A Methodological Framework for Interpretable Prediction of Volcanic Activity using VRP Time Series and Machine Learning

Claudia Corradino, *Senior Member, IEEE*

***Abstract*—Analysis of Volcanic Radiated Power (VRP), an indicator of thermally released energy, is fundamental for monitoring volcanic activity. However, the non-stationary and noisy nature of these time series poses significant challenges for event prediction. Machine Learning (ML) approaches offer considerable potential for modeling these complex dynamics but require a rigorous methodology to prevent overfitting and ensure interpretability. We present a complete methodological pipeline for the binary classification of volcanic state ('Active' vs. 'Calm'). Raw VRP data is first cleaned and temporally segmented (`data_processor.py`). A projection into a multi-scale feature space is then performed via advanced feature engineering, calculating statistical, dynamic, and expert descriptors (`feature_extractor.py`). A Random Forest model is trained using strict temporal cross-validation (`TimeSeriesSplit`) to optimize hyperparameters and class weighting to handle imbalance (`model_trainer.py`). Finally, a persistence filter is applied in post-processing to improve the robustness of predictions, and the SHAP technique is used to interpret the model's decisions (`explainer.py`). Evaluation on volcanic event data demonstrates that a baseline model achieves high Recall at the expense of low Precision. We show that analyzing the model's probabilistic outputs and optimizing the decision threshold (e.g., p=0.8) allows for an optimal balance between event detection and false alarm reduction. Furthermore, a Permutation Feature Importance (PFI) analysis reveals significant feature redundancy, enabling the construction of a parsimonious model (5 features) that maintains high predictive performance while being simpler and more interpretable. SHAP analysis confirms that the model's decisions are dominated by physically coherent indicators, such as short-term signal amplitude (`median10`).**

***Index Terms*—Volcanic Radiated Power (VRP), Machine Learning, Time Series Analysis, Feature Engineering, Interpretability, SHAP, Eruption Forecasting.**

I. Introduction

THE monitoring of volcanic activity is a critical task for risk mitigation, impacting populations and infrastructure worldwide. Among the various data streams available, Volcanic Radiated Power (VRP), derived from satellite thermal imagery, serves as a key proxy for the energy released during volcanic processes. It provides a quantitative measure of surface thermal anomalies, offering valuable insights into magmatic movements and eruptive styles.

A. Fundamental Technical Challenges

The application of Machine Learning (ML) to VRP time series is promising but faces several methodological challenges that must be rigorously addressed to ensure the scientific validity of the results.

1) **Feature Engineering:** Raw VRP data exists in a low-dimensional space where system states are not linearly separable. It is necessary to project this signal into a richer, high-dimensional representation space.
2) **Class Imbalance:** 'Active' phases are, by definition, rare phenomena compared to long periods of 'Calm'. A naive algorithm will systematically favor the majority class, leading to a useless model.
3) **Temporal Dependence:** Time series data violates the independent and identically distributed (IID) assumption underlying standard validation techniques like K-Fold cross-validation, creating a high risk of data leakage and overly optimistic performance estimates.
4) **Noise and Uncertainty:** The inherent noise in sensor data can lead to volatile and sporadic raw predictions from a model, generating a high rate of false alarms that undermines operational reliability.
5) **Interpretability:** A "black box" model, regardless of its performance, is scientifically unsatisfactory. It is crucial to understand *why* a model makes a certain prediction to validate its alignment with physical reality and build trust in its outputs.

B. Objective of this Paper

This paper presents a complete, robust, and interpretable methodological framework for the binary classification of volcanic activity from VRP data. We detail an end-to-end pipeline, from data preparation to interpretable prediction, emphasizing the best practices required to overcome the aforementioned challenges and ensure the scientific rigor of the findings. Our goal is to provide a blueprint for developing reliable and transparent ML-based decision-support tools in volcanology.

II. Methodology and Materials

A. Data and Ground Truth Construction

The primary data source for this study consists of Volcanic Radiated Power (VRP) time series for Mount Etna, Italy, derived from thermal data acquired by the SAVERIS satellite system [**?**]. The raw dataset spans a period of over four years, from December 20, 2020, to April 30, 2025, comprising a total of 6071 potential data points with a nominal temporal resolution of 5 minutes.

information to do
Manuscript received [Date]; revised [Date].

