What is Machine Learning Teaching Us? Explainable AI for Seismic Models

Michele Magrini[1], Francesco Marrocco[1]

Abstract
Understanding and predicting seismic events is a critical challenge in geophysics. Machine learning (ML) has emerged as a powerful tool for analyzing seismic data, yet the black-box nature of many ML models limits their interpretability and applicability in scientific contexts. This study investigates the integration of explainable AI techniques with ML models to classify seismic events and predict earthquake magnitudes, aiming to enhance both performance and interpretability.
For the classification task, a Convolutional Neural Network (CNN) was trained on seismic spectrograms to distinguish between foreshocks and aftershocks. SHapley Additive exPlanations (SHAP) analysis revealed distinct patterns in frequency and timing around P-wave arrivals, providing insights into the model's decisions and seismic processes. The CNN achieved high accuracy on station-specific datasets but struggled with generalization to diverse seismic stations, highlighting the importance of dataset diversity in training.
For magnitude prediction, a Random Forest Regressor was employed, achieving robust performance metrics with an R^2 score of 0.8834. SHAP analysis identified key spectral and temporal features, such as mean spectral power and P-wave travel times, as critical drivers of the model's predictions. These findings emphasize the role of explainable ML models in bridging the gap between data-driven methods and domain-specific insights.
This study not only demonstrates the potential of interpretable ML in seismic research but also outlines pathways for improving model generalizability and incorporating domain knowledge, paving the way for advancements in earthquake analysis and real-time monitoring systems.

Keywords
Earthquake Physics — Machine Learning — Explainability

[1] *Department of Mathematics, La Sapienza University of Rome, Rome, Italy*

Contents

Introduction

Understanding earthquakes and their effects is a key challenge in geophysics. ML has become an important tool for analyzing seismic data, as it can detect patterns and make predictions from complex datasets. However, many ML models are difficult to interpret, which makes it harder to trust and apply their results in scientific contexts.

In this project, titled *"What is Machine Learning Teaching Us? Explainable AI for Seismic Models"*, we focus on using ML models for two main tasks: classifying seismic events as either aftershocks or foreshocks, and predicting earthquake magnitudes. We also aim to improve the interpretability of these models using SHapley Additive exPlanations (SHAP) [1], a method for understanding the contributions of input features to the models' predictions.

The first experiment uses a CNN inspired by the work of Laura Laurenti et al. [2] to classify aftershocks and foreshocks. SHAP is applied to explain how the model makes its predictions and identify which features are most important.

The second experiment involves predicting earthquake magnitudes using a Random Forest regression model. We again use SHAP to explore the key factors influencing the

model's predictions and to provide insights into the relationships between features and earthquake magnitudes.

This project combines ML techniques with explainability tools to better understand how these models work and to explore what they can teach us about seismic processes.

1. CNN for the Classification of Foreshocks and Aftershocks

1.1 Methods

For the first task, we focused on the classification of seismic events into aftershocks and foreshocks. The dataset used in this experiment was downloaded from Zenodo [3], as referenced in the work of Laurenti et al. [2].

1.1.1 Dataset

The data consisted of three-channel waveforms representing 25 seconds of seismic activity: 5 seconds before the P-wave arrival and 20 seconds after. The three channels corresponded to the North-South (N-S), East-West (E-W), and vertical components of ground motion, capturing the multidirectional nature of seismic signals.

To prepare the data for input into the CNN, we converted the waveforms into three-channel spectrograms. Spectrograms were chosen because they provide a time-frequency representation of the signals, which can reveal patterns that are difficult to observe in raw waveform data. By visualizing the seismic data in this way, the CNN can leverage its ability to extract spatial and temporal features more effectively, as it would from an image.

We then applied log normalization to the spectrograms to enhance the contrast of features, particularly in the lower amplitude ranges. This step was crucial to making subtle variations in the data more distinguishable, aiding the model in identifying important features for classification. The normalized spectrograms were saved as RGB images in PNG format, preserving their three-channel structure and enabling their use as inputs for the CNN.

