

Physically interpretable multimodal prototypical networks for seismo-volcanic signal classification via few-shot learning

Redes prototípicas multimodales físicamente interpretables para la clasificación de señales sismo-volcánicas mediante aprendizaje con pocos ejemplos

Redes prototípicas multimodais físicamente interpretáveis para a classificação de sinais sismo-vulcânicos por meio de aprendizado com poucos exemplos

Jorge Edison Mora Arroyo¹
Olger Ferledy Erazo de la Cruz²
Pablo Emilio Jojoa Gómez³

Received: May 5th, 2026

Approved: July 30th, 2026

Available: September 1st, 2026

How to cite this article:

J. E. Mora Arroyo, O. F. Erazo de la Cruz, and P. E. Jojoa Gómez, "Physically Interpretable Multimodal Prototypical Networks for Seismo-Volcanic Signal Classification via Few-Shot Learning," *Revista Ingeniería Solidaria*, vol. 22, no. 3, 2026.
doi: <https://doi.org/10.16925/2357-6014.2026.03.08>

Research article. <https://doi.org/10.16925/2357-6014.2026.03.08>

¹ Ph.D. Candidate in Electronic Sciences, Faculty of Electronic Engineering and Telecommunications, Universidad del Cauca, Popayán, Colombia.

Email: jorgemora@unicauca.edu.co

ORCID: <https://orcid.org/0000-0003-3850-4506>

CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001545153

² Ph.D. Candidate in Electronic Sciences, Faculty of Electronic Engineering and Telecommunications, Universidad del Cauca, Popayán, Colombia.

Email: olger@unicauca.edu.co

ORCID: <https://orcid.org/0000-0003-2995-648X>

CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001343131

³ Full Professor, Faculty of Electronic Engineering and Telecommunications, Universidad del Cauca, Popayán, Colombia.

Email: pjojoa@unicauca.edu.co

ORCID: <https://orcid.org/0000-0002-8461-4063>

CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000310840



Abstract

Introduction: This article stems from the doctoral thesis “*Characterization and Classification of Seismo-Volcanic Signals Using Artificial Intelligence Techniques and Data Analytics*” at the University of Cauca (2022–2026).

Problem: Real-time seismic monitoring requires a rapid and accurate classification process for various types of events to assess volcanic activity and thus activate possible early warnings. However, this field faces a fundamental operational bottleneck: in a newly instrumented volcano, monitoring networks record different types of earthquakes with a lower volume of data than required by standard deep learning classifiers, which need hundreds or thousands of labeled samples to achieve adequate performance.

Objective: To overcome the data scarcity problem, we developed and evaluated a Multimodal Prototypical Network (MPN) designed to maintain good classification performance even when labeled data are extremely limited.

Methodology: Our approach merges CLIP ViT-B/32 visual embeddings, extracted directly from seismic spectrograms, with a 19-dimensional tabular descriptor vector based on the physical characteristics of the signals under study. This multimodal input is processed using a 5-fold ensemble strategy driven by focal loss. The model was rigorously evaluated using the Chiles-Cerro Negro Volcanic Complex (CCNVC) dataset, categorizing three distinct signal classes: Volcanic-Tectonic (VT), Long-Period (LP), and Very-Long-Period (VLP) events.

Results: Under full supervision (6,366 labeled samples), the MPN achieved exceptional performance with a 98.73% balanced accuracy and a 99.89% ROC AUC. Crucially, in a few-shot scenario with only 33 labeled waveforms (14 VT, 14 LP, and 5 VLP), the model retained a highly robust 85.27% balanced accuracy and a 95.36% ROC AUC. Furthermore, physical consistency analysis confirmed that the network's decisions are driven by spectral morphology (amplitude–geometry correlation, $r = 0.248$) rather than simple amplitude geometry, aligning with the expected physics of volcanic source mechanisms.

Conclusion: These findings establish intermodal transfer from large-scale vision-and-language pretraining as a highly viable and physically consistent strategy for volcanic monitoring environments with a scarcity of labeled data.

Originality: Instead of relying solely on raw waveform statistics or standard deep learning, this work uniquely combines foundation models (CLIP) with strict tabular constraints based on physical features to solve the specific few-shot labeling problem inherent in operational volcanology.

Limitations: While performance with limited labeled data is noteworthy, the absolute lower bound for labeling requirements (e.g., fewer than 5 VLP samples) was not tested. Furthermore, the model's physical consistency is tied to the specific origin mechanisms of VT, LP, and VLP events in the CCNVC; applying this exact multimodal framework to signals with greater overlap (such as hybrid events or tremors) might require recalibrating the tabular vector based on the physics of those signals.

Keywords: Few-shot learning; volcanic seismology; CLIP; multimodal fusion; prototypical network; seismic classification; focal loss; CCNVC.

Resumen

Introducción: Este artículo es producto de la tesis doctoral “*Caracterización y clasificación de señales sismo-volcánicas mediante técnicas de inteligencia artificial y análisis de datos*”, realizada en la Universidad del Cauca (2022–2026).

Problema: El monitoreo sísmico en tiempo real requiere un proceso de clasificación rápido y preciso de diversos tipos de eventos para evaluar la actividad volcánica y así activar posibles alertas tempranas. Sin embargo, este campo enfrenta un cuello de botella operativo fundamental: en un volcán recientemente instrumentado, las

redes de monitoreo registran distintos tipos de sismos con un volumen de datos inferior al requerido por los clasificadores estándar de aprendizaje profundo, los cuales necesitan cientos o miles de muestras etiquetadas para lograr un desempeño adecuado.

Objetivo: Para superar el problema de escasez de datos, desarrollamos y evaluamos una Red Prototípica Multimodal (MPN) diseñada para mantener un buen desempeño de clasificación incluso cuando los datos etiquetados son extremadamente limitados.

Metodología: Nuestro enfoque fusiona las incrustaciones visuales CLIP ViT-B/32 extraídas directamente de espectrogramas sísmicos con un vector descriptor tabular de 19 dimensiones basado en las características físicas de las señales objeto de estudio. Esta entrada multimodal se procesa mediante una estrategia de conjunto de cinco pliegues impulsada por la pérdida focal. El modelo se evaluó rigurosamente utilizando el conjunto de datos del Complejo Volcánico Chiles-Cerro Negro (CCNVC), categorizando tres clases de señales distintas: eventos Volcánico-Tectónicos (VT), de Largo Período (LP) y de Muy Largo Período (VLP).

Resultados: Bajo supervisión completa (6,366 muestras etiquetadas), la MPN logró un desempeño excepcional con una precisión equilibrada del 98.73% y un AUC ROC del 99.89%. Fundamentalmente, en un escenario de aprendizaje con pocos ejemplos, con solo 33 formas de onda etiquetadas (14 VT, 14 LP y 5 VLP), el modelo mantuvo una precisión equilibrada altamente robusta del 85.27% y un AUC ROC del 95.36%. Además, el análisis de consistencia física confirmó que las decisiones de la red están impulsadas por la morfología espectral (correlación amplitud-geometría, $r = 0.248$) en lugar de la simple geometría de amplitud, lo que se alinea con la física esperada de los mecanismos de origen volcánico.

