

Research

Volcano activity classification from synergy of EO data and machine learning: an application to Mount Etna volcano (Italy)

C. Petrucci¹ · G. Romoli¹ · A. Pignatelli¹ · E. Trasatti¹ · F. Zuccarello² · F. Greco² · M. Dozzo^{2,3} · G. Bilotta² · F. Spina^{2,4} · G. Ganci²

Received: 12 March 2025 / Accepted: 11 June 2025

Published online: 22 June 2025

© The Author(s) 2025 [OPEN](#)

Abstract

This study investigates the integration of Earth Observation (EO) data and Machine Learning (ML) techniques for classifying volcanic activity states at Mount Etna, one of the world's most active and monitored volcanoes. Using satellite data, including ground deformation, radiance, land surface temperature, sulfur dioxide emissions, and gravity anomalies, five volcanic activity states were identified: Quiet, Preparatory, Unrest, Eruption, and Cooling. Supervised ML algorithms, such as random forest, support vector machines, decision trees, and k-nearest neighbors, were employed to classify these states. Random forest achieved the highest accuracy, demonstrating its robustness for this application.

The study addresses challenges like temporal and spatial disparities and class imbalances through data preprocessing, ensuring a reliable dataset for training and validation. A k-fold cross-validation approach was used to evaluate model performance systematically. The results underline the potential of ML techniques combined with EO data for volcanic hazard monitoring, with implications for improving risk assessment and early-warning systems. This methodology, tested on a well-instrumented volcano like Mount Etna, provides a foundation for extending the approach to other less-monitored volcanoes.

These findings are one of the first attempts of integrating satellite data with Artificial Intelligence (AI) to enhance the accuracy of volcanic state predictions and mitigate risks associated with eruptions, while emphasizing the need for rigorous validation against well-documented case studies.

Keywords Machine learning · Earth observation data · Volcanic hazard · Optical data · SAR interferometry · Mount Etna

1 Introduction

The need for integrated and efficient volcano observation systems, with the capability of operating on a global scale, and including tools for producing different scenarios as eruptive conditions change, is a primary challenge for volcanic hazard assessments [1]. Satellite data has become a strong focus of global interest, offering valuable tools to study Earth and to improve physical models. Abundant datasets from multi-mission satellite remote sensing during recent years have provided an opportunity to improve not only the model estimates but also to update model parametrization strategies.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42452-025-07311-8>.

✉ A. Pignatelli, alessandro.pignatelli@ingv.it | ¹Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy. ²Istituto Nazionale di Geofisica e Vulcanologia, Catania, Italy. ³University of Palermo, Palermo, Italy. ⁴University of Catania, Catania, Italy.



Earth Observation (EO) data, encompassing spectral, spatial, and temporal information about the Earth's surface and atmosphere, is vital for monitoring environmental changes, managing natural resources, and mitigating the impacts of natural disasters [2, 3].

Moreover, the current evergrowing multi-source capability of orbiting instruments requires new analysis approaches for the synergic processing and joint interpretation of this large and heterogeneous amount of data in order to understand the volcanic dynamics from space [4].

Artificial Intelligence (AI) has emerged as a transformative technology in the realm of EO capabilities, revolutionizing the way we analyze and interpret data collected from satellite platforms. Particularly, Machine Learning (ML) and Deep Learning (DL) are uniquely suited to handle the large volumes of EO data, which often exhibit high dimensionality, heterogeneity, and complexity [5, 6]. These methods enable efficient processing and extraction of meaningful patterns from multispectral and hyperspectral imagery, time-series data, and LiDAR (Light Detection and Ranging) datasets, fostering advancements in several applications including volcano hazard monitoring [7].

ML algorithms can be categorized into three groups: supervised methods, unsupervised methods, and reinforcement learning [8–10]. The primary problem addressed by supervised methods is predicting a target variable (response or label) based on a set of input variables (predictors). For instance, in the context of medical diagnosis, the predictors might include clinical parameters such as temperature, blood pressure, and cholesterol levels, while the response could be the determination of whether the patient is ill or healthy. To achieve this result, an initial *training phase* involves supplying machines with datasets that contain both the input variables and the corresponding outcomes, referred to as *labeled data*. During this phase, the system autonomously learns to identify the relationship between the inputs and the outcomes and the primary role of the user is to furnish the labeled datasets. Once training is successful, the algorithm can independently predict outcomes for new, unseen datasets based on the learned relationships.

Unsupervised learning techniques are employed to identify patterns and structures within datasets that do not contain labeled responses. Unlike supervised learning, where the model is trained on input–output pairs, unsupervised learning operates solely on input data. The primary objective is to uncover the underlying distribution, associations, or clusters present in the data. Unsupervised learning techniques facilitate data exploration and provide valuable insights.

Reinforcement learning techniques are designed to perform a sequence of decisions by interacting with an environment. Unlike supervised and unsupervised learning, reinforcement learning is characterized by a feedback loop where the learning comes from the consequences of the actions, receiving rewards or penalties based on the outcomes. The primary goal is to develop a policy that maximizes the cumulative reward over time.

The application of ML techniques to volcano hazard monitoring from space encounters objective challenges due to the limited number of observed eruptions in temporal overlap with the most recent space missions. This problem is particularly relevant considering the great variability of observable phenomena and the different types of volcanoes around the world. While the integration of EO data and AI techniques can be an extremely powerful tool, it presents significant risks if it is not adequately tested and validated against well-monitored volcanoes [11].

Mt Etna, located on the eastern coast of Sicily, Italy, is one of the most active and intensively monitored volcanoes in the world. It is characterized by persistent eruptive activity, ranging from effusive lava flows to explosive paroxysms, making it a natural laboratory for volcanic research and hazard assessment [12]. Nowadays Mt Etna is equipped with a dense network of ground-based instruments such as broadband seismometers, tiltmeters, and GNSS (Global Navigation Satellite Systems) stations for detecting deformation, as well as gas sensors to measure volcanic emissions. Additionally, visual and thermal cameras are installed all around the volcano looking at its summit and periodic surveys are conducted with different instruments also mounted on drones [13].