This preprocessing approach ensured that the key features of the seismic signals were preserved and amplified, making it easier for the CNN to learn the distinctions between aftershocks and foreshocks.

1.1.2 Model Architecture

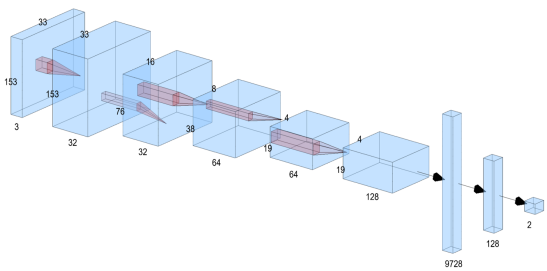

Figure 1. CNN Architecture

For the classification task, we implemented a straightforward CNN (Figure 1) to process the spectrogram images and classify seismic events as aftershocks or foreshocks. The architecture consists of three convolutional layers, each followed by batch normalization, ReLU activation, and max-pooling for the first two layers to reduce spatial dimensions.

The first convolutional layer takes the three-channel RGB spectrograms as input and outputs 32 feature maps. The second and third layers increase the number of feature maps to 64 and 128, respectively, while reducing spatial dimensions with a stride of 2 in the second layer.

After the convolutional layers, the feature maps are flattened into a one-dimensional vector and passed through two fully connected layers. The first fully connected layer has 128 units, and the second outputs the class probabilities for aftershock and foreshock.

1.1.3 SHAP

Figure 2. SHAP Example on Spectrogram (Red squares represent the pixels that contribute more to the classification)

To interpret the predictions of our CNN for classifying seismic events, we utilized SHAP (SHapley Additive exPlanations) values. SHAP is a model-agnostic explainability framework that assigns each feature an importance value, representing its contribution to the model's prediction. It leverages concepts from cooperative game theory, treating the prediction process as a game where features collaborate to generate the output.

How SHAP Works SHAP explains the prediction for a given input by computing Shapley values, which fairly distribute the difference between the model's output and a baseline value across all features. For image data, SHAP typically works by masking portions of the image and observing the changes in the model's output, quantifying how each pixel or region contributes to the prediction.

Application to Our Project In our study, SHAP was employed to analyze the CNN's predictions and highlight which parts of the spectrograms were most influential in distinguishing between aftershocks and foreshocks. Below, we describe the steps followed in our implementation:

- **Masking:** We used SHAP's *Image* masker with an *inpaint_telea* method to simulate the removal of image regions, allowing us to estimate their contributions to the predictions.
- **Explainer Setup:** We defined the SHAP explainer using our CNN as the predictive model, with the output labels ("Foreshock" and "Aftershock") explicitly provided.