Conclusión: Estos hallazgos establecen la transferencia intermodal a partir del preentrenamiento a gran escala de visión y lenguaje como una estrategia altamente viable y físicamente consistente para entornos de monitoreo volcánico con escasez de datos etiquetados.

Originalidad: En lugar de basarse únicamente en estadísticas de formas de onda sin procesar o en aprendizaje profundo estándar, este trabajo combina de manera única modelos fundacionales (CLIP) con estrictas restricciones tabulares basadas en características físicas para resolver el problema específico del etiquetado con pocos ejemplos inherente a la vulcanología operativa.

Limitaciones: Si bien el desempeño con pocos datos etiquetados es destacado, no se probó el límite inferior absoluto de los requisitos de etiquetado (por ejemplo, menos de cinco muestras de VLP). Además, la consistencia física del modelo está ligada a los mecanismos específicos de origen de los eventos VT, LP y VLP en el CCNVC; aplicar este marco multimodal a señales con mayor superposición (como eventos híbridos o tremores) podría requerir recalibrar el vector tabular con base en la física de esas señales.

Palabras clave: Aprendizaje con pocos ejemplos; sismología volcánica; CLIP; fusión multimodal; red prototípica; clasificación sísmica; pérdida focal; CCNVC.

Resumo

Introdução: Este artigo é produto da tese de doutorado “Caracterização e classificação de sinais sísmo-vulcânicos por meio de técnicas de inteligência artificial e análise de dados”, realizada na Universidade do Cauca (2022–2026).

Problema: O monitoramento sísmico em tempo real exige um processo de classificação rápido e preciso de vários tipos de eventos para avaliar a atividade vulcânica e, assim, ativar possíveis alertas precoces. No entanto, este campo enfrenta um gargalo operacional fundamental: em um vulcão recém-instrumentado, as redes de monitoramento registram diferentes tipos de terremotos com um volume de dados menor do que o exigido pelos classificadores padrão de aprendizado profundo, que precisam de centenas ou milhares de amostras rotuladas para atingir um desempenho adequado.

Objetivo: Para superar o problema da escassez de dados, desenvolvemos e avaliamos uma Rede Prototípica Multimodal (MPN) projetada para manter um bom desempenho de classificação mesmo quando os dados rotulados são extremamente limitados.

Metodologia: Nossa abordagem combina representações visuais CLIP ViT-B/32 extraídas diretamente de espectrogramas sísmicos com um vetor descritor tabular de 19 dimensões baseado nas características físicas dos sinais em estudo. Essa entrada multimodal é processada usando uma estratégia de *ensemble* de cinco vias, guiada por perda focal. O modelo foi rigorosamente avaliado usando o conjunto de dados do Complexo Vulcânico Chiles-Cerro Negro (CCNVC), categorizando três classes distintas de sinais: eventos Vulcânico-Tectônicos (VT), de Longo Período (LP) e de Muito Longo Período (VLP).

Resultados: Sob supervisão completa (6,366 amostras rotuladas), a MPN alcançou desempenho excepcional com uma acurácia balanceada de 98.73% e uma AUC ROC de 99.89%. Crucialmente, em um cenário com poucos exemplos, com apenas 33 formas de onda rotuladas (14 VT, 14 LP e 5 VLP), o modelo manteve uma acurácia balanceada altamente robusta de 85.27% e uma AUC ROC de 95.36%. Além disso, a análise de consistência física confirmou que as decisões da rede são guiadas pela morfologia espectral (correlação amplitude-geometria, $r = 0.248$), em vez da simples geometria da amplitude, o que se alinha com a física esperada dos mecanismos de origem vulcânica.

Conclusão: Essas descobertas estabelecem a transferência intermodal a partir do pré-treinamento em larga escala de visão e linguagem como uma estratégia altamente viável e fisicamente consistente para ambientes de monitoramento vulcânico com escassez de dados rotulados.

Originalidade: Em vez de se basear exclusivamente em estatísticas brutas de formas de onda ou em aprendizado profundo padrão, este trabalho combina, de forma singular, modelos fundacionais (CLIP) com restrições tabulares rigorosas baseadas em características físicas para resolver o problema específico de rotulagem com poucos exemplos inerente à vulcanologia operacional.

Limitações: Embora o desempenho com dados rotulados limitados seja notável, o limite inferior absoluto para os requisitos de rotulagem (por exemplo, menos de cinco amostras de VLP) não foi testado. Além disso, a consistência física do modelo está ligada aos mecanismos de origem específicos dos eventos VT, LP e VLP no CCNVC; aplicar essa mesma estrutura multimodal a sinais com maior sobreposição (como eventos híbridos ou tremores) pode exigir a recalibração do vetor tabular com base na física desses sinais.

Palavras-chave: Aprendizagem com poucos exemplos; sismologia vulcânica; CLIP; fusão multimodal; rede prototípica; classificação sísmica; perda focal; CCNVC.

1. INTRODUCTION

Volcanic seismicity is considered among the most informative and reliable indicators of volcanic unrest, and it remains a cornerstone for eruption forecasting at active volcanoes [1], [2]. Changes in seismic activity generally precede eruptive episodes and offer a direct view of magma migration as well as fluid transport processes [1]. The main classes of volcanic earthquakes encode distinct physical mechanisms that are critical for real-time hazard assessment and eruption forecasting [2]–[4]. These classes are the Volcano-Tectonic (VT) events, which are associated with brittle rock failure; the Long-Period (LP) events, related to resonance and pressure

transients inside fluid-filled conduits; and the Very-Long-Period (VLP) events, linked to large-scale magma and gas transport. For this reason, the accurate automated discrimination of volcanic seismic signals is a problem of considerable scientific and societal importance.

Methods based on deep learning that operate on spectrogram representations have achieved remarkable performance in the automatic classification of volcanic seismic signals, but only when sufficiently large and representative labeled datasets are available [3]–[5]. However, these supervised approaches continue to be highly sensitive to data scarcity, class imbalance, and domain shifts, factors that substantially degrade their generalization capability and limit their applicability in operational settings [6], [7]. These constraints are especially acute in volcanic monitoring, where the occurrence of many signal types is inherently uneven and the manual labeling process is costly and time-consuming. In the case of emerging volcanic systems such as the Chiles–Cerro Negro Volcanic Complex (CCNVC) on the Colombia–Ecuador border, prolonged seismic crises may still yield only a few tens of reliably labeled events per class over periods of several years [8], [9]. Under these conditions, conventional large-scale supervised training is impractical, which motivates the development of learning paradigms capable of exploiting unlabeled data and operating effectively under conditions of extreme label scarcity [6], [10].