A crucial feature to manage a volcanic crisis is the ability of volcanologists to promptly detect an impending eruption [14]. This is often affected by significant uncertainty, mainly for the difficulty in interpreting the monitoring signals in terms of the exact timing of a possible eruption even for heavily monitored volcanoes. Here we contribute to this problem, focusing on the states of Mt. Etna deduced from EO data and AI by using interpretations by volcanologists and ground-based data for a specific time window in a retrospective way. In particular, we identify five main states to classify the level of activity of a volcano: Quiet, Preparatory, Unrest, Eruption and Cooling. We employ a range of classical supervised ML techniques to infer one of these five states for a given time instant from physical parameters measured by satellites and conduct a comparative analysis of their performance to identify the most effective approach. The application to a well monitored volcano such as Mt. Etna provides useful hints to test the effectiveness of the overall methodology and its potential use to other volcanoes with limited ground instrumentation of the world.

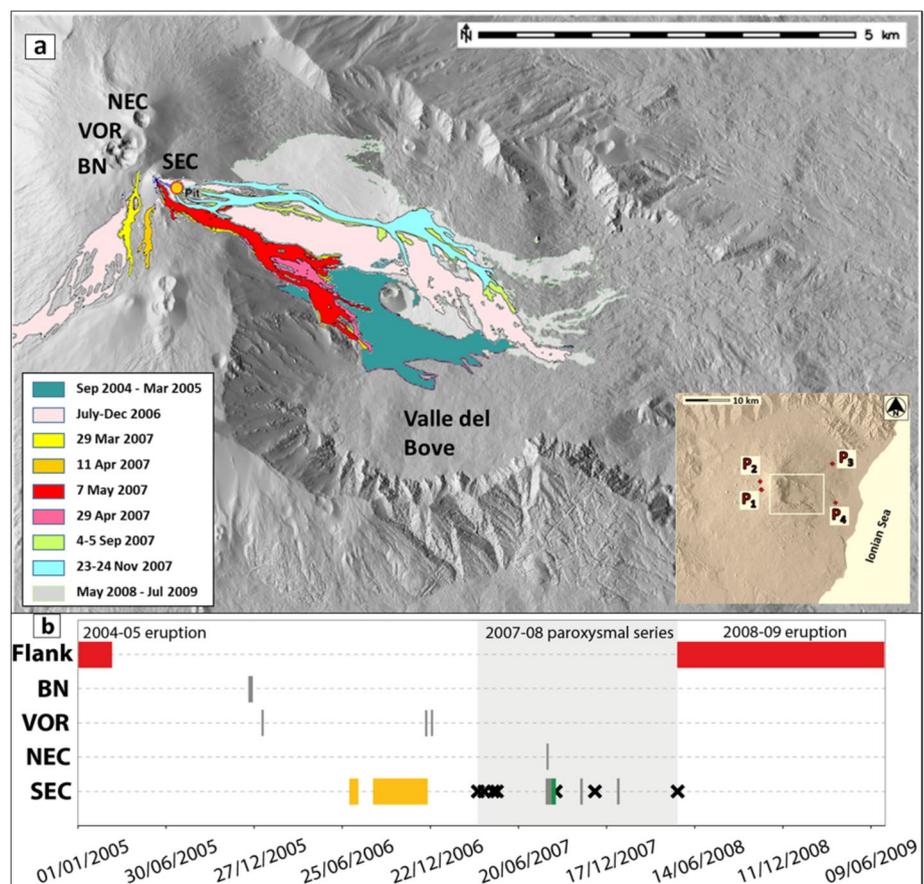
2 Materials and methods

2.1 Study area

Mt. Etna (Fig. 1), located in Italy, is one of the most active and comprehensively monitored basaltic volcanoes in the world. Its activity is characterized by persistent summit degassing and explosive eruptions, interspersed with recurrent flank eruptions [15, 16]. These flank eruptions pose the greatest hazard to the densely populated regions surrounding the volcano, as basaltic lava emissions occur from vents at lower elevations, increasing the likelihood of impacting inhabited areas. The summit of Mt. Etna hosts a complex of craters that are the focal point of persistent volcanic activity. These include the Northeast Crater (NEC), Bocca Nuova (BN), the Voragine (VOR), and the Southeast Crater (SEC) complex, which has evolved into multiple vents in recent years. Each of these craters exhibits distinct activity patterns, ranging from continuous degassing and fumarolic emissions to episodic strombolian and effusive eruptions. The SEC complex, in particular, has been the site of intense paroxysmal activity in recent decades, contributing to substantial modifications of the summit landscape. The eruptive dynamics of Mt. Etna are systematically monitored by the Istituto Nazionale di Geofisica e Vulcanologia—Osservatorio Etneo (INGV-OE). This monitoring utilizes a multidisciplinary network comprising thermal and visible cameras, seismic and infrasonic stations, tiltmeters, GPS (Global Positioning System) units, and strainmeters for ground deformation analysis, along with ultraviolet (UV) scanners to quantify SO₂ emissions from the craters. Additionally, satellite observations are routinely employed to provide sensitive data to complement ground-based monitoring efforts.

In this study, the 2005–2008 period was analysed and Mt. Etna volcanic states labeled taking into account the information available from INGV reports [17], recognizing five states of activity, namely “Quiet”, “Preparatory”, “Unrest”, “Cooling” and “Eruption”. During the selected period, in order to minimize the introduction of labeling errors, which could negatively impact the performance and reliability of the machine learning models, only time windows for which we had high confidence were labeled; this is why some intervals are not associated with any defined volcanic state. At

Fig. 1 **a** Map of Mt. Etna including the lava flows emitted during the 2005–2009. NEC=North East Crater, VOR=Voragine crater, BN=Bocca Nuova crater, SEC=South East Crater. The inset reports the location of the pixels used for the Synthetic Aperture Radar data. **b** Volcanic activity of Mt. Etna during 2005–2009, the time window analyzed in the present study. Grey = ash emission, green = Strombolian activity, X = lava fountain, orange = subterminal eruption, red = flank eruption



the end of the labeling process, we identified a total of 298 days characterized by eruptive activity, 47 days labeled as “Quiet”, 6 days as “Cooling”, 26 as “Unrest”, and 4 days as “Preparatory”.