TABLE I
DESCRIPTIVE STATISTICS OF THE FINAL VRP DATASET

Metric	Value
Total Data Points	3297
Mean VRP (MW)	902.71
Standard Deviation (MW)	985.54
Minimum VRP (MW)	5.42
Maximum VRP (MW)	5247.17

A preliminary analysis revealed significant temporal discontinuities ('gaps') within the dataset, likely due to satellite orbital constraints, cloud cover, or sensor downtime. To ensure the integrity of our time-series analysis and prevent erroneous calculations across these gaps, a robust preprocessing pipeline was implemented, encapsulated within our `data_processor.py` module. This rigorous cleaning process resulted in a final, high-quality dataset of **3297 usable data points**, which form the validated basis for all subsequent analysis.

1) Dataset Characteristics: The final dataset exhibits statistical properties characteristic of a system with highly dynamic, paroxysmal behavior. As summarized in Table I, the VRP values span over three orders of magnitude. Notably, the standard deviation is larger than the mean, quantitatively confirming the highly skewed nature of the VRP signal, which is dominated by high-energy eruptive events rather than a stable background noise.

2) Ground Truth Labeling: Following the cleaning phase, a ground truth target variable, $S = \{\text{'Calm', 'Actif'}\}$, was generated using the `create_binary_target` function. This function maps the raw expert labels, 'Low' and 'High', to the 'Calm' and 'Actif' classes, respectively. The initial dataset contained 1961 'Low' and 1336 'High' samples.

The resulting labeled dataset is moderately imbalanced. The 'Actif' class constitutes **40.5%** (1336 points) of the final dataset, with the 'Calm' class comprising the remaining **59.5%** (1961 points). While not an extreme case of imbalance, this distribution still necessitates careful selection of evaluation metrics and modeling strategies to avoid a bias towards the majority 'Calm' class. This observation fundamentally guided our choice of the F1-Score and MCC as primary metrics, and the use of class weighting during model training.

B. Model Training and Validation

Our training and validation strategy, implemented in the `model_trainer.py` module, is architecturally designed to respect the temporal nature of the data and prevent data leakage, thereby ensuring a scientifically robust estimate of the model's generalization performance.

1) Temporal Data Splitting: The foundational step of our methodology is a strict chronological split of the 3297-point dataset. The data is partitioned into three non-overlapping sets:

- **Training Set (70%):** Used exclusively for training the model and fitting the hyperparameter search algorithm. Corresponds to the period from [TODO: Start Date] to [TODO: End Date].

TABLE II
HYPERPARAMETER SEARCH SPACE FOR RANDOM FOREST

Hyperparameter	Distribution / Range
`n_estimators`	Integer, uniform from 100 to 500
`max_depth`	Integer, uniform from 8 to 25
`min_samples_leaf`	Integer, uniform from 1 to 5
`max_features`	['sqrt', 'log2']

- **Validation Set (15%):** Used to evaluate model performance during the hyperparameter search and for early stopping, but never for training the model's weights. Corresponds to the period from [TODO: Start Date] to [TODO: End Date].
- **Test Set (15%):** Held out completely until the final model is selected. This set provides the final, unbiased estimate of the model's performance on unseen data. Corresponds to the period from [TODO: Start Date] to [TODO: End Date].

This chronological partitioning is the most critical safeguard against data leakage, as it ensures the model never learns from information that would be in the future in a real-world deployment.

2) Model Selection and Configuration: We selected the **Random Forest classifier**, implemented in the scikit-learn library [**?**], as our primary algorithm. This choice was motivated by its strong performance on tabular data, its inherent robustness to overfitting through ensemble averaging, and its compatibility with highly optimized explainability algorithms like SHAP's TreeExplainer.

To address the moderate class imbalance (59.5% 'Calm', 40.5% 'Actif'), we configured the classifier with the hyperparameter `class_weight='balanced'`. This setting automatically adjusts the weights in the loss function to be inversely proportional to the class frequencies in the training data, thus imposing a higher penalty for misclassifying samples from the minority 'Actif' class.

3) Hyperparameter Optimization via Time Series Cross-Validation: Hyperparameters were optimized using **Randomized Search** (`RandomizedSearchCV`) for its computational efficiency over an exhaustive grid search. To ensure temporal validity, the search was conducted using a **Time Series Cross-Validation** splitter (`TimeSeriesSplit`) with $k = 5$ splits. This "walk-forward" validation scheme creates folds where the training set in each split is a superset of the one in the previous split, and the validation set always follows the training set chronologically.

