification of models on different sets of data, DMMF makes the classification based on the feature fusion of different groups of data in a hierarchical fashion.

There are some remarks on the design of DMMF. First, DMMF is built on the top of Volnet, hence takes advantages of deep features extracted from VolNet which is powerful to discriminate between classes. Second, DMMF does not directly concatenate all features at once, but it fuses the modalities in some different modules in a hierarchical fashion. Intuitively, when related modalities are fused together, the fusion will be more robust to noise as relevant modalities tend to complement with each other.

4 Experiments

4.1 Evaluation Metrics

Because of class imbalance problem, we use two type of F-score to do the evaluation. The first metric is F-score of the explosive class (F_E) which is the minority class in VAR. This metric measures the performance of the model in terms of minor but important class. The second metric is the unweighted average F-score (F_{avg}) of the explosive class and the not explosive class. Unweighted F-score is the general metric of the model and the contribution of each class to the score is equal.

4.2 Experiment 1: VAR on a Single Data

In this part, we conduct VAR experiments on each sensor data separately and compare the performance of proposed VolNet with the baseline models. In this experiment, we firstly show the effectiveness of deep feature representation in term of accuracy, and secondly to get insight into the best sensor for VAR.

The data is obtained from Sakurajima volcano as shown in Sect. 2. We use the sliding window technique to segment the time series into sequences. The sliding window length of raw data is 101 min and the sliding step is 10 min in our experiments. The choice of window length is based on the average sum of the inflation and deflation time of all eruptions, which is the domain knowledge from volcano experts. The sliding step of 10 min is chosen to reduce the chance of missing important patterns. In this paper, we use the first order differences of the series $\Delta x_i = x_{i+1} - x_i$ for all experiments. This is based on the characteristics of volcanic sensor signal where the change of two consecutive data points is important. Preliminary experimental result shows that all models using first order difference data outperformed the ones using raw data. Due to using first order differences, the length of each data sample now is 100 data points with the time interval of one minute.

We label each sequence into two classes: *explosive* and *not explosive*. If there is at least one eruption between the starting time and ending time of the sequence, we set it explosive, otherwise not explosive. This strategy considers the importance of the sequence with different numbers of eruption equally, but in fact, it is not common to have a sequence with more than one eruption within its time period. The total number of extracted sequences for eight years from 2009 to 2016 is around 400,000. The ratio of explosive class over not explosive class is approximately 1:10. For each year, we split part of dataset for testing to make sure test set covers all the dataset. In total, we obtain 50,000 sequences for testing and the rest 350,000 for training. There is no overlapping about time between test set and the training set to make sure the accuracy is reliable. We use validation set to pick up the hyper parameters of the models. Validation set is 20% of training data.

We do the experiments on proposed VolNet and the following baseline models:

- VolNet: The network architecture is shown in Fig. 3.
- 1 Nearest Neighbor with Euclidean distance (1NN-ED): Even though 1NN-ED is simple, it is considered one of the best techniques in time series classification [10]. First order difference time series data with normalization is the input of the model.
- 1 Nearest Neighbor with Dynamic Time Warping distance (1NN-DTW): Same as 1NN-ED, but the distance metric is DTW instead. 1NN-DTW is very effective in many time series classification due to its flexibility for computing distance. It also achieves state-of-the-art performance on the task of time series classification together with 1NN-ED [11]. One disadvantage of 1NN-DTW is that testing time is extremely slow, so we only test on 5% of data. We run the test multiple times and take average score. Running in all dataset takes months to finish.
- Means and Variance (MV): Mean and Variance show advantages in some time series classification task [13]. We use the mean and variance of the time series as the features and do the classification using 1NN-ED.
- Symbolic Aggregate Approximation - Vector Space Model (SAX-VSM): Unlike Nearest Neighbor which is distance-based approach, SAX-VSM is a well-known and effective feature-based approach for time series classification [12]. The input is also first order difference of sensor data.
- Support Vector Machine (SVM): The support vector machine with radial basis function (RBF) kernel is used as a classifier [14]. RBF kernel is chosen

Table 1. The results of VolNet and baselines for each type of data. The best model for each data is shown in bold.