At Mt. Etna summit, passive degassing is almost persistent. Also, the volcanic tremor signal is persistently recorded from seismic sensors. This reflects that a true “Quiet” state (i.e., absence of or negligible volcanic signals) is never attained at Mt. Etna. Thus, we chose to label “Quiet” the background periods characterized by a low degassing regime with no emission of volcanic material from the craters.

The “Preparatory” state identifies moments where seismic swarms are recorded underneath the volcanic edifice up to a depth of 10–20 km b.s.l.. Such seismic swarms usually anticipate a major eruptive phase (e.g., [18]), although any signs of activity may be not recorded during the occurrence of earthquakes.

The “Unrest” state recognizes periods characterized by an increase of monitoring signals such as volcanic tremor, ground deformation, impulsive degassing at summit, including occasional emission of volcanic ash or weak explosions from the craters. Although from a volcanological point of view any emission of volcanic material is classified as “eruption”, we preferred to assign the label “Eruption” to identify the major phases of the eruptive activity (i.e., energetic Strombolian activity, lava fountaining episode, emission of lava flows from summit craters, flank eruptions).

Moreover, the “Cooling” state is used to describe the periods following an emplacement of lava fields that are not characterized by new emission, although thermal anomaly is still recorded from satellites due to the cooling of the lava fields. This state is particularly significant for lava fountaining episodes, where huge amounts of lava flows are emitted in a few hours.

Finally, “Eruption” states were identified referring to the following events. The selected period was characterized by eruptive activity both at summit craters and at flanks of the volcanic edifice (Fig. 1). In 2005, the ongoing eruption started in September 2004 from a radial fissure opened between 3000 and 2350 m inside the Valle del Bove produced a lava field which expanded to an elevation of 1670 m [19]. Emission of lava flows ended in the early March 2005, and the activity shifted to passive degassing at summit crater, with occasional ash emissions from the central crater (BN and VOR) between the end of 2005 and beginning of 2006.

The eruptive activity resumed at the summit on the night of 14 July 2006, where a sub-terminal fracture opened at the base of the eastern flank of SEC emitting lava flows from two vents [20]. The eruption ended on 23 July 2006, and started again in late August 2006 in the same area, with emission of short lava flows and strombolian activity. Starting from 13 October 2006 various effusive fissures opened in different areas around the SEC and central crater, with emission of lava flows towards the Valle del Bove from a vent located at 2800 m and towards south-southwest from a fracture opened at the south flank of the central crater. Two more energetic phases of the eruption occurred on 16 November 2006, when collapse of the east flank of the SEC generated a PDC (Pyroclastic Density Current) which propagated for 1.2 km eastward [21], and on 24 November where an increase of the explosive activity at SEC led to a formation of an eruption column [22]. The eruption definitely ended on 15 December 2006.

A new phase of eruptive activity started in 2007, with a series of paroxysmal eruptions between 29 March 2007 and 10 May 2008 [23]. Each episode was characterized by an initial phase of ash emission and Strombolian explosions which gradually increased culminating in lava fountaining and formation of eruptive columns and emission of lava flows from the SEC, and from a pit crater located on the east flank of SEC from 4 September 2007. At the end of the last lava fountaining episode, a seismic swarm was recorded which culminated in a flank eruption that started on 13 May 2008 [24]. The initial phase of the eruption was characterized by explosive activity from a fissure located at east of the summit craters in NorthNorthwest—SouthSoutheast direction at 2800 m.

2.2 Satellite-derived datasets

We analyze different satellite data coming from various sources and covering a wide range of information useful for monitoring and analyzing volcanic activity. The data includes radiance, ground deformation, Land Surface Temperature (LST), daily SO₂ flux, and gravity anomalies. These anomalies can provide insights into magma migration and changes in the subsurface volcanic system. Although ash emission products are available, we decided not to include them in this study due to some limitations in data validation and reliability, which could affect the robustness of our analysis.

2.2.1 Ground deformation

Ground deformation is measured using SAR (Synthetic Aperture Radar) active sensors aboard satellites. SAR satellites operate in a sun-synchronous polar orbit at approximately 800 km altitude. Considering the time period of the present

analysis, the ENVISAT mission from the European Space Agency (ESA) has been exploited. The InSAR (Interferometric SAR) data analysis provides information on the surface deformation with a variable spatial resolution (from tens to hundreds of meters) and, in this case, a temporal resolution of about 35 days (reduced to 6 days with the modern Sentinel-1 satellites). For improved analysis, both ascending and descending orbit images are used, allowing to better constrain the vertical and east–west surface movements.

In this work, we employ the multi-temporal InSAR products available at the geo-data portal of the MASE (Italian Ministry of Environment and Energy Security). The InSAR technique adopted in the processing is based on the Persistent Scatterers [25]. The ascending orbit dataset consists of 61 images acquired along track 129 collected between 22/01/2003 and 14/07/2010. The descending orbit dataset consists of 50 images acquired along track 222 from 9/04/2003 to 16/06/2010. The results of the InSAR processing are the mean ground velocity and the displacement time series along the two orbits' Line Of Sight, LOS (about 23° from the vertical). The mean ground velocity (Supplementary fig. S1) shows complex deformation processes taking place at Mt. Etna during 2003–2010. The overall negative LOS signal of the upper part of the volcano, considering that it has the same sign in both orbits, is a subsidence and can be reasonably attributed to the deflation of the underlying magma chamber after the 2001–2003 intense eruptive activity (pixels P1 and P2). Also, the InSAR data clearly evidence two well-known structural phenomena affecting the eastern flank of Mt. Etna. One is the slip along the Pernicana Fault located on the North (P4). This fault is also the northern detachment surface of the southeastern flank of Mt. Etna. This flank is sliding toward East, confirmed by the opposite signs of the InSAR velocities (identifying an East–West movement due to the opposite orbits of the satellites, pixel P3).

The InSAR datasets needed post processing to be AI-ready for the ML operations. Two main aspects are considered. First, the InSAR time-series dataset has been sliced to fit the temporal window of 2005–2008. The selected dates are 02/03/2005–25/03/2008 for the ascending orbit and 13/04/2005–07/05/2008 for the descending orbit. As a second step, downsampling of the nearly 750000 and 440000 pixels of the ascending and descending orbits was carried out. The downsampling mask consists of a central area on the summit craters with a sampling step of 200 m, increasing to 400 m and 1000 m progressively (Figure S1). During this time window, Mt. Etna showed inflation of the summital area, very close to the highest elevations reached without losing spatial coherence. The Pernicana fault and the south-eastern flank show the eastern sliding as previously described.

