UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Application of Transformer Based Machine Learning Models for the Classification of Seismic Signals

Kevin Daniel Huertas García

Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniería en Ciencias de la Computación

Quito, 16 de diciembre de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Application of Transformer Based Machine Learning Models for the Classification of Seismic Signals

Kevin Daniel Huertas García

Nombre del profesor, Título académico **Noel Pérez, Ph.D.**

Quito, 16 de diciembre de 2024

© DERECHOS DE AUTOR

Nombres y apellidos: Kevin Daniel Huertas García

Código: 00213278

Cédula de identidad: 1722344965

Lugar y fecha: Quito, 16 de diciembre de 2024

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

La detección y clasificación de eventos sísmicos volcánicos son fundamentales para el monitoreo de la actividad volcánica y la mitigación de desastres naturales. La identificación temprana de señales sísmicas precursoras permite emitir alertas oportunas, con el potencial de salvar vidas en zonas amenazadas por volcanes activos. En este trabajo, proponemos un método automático de clasificación de eventos sísmicos basado en modelos tipo transformer, con el fin de maximizar la clasificación de señales volcánicas en bruto, evitando cualquier tipo de preprocesamiento de datos y retrasos en la clasificación.

El método propuesto se entrenó y validó mediante una estrategia de validación cruzada en el conjunto de datos MicSigv1. El mejor modelo alcanzó una precisión, una puntuación F1 y un valor AUC promedio de 96.4%, 96.5%, y 97.9%, respectivamente, en el conjunto de prueba. El alto rendimiento obtenido en múltiples métricas demuestra la efectividad de la metodología para la clasificación de eventos sísmicos volcánicos, evidenciando su potencial para futuros sistemas de monitoreo en tiempo real.

Palabras clave: Clasificación de señales volcánicas, clasificación de eventos sísmicos, modelos transformer, deep learning, procesamiento de señales, sismología.

ABSTRACT

The detection and classification of volcanic seismic events are crucial for monitoring volcanic activity and mitigating natural disasters. Early identification of precursor seismic signals enables timely warnings, potentially saving lives in areas threatened by active volcanoes. We propose an automatic seismic events classification method based on transformer models to maximize the volcanic raw signal classification, avoiding any data preprocessing and delayed classification. The proposed method was trained and validated using a cross-validation strategy on the MicSigv1 dataset. The best model achieved an average accuracy, F1-score, and AUC scores of 96.4%, 96.5%, and 97.9%, respectively, on the test set. The high performance across multiple metrics indicated the effectiveness of the proposed method for volcano seismic events classification, demonstrating its potential for future real-time monitoring systems.

Key words: Volcano signal classification, seismic events classification, transformer models, deep learning, signal processing, seismology.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Introduction

The detection and classification of volcanic seismic events are essential for monitoring volcanic activity and mitigating natural disasters (Ibáñez, 2022; Pérez et al., 2020). Active volcanoes pose a significant threat to nearby populations, and early identification of precursor seismic signals can save lives by enabling early warnings (Ibáñez, 2022; Lara-Cueva et al., 2016). However, manual analysis of large volumes of seismic data is a laborious process prone to human error (Malfante et al., 2017; Oliveira et al., 2020). Moreover, volcanic seismic signals exhibit complex patterns and significant variations, making precise classification difficult (Pérez et al., 2020).

Some machine learning-based approaches have been proposed using shallow and deep learning models. For example, shallow learning methods have employed feature-based techniques and traditional classifiers. Pérez et al. (Pérez et al., 2020) used Artificial Neural Networks (ANN) and Random Forest (RF) models that relied on precomputed features, achieving accuracies of 95% and 93%, respectively. Similarly, traditional statistical techniques such as Gaussian Mixture Models (GMM) or Support Vector Machines (SVM) demand careful feature extraction and selection, increasing the computational complexity (Malfante et al., 2017; Pérez et al., 2020). Unsupervised approaches like the BFR clustering algorithm used by Duque et al. (Duque et al., 2020) achieved an accuracy of 88%, yet still relied on transformed data representations.