	Seismic energy		Maximum amplitude		Radial strain		Tangential strain	
	F_{avg}	F_E	F_{avg}	F_E	F_{avg}	F_E	F_{avg}	F_E
1NN-ED	82.0	67.4	80.4	64.5	70.4	45.3	70.6	45.8
1NN-DTW	82.6	68.3	82.8	69.0	72.8	50.0	57.2	20.6
MV	79.8	63.3	80.5	64.6	58.5	26.9	59.8	28.8
SAX-VSM	74.3	43.4	75.0	44.7	70.6	35.2	71.7	37.0
SVM	85.7	74.2	57.3	18.7	68.6	41.3	72.3	48.4
MLP	84.2	71.6	83.3	70.2	76.0	56.4	76.4	57.2
LSTM	86.0	74.7	87.5	77.4	79.4	62.9	78.9	62.2
VolNet	**92.8**	**87.7**	**93.2**	**88.9**	**89.9**	**88.5**	**92.0**	**86.1**

because it shows the best results among all kernels. The input of the model is the first order difference data and the hyper parameter is carefully tuned using validation set.

- Multilayer Perceptron (MLP): MLP is an effective technique for time series classification [15]. The input is also first order difference sensor data. The architecture of the MLP includes one input layer, two hidden layers, and one output. The architecture and the number of neurons are optimized using validation set. We include this method to show the effectiveness of the feature learning from VolNet.
- Long Short-Term Memory (LSTM): LSTM is well-known to deal with sequential data. The architecture of LSTM includes two layers with the dimension of hidden state is 64. The input is also first order difference of sensor data.

The results of VolNet and the baselines for VAR are shown in Table 1. From the results, VolNet works well on different types of data and consistently outperforms all other baseline models on two evaluation metrics. From the accuracy of all models, we can see that seismic energy and maximum amplitude are the two best for VAR according to the experiments. The accuracy between models are quite different. VolNet is the best model among all models. LSTM is the second best model. The fact that VolNet which is built on CNN works better than LSTM in this case may suggest that the shape of the sequence is more important than the dependency of data through time. MLP also gains good accuracy, but much worse than VolNet. Other baseline models are quite unstable as they only work well on some data. For example, SVM works very well on seismic energy data, but when it comes to maximum amplitude data, it becomes the worst model with very low F_E. Both distance-based and feature-based methods like 1NN-ED, 1NN-DTW, MV and SAX-VSM did not work well on this task. This suggests that the raw signal and handcrafted feature extraction are not as good as deep automatic feature extraction from VolNet.

4.3 Experiment 2: Comparison with Volcano Experts

Volcano experts always want to predict the volcano status by using the sensor data. They try to understand the pattern of explosive eruption extracted from the sensor data. So far, the best way for volcano experts to recognize the eruption at Sakurajima volcano is using tangential strain. The pattern of eruption is gradually increase of tangential strain and then suddenly decrease. The point of eruption is usually the starting point of the decrease [22]. We would like to compare the model from expert and our proposed VolNet in the task of VAR.

We implement the expert model to detect explosive eruption using tangential strain. The dataset for this experiment is different from the previous experiments due to special condition of expert model. The experts need to observe the sensor data prior and after the eruption. The way we create dataset is as follows. For *explosive eruption class* we segment the sequences which has the explosive eruption at the position 80 of a sequence with length 100. This is based on the average inflation length of an eruption is about 80 min and the average deflation length is about 20 min. For *not explosive class* we use all the sequences which do not have any eruption.

Because the common pattern of eruption is an increase of tangential strain until the explosive point and then decrease of tangential strain after the eruption, we can divide the sequence into two parts: inflation (before explosive point) and deflation (after the explosive point). In the case of eruption, if we calculate the first order difference of the values in the inflation part, the amount of change will be positive. And in the case of deflation part, the amount of change will be negative. If there is no eruption, the amount of change in both inflation and deflation will not follow that rule. We call the amount of change in the inflation part is accumulated inflation, and that amount in the case of deflation is accumulated deflation. Expert model classifies the status of volcano based on the accumulated inflation and accumulated deflation.