2.2.2 Radiance

Spectral radiance data, acquired by the SEVIRI (Spinning Enhanced Visible and InfraRed Imager) sensor, onboard Meteosat Second Generation satellites, provides observations across 11 bands ranging from optical to thermal infrared with a spatial resolution of approximately 3 km at nadir and a temporal resolution of 15 min. These frequent, high-resolution data are particularly valuable for studying volcanic phenomena. In terms of thermal emissions, SEVIRI data allow for the detection and tracking of hotspots, enabling the assessment of lava flows and surface temperature anomalies, as demonstrated by [26]. Additionally, SEVIRI has proven effective in monitoring volcanic plumes, including ash and gas emissions, providing insights into plume dynamics as highlighted by [27]. This capability supports hazard assessment and informs aviation safety during eruptive events. We here extracted the time series of calibrated radiances in a box centered in Mt. Etna summit of 19×11 pixels covering an area of 62×48 km².

2.2.3 Land surface temperature

To take into account the temperature of the volcanic edifice, Land Surface Temperature (LST) was measured using MODIS (Moderate-Resolution Imaging Spectroradiometer) sensors installed on the polar-orbiting satellites AQUA and TERRA, with a spatial resolution of 1 km at nadir and a temporal resolution of about 1–2 daily passages for each point on Earth. These data, already corrected in terms of cloud cover, were extrapolated and processed over Mt. Etna volcano for the period 2005–2009, exploiting the open-source Google Earth Engine computing platform, which provides the entire MODIS collection. Daytime and nighttime data were considered in order to extract surface temperature values for all the pixels within a defined area around Etna. To exclude the contribution from volcanic activity, i.e., lava flows or strombolian activity, the summit area and part of the Valle del Bove, located in the eastern flank of the volcano, were excluded from the analysis, as their temperature could lead to an alteration of the results. From the average between daytime and nighttime data, the maximum temperature difference was calculated, considering the median of the lowest 1% and the highest 10% of values. The seasonality was removed from the time series by subtracting the values from non-volcanic

areas on the northern and southern slopes of the volcano [28]. The resulting data were filtered, doing a 60-day moving average of the difference between the area around the volcano and the surrounding area.

2.2.4 Sulphur dioxide

The daily SO₂ flux data, collected by the OMI (Ozone Monitoring Instrument) sensor aboard the polar-orbiting satellite AURA, is crucial for monitoring volcanic gas emissions. OMI, which follows a sun-synchronous orbit at approximately 705 km altitude with a spatial resolution of about 13 × 24 km per pixel, provides daily global coverage. SO₂ emissions are a key indicator of volcanic activity, as the release of large amounts of gas is often associated with eruptive episodes and magma ascent. OMI is a nadir-viewing imaging spectrograph that measures atmosphere-backscattered sunlight in the ultraviolet–visible range from 270 to 500 nm with a spectral resolution of about 0.5 nm [29].

Volcanic SO₂ fluxes have been retrieved exploiting different methods, such as the traverse method [30–32], which involves measuring SO₂ concentrations along a specific path across the plume and calculating the flux based on the measured concentration and plume geometry, and the Delta-M method [33], which is based on the differential absorption of solar radiation by SO₂. To calculate SO₂ fluxes atmospheric corrections are required for the attenuating effect of clouds and particles that may affect the measurements. Here we employed the OMSO2 Level 2G (0.125° × 0.125°) global dataset to estimate total SO₂ in a box of 500 × 500 km around Etna summit.

2.2.5 Gravity field

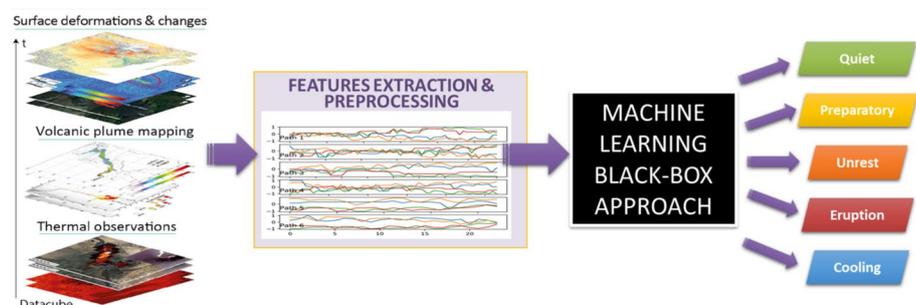
Time-variable gravity measurements from the Gravity Recovery and Climate Experiment (GRACE; 2002–2017) data have opened new possibilities for studying large-scale mass redistribution and transport in the Earth system, including the hydrosphere, ocean, cryosphere, and solid Earth [34]. We here use the time-variable GRACE satellite gravimetry, with the aim of investigating the capabilities of this approach to study the dynamics of Mt. Etna volcano over time-scales of months to years. In particular we focus on the three-year period (2005–2008) satellite gravity data collected in a sun-synchronous polar orbit at approximately 500 km altitude, with a spatial resolution of 300 km² and a temporal resolution of about one month. The CNES/GRGS RL05 Truncate Singular Value Decomposition (TSVD) monthly gravity anomalies data were chosen [35]. The time series of gravity field models are expressed in normalized spherical harmonic coefficients. Taking into account GRACE resolution, a positive trend from 2005 to 2008 was observed in the region of Mt. Etna, which could represent volcano-scale variations, most likely attributed to hydrological and volcanological effects. Overall, results have important implications for the volcanic hazard assessment and encourage the joint use of space platforms (also including the Next Generation Gravity Mission; NGGM) and terrestrial data at volcanoes where bulk mass redistributions may develop over a long-term interval.

2.3 Methodology

In this study, we carry out a preliminary evaluation to assess whether EO data combined with AI techniques is capable of characterizing the state of an active volcano and potentially predicting its future behavior (Fig. 2).

The EO data employed for volcanic monitoring includes data described in the paragraph 2.2, i.e., SAR data to track ground deformation, multispectral data in the mid-infrared and thermal bands to monitor respectively high-temperature events and the temperature of the volcanic edifice, SO₂ data to provide insights into degassing magma, and gravity data.