On the other hand, deep learning methods have also been explored to address these challenges. Convolutional Neural Networks (CNNs) have been particularly prominent due to their ability to process spectrogram representations of seismic signals (Salazar et al., 2020; Calderón et al., 2020). Salazar et al. (Salazar et al., 2020) introduced DCNN-based architectures, achieving 95% accuracy and 98% AUC, while Calderón et al. (Calderón et al., 2020) reported a 94% accuracy and 94% AUC. Furthermore, hybrid CNN-RNN models (Salazar et al., 2020) have attempted to capture temporal dependencies, though they often require extensive preprocessing and careful feature engineering. Overall, these traditional pipelines—both shallow and deep—generally depend on transforming the original signal space, computing and selecting features, and then classifying, which can be memory-intensive and time-consuming.

The reliance on these extended workflows suggests that a more direct approach to analyzing raw seismic signals could improve efficiency. The use of time-series data makes methods capable of effectively capturing long-range dependencies and temporal patterns, which are particularly desirable. In this regard, transformer models have recently been explored for classifying seismic signals (Suzuki et al., 2024; Rinaldi et al., 2022). Unlike recurrent neural networks (RNNs) or long-short-term memory neural networks (LSTMs), which suffer from vanishing gradients and limited context understanding, transformers use powerful self-attention mechanisms to capture long-range dependencies and complex patterns in data (Vaswani et al., 2017). This capability is critical for analyzing raw seismic time series, where patterns over extended periods are crucial. By directly handling raw signals without extensive preprocessing and feature computation stages, transformers can potentially streamline the classification process and mitigate delays associated with traditional workflows.

For example, Rinaldi et al. (Rinaldi et al., 2022) proposed a transformer-based classification system for volcanic seismic signals, combining convolutional layers, residual connections, LSTM, and multi-headed self-attention (MHSA) parts. While their model achieved an accuracy of 96.4% on filtered data and showed potential in analyzing raw seismic data without the need for extensive preprocessing, it came with significant computational costs due to the inclusion of multiple convolutional layers and LSTM in its architecture.

Therefore, we propose an automatic seismic events classification method based on transformer models to maximize the volcanic raw signal classification, avoiding any data preprocessing and delayed classification. The main contribution is to develop a model that efficiently captures temporal dependencies in seismic signals through a self-attention mechanism that analyzes and prioritizes the importance of each part of the data regardless of temporal proximity, which is crucial in seismic signal analysis where significant relationships may exist between distant events.

Materials and Methods

A. Database

This study uses the MicSigV1 dataset, available in the ESeismic repository, which contains 1,187 seismic records captured at the Cotopaxi volcano in Ecuador by the VC1 and BREF stations. The seismic events are classified into five categories: Long Period (LP) with 1,044 samples, Volcano-Tectonic (VT) with 101, Regional (REG) with 27, Hybrid (HB) with 8, and Icequake (ICE) with 7. This database is available at: https://www.igepn.edu.ec/eseismic_web_site/index.php

B. Proposed method

We propose an automatic seismic events classification method based on transformer models to maximize the volcanic raw signal classification, avoiding any data preprocessing and delayed classification. It aims to capture temporal dependencies in seismic signals through the implemented self-attention mechanism regardless of temporal proximity. This is crucial in seismic signal analysis, where significant relationships may exist between distant events, as shown in Fig. 1.