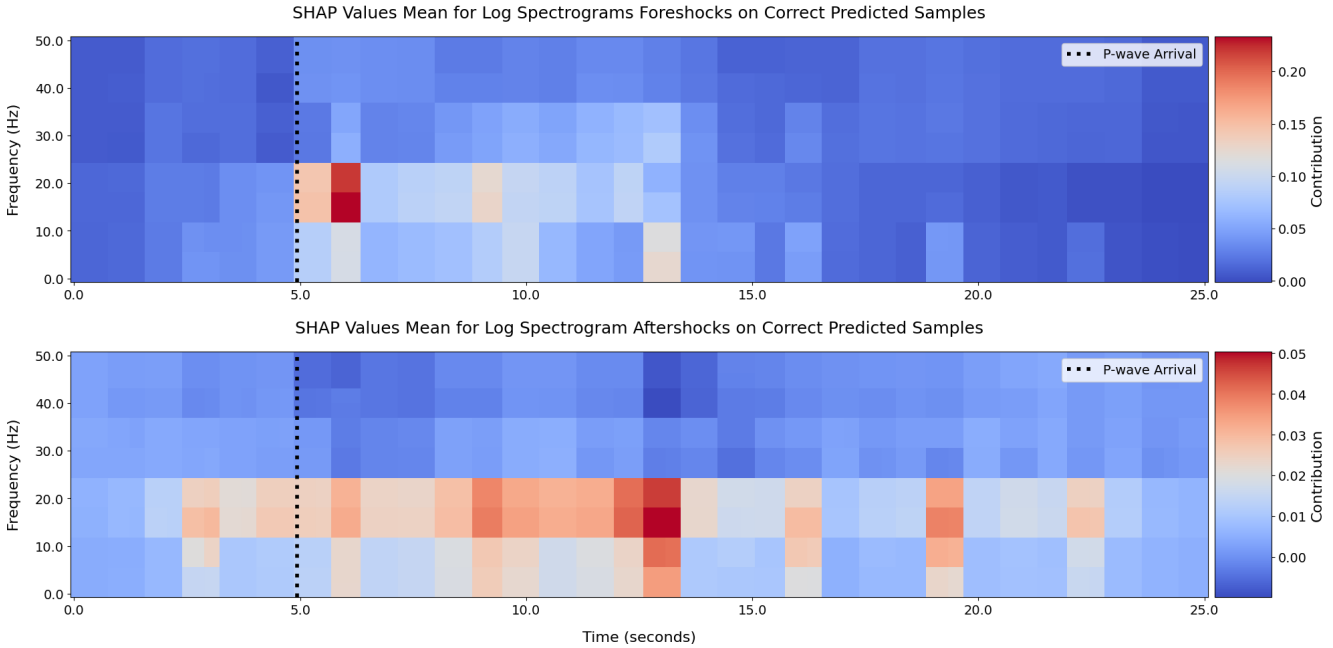

Figure 3. Average SHAP Values on NRCA Model

- **SHAP Value Computation:** SHAP values were computed for a subset of test images, with the evaluations limited to 1,000 steps for efficiency. The grayscale SHAP values were normalized and aggregated to create visual explanations.

To analyze model behavior further, we averaged SHAP values separately for correctly classified foreshocks and aftershocks. This revealed characteristic patterns in the spectrograms that the model relied upon for classification. Specifically, regions of high SHAP importance highlighted time-frequency segments where the seismic signals diverged most significantly between the two classes.

By applying SHAP, we gained insights into the inner workings of our CNN, identifying the key features that influenced its decisions. This understanding not only helped validate the model's predictions but also offered a step toward bridging the gap between machine learning and seismological interpretation.

1.2 Model Evaluation

1.2.1 NRCA Dataset

We trained our CNN2D model on only the NRCA shocks (5862 samples in the training set), similar to Laura Laurenti's approach. Our model achieved an accuracy of 0.9957, slightly better than Laura's CNN1D model on waveforms, which achieved 0.9919.

1.2.2 Full Dataset

We trained the CNN2D model on all available data from multiple stations (FDMO, T1216, MC2, MMO1, NRCA, T1212, T1213, T1214, T1244), resulting in 55617 samples in the training set. The model achieved an accuracy of 0.9575 on this comprehensive dataset.

Additionally, we attempted to predict the full dataset using only the NRCA model, which resulted in a significantly lower accuracy of 0.5019, highlighting the differences in characteristics between the Norcia case and other areas.

1.3 SHAP Results

In this section, we present the analysis of SHAP values obtained from the application of the SHAP explainer on approximately 300 correctly predicted samples for each of the two datasets. The SHAP explainer allows us to interpret the contribution of different features to the classification decisions of the model. The results are presented separately for the NRCA Dataset and the Full Dataset.

1.3.1 NRCA Dataset

In the NRCA Dataset explainer's plot, we observe that the most important values for Foreshock Classification (with more than 0.2 SHAP score) are concentrated in instants after the P-wave arrival, with frequencies between 10Hz and 25Hz. Meanwhile, the features most important for Aftershocks are more uniform in time (from 2.5s to 22.5s) in the frequency bins between 10Hz and 25Hz, with a maximum SHAP score of only 0.05 observed around 13 seconds (see Figure 3).