Let say E is the set of sequences having eruption, and NE is the set of sequences which do not have eruption. For each sequence in E, we calculate the accumulated inflation and accumulated deflation. Then we compute the mean of accumulated inflations and the mean of accumulated deflations over set E. Intuitively, a sequence having the accumulated inflation and deflation near to these means has a high chance to have an eruption. We also do the same calculation for set NE to find the mean of accumulated inflation and deflation for the case of no explosive eruption. In testing phase, given a sequence, we first calculate accumulated inflation and deflation. The sequence will be assigned explosive label if both difference in inflation and deflation are nearer to the explosive case than the not explosive case. Otherwise, we set it not explosive. We use Piecewise Aggregate Approximation (PAA) [9] to transform the time series before calculating the first order difference. This transformation can help to remove the small variations in the sequence. The window size of PAA is decided using validation set. In our experiment, the optimal window size is 4. Table 2 shows the parameters for expert model. We can clearly see that in the case of explosive eruption, accumulated inflation is positive and accumulated deflation is negative.

Table 2. The obtained parameters of expert model

	Explosive eruption	Not explosive eruption
Mean accumulated inflation	0.367	0.019
Median accumulated inflation	0.376	0.023
Standard deviation accumulated inflation	0.569	0.353
Mean accumulated deflation	−2.041	0.038
Median accumulated deflation	−1.622	0.052
Standard deviation accumulated deflation	2.319	0.780

Table 3. The result of expert model and VolNet using tangential strain. The better model is shown in bold.

	F_{avg}	F_E
Expert model	75.1	46.3
VolNet	**86.5**	**73.3**

The results of expert model and our VolNet are shown in Table 3. VolNet outperforms expert model with a wide margin. This once again confirms that deep feature extraction from VolNet is much more powerful than handcrafted feature extraction.

4.4 Experiment 3: Multimodal Fusion

In this part, multiple sensors are used in the experiment. We would like to firstly show the effectiveness of multimodal fusion in term of accuracy compared with the best results obtained using only one sensor, and secondly to show the effectiveness of DMMF over other fusion strategies. We use the same dataset with Experiment 1 and run the experiments on proposed DMMF and the following models:

- Proposed DMMF: The architecture is shown as Sect. 3.4. In the fusion step, modalities are grouped into two modules called “module of strain data” including tangential strain and radial strain, and “module of seismic data” including maximum amplitude and seismic energy. This is based on related sensors from domain knowledge.
- Best model without fusion: We copy the best results with one data from Experiment 1.
- Early fusion: The first convolutional layer accepts data from all sensors using the kernel of $d \times k$, where 4 is the number of sensors and k is the kernel size. In this experiment, $d = 4$ and $k = 5$.
- One-time fusion: The features extracted from different modalities are fused at one step.

The results of fusion models for multiple sensors and the best model with no fusion are shown in Table 4. From the results, proposed DMMF consistently outperforms all other models in both evaluation metrics. Specifically, compared with the best results obtained from one sensor, DMMF improves the performance by 3% and 4.2% on average for F_{avg} and F_E respectively. This proves the effectiveness of fusion model on improving the performance for VAR when more data is used. Early fusion does not improve the accuracy, but worsen the overall accuracy compared with the best model with no fusion. This suggests that fusion strategy is important to improve the accuracy. In contrast, one-time fusion improves the accuracy more than 1% compared to no fusion. However, compared with one-time fusion, DMMF gains an improvement with F_{avg} and F_E increased by 1.9% and 3.7% on average respectively. This result supports our assumption that hierarchical fusion is better than one-time fusion in term of feature learning for VAR.

Table 4. The results of proposed DMMF and compared models. The best model is shown in bold.