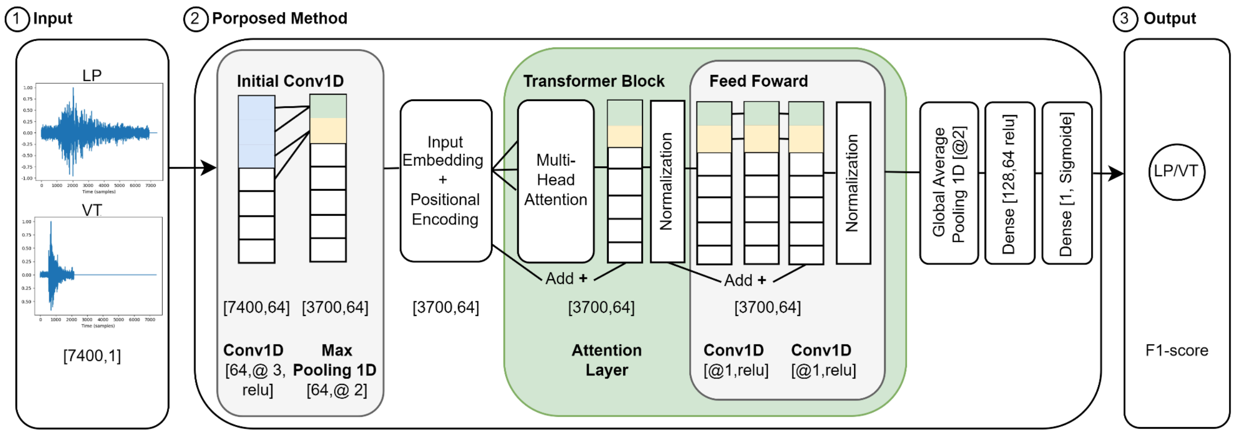

Figure 1. Transformer-based architecture of the proposed method

From Fig. 1, it is possible to observe one of the proposed single transformer-based architectures. It comprises several connected parts, such as the initial convolutional layer that applies a Conv1D operation with 64 filters, a kernel size of 3, and a padding type same to guarantee the input signal meets the input dimension requirement of the model. The MaxPooling1D layer with a pool size of 2 reduces the computed feature maps by half, condensing relevant information and lowering the computational load for the subsequent stages.

Then, a learnable positional embedding is incorporated before the multi-head attention system to identify the relative position of each data point within the sequence, which is essential for capturing temporal relationships in parallel environments. Subsequently, the multi-head attention module of the transformer block computes the queries (Q), keys (K), and values (V) through linear projections of the input data. Having separate linear transformations for Q, K, and V enables the model to learn different representations for querying, attending, and aggregating information from the sequence, enhancing its ability to model complex dependencies (Vaswani et al., 2017) while facilitating gradient propagation during training without degradation problems (Ahmed et al., 2022). The last part of this block uses a feed-forward module with two Conv1D layers with kernel sizes of 1 and ReLU activation functions, introducing non-linearity and flexibility into the model without altering the original data dimensions (He et al., 2016). Additionally, regularization mechanisms such as dropout are applied to mitigate overfitting and improve the model's generalization.

Lastly, the GlobalAveragePooling1D layer summarizes the information in a global vector that feeds two dense hidden layers with 128 and 64 ReLU neurons, respectively, and an output layer with one sigmoid neuron for the binary classification.

We extended the proposed single architecture to create the dual transformer architecture that augments one transformer block. It should be noted that for both transformer architectures, the initial and feed-forward convolutional layers vary in the intervale 1 … 2 and 1 … 3, respectively. Also, a single max pooling layer is used in all configurations, whether there are one or two initial convolutional layers. These configurations will allow experimenting with 12 transformers (six per transformer block).

C. Experimental setup

1) Dataset creation

For this work, only the LP and VT events were selected, as they represent the categories with the largest number of samples in the dataset. A total of 1,123 samples were utilized, with 80\% allocated for training (898 samples) and 20\% reserved for testing (225 samples). This division supports the creation and evaluation of a reliable classification model.

2) Signal processing

The selected signals underwent a preprocessing stage that included duration adjustment. The decision to fix all signals to a duration of 74 seconds was based on statistical analysis of their durations. A box plot analysis revealed that for both LP and VT events, the upper whisker extends to 74 seconds, effectively separating the outliers. Outlier data beyond this duration were excluded as they may represent noisy or anomalous data. By selecting this duration, we ensured that most of the data could be used without any filtering or signal modification, allowing us to work with the rawest data possible. Signals shorter than 74 seconds were adjusted by padding, adding zero values at the end of each signal until reaching 7400 samples. This procedure guarantees uniformity in signal length, which is crucial for consistent input into the model (Suzuki et al., 2024). Finally, each signal was normalized to a range between 0 and 1, standardizing their amplitudes for analysis in the model.