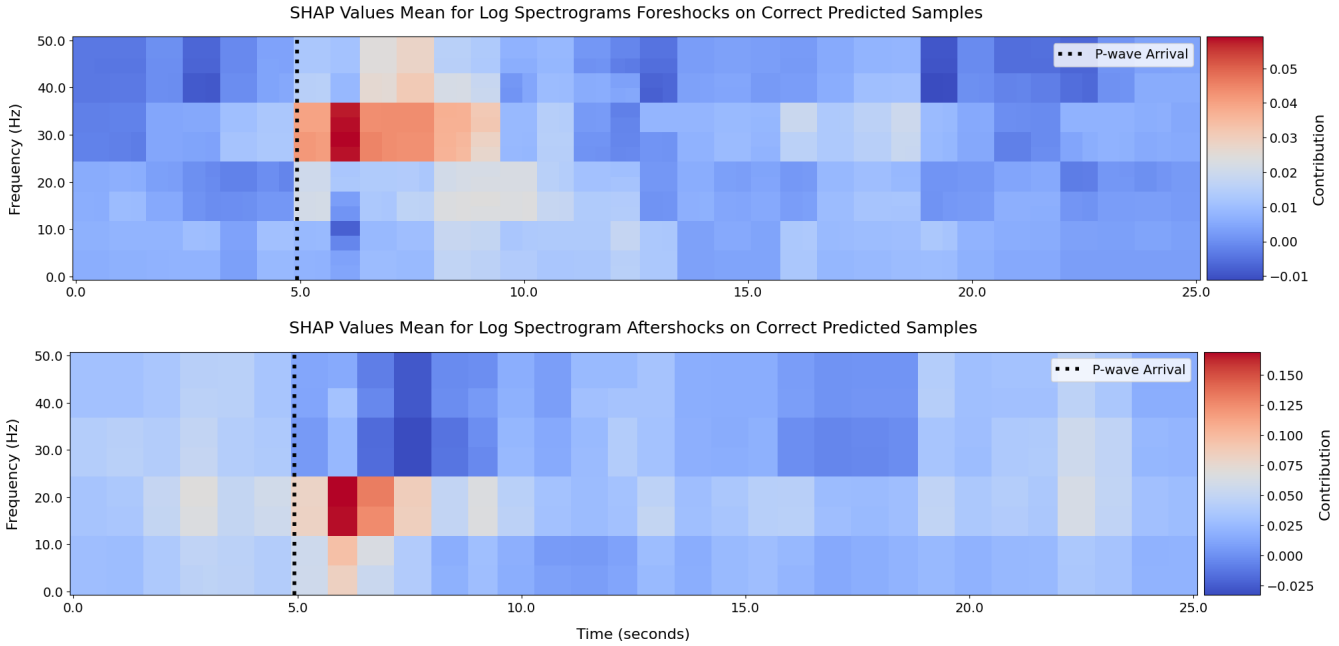

Figure 4. Average SHAP Values on Full Dataset Model

1.3.2 Full Dataset

In the Full Dataset explainer's plot, we observe that the most important values for Foreshock Classification (with more than 0.05 SHAP score) are concentrated in instants after the P-wave arrival, with frequencies between 25Hz and 40Hz. Meanwhile, the features most important for Aftershocks are concentrated in the same instants after the P-wave, but with frequencies between 10Hz and 25Hz, with a maximum SHAP score of 0.15 (see Figure 4).

1.3.3 Comparisons

The SHAP explanations for the NRCA Dataset and the Full Dataset reveal distinct patterns in how the model identifies important features for classifying foreshocks and aftershocks. These differences are particularly evident in the frequency ranges and the distribution of feature importance over time.