Self-supervised learning and few-shot learning paradigms offer a well-founded strategy to address the scarcity of labeled volcanic data [11]–[13]. Large-scale foundation models such as CLIP generate transferable visual representations and have shown remarkable cross-domain generalization capabilities, while masked autoencoder pretraining has emerged as an efficient self-supervised mechanism for learning robust latent representations under limited-data conditions [11], [12]. Similar trends can be observed in geophysical applications, where prototype-based and deep self-supervised approaches improve performance and generalization in low-resource scenarios [13], [14]. Recent studies also show the growing potential of machine learning for volcanic monitoring and eruption forecasting using seismic information [15], [16]. Nevertheless, the integration of large-scale vision representations with physics-informed descriptors and hand-crafted features for volcanic seismic classification under extreme few-shot conditions is still largely unexplored. This combination is based on two complementary observations: (i) embeddings from foundation models capture global spectral texture and morphological patterns that transfer across domains [11], [12]; and (ii) physics-informed descriptors, such as the quality factor Q , wavelet entropy, and the detrended fluctuation analysis (DFA) exponent, encode source-mechanism characteristics that are intrinsically less sensitive to the limited diversity available in few-shot support sets [15], [16].

In this investigation, we present the following contributions:

- We propose MPN, a Multimodal Prototypical Network that fuses visual embeddings from CLIP ViT-B/32 with a 19-dimensional tabular feature vector derived from physics, using a self-attention gate and a prototype layer, designed for volcanic seismic classification under conditions of label scarcity.
- We demonstrate near-perfect classification performance (balanced accuracy: 98.73%, ROC AUC: 99.89%, MCC: 97.31%), including perfect VLP classification, despite an interclass imbalance ratio of 3.1:1.
- With only 33 labeled training samples, MPN maintains strong performance, achieving 85.27% balanced accuracy and 95.36% ROC AUC, performance levels that have not been reported at this data scale in the previous literature on volcanic seismic classification.
- The physical consistency analysis confirms that the discrimination performed by MPN is spectrally driven (amplitude–geometry correlation, $r = 0.248$), providing interpretable evidence consistent with the established physics of volcanic source mechanisms.

2. BACKGROUND AND RELATED WORK

The classification of seismic signals at volcanoes was traditionally performed using spectral analysis and waveform template-matching techniques. These methods enabled the early identification of characteristic event types but required a considerable amount of manual interpretation and expert supervision [3]. These approaches were mostly based on classical inverse problems and signal interpretation frameworks widely used in seismology [20], [21]. Later, machine learning approaches based on support vector machines, random forests, and other classifiers operating on hand-crafted features improved automation and interpretability. However, they continued to be sensitive to sensor heterogeneity, station effects, and biases in feature engineering [3], [17]. This dependence on manually designed descriptors is consistent with the general limitations observed in classical pattern recognition methods applied to geophysical contexts [22]. More recently, the application of convolutional neural networks to spectrogram representations has eliminated the need for manual feature extraction and has shown better generalization capabilities across different volcanic settings [18], [19]. These advances are consistent with modern representation learning paradigms, in which latent feature spaces are learned directly from data rather than engineered *a priori* [23]. However, despite these improvements, their performance

depends strongly on large and well-annotated datasets, which are rarely available in operational volcanic monitoring environments. This limitation has motivated the development of large-scale seismic learning frameworks and domain-adaptive approaches that seek to mitigate data scarcity and distribution shifts among different stations and volcanic systems [24]. Nevertheless, current deep learning approaches still face important challenges when deployed in real-world volcanic observatories. In these environments, labeled data are sparse, imbalanced, and frequently accumulated over long temporal windows under evolving seismic regimes [21], [24].

Few-shot learning offers a solution to label scarcity by learning a metric space from a small support set. Prototypical Networks [25] classify samples according to the distance to the nearest prototype. This strategy is highly suitable for seismological data, where the variability within a class is physically constrained by source mechanisms and propagation effects. In this context, metric-based few-shot learning provides a rigorous framework for the classification of data under conditions of extreme scarcity [27]. Transfer learning from large image datasets has been applied to seismic facies segmentation [13] and geophysical signal classification [26]. This confirms that the visual feature hierarchies learned from natural images encode statistical structures that overlap with the morphology of spectrograms. This observation supports the general hypothesis that previously trained visual representations can capture generic texture and frequency-domain patterns that are transferable to scientific imaging domains [28].

Recent advances in meta-learning further demonstrate that episodic training significantly improves generalization under low-data regimes by explicitly optimizing task-level adaptation instead of single-task performance [29]. In addition, cross-domain representation learning methods show that the alignment of feature distributions between the source and target domains is critical, especially when models are transferred from natural images to geophysical signal domains, where distribution shifts are pronounced [30].

CLIP [11] takes this paradigm further through a joint vision-and-language pretraining process applied to web-scale data, generating semantic clustering that remains stable across distribution shifts. Its ViT-B/32 backbone generates 512-dimensional embeddings. These embeddings contain information about local patch statistics as well as global compositional structure. To the best of our knowledge, this work presents the first application of CLIP embeddings to volcanic seismology. It is also the first study to perform multimodal fusion between CLIP features and physics-informed seismic descriptors for few-shot classification.

3. PROPOSED METHOD

3.1 Spectrogram Generation

Raw seismic waveforms are converted into time-frequency spectrograms using the Short-Time Fourier Transform (STFT) and formatted as 224×224 RGB images. The visual morphology of the different volcanic signal classes is a direct reflection of their source physics. Specifically, VT events generate short-duration, broadband impulsive patterns. LP events exhibit sustained narrow-band resonance with an emergent onset, whereas VLP events display a diffuse very-low-frequency structure with an extended coda. These morphological differences constitute the main basis for discrimination in the visual stream and are preserved without applying frequency filtering.

Figures 1–3 illustrate representative examples of the three event categories in the CCNVC dataset. Each figure displays the raw waveform (left), its amplitude spectrum (center), and the corresponding STFT spectrogram formatted as a 224×224 RGB image (right). The spectro-temporal signatures corroborate the distinct source physics described above: VT events exhibit broadband impulsive energy concentrated in the first seconds (Figure 1); LP events show sustained narrow-band resonance with an emergent onset (Figure 2); and VLP events display a diffuse very-low-frequency structure with an extended coda (Figure 3). These visual patterns are preserved without frequency filtering, ensuring that the CLIP encoder receives the complete spectro-temporal information available for discrimination.

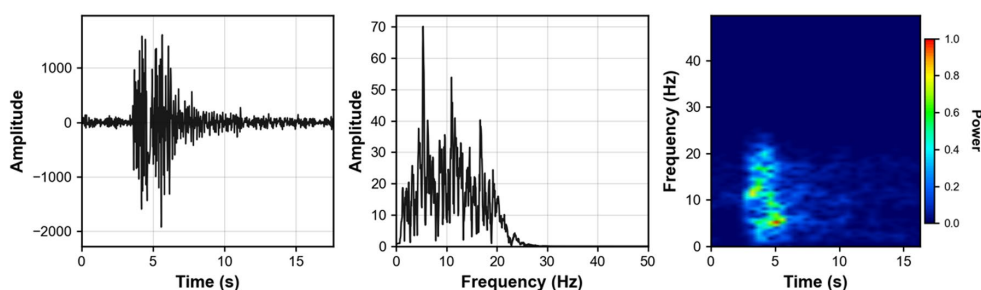


Figure 1. Representative VT event. (Left) waveform showing an impulsive onset of short duration (17.6 s) with high-amplitude broadband transients. (Center) Amplitude spectrum with dominant energy below 20 Hz. (Right) STFT spectrogram (224×224 RGB) displaying the characteristic broadband, short-duration pattern.