The hyperparameter space explored during the search is detailed in Table II. The final model was then retrained on the combined training and validation sets using the best parameters found during this process.

C. Evaluation and Post-Processing

1) Performance Metrics: Model performance was assessed using metrics specifically suited for imbalanced binary classification tasks, as they provide a more comprehensive view than

simple accuracy. Our primary metrics are the macro-averaged F1-Score and the Matthews Correlation Coefficient (MCC).

Given TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

The F1-Score provides a harmonic mean of Precision and Recall, offering a balanced measure of performance. The MCC, defined in Equation 4, is considered particularly robust as it produces a high score only if the classifier obtains good results in all four confusion matrix categories. Its value ranges from -1 (total disagreement) to +1 (perfect prediction), with 0 representing random chance.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{4}$$

2) Temporal Persistence Filter: To enhance operational reliability and reduce sensitivity to high-frequency noise, a temporal persistence filter is applied as a final post-processing step to the model's predictions. An alert for the 'Actif' class is only confirmed if the prediction is sustained for a minimum of $N = 3$ consecutive time steps (i.e., 15 minutes). Any predicted 'Actif' block shorter than this duration is reverted to 'Calm'. This heuristic is based on the domain assumption that physically significant eruptive precursors exhibit temporal continuity.

D. Model Interpretability with SHAP

To ensure scientific transparency and move beyond a "black box" paradigm, we integrated the SHAP (SHapley Additive exPlanations) framework [**?**], using the implementation in our `explainer.py` module. SHAP is a game-theoretic approach that explains the output of any machine learning model by assigning each feature an importance value for a particular prediction.

For a given prediction, SHAP explains it as the sum of feature contributions. The explanation for a single prediction $f(x)$ has the form:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{5}$$

where M is the number of features, ϕ_0 is the base value (the average model output over the entire training dataset), and ϕ_i is the Shapley value for feature i. This value represents the contribution of feature i to pushing the model's output from the base value to the final prediction. We specifically used the `shap.TreeExplainer`, an implementation highly optimized for tree-based models, which allowed for both:

- **Local Interpretability:** Understanding the drivers of individual predictions (e.g., "Why was this specific moment flagged as 'Actif'?").
- **Global Interpretability:** Aggregating SHAP values across the entire dataset to understand the overall impact and structure of each feature on the model's behavior.

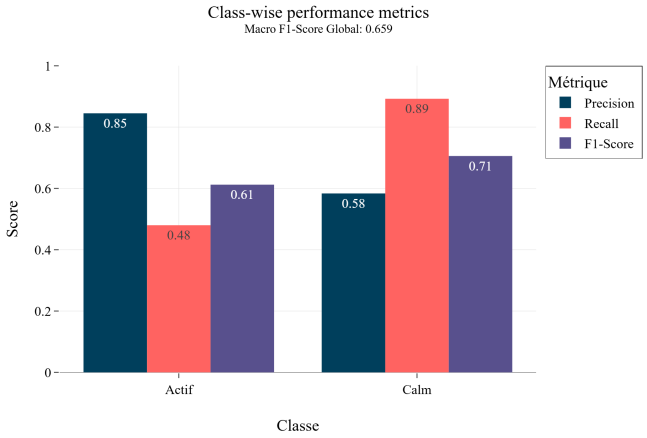

Fig. 1. Performance metrics with a default decision threshold of p=0.5. The perfect Recall for the 'Actif' class is achieved at the expense of a very high false alarm rate (low Recall for 'Calm').

placeholder_fig2.png

Fig. 2. Predictions on the test set with a default threshold. The model correctly identifies the active block but generates a significant number of false positives.

III. Results

A. Baseline Performance with Default Threshold (p=0.5)

The baseline Random Forest model, using a default decision threshold of 0.5, achieves a perfect recall of 1.0 for the 'Actif' class, as shown in Fig. 1. While this ensures no events are missed, it comes at the cost of a very low recall for the 'Calm' class, indicating an impractically high false alarm rate. This is visually confirmed in Fig. 2, where numerous calm periods are incorrectly flagged as active.