In the NRCA Dataset, the model's focus is more localized, with foreshocks showing high importance in the 10 Hz to 25 Hz range shortly after the P-wave arrival. Aftershock-related features, while more evenly distributed in time, also fall within the same frequency range but with generally lower contributions. By contrast, the Full Dataset introduces a broader and more diverse feature representation. Foreshocks in this dataset shift toward higher frequencies (25 Hz to 40 Hz), while aftershocks remain concentrated in the 10 Hz to 25 Hz range but with stronger contributions compared to the NRCA Dataset.

These differences likely stem from the diversity of seismic stations included in the Full Dataset. Variations in signal characteristics, such as noise levels, station-specific site effects, and regional differences in seismic activity, could influence the learned patterns. This additional variability forces the model to generalize more, potentially diluting the sharp distinctions observed in the NRCA-only analysis. Furthermore, the larger dataset may enhance the model's ability to identify subtle features, which could explain the higher SHAP scores for aftershocks in the Full Dataset.

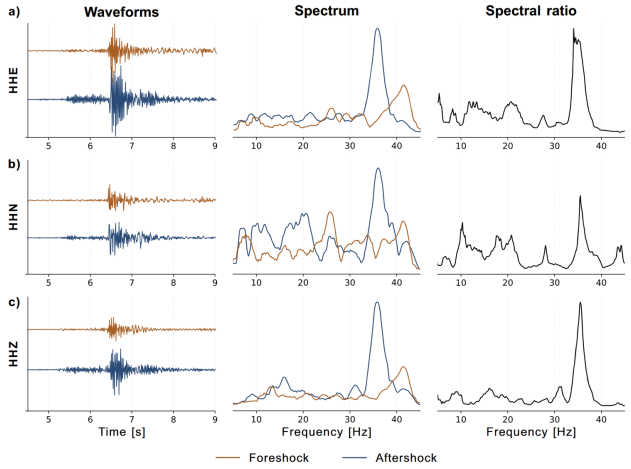

Figure 5. Comparisons of Spectra of co-located foreshocks and aftershocks to access time dependant changes in crustal properties. (Figure 6 from [2])

Laura's study (Figure 5) highlights a frequency difference in the waveforms of foreshocks and aftershocks in the NRCA dataset, with foreshocks typically exhibiting higher frequency components compared to aftershocks. Our NRCA explainer aligns with this observation, as the SHAP values for foreshocks concentrate in the 10 Hz to 25 Hz range, which is higher relative to aftershocks.

1.4 Final Observations

This study explored the use of a CNN and SHAP explainability tools to classify seismic events into foreshocks and aftershocks. Our experiments highlighted differences in feature importance between models trained on the NRCA Dataset and those trained on the Full Dataset, revealing insights into the challenges of generalizing across diverse seismic stations.

The NRCA-specific model demonstrated high accuracy and sharply localized feature importance patterns, with foreshocks relying heavily on features shortly after the P-wave arrival and within the 10 Hz to 25 Hz frequency range. In contrast, the Full Dataset model exhibited more generalized patterns, with foreshock-related features shifting toward higher frequencies (25 Hz to 40 Hz). This broader representation suggests the model adapts to the diverse signal characteristics introduced by including multiple stations, such as variations in noise, site effects, and regional seismic activity.

However, the poor performance of the NRCA-trained model when applied to the Full Dataset (accuracy of 0.5019) highlights the limited generalizability of models trained on a single station. This result underscores the importance of accounting for station-specific effects and dataset diversity when building models intended for broader applications.

In conclusion, while station-specific models like the NRCA CNN can capture precise local patterns, their utility diminishes when applied to diverse datasets. The Full Dataset model, despite its lower accuracy, offers a more generalized and adaptable solution. Future work could explore strategies to balance local precision and generalization, such as domain adaptation techniques or incorporating station metadata into model training.

2. Random Forest Regressor for Magnitude Prediction

2.1 Methods

One particularly important experiment conducted within the context of the Norcia region delves deeper into identifying factors that influence earthquake magnitude. By analyzing data from spectrograms and metadata, this study aims to determine which features hold the greatest relevance in understanding and predicting the magnitude of seismic events, as well as how these features contribute to the estimation process.