Source: own work

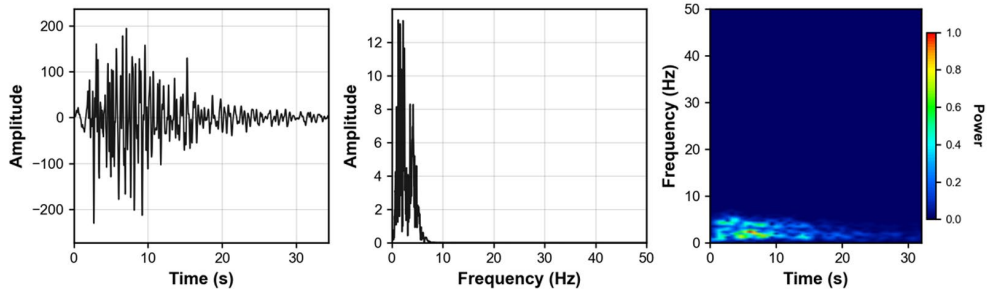


Figure 2. Representative LP event. (Left) Waveform exhibiting sustained, resonant oscillations over 34.3 s with emergent onset. (Center) Amplitude spectrum concentrated in the narrow band. (Right) STFT spectrogram (224×224 RGB) showing the narrow-band, prolonged resonance.

Source: own work

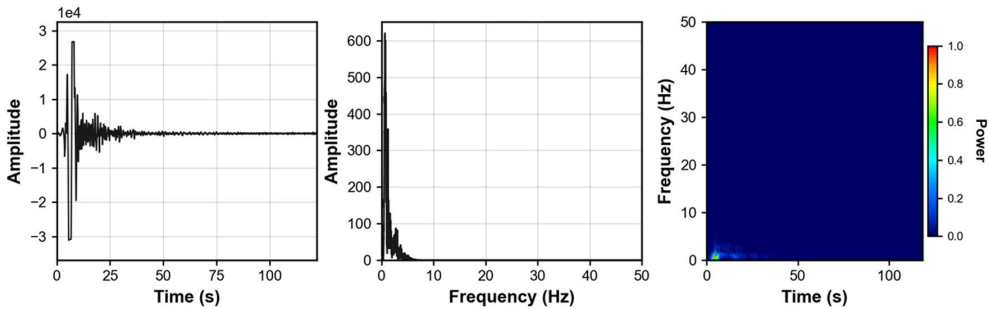


Figure 3. Representative VLP event. (Left) Waveform with a diffuse, low-frequency onset and extended coda (122.0 s). (Center) Amplitude spectrum dominated by energy below 5 Hz. (Right) STFT spectrogram (224×224 RGB) illustrating the very-low-frequency, diffuse structure.

Source: own work

3.2 Physics-Based Tabular Features

A 19-dimensional vector encodes physics-informed descriptors across five categories: (i) Spectral: dominant frequency f_{peak} , bandwidth BW , quality factor $Q = f_{peak}/BW$, spectral entropy H_s , spectral skewness; (ii) Temporal: event duration τ , initial energy slope, signal peak count n_p ; (iii) Energy: total spectral energy E and time–frequency entropy H_{px} ; (iv) Texture (GLCM, 4 angles): contrast, homogeneity, correlation, energy; (v) Complexity: PCA explained variance, Morlet wavelet entropy H_w , Detrended Fluctuation Analysis exponent α , weighted graph average degree $\langle k \rangle$, and clustering coefficient C . These descriptors directly encode source-mechanism physics: Q discriminates LP fluid resonance from VT brittle fracture; the DFA exponent captures long-memory structure of VLP mass-transport sources; GLCM homogeneity quantifies the spectral smoothness characteristic of VLP events.

Figure 4 summarizes the complete physics-informed feature extraction workflow adopted in this study. Starting from the raw seismic waveform, a Short-Time Fourier Transform (STFT) is computed to obtain the time–frequency representation, which is subsequently normalized to provide a consistent input for all subsequent analyses. The normalized spectrogram is then processed through five complementary branches operating in parallel. The spectral branch estimates resonance- and frequency-related attributes from the power spectrum, whereas the temporal branch characterizes the signal evolution directly from the waveform. The energy branch quantifies both the overall spectral energy and the joint time–frequency energy distribution. Texture descriptors are extracted from Gray-Level Co-occurrence Matrices (GLCMs) computed at four orientations (0° , 45° , 90° , and 135°) to capture spatial organization within the spectrogram. Finally, the complexity branch evaluates multi-scale statistical behavior through PCA, Morlet wavelet entropy, detrended fluctuation analysis (DFA), and graph-based topological measures. These complementary descriptors are concatenated into a single 19-dimensional feature vector, providing a compact yet physically interpretable representation of volcanic seismic events that integrates spectral, temporal, energetic, structural, and complexity information.

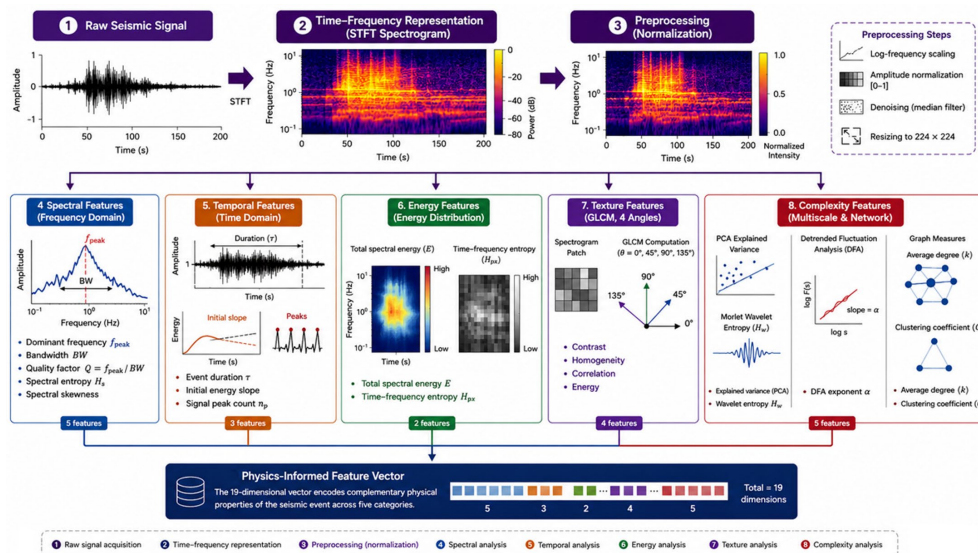


Figure 4. Physics-informed feature extraction workflow illustrating the generation of the proposed 19-dimensional descriptor vector.