3) Training and test sets

Initially, the dataset was split into 80% for training and 20% for testing. The training set was then used in a stratified 10-fold cross-validation scheme to determine the best-performing model. The best model was selected by identifying the fold with the highest evaluation metrics. For this purpose, a model checkpoint callback was implemented, which saved the model whenever the validation loss reached its minimum, ensuring that the best model was retained for further evaluation. After selecting the optimal model, this model was applied to the previously unseen 20% test set, allowing us to evaluate its performance on new data and verify its generalization capability.

To evaluate the model, a stratified 10-fold cross-validation scheme was implemented, ensuring consistent evaluation by calculating average metrics over multiple partitions of the training set. This method divides the data into subsets, using one as the validation set while training the model on the remaining subsets, and repeating the process across all possible combinations (Salazar et al., 2020). This approach guarantees that the original class proportions are reflected in both partitions, reducing classification bias and maximizing the model's generalization, as suggested in previous works (Salazar et al., 2020).

4) Model configuration

All models were trained for a maximum of 100 epochs, with early stopping implemented to halt training after 20 epochs of no improvement in validation loss. We chose a batch size of 16 to balance computational efficiency and stable convergence. To optimize the parameters, we employed the Adam optimizer with a fixed learning rate of 1e-4, taking advantage of its adaptive mechanisms for stable and effective updates (Kingma & Ba, 2015). Additionally, we used a ReduceLROnPlateau scheduler to halve the learning rate whenever validation loss did not improve after 4 consecutive epochs.

To combat overfitting and enhance generalization, we applied L2 regularization (0.01) to the final dense layers and introduced a 0.1 Dropout rate following activation layers (Srivastava et al., 2014). The multi-head attention mechanism was configured with 2 heads, each processing input features in subspaces of dimension 64. Additionally, all Conv1D layers were configured with 64 filters to ensure dimensional consistency across layers.

5) Assessment metrics and selection criteria

The classification performance of the proposed model was evaluated using the average results of a 10-fold cross-validation. Metrics considered include F1-Score, Accuracy, Precision, Recall, Area Under the Curve. Additionally, the Wilcoxon test will be applied to identify the model with the smallest significant performance difference across the evaluated metrics.

The F1-score metric was prioritized due to the significant class imbalance in our dataset, with 1044 samples for the LP class and 101 samples for the VT class. The F1-score provides a more informative measure than accuracy by considering both precision and recall. It is particularly suitable for imbalanced datasets as it balances the trade-off between false positives and false negatives, ensuring that the model's performance is accurately reflected despite the uneven distribution of classes (He & Garcia, 2009). Subsequently, the F1-score will be used to compare the model with the highest F1-score against the others, selecting the one with the least significant difference and lower computational complexity in terms of algorithmic performance.

Results and Discussion

A. Performance evaluation on the training set

Table 1. Average Performance Metrics for Different Model Configurations

Transformer Blocks	Conv1D Initial	FF Conv1D Blocks	F1-Score avg±std	Wilcoxon α = 0.05	Accuracy avg±std	Precision avg±std	Recall avg±std	AUC avg±std
1	1	**1**	**0.959±0.021**	**p=0.359**	**0.959±0.021**	**0.960±0.022**	**0.959±0.021**	**0.970±0.036**
		2	0.957±0.016	p=0.185	0.957±0.017	0.960±0.015	0.957±0.017	0.982±0.013
		3	0.964±0.011	p=0.476	0.964±0.011	0.966±0.011	0.964±0.011	0.980±0.017
	2	1	0.962±0.025	p=0.557	0.964±0.022	0.965±0.021	0.962±0.022	0.974±0.025
		2	0.963±0.021	p=0.557	0.963±0.019	0.965±0.022	0.963±0.019	0.973±0.028
		3	0.955±0.024	p=0.126	0.958±0.020	0.957±0.021	0.958±0.020	0.973±0.035
2	1	1	0.961±0.018	p=0.160	0.961±0.019	0.963±0.018	0.961±0.019	0.973±0.027
		2	0.962±0.015	p=0.375	0.962±0.015	0.964±0.015	0.962±0.015	0.976±0.022
		3	0.959±0.021	p=0.359	0.959±0.021	0.960±0.021	0.959±0.021	0.980±0.016
	2	1	0.965±0.026	p=0.695	0.967±0.024	0.965±0.027	0.967±0.024	0.973±0.034
		2	0.956±0.019	p=0.275	0.957±0.020	0.958±0.019	0.957±0.020	0.970±0.029
		3	**0.969±0.020***	-	0.969±0.019	0.970±0.020	0.969±0.019	0.978±0.027