2.1.1 Dataset

The dataset used for this experiment was sourced from Zenodo, consistent with the dataset referenced in the work of Laurenti et al. [3]. For this specific study, we concentrated exclusively on data related to seismic activity in the Norcia region. The dataset included a combination of features extracted from waveforms alongside associated metadata, offering a comprehensive foundation for model training and analysis.

2.1.2 Model

Using this data, a *Random Forest Regressor* with 100 trees was trained to predict the magnitude of earthquakes. The model's performance was evaluated using key regression metrics, yielding the following results:

1. **Mean Absolute Error (MAE):** 0.1052
2. **Mean Squared Error (MSE):** 0.0212
3. **Root Mean Squared Error (RMSE):** 0.1457
4. **R^2 Score:** 0.8834
5. **Explained Variance:** 0.8835

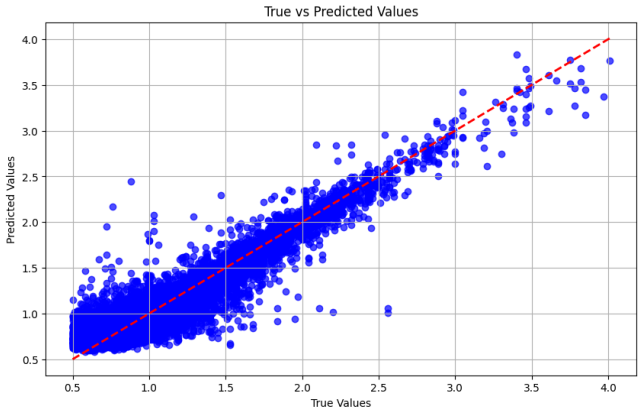

Figure 6. Prediction precision visualization

The results demonstrate that the model achieved a high level of accuracy, as evidenced by the R^2 score and the low error metrics.

The high Explained Variance score indicates that the model effectively captures the variance in the dataset, suggesting that the selected features are strongly correlated with earthquake magnitude. Additionally, the low MAE and RMSE values underline the model's ability to make reliable predictions, even when applied to a complex and potentially noisy dataset.

2.1.3 SHAP

Similarly to the case discussed in section 1.1.3, we utilized the SHAP library to gain insights into the behavior of the Random Forest model. However, it is important to highlight that the nature of the models differs significantly, as they perform completely different tasks (regression vs classification). In particular, the methodology and considerations specific to regression are detailed below:

- **TreeExplainer Creation**: Similar to the explainer setup in the classification case, we instantiate a *TreeExplainer* object. This object generates a representation of the tree structure used in the model, enabling SHAP to compute the contributions of individual features.
- **SHAP Value Computation**: The SHAP values are computed using the representation provided by *TreeExplainer*. For regression tasks, this involves evaluating each feature and determining its weight in the final prediction. Notably, in regression tasks, SHAP values are stored in a matrix format where rows correspond to samples and columns correspond to features used in the model.

Understanding these methodological differences clarifies that, in this case, the use of masking is unnecessary. Unlike tasks involving images (where masking helps identify the most relevant regions of an image), the primary objective in this regression scenario is to understand how each feature influences the final prediction. This approach directly interprets feature contributions without requiring masking or spatial relevance analysis.

2.2 SHAP evaluation

We now turn our attention to the study of the relevance of individual features, examining their importance on a case-by-case basis. While many features do not correspond directly to easily interpretable quantities, it is still possible to attempt to assign meaning and interpretation to each of them.

Figure 5, presented below, provides a visual representation of how each individual feature contributes to the final prediction. Each feature is represented using a color scale to indicate its magnitude, while its position along the horizontal axis reflects the extent of its influence on the model's predictions.

This analysis allows us to delve deeper into understanding the model's decision-making process, shedding light on which features are most critical and how they interact in determining the predicted magnitude of seismic events. While some patterns may emerge clearly, others may require further domain-specific investigation to interpret their significance fully.