Fig. 2 Workflow of the methodology developed



In this analysis, supervised machine learning techniques are applied, with labeling performed by expert volcanologists and validated through multiple ground-based observations. Tests are also conducted with random labeling to ensure that the algorithms do not learn indiscriminately, revealing significantly reduced performance in such cases. Various common machine learning methods are tested, and their performance is evaluated systematically. A more detailed pipeline plot of the methods used is presented in supplementary material as well as a pseudocode of data processing workflow.

2.3.1 Preprocessing

The collected data are particularly inhomogeneous, therefore we need to pre-process them to apply the ML analysis. The main issues are: different time sampling (minutes to days/weeks), different ground spatial sampling (point measurements to hundred of meters), and Class Imbalance (CI). The latter refers to a significant disproportion in the number of data points for each class. Regarding the CI, about 32000 data points are labelled as “Eruption State” while only 50 data points are labelled as “Preparatory State”. CI is a common challenge in classification tasks, as it causes ML models to be biased towards the class with a larger number of data points leading to poor classification performance [36].

A set of data containing examples of both inputs—features described in paragraph 2.2—and correct outputs—labels described in paragraph 2.1—is required. Initially, we had a time series for each feature examined, sourced from various sensors mounted on different satellites, each possessing its own distinct characteristics. More specifically, each sensor has its own temporal and spatial resolution for data acquisition, resulting in varying time measurements for different features from distinct areas (see paragraph 2.2). For example, the SEVIRI sensor provides a radiance measurement every 15 min, while measurements of gravity anomalies have a temporal resolution of approximately one month. To address the inhomogeneity of the data and to obtain a consistent dataset with measurements of each feature corresponding to a specific time instant, we choose a common temporal resolution and, where needed, generate synthetic data utilizing appropriate interpolation methods. More precisely, the SEVIRI temporal resolution of 15 min has been chosen, which means that we retain all the available radiance data. Hence, to generate synthetic data for the remaining four features the linear interpolation method has been selected.

The method essentially interpolates between two consecutive data points by assuming a linear variation within the interval. The computation is then carried out by assigning intermediate points the values inferred from the corresponding straight line. We acknowledge that interpolating to achieve a 15-min sampling rate from monthly data may be considered an overly simplistic approximation. However, the validity of this approach is supported by the fact that the final accuracy values remain high. To perform a more complete analysis we also used the previous neighbor interpolation, where the value of each feature is held constant throughout the time interval. We applied the machine learning methods to the datasets obtained from both interpolation approaches and found that the highest accuracies were achieved with the linear method. For this reason, as discussed in Sect. [Methodology](#), it was selected as the one adopted.

Once a consistent dataset, comprising both real and interpolated data, is generated, we address the CI problem by applying a random undersampling strategy: we select a fixed number of instances from each class—specifically, 350 data points. This reduces the considerable disproportion between classes while retaining a sufficiently large and balanced dataset to train the ML models effectively.

Beyond random undersampling, other well-known strategies for handling class imbalance include Synthetic Minority Oversampling TEchnique (SMOTE) and cost-sensitive learning. We experimented with SMOTE (see results in Table S9). Specifically, we applied a combination of under-sampling and over-sampling techniques to construct a balanced dataset. The majority classes—“Eruption”, “Quiet”, and “Unrest”—were each under-sampled to 350 randomly selected data points. In contrast, the minority class, “Preparatory”, originally represented by only 50 data points, was augmented using the SMOTE, which generates synthetic instances based on a k-nearest neighbors approach [37, 38]. Although the original dataset contained thousands of instances labelled as “Eruption”, “Quiet”, and “Unrest”, we deliberately limited the number of samples per class to 350. This decision was motivated in light of findings showing that generating a large number of synthetic samples from a small initial minority class compromises dataset quality and deteriorates classifier performance [39]. As one can see in the supplementary material, SMOTE approach did not lead to substantial changes in model performance.

On the other hand, cost-sensitive learning—while theoretically appealing—was not adopted, because, as described in details in [40], it requires defining specific misclassification costs. In our case, no reliable information was available to determine such costs, and the risk of assigning arbitrary or misleading values was deemed too high.

2.3.2 ML techniques

Our study aims at estimating the state of a volcano using the satellite-derived parameters. To achieve this, we need a dataset that includes these parameters as predictors and the corresponding volcano states as labels. In this context, a supervised ML technique is more suitable to automatically learn how to predict a volcano's state based on other parameters. This learning phase, or training, results in the creation of a predictive model. Once the model is trained, it can forecast the volcano's state for new records containing the same predictors. To assess the model's ability to generalize (i.e., its capacity to fit and respond correctly to previously unknown data) and its prediction accuracy, a portion of the dataset is used to test the model once it has been trained. In this second phase of the learning process known as the *test phase*, the model's predictions on new data are compared with the actual labels. We quantify the comparison by the *accuracy*, defined as the ratio of correctly predicted results to the total test data, expressed as:

$$\text{Accuracy} = \frac{\text{matched results}}{\text{total test data}} \quad (1)$$

While accuracy is used as the primary evaluation metric in this study, we acknowledge that other metrics, such as precision, recall, and F1-score, are often valuable in assessing model performance, particularly in binary classification tasks. However, given the multiclass nature of this problem, accuracy provides a more straightforward and interpretable measure of overall performance.

Another tool generally used to evaluate the performance of a ML algorithm is the *confusion matrix*. It is a square matrix C , where the dimension is the number of classes in the problem. Columns typically represent the predicted classes, while rows represent the actual classes (or vice versa). Each entry c_{ij} in the matrix indicates the number of instances belonging to the i -th class that the algorithm predicted as the j -th. Therefore, the diagonal elements of the confusion matrix represent the instances that are correctly classified by the algorithm, while off-diagonal elements correspond to misclassifications.

In order to further assess the algorithm's performance, another common procedure called *k-fold cross validation* [41] is used. The data is divided into k subsets (folds), then the model is trained on $k - 1$ folds and tested on the remaining fold. This process is repeated k times, each time using a different fold for testing. The final model accuracy is the average of the results from each fold, helping to reduce overfitting and variance from a single train-test split.