Note. Conv1D:Convolutional 1D layer; FF: feed-forward; Metrics avg±std over ten folds; F1-Score with * is the Wilcoxon pivot; Wilcoxon: test at α=0.05;

The results in Table 1 show that models integrating initial Conv1D layers and multiple feed-forward Conv1D layers can achieve high classification performance across all metrics. For instance, within the group with one Transformer block, the model featuring one initial Conv1D layer and three feed-forward Conv1D layers achieved an F1-score of 0.964. Similarly, in the group with two Transformer blocks, the model with two initial Conv1D layers and three feed-forward Conv1D layers reached an F1-score of 0.969.

However, increasing the number of Transformer blocks from one to two did not lead to substantial improvements in performance. While the two-block models achieved slightly better results, the gain was minimal, as illustrated by the top F1-scores of 0.964 (one Transformer block) and 0.969 (two Transformer blocks).

To statistically assess performance differences between models, we conducted the Wilcoxon signed-rank test with a significance level of p = 0.05. After applying the test, no significant difference in performance was found between the top-performing models with one and two Transformer blocks. Specifically, the model with one Transformer block, one initial Conv1D layer, and one Conv1D layer in the feed-forward block did not show a statistically significant difference compared to more complex architectures.

Given the lack of significant performance improvement and considering computational efficiency, we selected this simpler model as our optimal architecture. Its reduced complexity lowers computational cost, making it more suitable for practical applications without compromising accuracy.

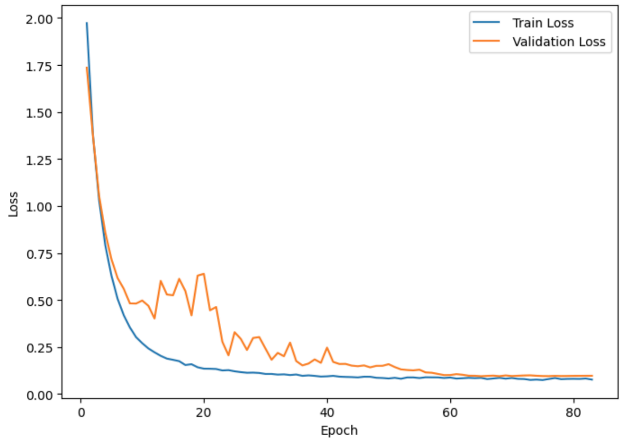

Figure 2. Average Loss Curves for the Selected Model (1 Transformer Block, 1 Initial Conv1D Layer, 1 Conv1D Feed-Forward Layer) on the training set

The selected model's loss curves exhibit consistent optimization behavior, as shown in Fig. 2. Both the training and validation loss curves display a general downward trend across epochs, indicating effective learning. However, there are noticeable spikes in the validation loss, particularly during the initial training epochs. These fluctuations are likely due to the imbalance in the dataset, which can cause temporary instability in the learning process. Despite these spikes, the model demonstrates strong convergence as training progresses, with both losses stabilizing effectively in later epochs. This behavior underscores the model's capacity to generalize well, despite the challenges posed by the data imbalance.

B. Performance evaluation on the test set

To evaluate the generalization capability of the proposed model, we tested it on a held-out test set comprising 20% of the dataset, which had been separated prior to the cross-validation process. This approach ensures that the test set remains entirely unseen during both training and validation, providing an unbiased assessment of the model's performance.