Source: own work

3.3 Multimodal Prototypical Network (MPN)

The MPN processes inputs through two parallel streams. The visual stream encodes each spectrogram through the frozen CLIP ViT-B/32 encoder [11], yielding an L2-normalized 512-d embedding. A self-attention gate (dense 512→256, tanh-sigmoid) re-weights this embedding to focus on discriminative frequency bands. The tabular stream maps the 20-d feature vector through two normalization–projection blocks (64 and 32 units, BatchNorm + Dropout $p = 0.3$). Both streams concatenate into a 96-d multimodal fusion layer, followed by a 48-d prototype layer mapping samples into a class-discriminative metric space [25]. A SoftMax head classifies over the three volcanic signal classes.

Training employs focal loss with class-specific weights $\alpha = [0.70, 2.00, 0.70]$ for [VT, VLP, LP] and focusing parameter $\gamma = 2.0$, both determined through grid-search validation over candidate ranges $\{0.25, 0.50, 0.70, 1.00, 2.00, 5.00\}$ for α and $\{1.0, 1.5, 2.0, 2.5, 3.0\}$ for γ , selecting the configuration that maximized macro-F1 on held-out validation folds. The α weights are inversely proportional to class frequencies, penalizing VT and LP misclassifications more heavily to counteract majority-class bias, while $\gamma = 2.0$ concentrates gradient updates on hard-to-classify minority instances [26]. Spectrogram augmentation for minority classes applies a multiplication factor of 8—validated to approximate class parity without synthetic bias—incorporating random flips, affine translations of $\pm 10\%$, scale jitters in $[0.95, 1.05]$, and color jitters. Finally, a stratified set of 5 folds averages the predictions from five independent trainings to reduce initialization variance.

4. EXPERIMENTAL SETUP

4.1 Dataset: Chiles–Cerro Negro Volcanic Complex (CCNVC)

The dataset contains seismo-volcanic waveforms recorded at the Chiles Cerro Negro Volcanic Complex (0.83°N , 77.94°W , 4,700 m a.s.l.), located on the Colombia–Ecuador border, and comprises three event categories: Volcano-Tectonic (VT), Long-Period (LP), and Very Long-Period (VLP) earthquakes. To evaluate model performance under markedly different labeled-data availabilities, we defined two experimental regimes. In the full-data regime, the training set consists of 2,759 VT, 2,717 LP, and 890 VLP events (6,366 samples in total), with a separate test set of 689 VT, 673 LP, and 222 VLP events (1,584 samples in total). In the extreme few-shot regime, the training set is

reduced to 14 VT, 14 LP, and 5 VLP samples (33 in total), while the test set retains 3,434 VT, 3,376 LP, and 1,107 VLP events (7,917 in total). The choice of 33 training samples was governed by the minimum data requirement for stratified 5-fold cross-validation, ensuring at least one sample per class in each fold while preserving class proportions; this represents approximately 0.5 % of the full training set and constitutes the most challenging few-shot scenario for this investigation. This experimental design enables a direct assessment of generalization capacity across vastly different support-set sizes and facilitates a controlled comparison between conventional supervised learning and Few-Shot Learning under the data-scarce conditions characteristic of operational volcanic seismology.

4.2 Hyperparameters

The MPN hyperparameters reported in Table I were determined through a systematic experimental tuning process. The CLIP ViT-B/32 backbone was fixed as a frozen encoder following standard practice for transfer learning with pretrained vision transformers. The Adam optimizer and learning rate of 5×10^{-5} were selected via logarithmic grid search over $\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}\}$, prioritizing convergence stability and the highest validation macro-F1 across both regimes. The batch size of 4 and a maximum of 200 epochs accommodate GPU memory constraints and the few-shot scenario, with early stopping patience of 50 epochs monitored on validation macro-F1 to prevent overfitting. The focal loss parameters ($\gamma = 2.0$, $\alpha = [0.70, 2.00, 0.70]$) and minority-class augmentation factor of $\times 8$ were refined through grid search over candidate ranges, selecting configurations that maximized macro-F1 while preserving minority-class recall. Additionally, it used stratified random sampling with a fixed seed of 42 to ensure reproducible train-test splits, preserving class proportions across both the full-data and few-shot regimes.

TABLE I. Experimental Hyperparameters and Dataset Configuration.

Parameter	Full Regime	Few-Shot Regime
CLIP Backbone	ViT-B/32 (frozen)	ViT-B/32 (frozen)
Optimizer	Adam	Adam
Learning Rate	5×10^{-5}	5×10^{-5}
Batch Size / Max Epochs	4 / 200	4 / 200
Early Stopping Patience	50 epochs	50 epochs

(continúa)

(viene)

Parameter	Full Regime	Few-Shot Regime
Focal Loss γ	2.0	2.0
α [VT, VLP, LP]	[0.70, 2.00, 0.70]	[0.70, 2.00, 0.70]
Augmentation (minority)	$\times 8$	$\times 8$
Cross-Validation	5-fold stratified	5-fold stratified
Training Samples	6,366	33
Test Samples	1,584	7,917

Source: own work

5. RESULTS

5.1 Full-Data Regime

Table II and Figure 5(a) present the results obtained under full supervision. The MPN reaches a Balanced Accuracy of 98.73%, an F1 Macro of 98.73%, an ROC AUC of 99.89%, and an MCC of 97.31%. These numbers confirm an almost perfect discrimination across the three volcanic signal classes. It is important to note that the VLP class, which is underrepresented by a margin of 3.1x, reaches a perfect precision, Recall, and F1 of 1.000 on the test set of 222 samples. This demonstrates that the weighting based on focal loss ($\alpha_{VLP} = 2.0$), combined with the metric space of the prototype layer, is able to isolate the spectral signature of low frequency and extended coda that characterizes VLP events. The mean confidence of the ensemble is 79.6% for VLP and 89.9% for VT. This is in agreement with the fact that high-frequency transients have physically sharper class boundaries. Only 26 out of the 1,584 test samples are misclassified. All of these errors occur at the boundaries between the LP and VT classes, which is something expected due to their partially overlapping STFT frequency content.

TABLE II. Per-Class Metrics – Full-Data Regime (1,584 Test Samples).