To implement the k -fold cross-validation method, the following procedure is adopted:

Segmentation by Volcano State: the entire dataset timeline is divided into distinct time segments, each corresponding to a specific volcano state. A change in the volcano's state marked the start of a new segment.

Division into Equal Portions: each segment is further divided into five equal chronological portions. For example, the first portion of each segment consisted of the first 20% of its chronological data, the second portion the next 20%, and so forth.

K-Fold Cross-Validation Process: for each k -fold iteration, one portion from every segment is selected as the test set, while the remaining portions are used for training. For instance: in the first iteration, the first portion of each segment served as the test set; in the second iteration, the second portion of each segment was the test set, and so on, until all five portions had been used as test sets across the iterations.

The approach outlined above, where the data are split into sequential, non-overlapping portions (like fivefold cross-validation), performs better than random choice for several reasons:

Preserves temporal structure: if data have an inherent order (e.g., time series or sequential data), random splits might disrupt temporal relationships, leading to an unrealistic model evaluation. Sequential splits maintain this natural order.

Avoids data leakage: when using a random choice for splitting, there is a higher risk of training data leaking into the test set, if related data points (like those close in time or with dependencies) are included in both sets. Sequential splits prevent such cross-contamination.

Balanced coverage: by ensuring each portion is used as a test set exactly once, every part of the data is tested in a systematic way. Random splits might lead to uneven coverage or over-representation of certain data points in training, making the evaluation less reliable.

In summary, this method ensures that the model is tested in a controlled, structured way, especially when data order matters or there's a risk of data leakage. The ML techniques that we adopt and cross-check are all well known in the literature [42–44]. Below, we provide a quick overview of them.

Decision Tree: it involves constructing a tree-like model based on variable values. Essentially, the algorithm identifies "decision points" known as nodes, where data are divided into subgroups. For instance, with a continuous variable,

the algorithm may determine a threshold that separates the data into those with values above and those with values below this threshold. The tree is composed of multiple sequential nodes until the data points are ultimately classified into a specific label class.

Random Forest: it builds on the decision tree method and aims to improve the generalization of decision tree outcomes. The core concept involves generating multiple trees, each using a subset of the data and variables from the entire dataset. The class of a record is determined by identifying the label predicted by the majority of these trees.

K-Nearest Neighbors (KNN): it represents each record in a Cartesian space, where each variable value acts as a coordinate. Essentially, each record is treated as a point in an n -dimensional space. The class of an unlabeled point is determined by the most common class among its k nearest points. KNN with $k = 1$ (knn1) means that the class of an unlabeled record is the same as its nearest labeled training point. When more than one nearest point is considered, the majority class among these neighbors determines the class of the unlabeled point.

Support Vector Machine (SVM): it also uses a Cartesian space representation but, instead of focusing on nearest points as KNN, the algorithm seeks to identify separating hyperplanes. These hyperplanes divide the space in such a way that an unknown record is classified based on which side of the hyperplane it falls on.

Naive Bayes: it works by constructing probability distributions for each variable based on the training data, representing the likelihood of belonging to each class. This method is particularly useful for small datasets. When a new record needs to be classified, the Bayes' theorem is applied to calculate the probability of the record belonging to each class. The class with the highest probability is then selected as the classification for the new record.

Discriminant Analysis: it works by finding a mathematical model, known as the *discriminant function*, that best separates classes based on their features. There are several types of discriminant analysis. In this case, we used *linear discriminant analysis*, which identifies a linear combination of features to separate the classes.

The choice of using these machine learning techniques instead of neural networks is primarily driven by the nature of the data. Specifically, we are working with a structured dataset in tabular form. Several empirical studies have been conducted to compare deep learning and traditional machine learning algorithms for solving problems involving tabular data, showing that, for now, no deep learning technique seems to outperform decision tree-based machine learning algorithms [45, 46].

Neural networks excel in handling unstructured data, such as images, text, or audio, where complex feature extraction is required. However, for structured data, tree-based models like decision Trees and random forest are typically more effective because they can capture non-linear relationships, handle missing values, and provide feature importance insights.

Additionally, these models are computationally more efficient than deep learning methods, requiring less data and training time while still achieving high accuracy. In fact, given the high accuracy levels obtained with these traditional models, the use of neural networks would have provided no significant advantage, making their added complexity and computational cost unjustified.

Finally, the achieved accuracies are so high that the neural network approach becomes unnecessary.

For these reasons, we have opted for the above mentioned classic ML techniques, ensuring a balance between performance, interpretability, and computational efficiency for our structured dataset.

We conducted our analysis on a multitude of satellite data acquired over Mt. Etna (paragraph 2.2) using all of these supervised ML techniques. Considering that we employed KNN in six different ways using $k = 1, 3, 5, 7, 9$ and 11 , we applied a total of eleven different algorithms.

Hyperparameter tuning is another common and often essential practice when working with ML techniques, as it can significantly impact the performance of a model. A comprehensive discussion of the distinction between parameters and hyperparameters, along with a general overview of hyperparameter selection, can be found in [47]. As these details are beyond the scope of this study, we provide only a brief summary here. Parameters in ML are the values that a model learns directly from the data during training, such as those determining how predictions are made. Hyperparameters, on the other hand, are predefined settings that influence the training process—for instance, in a random forest model, the number of trees or the number of variables included in each tree. While careful hyperparameter tuning is often critical for optimizing performance, in this study, as it will be shown in the next sections, the achieved accuracies were so high (reaching 1.0 in one case) that further tuning was deemed unnecessary.

3 Results

3.1 Satellite data analysis and preprocessing for ML

As previously discussed, there are significant differences in the temporal sampling of the time series. Table 1 provides a summary of the acquisition time steps for all features. To reconcile these differences, we adopted the minimum sampling rate of 15-min intervals as the common time step across all features and linearly interpolated the data to generate synthetic values, enabling the construction of an appropriate dataset for our analysis.

Figure 3 illustrates data time series after preprocessing. To ensure that the feature plots are readable and the temporal dynamics are understandable, a specific time interval was chosen for each feature, balancing the level of detail with overall clarity.