The selected model, configured as 1 Transformer Block, 1 Initial Conv1D Layer, and 1 Conv1D Feed-Forward Layer, was chosen based on its performance during training and validation. During the training phase, the model with the lowest validation loss from each fold was saved, resulting in multiple models across the folds. For the test evaluation, we selected the model from the fold with the best overall validation metrics and applied it to the test set.

Table 2. Performance Metrics of the Best Model from the Selected Architecture During Training

F1-Score	**Accuracy**	**Precision**	**Recall**	**AUC**
0.965	0.96.4	0.966	0.964	0.979

The evaluation results on the test set are summarized in Table 2. The model achieved an F1-score of 0.965, along with consistent metrics across accuracy, precision, recall, and AUC. These results are closely aligned with the training metrics reported in Table 1, reflecting the model's ability to generalize well to unseen data.

The consistency between the training and test metrics demonstrates the robustness of the proposed model, highlighting its ability to maintain high classification performance without overfitting. This level of generalization reinforces the model's suitability for real-world applications, such as seismic signal monitoring and classification tasks, where reliability on unseen data is critical.

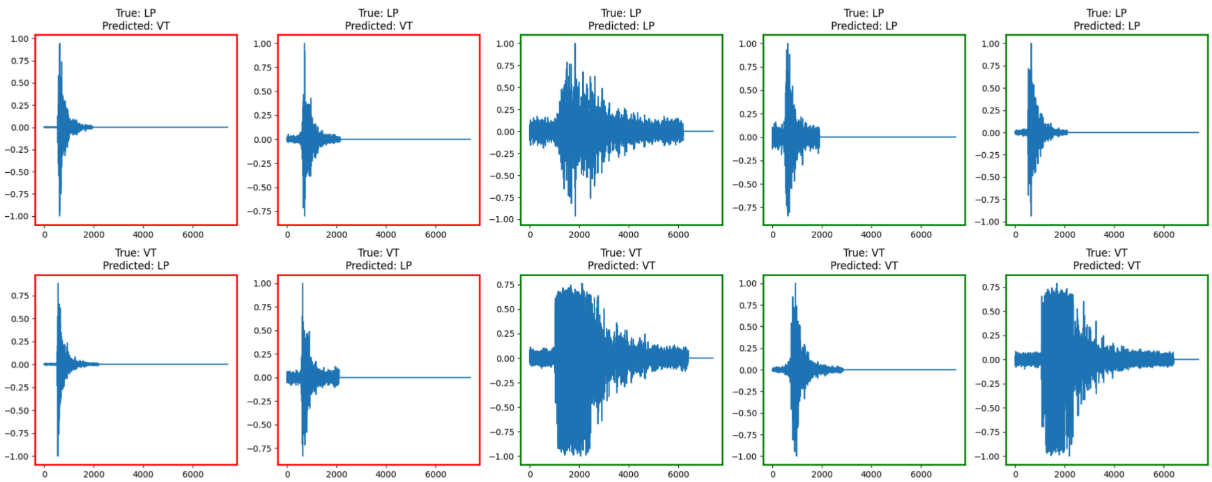

Figure 3. Classification Visual Results of the Selected Model on the Test Set

As illustrated in Fig. 3, a sample of correctly and incorrectly classified signals is presented. It is evident that the signals misclassified by the model share very similar shapes and temporal structures, making them challenging to distinguish. This overlap in features contributes to the model's difficulty in accurately classifying these instances, underscoring the complexity of the task.

C. State of art based comparison

The performance of the proposed Transformer-based model is evaluated alongside state-of-the-art methods for seismic signal classification, as summarized in Table 3.