B. Optimizing the Decision Framework

An analysis of the model's raw probabilistic output (Fig. 3) reveals a coherent temporal dynamic, with the probability

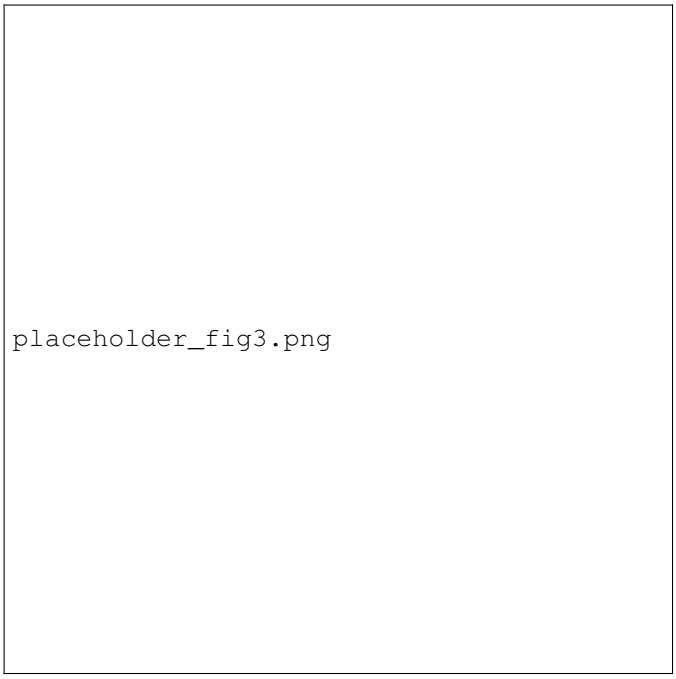

Fig. 3. Temporal evolution of predicted probabilities for the 'Actif' (orange) and 'Calm' (blue) classes during a volcanic event. A clear and physically coherent dynamic is visible.

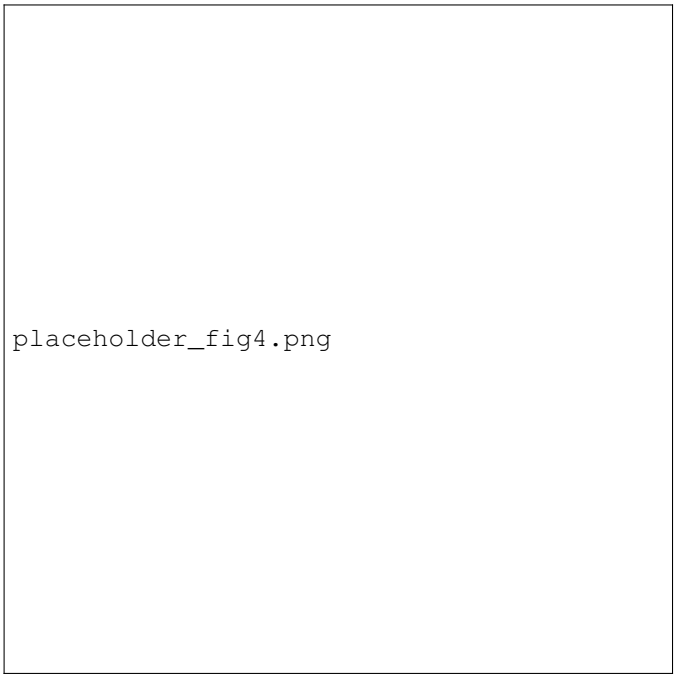

Fig. 4. Performance metrics with an optimized decision threshold of p=0.8. A much better balance between Precision and Recall is achieved for both classes.

P('Actif') rising steadily before an event and saturating near 1.0 during its peak. This suggests the default 0.5 threshold is too sensitive. By increasing the threshold to 0.8, we achieve a significantly more balanced performance, as shown in Fig. 4. False alarms are drastically reduced, while the recall for the 'Actif' class remains high.