3.2 ML results

The results of each applied method are summarized in Table 2, expressed in terms of accuracy as described in the previous section. We used a fivefold cross-validation procedure, meaning each method has been applied five times employing five different pairs of training and test datasets. As one can see the best method is random forest showing an accuracy of 1 for three folds.

For completeness, Fig. 4 illustrates the performance of the random forest classifier across fivefold cross-validation. The diagonal elements represent the mean percentage of instances correctly classified, while the off-diagonal elements correspond to the mean percentage of misclassification. The confusion matrices resulting from the other ML methods are provided in the supplementary material, Fig. S3-S12. As stated in Sect. [Methodology](#), we also applied the same techniques using previous neighbor data interpolation to generate data for ground deformation, gravity anomaly, and daily SO₂ concentration to obtain a dataset with a 15-min resolution. Although the resulting accuracy values remained high, they were lower than those obtained with the dataset created using linear interpolation. These findings are available in the supplementary material.

4 Discussion and conclusions

In this study, we applied several widely used supervised machine learning techniques to classify volcanic eruption states using satellite data, employing k -fold cross-validation with $k = 5$ to assess their performances. As shown in Table 2, all methods achieved high levels of accuracy, reaching in the majority of the cases an accuracy of at least 0.9 and with none falling below 0.6. Although the random forest classifier is the only method to achieve an accuracy of 1 three times, it cannot be conclusively stated that it outperforms the other methods in a more general context.

First, it is important to note that random forest is an ensemble method, which is a technique that combines multiple models—in this case, decision trees—to improve the reliability of the model and reduce overfitting. Therefore, we hypothesize that minor errors may be introduced by the specific dataset used to train our models. These errors appear to be mitigated by the use of an ensemble method such as random forest.

Table 1 Acquisition time steps and acquisition time interval for all the features employed in our analysis

Feature	Acquisition Time Step	Acquisition Time Interval
Spectral Radiance	15 min	01/01/2005–07/08/2008
Gravity Anomaly	1 month ca	16/03/2005–16/05/2008
Ground Deformation Ascending Orbit	35 days ca	02/03/2005–25/03/2008
Ground Deformation Descending Orbit	35 days ca	13/04/2005–07/05/2008
SO ₂ Daily Concentration	2–3 days	15/03/2005–30/05/2008
Land Surface Temperature	12 h ca	01/01/2005–30/05/2008

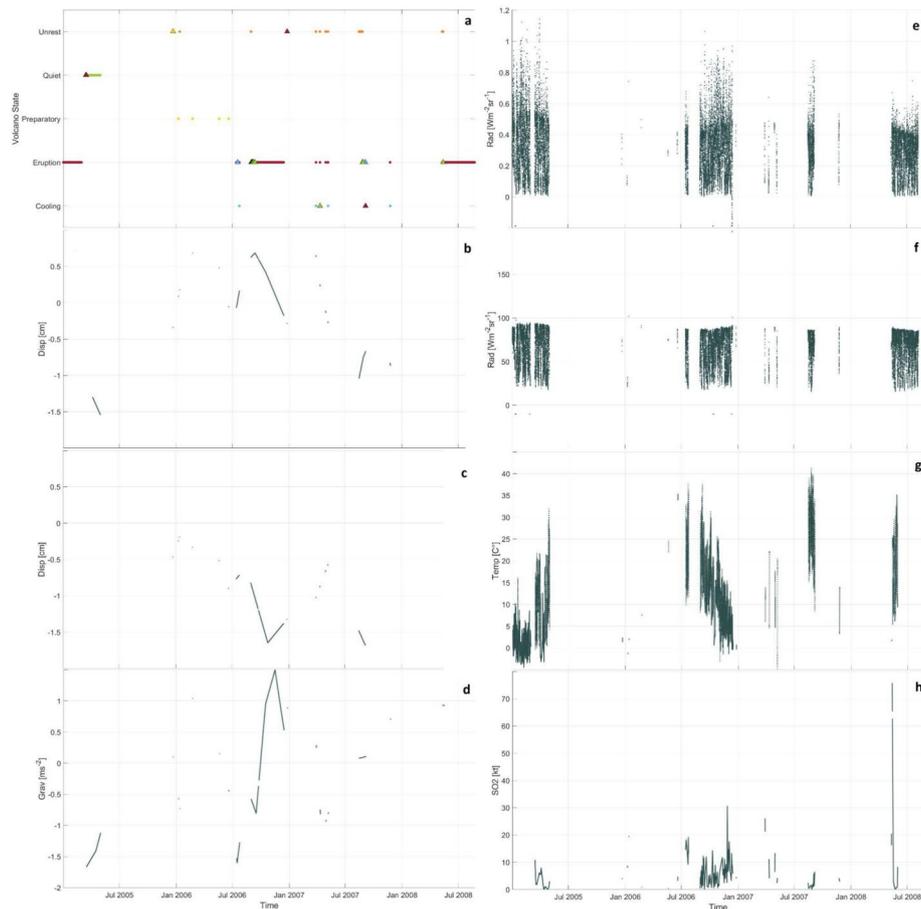


Fig. 3 **a** Time series of the labelled volcanic states of Mt Etna from January 2005 to August 2008; points misclassified by the Random Forest model have been marked as triangles. The perceived overlap arises from the fact that some states have a very short duration—almost point-like in time—while others consist of sequences of closely spaced points, which appear as continuous lines. This visual representation can give the impression of temporal coincidence. **b** Ground deformation time series measured at the point P1 (see Fig. 1) by ENVISAT from April 2005 to November 2007 in ascending orbit. **c** The same as (b), but for ENVISAT descending orbit. **d** Gravity anomalies measured by the GRACE satellites from October 2006 to September 2007. **e** Radiance values measured during January 2005 by SEVIRI's fourth channel. **f** the same as (e), but for SEVIRI's ninth channel. **g** Land surface temperature time series measured by MODIS from January 2005 to February 2005; **h** daily SO₂ flux data measured by OMI January 2005 to May 2008