Table 3. Performance Metrics of the Best Model from the Selected Architecture During Training

Method	Samples	Features	Spectrogram	ACC (%)
ANN [18]	914	6	No	97
DT [18]	914	3	No	96
ANN [2]	637	17	No	95
RF [2]	637	17	No	93
Linear SVM [3]	914	5	No	97
ANN [19]	1,033	8	No	94
HMM [20]	512	39	No	90
GMM [21]	667	2	No	94
DCNN1 [7]	1,108	–	Yes	95
CNN3 [8]	668	–	Yes	94
Proposed Model	1,145	–	No	**96**

Note. ACC - accuracy; *values rounded to the closest integer

Although a direct statistical comparison with certain existing methods, such as those presented in (Rinaldi et al., 2022), is not feasible due to varying experimental conditions, we have compared our model based on the accuracy (ACC) scores reported in other studies, as shown in Table 3.

As observed, the proposed Transformer based model achieves an accuracy of 96%, which is on par with or exceeds many existing models in the literature. Models such as ANN and Linear SVM (Lara-Cueva et al., 2016; Curilem et al., 2009; Pérez et al., 2020; Duque et al., 2020) have achieved high accuracies around 94-97%, while traditional methods like

HMM (Benítez et al., 2007), GMM (Venegas et al., 2019) and CNN-based (Salazar et al., 2020; Calderón et al., 2020) achieved similar accuracies.

The proposed model demonstrates robust performance within the range of state-of-the-art methods, indicating its effectiveness in accurately classifying volcanic seismic events

Table 3 highlights the comparative performance of the proposed model and state-of-the-art methods. Notably, the Transformer-based model achieves similar or superior performance on a smaller labeled dataset, emphasizing its effectiveness in learning complex temporal and spatial patterns from limited data

Conclusions and Future Work

In conclusion, we developed a model that effectively combines transformer architectures with the feature extraction capabilities of CNNs to classify volcanic seismic signals. By experimenting with different configurations, we identified a setup that achieved high performance, with a F1-score of 96.5% and an AUC of 97.9%. The use of k-fold cross-validation confirmed the robustness and generalizability of the model across different data partitions. The results demonstrate that integrating transformers enhances the capture of temporal dependencies, making it possible to analyze seismic signals directly from their raw form. This direct utilization of unprocessed input data reduces the need for extensive feature extraction and complex preprocessing steps, ultimately decreasing computational overhead and potentially improving the model's adaptability to various conditions

Future work will focus on incorporating additional layers for sequential analysis to further improve the model's ability to capture complex temporal patterns.Exploring more attention mechanism and larger datasets may enhance performance and generalization. Implementing this model in real-time monitoring systems will be a significant step towards practical application in volcanic activity surveillance.

Acknowledgment

Authors thank to the Applied Signal Processing and Machine Learning Research Group of USFQ for providing the computing infrastructure (NVidia DGX workstation) to implement and execute the developed source code. Publication of this article was funded by the Academic Articles Publication Fund of Universidad San Francisco de Quito USFQ.

References

Ibáñez, J. M. (2022). El uso de la Inteligencia Artificial (IA) para producir erupciones volcánicas. *Pasaje a la Ciencia.* Recuperado de https://pasajealaciencia.es/wp-content/uploads/2022/09/Numero-23-2022-comprimido.pdf#page=37

Pérez, N., Venegas, P., Benítez, D., Lara-Cueva, R., & Ruiz, M. (2020). A new volcanic seismic signal descriptor and its application to a data set from the Cotopaxi volcano. *IEEE Transactions on Geoscience and Remote Sensing.* https://doi.org/10.1109/TGRS.2020.2976896

Lara-Cueva, R. A., Benítez, D. S., Carrera, E. V., Ruiz, M., & 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*, 1–11. https://doi.org/10.1109/TGRS.2016.2559440

Malfante, M., Dalla Mura, M., Mars, J. I., & Métaxian, J. P. (2017). Machine Learning for Automatic Classification of Volcano-Seismic Signatures. En *Proceedings of the 25th European Signal Processing Conference (EUSIPCO)* (pp. 2393–2397). EURASIP.