	F_{avg}	F_E
Best model without fusion	93.2	88.9
Early fusion	91.0	83.5
One-time fusion	94.3	89.4
DMMF	**96.2**	**93.1**

To further confirm the effectiveness of modular fusion, we run some experiments with some combinations of modalities. Some combinations are:

- Seismic module fusion: Fusion with seismic energy and maximum amplitude.
- Strain module fusion: Fusion with tangential strain and radial strain.
- Maximum Amplitude Tangential Strain fusion: Fusion with maximum amplitude and tangential strain.
- Change group fusion: DMMF fusion with architecture of grouping are {seismic energy, tangential strain} and {maximum amplitude, radial strain}.

The results of experiment are shown in Table 5. We can see that the combination of seismic energy and maximum amplitude (seismic energy data) are better than using seismic energy or maximum amplitude alone. The same thing can apply to strain module fusion. However, when we combine maximum amplitude and tangential strain which belong to two different group of data, the accuracy goes down. This suggests that it is important to consider smart grouping when performing fusion because the relevant modalities in each module can complement each other and improve the accuracy. In change group fusion, we build an architecture exactly the same with DMMF, but try to change the group of data into {seismic energy, tangential strain} and {maximum amplitude, radial strain} which is against data properties. Compared with DMMF, the accuracy of change group fusion goes down. This suggests that the effectiveness of DMMF

Table 5. The results of proposed DMMF and the models with different combinations of modalities. The best model is shown in bold.

	F_{avg}	F_E
Seismic module fusion	95.1	91.1
Strain module fusion	93.8	88.6
Maximum Amplitude Tangential Strain fusion	92.9	87.0
Change group fusion	94.8	90.5
DMMF	**96.2**	**93.1**

is due to smart grouping, not due to deeper architecture because grouping differently worsen the performance. In general, the effectiveness of DMMF comes from hierarchical fusion and smart grouping.

5 Related Work

In this section, we briefly review some related work on volcano activity study and multiple sensor fusion. There is some work using sensor signals from the volcanic monitoring system for volcano-related events. Noticeably, as in [16], the authors applied neural network on seismic signals to classify the volcano events such as landslides, lightning strikes, long-term tremors, earthquakes, and ice quakes. The same research purpose was conducted by [18], but the methodology is based on hidden Markov model instead. The authors in [19] combined seismic data from multiple stations to improve the accuracy of the classifier. One common point of these papers is that the classes of the task are tremors, earthquakes, ice quakes, landslides, and lightning strikes. The concern about the explosive and not explosive status of the volcano is ignored in these work. Our work focuses on the classification of this class using the volcanic sensor data. The closet work to ours is [17]. The author tries to classify the time series of seismic signals to classify the volcano statuses using Support Vector Machine, still the methodology is quite simple and the accuracy is not high. To the best of our knowledge, our work is the first attempt to employ deep neural network for effective VAR on the explosive status of the volcano.

Multimodal deep learning has been successfully applied for many applications like speech recognition, emotion detection [20,21]. In these applications, the modalities are obtained from audio, video, and images. Human activity recognition is one of the most popular applications which uses multiple sensor data for the classification of human activity [6,8]. In these work, CNN is used on multichannel time series. However, in the fusion step, the authors ignored the properties of the modalities and fused all the modalities in one step. Our research considers the properties of the modalities in the fusion step and fuses the modalities in different modules in a hierarchical fashion.

6 Conclusion

In this paper, we demonstrated the advantages of deep architecture based on CNN for VAR. We proposed VolNet and a simple but effective fusion model DMMF which uses VolNet as a building block. DMMF adapts the modality properties to build the deep architecture and form the fusion strategy. The idea is that relevant modalities should be fused together before they will be fused with other less relevant modalities. The key advantages of DMMF are: (1) take advantages of deep non-handcrafted feature extraction and hence powerful to discriminate between classes, (2) relevant modalities in the same module complements with each other and hence is able to deal with noise data. With the extensive experiments, we demonstrated that DMMF consistently outperforms other compared models. This shows the ability of DMMF on combining multiple sensors into one model and the advantages of modular fusion. Moreover, DMMF is not only limited to VAR as it also has the potential to apply for other tasks that require multimodal fusion.