This analysis is particularly valuable as it enables us to adopt an approach distinct from the conventional "black-box" perspective often associated with machine learning models. Unlike traditional methods, this approach explicitly highlights the contribution of each feature to individual predictions.

While the exact meaning of certain features may not always be immediately clear, this process represents an essential step toward a deeper understanding of the mechanisms driving the predictions. It moves us closer to unraveling the inner workings of the deep learning methods employed, bridging the gap between model complexity and interpretability.

By visualizing and quantifying the influence of each feature, researchers and practitioners can better trust and refine these models, especially in critical applications such as seismic prediction. This transparency not only aids in improving model performance but also facilitates the development of more interpretable and accountable machine learning systems.

2.2.1 SHAP summary plot

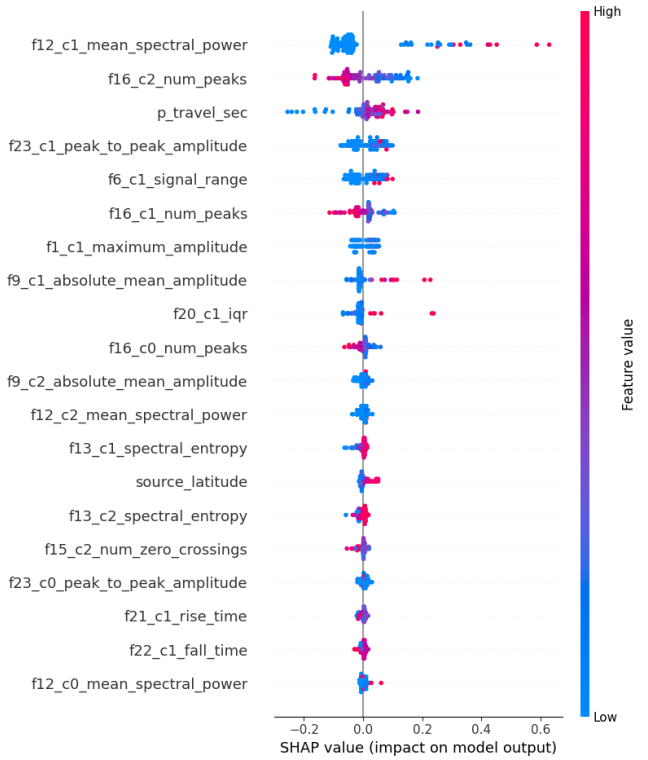

Figure 7. Regression Model features summary plot

To better understand the feature names, here is how to read it: $f\{feature\_index\}\_c\{channel\_number\}\_\{feature\_name\}$

Where:

- Channel 0 is for N-S movement
- Channel 1 is for E-W movement
- Channel 2 is for vertical movement

From the analysis of the provided SHAP summary plot, we can derive the following observations:

- The plot clearly highlights that certain features, such as *f12_c1_mean_spectral_power*, *f16_c2_num_peaks*, and *p_travel_sec*, have a significant impact on the model's predictions. These features demonstrate a wide range of SHAP values, indicating their strong influence.
- The feature *p_travel_sec* appears to have a direct proportional relationship with the model's output. Higher feature values (marked in pink) correspond to higher SHAP values, implying a positive contribution to the prediction.
- Conversely, features such as *f12_c1_mean_spectral_power* and *f16_c2_num_peaks* show an inverse proportional relationship. Higher feature values in these cases tend to negatively impact the model's output, as evidenced by the blue values clustering towards negative SHAP values.
- Other features, such as *f23_c1_peak_to_peak_amplitude*, *f6_c1_signal_range*, and *f1_c1_maximum_amplitude*, also play substantial roles in determining the model's behavior, but their effects are less pronounced compared to the top three mentioned.
- Lower-ranked features like *f12_c0_mean_spectral_power* and *f22_c1_fall_time* have relatively limited influence, as shown by the smaller spread of SHAP values around zero.
- The analysis highlights the dominance of spectral features (*mean_spectral_power*, *peak_to_peak_amplitude*, *signal_range*) and temporal characteristics (*p_travel_sec*, *rise_time*, *fall_time*) in driving the model's predictions.