Oliveira, S. S., Canuto, A. M. P., Carvalho, B. M., & Kreutz, M. E. (2020). Machine learning based seismic region classification. En *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. https://doi.org/10.1109/IJCNN48605.2020.9206961

Salazar, A., Arroyo, R., Pérez, N., & Benítez, D. (2020). Deep-Learning for Volcanic Seismic Events Classification. En *2020 IEEE Latin American Conference on Computational*

Intelligence (LA-CCI). IEEE. Recuperado de https://www.ieee.org/publications/ieee-xplore.html

Calderón, F., Benítez, D. S., Pérez, N., & Riofrío, D. (2020). Volcanic seismic event classification based on CNN architectures. En *2020 IEEE ANDESCON* (pp. 1–6). IEEE. https://doi.org/10.1109/ANDESCON50619.2020.9272051

Duque, A., Pérez, N., González, K., & Benítez, D. S. (2020). Categorizing Volcanic Seismic Events with Unsupervised Learning. En *2020 IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI)* (pp. 1–6). IEEE. https://doi.org/10.1109/ColCACI50549.2020.9247854

Suzuki, Y., Yukutake, Y., Ominato, T., Yamasaki, M., & Kim, A. (2024). Volcanic Earthquake Classification using Transformer Encoder and Its Interpretability Evaluation. *Manuscript submitted for publication in Geophysical Research Letters.* Recuperado de https://essopenarchive.org/doi/full/10.1002/essoar.857036

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. En *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Recuperado de https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. En *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400.*

Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST: Audio Spectrogram Transformer. *arXiv preprint arXiv:2104.01778.* Recuperado de https://arxiv.org/abs/2104.01778

Zhang, X., Wang, X., Zhang, Z., & Wang, Z. (2023). CNN-Transformer for Microseismic Signal Classification. *Electronics, 12*(11), 2468. https://doi.org/10.3390/electronics12112468

Rinaldi, A., Mora Stock, C., & Bravo Roman, C. (2022). Classifying Volcanic Seismic Signals Using Self-Attention: A Transformer-Based Classification System for Volcanic Seismic Signals. *Preprint.* Recuperado de https://anthonyprinaldi.github.io/files/VolcanicTransformer.pdf

Ahmed, S., Nielsen, I. E., Tripathi, A., Siddiqui, S., Rasool, G., & Ramachandran, R. P. (2022). Transformers in Time-series Analysis: A Tutorial. *arXiv preprint arXiv:2205.01138.*

Liu, M., Ren, S., Ma, S., Jiao, J., Chen, Y., Wang, Z., & Song, W. (2021). Gated Transformer Networks for Multivariate Time Series Classification. *arXiv preprint arXiv:2103.14438.*

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv preprint arXiv:1803.01271.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. En *Advances in Neural Information Processing Systems* (pp. 5998–6008).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research, 15*(56), 1929–1958.

Smith, L. N. (2017). Cyclical Learning Rates for Training Neural Networks. En *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 464–472). https://doi.org/10.1109/WACV.2017.58

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR).* Recuperado de https://arxiv.org/abs/1412.6980

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research, 15*(56), 1929–1958. Recuperado de http://jmlr.org/papers/v15/srivastava14a.html

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Lara-Cueva, R., Carrera, E. V., Morejon, J. F., & Benítez, D. (2016). Comparative analysis of automated classifiers applied to volcano event identification. En *2016 IEEE Colombian Conference on Communications and Computing (COLCOM)* (pp. 1–6). IEEE. https://doi.org/10.1109/ColComCon.2016.7516377

Benítez, M. C., Ramírez, J., Segura, J. C., Ibáñez, J. M., Almendros, J., García-Yeguas, A., & Cortés, G. (2007). Continuous HMM-based seismic-event classification at Deception

Island, Antarctica. *IEEE Transactions on Geoscience and Remote Sensing, 45*(1), 138–146.

Curilem, G., Vergara, J., Fuentealba, G., Acuña, G., & 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), 1–8.

Venegas, P., Pérez, N., Benítez, D., Lara-Cueva, R., & Ruiz, M. (2019). Combining filter-based feature selection methods and Gaussian mixture model for the classification of seismic events from Cotopaxi Volcano. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12*(6), 1991–2003.