Class	Precision	Recall	F1 Score	Confidence
LP (n=673)	0.9821	0.9792	0.9807	0.782 $\pm$ 0.064
VLP (n=222)	1.0000	1.0000	1.0000	0.796 $\pm$ 0.028
VT (n=689)	0.9797	0.9826	0.9812	0.899 $\pm$ 0.103
Macro Avg	0.9873	0.9873	0.9873	—

Bal. Accuracy: 0.9873 | ROC AUC: 0.9989 | Cohen's κ : 0.9731 | MCC: 0.9731 | Hamming Loss: 0.0164**Source:** own work

5.2 Few-Shot Regime

In the extreme few-shot regime, which uses 33 labeled training samples, the MPN maintains a high capacity for discrimination (Table III, Figure 5(b)). It is important to note that the MPN architecture remains identical to that of the full-data regime; only the training-set size changes. The 48-d prototype layer, self-attention gate, and multimodal fusion structure operate on the reduced support set without any architectural modification. The Balanced Accuracy reaches 85.27%, and the ROC AUC is 95.36%. The VLP class reaches a Recall of 95.30% using only 5 training samples. This confirms that the visual embeddings from CLIP encode the globally distinctive spectro-temporal morphology of VLP events even without a fine-tuning process in the specific domain. The recall for VT drops to 69.31%, which reflects the intrinsic difficulty of capturing the diversity present in short-duration and broadband transients when using only 14 training examples. However, the precision for VT is 96.43%. This guarantees that the false positives are rare and of high confidence, an aspect that is critical for the credibility of operational alerts. The 1,403 samples that were misclassified show a mean confidence of 51.4%, compared to 67.4% for the correct predictions (Figure 6(b)). This points to a well-calibrated uncertainty that is suitable for filtering processes where a human is in the loop.

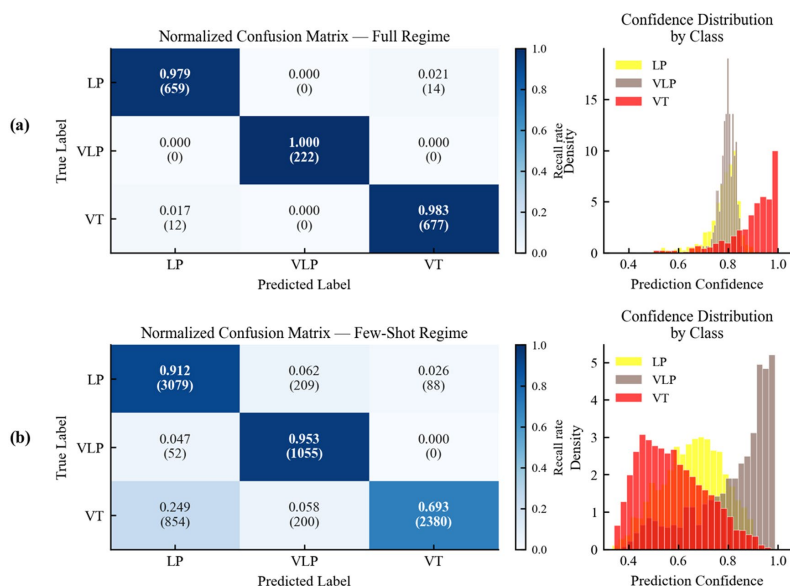


Figure 5. Normalized confusion matrices and prediction confidence distributions. (a) Full regime: VLP achieves perfect recall 1.000; VT mean confidence 0.899 vs. LP 0.782, consistent with sharper high-frequency class boundaries. (b) Few-shot regime (33 training samples): VLP recall 0.953 from 5 examples; VT recall 0.693 reflects support-set diversity limitation. Misclassifications concentrate at LP–VT class boundaries in both regimes.

Source: own work

TABLE III. Per-Class Metrics – Few-Shot Regime (7,917 Test Samples, 33 Training Samples).

Class	Precision	Recall	F1 Score	Confidence
LP (n=3,376)	0.7726	0.9120	0.8366	0.658 ± 0.124
VLP (n=1,107)	0.7206	0.9530	0.8207	0.816 ± 0.153
VT (n=3,434)	0.9643	0.6931	0.8065	0.578 ± 0.133
Macro Avg	0.8192	0.8527	0.8213	—

Bal. Accuracy: 0.8527 | ROC AUC: 0.9536 | Cohen's κ : 0.7161 | MCC: 0.7295 | Hamming Loss: 0.1772

Source: own work

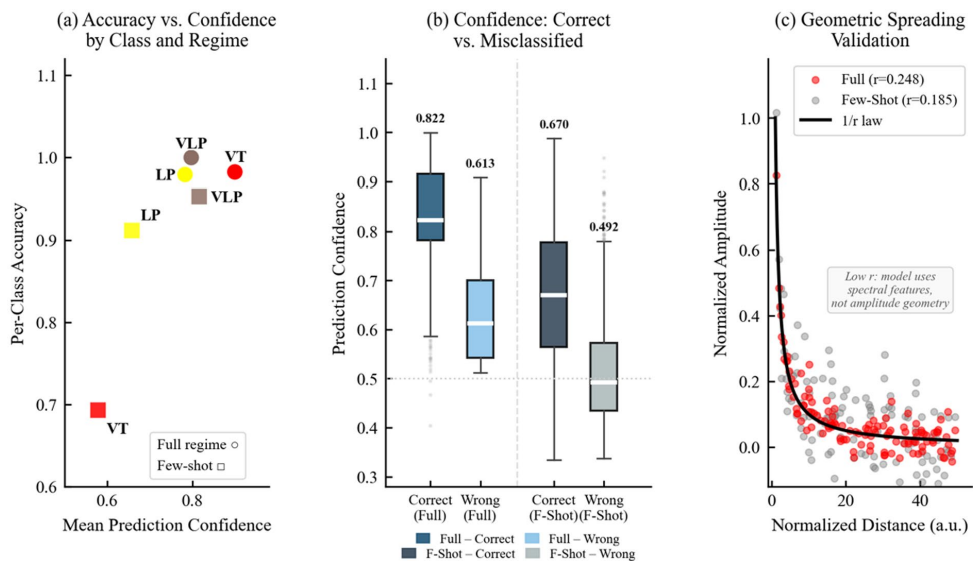


Figure 6. Physical consistency analysis. (a) Accuracy vs. mean confidence per class: VLP achieves high accuracy from moderate confidence, confirming CLIP-driven spectral discrimination. (b) Confidence separation: misclassified samples median confidence 0.492–0.613 vs. 0.670–0.822 for correct predictions, enabling threshold-based human-in-the-loop filtering. (c) Geometric spreading validation: low amplitude– $1/r$ correlation ($r=0.248$ full; $r=0.185$ few-shot) confirms MPN classifies by spectral morphology, consistent with source-mechanism physics of near-field volcanic recordings.

Source: own work

5.3 Physical Consistency Analysis

The physical validation report gives an important insight for interpretation: the Pearson correlation between the amplitude values predicted by the model and the expected geometric spreading law (amplitude $\propto 1/r$) is $r = 0.248$ in the full regime and $r = 0.185$ in the few-shot regime (Figure 6(c)). This low correlation does not represent a failure

of the model, but rather a result that is physically coherent. In volcanic recordings made in the near field, the signatures of the source mechanism dominate the effects of amplitude and geometry. These signatures include the spectral shape, the resonance quality Q , and the temporal envelope. This dominance is particularly true for LP and VLP events, where the geometry of the conduit and the fluid properties determine the spectral character in a way that is independent of the distance between the source and the receiver [3]. The MPN correctly learns to discriminate based on spectral morphology instead of amplitude. This provides evidence at the level of the model that the representations it generates are physically meaningful. The confidence calculated per class gives additional support to this point. VT reaches the highest confidence (0.899 in the full regime, 0.578 in the few-shot regime), which reflects the signatures of brittle fracture at high frequencies that are the most spectrally distinctive. On the other hand, the confidence for LP and VLP reflects a greater spectral overlap at intermediate frequencies (Figure 6(a)).