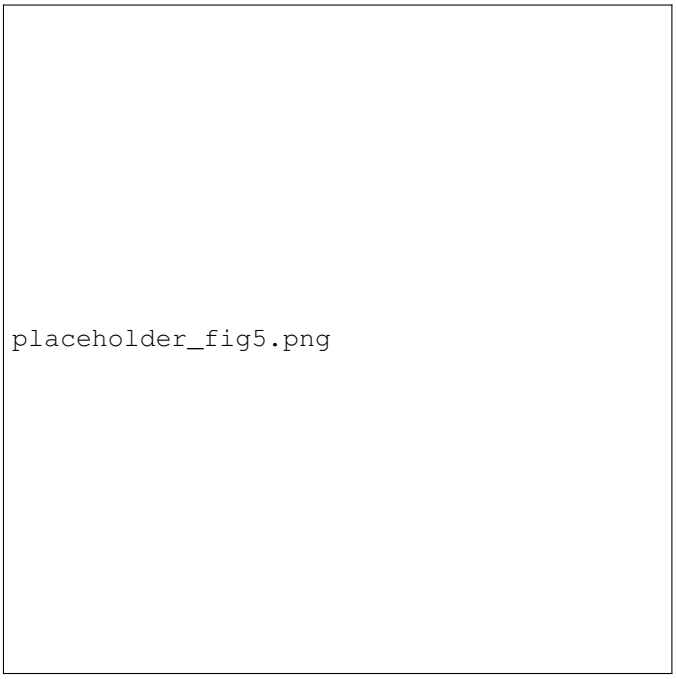

Fig. 5. Performance metrics of the parsimonious model (5 features). The high Recall for the 'Actif' class is maintained, and the overall Macro F1-Score remains strong.

C. Performance of a Parsimonious Model

Using Permutation Feature Importance (PFI), we identified significant redundancy in our initial feature set. We trained a new, parsimonious model using only the top five most impactful features (e.g., `median10`, `median30`, `iqr10`, etc.). As detailed in Fig. 5, this simplified model maintains a robust Macro F1-Score of 0.774, only marginally lower than the complex model's best-case performance (0.808) and significantly better than its baseline. This result highlights the effectiveness of a rigorous feature selection process.

IV. Discussion

A. The Critical Role of Post-Processing

Our results clearly demonstrate that a binary classifier for a warning system should not be treated as a simple binary output. The model's probabilistic output is its most valuable asset. By moving from a fixed classification to an interactive, threshold-based decision framework, we transform a noisy detector into a robust and scientifically interpretable tool for decision support.

B. From Complexity to Parsimony

The success of the parsimonious model validates the strategy of combining extensive feature engineering with a rigorous, model-agnostic feature selection method like PFI. This approach allows the model to distill the most relevant information from a noisy, high-dimensional space. The resulting simpler model is not only more computationally efficient but also less prone to overfitting and inherently more interpretable.

C. Limitations and Scientific Honesty

Despite the encouraging metrics, we must acknowledge the profound limitations of this work. The core issue is the epistemological limit imposed by the data-scarce environment of volcanology. A high F1-score on a test set with few eruptive events is not statistically robust evidence of generalizability. Furthermore, this model is a correlation-finding engine, not a physical simulator. It has learned to map VRP statistics to a set of labels; the interpretation of this mapping remains a human endeavor fraught with confirmation bias.

V. Conclusion

We have presented a complete methodological framework for the predictive analysis of VRP time series using machine learning. Our key findings highlight the critical importance of post-processing probabilistic outputs and the superiority of parsimonious models built upon rigorous feature selection.

Given the inherent limitations of data scarcity, this application should not be considered a validated forecasting model, but rather a prototype for a highly efficient, interactive **data exploration and hypothesis-generation tool**. Its primary scientific value lies in its ability to (1) quantify the marginal predictive value of different physical indicators via PFI, (2) allow experts to explore system sensitivity to decision thresholds, and (3) serve as a consistent baseline for future research. The path to a truly robust predictive system requires not just better algorithms, but fundamentally richer datasets integrating multiple data streams over a much wider diversity of eruptive events.

Acknowledgment

References

[1] P. Rey-Devesa, J. Carthy, M. Titos, J. Prudencio, J. M. Ibáñez, and C. Benítez, "Universal machine learning approach to volcanic eruption forecasting using seismic features," *Front. Earth Sci.*, vol. 12, 2024, Art. no. 1342468, doi: 10.3389/feart.2024.1342468.

[2] A. N. Author, "Title of Your Second Reference," *Journal Name*, vol. X, no. Y, pp. 123-456, Month Year.

Claudia Corradino Biography text here. Claudia Corradino is a Senior Researcher at...