2.3 Final Observations

The application of a Random Forest Regressor to predict earthquake magnitudes in the Norcia region demonstrates promising results, with the model achieving high accuracy and robust performance metrics. The integration of SHAP analysis provides critical insights into the model's behavior, highlighting the most influential features and their contributions to the predictions. Features such as mean spectral power, number of peaks, and travel times emerge as key drivers, underscoring the relevance of both spectral and temporal characteristics in seismic prediction.

This study not only validates the efficacy of using machine learning for magnitude prediction but also underscores the importance of interpretability in such applications. By visualizing feature contributions, researchers can better understand the model's decision-making process, paving the way for further refinements and more reliable deployment in real-world scenarios. Ultimately, this approach bridges the gap between model performance and transparency, offering a valuable framework for advancing seismic analysis and prediction methodologies.

3. Conclusions

This study explored the integration of ML techniques with explainability tools to enhance our understanding of seismic processes, focusing on two primary tasks: classifying seismic events as foreshocks or aftershocks using a CNN and predicting earthquake magnitudes through a Random Forest Regressor.

For the classification task, we demonstrated that CNN models trained on station-specific datasets, such as the NRCA dataset, achieved high accuracy by leveraging localized patterns in spectrograms. However, their performance was significantly reduced when applied to a broader dataset, underscoring the challenge of generalizing across diverse seismic stations. SHAP analysis revealed that foreshocks and aftershocks exhibit distinct patterns in frequency and timing, particularly around P-wave arrival, highlighting the potential for these models to uncover meaningful physical insights. Despite the limitations in generalizability, the integration of SHAP values provided valuable transparency, enabling a deeper understanding of the model's decisions.

In the magnitude prediction task, the Random Forest Regressor exhibited strong performance metrics, demonstrating the capability of ML models to capture complex relationships between seismic features and earthquake magnitudes. SHAP analysis identified key contributors such as spectral power, peak counts, and P-wave travel times, further validating the importance of both spectral and temporal features. This approach not only quantified feature importance but also improved interpretability, offering a pathway to bridge ML methodologies with domain expertise in seismology.

Our findings emphasize the critical role of explainability in enhancing trust and usability of ML models in scientific contexts. While station-specific models can provide precise local insights, their application to diverse datasets requires careful consideration of generalization strategies, such as incorporating domain adaptation techniques or metadata-informed training. Similarly, regression models for magnitude prediction benefit greatly from transparent feature analysis, aiding both validation and refinement.

In conclusion, the combination of ML techniques and explainability tools offers a powerful framework for advancing seismic research. Future work should aim to balance precision and generalizability, explore methods for integrating additional geophysical knowledge into model training, and extend these approaches to real-time seismic monitoring and early warning systems. This study demonstrates that interpretable ML models are not only tools for prediction but also valuable instruments for uncovering new insights into the physics of earthquakes.

Acknowledgments

We would like to express our gratitude to Laura Laurenti and Gabriele Paoletti for their studies and the seminar they conducted during our course, which greatly enriched our understanding of the intersection between machine learning and seismic phenomena.

Our thanks also go to Professor Chris Marone for his lessons and guidance, which have been instrumental in shaping our knowledge and approach throughout this project.

Thank you all for inspiring us and helping us navigate this fascinating field.

References

[1] S. Lundberg. Welcome to the shap documentation, 2018.

[2] L. Laurenti, G. Paoletti, E. Tinti, F. Galasso, C. Collettini, and C. Marone. Probing the evolution of fault properties during the seismic cycle with deep learning. *Nature Communications*, 15, 2024.

[3] G. Paoletti. D-set: Probing the evolution of fault properties during the seismic cycle with deep learning, 2024.