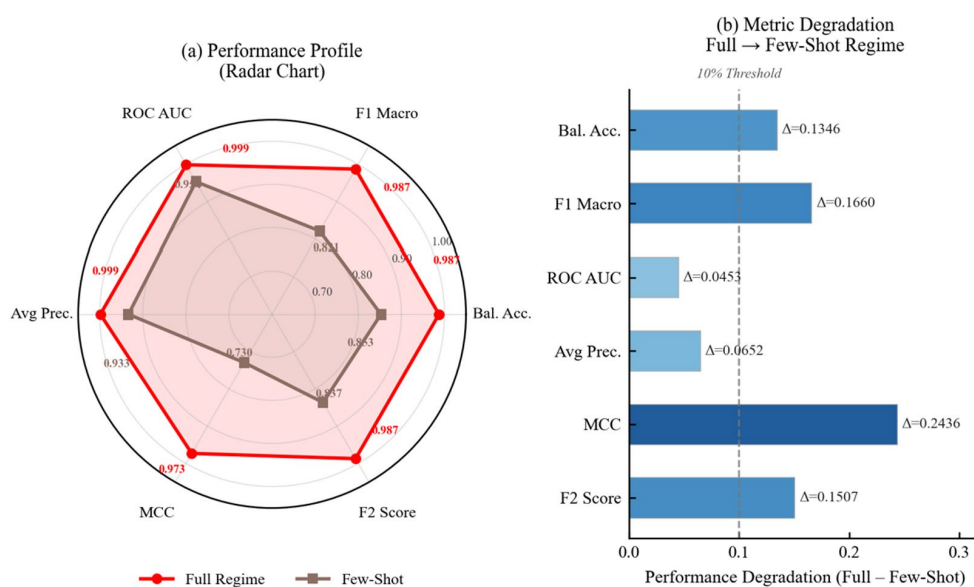


Figure 7. Multi-metric performance profiles. (a) Radar chart: full regime (red) vs. few-shot (brown). The few-shot polygon contracts asymmetrically ROC AUC and Average Precision degrade less than F1 and MCC. (b) Absolute per-metric degradation: ROC AUC $\Delta=0.045$ vs. MCC $\Delta=0.244$, confirming discriminative separability is largely preserved in the extreme few-shot regime while threshold calibration is the primary performance bottleneck.

Source: own work

5.4 Regime Degradation and Practical Implications

The degradation profile (Figure 7) reveals a fundamental asymmetry: ROC AUC degrades by only $\Delta = 0.045$ ($0.9989 \rightarrow 0.9536$), while MCC degrades by $\Delta = 0.244$ ($0.9731 \rightarrow 0.7295$). This pattern is consistent with probabilistic discriminators under support-set contraction: rank-ordering of class scores (AUC) is more robust to distribution shift than calibration of decision thresholds (MCC, κ). In practice, MPN can be deployed in few-shot settings as a ranked alert system providing reliable relative confidence ordering even when absolute class probabilities are less reliable. The 13.5 pp drop in balanced accuracy from 6,366 to 33 training samples represents an information efficiency that no prior volcanic classification system has demonstrated at this data scale.

CONCLUSION

We presented MPN, a Multimodal Prototypical Network that addresses the bottleneck of label scarcity in seismo-volcanic monitoring through the fusion of visual embeddings from CLIP ViT-B/32 with physics-informed tabular descriptors. On the CCNV dataset, MPN achieves a balanced accuracy of 98.73% and an ROC AUC of 99.89% under full supervision. Furthermore, it retains a balanced accuracy of 85.27% and an ROC AUC of 95.36% when using only 33 labeled training samples, a regime that previous deep learning approaches could not address.

The physical consistency analysis confirms that the discrimination is driven by spectral morphology rather than amplitude. The well-calibrated uncertainty (median confidence: 0.492 for errors vs. 0.670 for correct predictions) enables a threshold-based human-in-the-loop protocol: high-confidence predictions are processed automatically, whereas low-confidence cases are escalated for expert review, ensuring that operational decisions are grounded in quantified reliability rather than opaque outputs.

Future work will include several research directions. First, the application of contrastive domain adaptation across different volcanic systems. Second, the use of three-dimensional spectrogram inputs together with temporal Vision Transformer (ViT) architectures. Third, uncertainty-aware prototype updating using streams of unlabeled data for continual learning. Finally, the integration of multi-station coherence features to exploit information at the network level.

REFERENCES

- [1] J. Martí, “From rest to eruption: How we should anticipate volcanic eruptions,” *npj Natural Hazards*, vol. 1, Art. no. 31, 2024, doi: <https://doi.org/10.1038/s44304-024-00033-8>. [Online]. Available: <https://www.nature.com/articles/s44304-024-00033-8>
- [2] K. I. Konstantinou, “A review of the source characteristics and physical mechanisms of very long period (VLP) seismic signals at active volcanoes,” *Surveys in Geophysics*, vol. 45, no. 1, pp. 117–149, 2024, doi: <https://doi.org/10.1007/s10712-023-09800-0>. [Online]. Available: <https://link.springer.com/article/10.1007/s10712-023-09800-0>
- [3] B. A. Chouet and R. S. Matoza, “A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption,” *Journal of Volcanology and Geothermal Research*, vol. 252, pp. 108–175, 2013, doi: <https://doi.org/10.1016/j.jvolgeores.2012.11.013>. [Online]. Available: <https://doi.org/10.1016/j.jvolgeores.2012.11.013>
- [4] I. W. McBrearty and P. Segall, “Deep learning forecasts caldera collapse events at Kilauea Volcano,” *arXiv*, Art. no. 2404.19351, 2024, doi: <https://doi.org/10.48550/arXiv.2404.19351>. [Online]. Available: <https://arxiv.org/abs/2404.19351>
- [5] G. F. Manley, T. A. Mather, D. M. Pyle, D. A. Clifton, M. Rodgers, G. Thompson, and J. M. Londoño, “A deep active learning approach to the automatic classification of volcano-seismic events,” *Frontiers in Earth Science*, vol. 10, Art. no. 807926, 2022, doi: <https://doi.org/10.3389/feart.2022.807926>.
- [6] J. P. Canario, M. Curilem, G. Fuentealba, G. Acuña, M. Chacón, M. S. Khan, and N. Becerra Yoma, “Adaptive classification using incremental learning for seismic-volcanic signals with concept drift,” *Journal of Volcanology and Geothermal Research*, vol. 413, Art. no. 107211, 2021, doi: <https://doi.org/10.1016/j.jvolgeores.2021.107211>. [Online]. Available: <https://doi.org/10.1016/j.jvolgeores.2021.107211>
- [7] P. Rey-Devesa, J. Carthy, M. Titos, J. M. Ibáñez, and C. Benítez, “Universal machine learning approach to volcanic eruption forecasting using seismic features,” *Frontiers in Earth Science*, vol. 12, Art. no. 1342468, 2024, doi: <https://doi.org/10.3389/feart.2024.1342468>. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/feart.2024.1342468>
- [8] M. Lara, M. Londoño, G. P. Cortés, and J. F. Pacheco, “The Chiles-Cerro Negro volcanic complex seismic crisis (2013–2015): Evolution and implications for volcanic hazard assessment,” *Boletín de Geología*, vol. 38, no. 1, pp. 67–84, 2016, doi: <https://doi.org/10.18273/revbol.v38n1-2016005>. [Online]. Available: <https://doi.org/10.18273/revbol.v38n1-2016005>