Acknowledgments. We would like to thank Osumi Office of River and National Highway, Kyushu Regional Development Bureau, MLIT for providing volcanic sensor datasets.

References

1. Sparks, R.S.J.: Forecasting volcanic eruptions. Earth Planetary Sci. Lett. **210**(1), 1–15 (2003)
2. Palaz, D., Collobert, R.: Analysis of CNN-based speech recognition system using raw speech as input. EPFL-REPORT-210039. Idiap (2015)
3. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
4. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 (2015)
5. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
6. Yang, J., Nguyen, M.N., San, P.P., Li, X., Krishnaswamy, S.: Deep convolutional neural networks on multichannel time series for human activity recognition. In: IJCAI, pp. 3995–4001 (2015)
7. Ordóñez, F.J., Roggen, D.: Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1), 115 (2016)
8. Zheng, Y., Liu, Q., Chen, E., Ge, Y., Zhao, J.L.: Time series classification using multi-channels deep convolutional neural networks. In: Li, F., Li, G., Hwang, S., Yao, B., Zhang, Z. (eds.) WAIM 2014. LNCS, vol. 8485, pp. 298–310. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08010-9_33
9. Senin, P., Malinchik, S.: SAX-VSM: interpretable time series classification using SAX and vector space model. In: 2013 IEEE 13th International Conference on Data Mining (ICDM), pp. 1175–1180. IEEE (2013)

10. Batista, G.E., Wang, X., Keogh, E.J.: A complexity-invariant distance measure for time series. In: Proceedings of the 2011 SIAM International Conference on Data Mining, pp. 699–710. Society for Industrial and Applied Mathematics (2011)
11. Xi, X., Keogh, E., Shelton, C., Wei, L., Ratanamahatana, C.A.: Fast time series classification using numerosity reduction. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 1033–1040. ACM (2006)
12. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. Knowl. Inf. Syst. **3**(3), 263–286 (2001)
13. Bulling, A., Blanke, U., Schiele, B.: A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput. Surv. (CSUR) **46**(3), 33 (2014)
14. Cao, H., Nguyen, M.N., Phua, C., Krishnaswamy, S., Li, X.: An integrated framework for human activity classification. In: UbiComp, pp. 331–340 (2012)
15. Koskela, T., Lehtokangas, M., Saarinen, J., Kaski, K.: Time series prediction with multilayer perceptron, FIR and Elman neural networks. In: Proceedings of the World Congress on Neural Networks, pp. 491–496. INNS Press, San Diego (1996)
16. Ibs-von Seht, M.: Detection and identification of seismic signals recorded at Krakatau volcano (Indonesia) using artificial neural networks. J. Volcanol. Geothermal Res. **176**(4), 448–456 (2008)
17. Malfante, M., Dalla Mura, M., Metaxian, J.-P., Mars, J.I., Macedo, O., Inza, A.: Machine learning for Volcano-seismic signals: challenges and perspectives. IEEE Sig. Process. Mag. **35**(2), 20–30 (2018)
18. Benítez, M.C., et al.: Continuous HMM-based seismic-event classification at Deception Island Antarctica. IEEE Trans. Geosci. Remote Sens. **45**(1), 138–146 (2007)
19. Duin, R.P.W., Orozco-Alzate, M., Londono-Bonilla, J.M.: Classification of volcano events observed by multiple seismic stations. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 1052–1055. IEEE (2010)
20. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 689–696 (2011)
21. Kahou, S.E., et al.: Emonets: multimodal deep learning approaches for emotion recognition in video. J. Multimodal User Interfaces **10**(2), 99–111 (2016)
22. Iguchi, M., Tameguri, T., Ohta, Y., Ueki, S., Nakao, S.: Characteristics of volcanic activity at Sakurajima volcano's Showa crater during the period 2006 to 2011 (special section Sakurajima special issue). Bull. Volcanol. Soc. Japan **58**(1), 115–135 (2013)