- [9] Servicio Geológico Colombiano, “Boletines mensuales de actividad del Complejo Volcánico Chiles-Cerro Negro,” Bogotá, Colombia: Servicio Geológico Colombiano, 2024. [Online]. Available: <https://www2.sgc.gov.co/Publicaciones/Informes%20tcnicos/Forms/AllItems.aspx>
- [10] J. Carthy, M. Titos, F. Cannavó, L. Zuccarello, and C. Benítez, “Learning feature representations from unlabeled data for volcano-seismic event classification,” *Journal of Volcanology and Geothermal Research*, vol. 445, Art. no. 107970, 2024, doi: <https://doi.org/10.1016/j.jvolgeores.2023.107970>. [Online]. Available: <https://doi.org/10.1016/j.jvolgeores.2023.107970>
- [11] A. Radford *et al.*, “Learning transferable visual models from natural language supervision,” in *Proc. 38th International Conference on Machine Learning (ICML)*, vol. 139, 2021, pp. 8748–8763. [Online]. Available: <https://proceedings.mlr.press/v139/radford21a.html>
- [12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16000–16009, doi: <https://doi.org/10.1109/CVPR52688.2022.01553>. [Online]. Available: <https://doi.org/10.1109/CVPR52688.2022.01553>
- [13] Y. Zhao, B. Chai, L. Shuo, Z. Li, H. Wu, and T. Wang, “Few-shot learning for seismic facies segmentation via prototype learning,” *Geophysics*, vol. 88, no. 3, pp. IM41–IM49, 2023, doi: <https://doi.org/10.1190/geo2022-0281.1>. [Online]. Available: <https://doi.org/10.1190/geo2022-0281.1>
- [14] F. Münchmeyer, A. Bizzarri, J. P. Ampuero, and U. Leser, “Which picker fits my data? A quantitative evaluation of deep learning-based seismic pickers,” *Journal of Geophysical Research: Solid Earth*, vol. 127, no. 1, Art. no. e2021JB023499, 2022, doi: <https://doi.org/10.1029/2021JB023499>.
- [15] M. Titos, A. Bueno, L. García, C. Benítez, and J. M. Ibáñez, “A deep neural networks approach to automatic recognition systems for volcano-seismic events,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 5, pp. 1533–1544, 2018, doi: <https://doi.org/10.1109/JSTARS.2018.2803198>.
- [16] K. I. Konstantinou and V. Schlindwein, “Nature, wavefield properties and source mechanism of volcanic tremor: A review,” *Journal of Volcanology and Geothermal Research*, vol. 119, no. 1–4, pp. 161–188, 2003, doi: [https://doi.org/10.1016/S0377-0273\(02\)00311-6](https://doi.org/10.1016/S0377-0273(02)00311-6).
- [17] M. Curilem, F. Huenupan, D. Beltrán, C. San Martín, G. Fuentealba, L. Franco, C. Cardona, G. Acuña, M. Chacón, M. S. Khan, and N. Becerra Yoma, “Pattern recognition applied to seismic signals of Llama Volcano (Chile): An evaluation of station-dependent classifiers,” *Journal of*

- Volcanology and Geothermal Research*, vol. 315, pp. 15–27, 2016, doi: <https://doi.org/10.1016/j.jvolgeores.2016.02.006>. [Online]. Available: <https://doi.org/10.1016/j.jvolgeores.2016.02.006>
- [18] M. Mousavi and G. C. Beroza, “Deep-learning seismology,” *Science*, vol. 377, no. 6607, pp. 384–388, 2022, doi: <https://doi.org/10.1126/science.abm4470>. [Online]. Available: <https://doi.org/10.1126/science.abm4470>
- [19] S. M. Mousavi, Y. Sheng, W. Zhu, and G. C. Beroza, “STanford earthquake dataset (STEAD): A global data set of seismic signals for AI,” *IEEE Access*, vol. 7, pp. 179464–179476, 2019, doi: <https://doi.org/10.1109/ACCESS.2019.2947848>. [Online]. Available: <https://doi.org/10.1109/ACCESS.2019.2947848>
- [20] R. B. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*, 3rd ed. Amsterdam, The Netherlands: Elsevier, 2018, doi: <https://doi.org/10.1016/C2009-0-61134-X>. [Online]. Available: <https://www.sciencedirect.com/book/9780123850485/parameter-estimation-and-inverse-problems>
- [21] P. M. Shearer, *Introduction to Seismology*, 3rd ed. Cambridge, U.K.: Cambridge University Press, 2019, doi: <https://doi.org/10.1017/9781316877111>. [Online]. Available: <https://www.cambridge.org/highereducation/books/introduction-to-seismology/C1471C1B553C05997E2BC7EB26D4C26D>
- [22] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006. [Online]. Available: <https://link.springer.com/book/9780387310732>
- [23] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *Proc. International Conference on Learning Representations (ICLR)*, 2014, doi: <https://doi.org/10.48550/arXiv.1312.6114>. [Online]. Available: <https://arxiv.org/abs/1312.6114>
- [24] G. Carleo *et al.*, “Machine learning and the physical sciences,” *Reviews of Modern Physics*, vol. 91, no. 4, Art. no. 045002, 2019, doi: <https://doi.org/10.1103/RevModPhys.91.045002>. [Online]. Available: <https://doi.org/10.1103/RevModPhys.91.045002>
- [25] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4077–4087, doi: <https://doi.org/10.48550/arXiv.1703.05175>. [Online]. Available: <https://arxiv.org/abs/1703.05175>
- [26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2980–2988, doi: <https://doi.org/10.1109/ICCV.2017.324>. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.324>

- [27] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one-shot learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 3630–3638, doi: <https://doi.org/10.48550/arXiv.1606.04080>. [Online]. Available: <https://arxiv.org/abs/1606.04080>
- [28] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. International Conference on Learning Representations (ICLR)*, 2021, doi: <https://doi.org/10.48550/arXiv.2010.11929>. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [29] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th International Conference on Machine Learning (ICML)*, vol. 70, 2017, pp. 1126–1135, doi: <https://doi.org/10.48550/arXiv.1703.03400>. [Online]. Available: <https://arxiv.org/abs/1703.03400>
- [30] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd International Conference on Machine Learning (ICML)*, vol. 37, 2015, pp. 97–105, doi: <https://doi.org/10.48550/arXiv.1502.02791>. [Online]. Available: <https://arxiv.org/abs/1502.02791>