Table 2 Accuracy achieved for each k-fold iteration with the ML methods adopted

Method	Accuracy K-Fold 1	Accuracy K-Fold 2	Accuracy K-Fold 3	Accuracy K-Fold 4	Accuracy K-Fold 5
Decision Tree	0.8378	0.9963	0.9963	0.9963	0.6515
Random Forest	0.8687	0.9963	1.0000	1.0000	1.0000
Support Vector Machine	0.8108	0.9522	0.9596	0.9669	0.9697
Naïve Bayes Classifier	0.6178	0.7096	0.9154	0.9816	0.9659
KNN with 1 neighbour	0.8494	0.9632	0.9816	0.9669	0.9962
KNN with 3 neighbours	0.8571	0.9669	0.9816	0.9706	0.9848
KNN with 5 neighbours	0.8610	0.9265	0.9890	0.9632	0.9848
KNN with 7 neighbours	0.8764	0.9154	0.9743	0.9596	0.9735
KNN with 9 neighbours	0.8842	0.9044	0.9706	0.9559	0.9773
KNN with 11 neighbours	0.8842	0.9044	0.9632	0.9485	0.9735
Linear Discriminant Analysis	0.8069	0.9559	0.9596	0.9669	0.9583

Fig. 4 Confusion matrix of the random forest classifier obtained over fivefold cross-validation. The diagonal elements represent the percentage of instances correctly classified

True Class	Predicted Class				
	Cooling	Eruption	Preparatory	Unrest	Quiet
Cooling	96.91%	2.7%			0.39%
Eruption		94.72%			5.28%
Preparatory			100%		
Unrest		2.29%		97.71%	
Quiet		0.29%			99.71%

A very important point to discuss is the use of the interpolation to force very different resolution data to be homogeneous. It is acknowledged that applying linear interpolation to low-frequency data introduces assumptions that may not fully capture the true dynamics of volcanic activity. Nevertheless, a trade-off was sought to enable data alignment while preserving meaningful classification performance. The results show accuracies high enough to validate this process. Moreover the results of nearest neighbour and linear interpolation strategies were compared, showing that linear interpolation led to higher classification accuracies. However, it is also recognized that, for operational monitoring systems, nearest-neighbor interpolation may represent a more appropriate and robust choice. This issue is therefore identified as an important direction for future work, particularly in scenarios involving real-time volcanic surveillance and high-frequency sensing applications.

As outlined in Sect. [ML results](#), previous neighbor data interpolation has been applied to derive 15-min resolution datasets for ground deformation, gravity anomaly, and daily SO₂ concentrations. While this interpolation method shows lower accuracies than those obtained with data generated from linear interpolation, it is better suited for real-time monitoring scenarios, especially when different data types and sampling intervals must be made compatible and organized in a unified tabular format. In this study, we focused on analyzing time series data from the 2005–2008 period to identify the most effective method for classifying volcanic activity states. Since real-time classification requires additional considerations and analysis, and is not the primary focus of this work, we plan to explore it more thoroughly in future research.

Finally, interpolating low-temporal-resolution features introduces a potential risk of bias, as the filled values may not fully capture the true temporal dynamics of the data. Nonetheless, these features often contain valuable information that can enhance model performance.

To assess their impact, we conducted an ablation test by removing the interpolated variables entirely. Remarkably, even without these predictors, the model maintained a solid accuracy of approximately 73%. This result demonstrates the robustness of the remaining feature set and confirms the model's capacity to generalize effectively, even in a more constrained scenario.

The fact that accuracy climbs to around 97% when including the interpolated features highlights their added value. However, it's important to emphasize that the 73% baseline represents a strong foundation. It shows that, despite excluding two potentially informative variables, the model retains a high level of predictive skill. This suggests that while the interpolated variables enhance performance, they are not indispensable for achieving meaningful results.

In summary, the interpolated features offer a substantial performance boost—with minimal risk of overfitting—while the model's strong performance without them underscores its inherent reliability and the quality of the underlying signal in the remaining data.

Furthermore, the importance of each feature for the random forest classifier was computed using the Out-Of-Bag (OOB) Error. The complete results are presented in Table S7. Specifically, the importance of each feature was estimated by the increase in the OOB Error when the values of that feature were randomly shuffled, thereby breaking the relationship

between the predictor and the response. This measure was computed for every tree in the random forest, averaged over the entire ensemble, and then normalized by dividing by the standard deviation of the delta errors across trees.

Among the features with the highest importance value are those with low-frequency sampling. In particular, ground displacement—specifically, the time series of Point 3 observed from the ascending orbit—and gravity anomaly, with importance values of 1.03 and 1.66, respectively. These are the only features having an importance value greater than 1, revealing their strong influence on predicting the correct volcanic state since the average increase of error is larger than the standard deviation of that increase across all trees in the ensemble. SO₂ daily concentration, with a sampling rate of 2–3 days, also showed a notable importance value of 0.84. These are followed by temperature, with a sampling rate of 12 h and showing an importance value of 0.6, and the seventh and eighth SEVIRI acquisition channels for radiance values, with a sampling rate of 15 min, showing importance values of 0.57 and 0.56, respectively.

These results are consistent with the observed decrease in classification accuracy upon exclusion of low-frequency data.

Regarding the labeling of the data, to strongly validate our dataset and to ensure that the volcanic states were correctly identified, the entire ML analysis was also performed using a dataset with randomly shuffled labels. The accuracies resulting from this analysis are shown in Table S8. The highest accuracy was achieved by the random forest with a value of 0.2867. This shows that the ML algorithms were not able to find any meaningful correlation between predictors and labels, effectively selecting a response almost at random from the possible outputs.

To further analyze and interpret the results obtained performing our analysis with multiple classifiers, we conducted statistical tests to evaluate the accuracy data (Table 2) for each ML method. Since k-fold cross-validation provides five accuracy values for each method, these can be considered as a sample of size five from the general accuracy distribution. In this context, inferential statistics can be used to determine whether the accuracy distributions of the methods differ significantly.

In order to select the statistical test to be used, normality and homoscedasticity (variances equality) tests have to be performed. The tests used for normality distribution checks are generally Shapiro–Wilk, Anderson–Darling and Kolmogorov–Smirnov [48–50]. Usually the Kolmogorov–Smirnov test is used when a large amount of data is available: more than 50 data points per group [51]. Since we have only five data points for each group, we tested our data with the Shapiro–Wilk and Anderson–Darling tests. Both these tests revealed that the samples are not normally distributed. Levene’s and Bartlett’s tests revealed there is also an absence of homoscedasticity in our data [52, 53]. Consequently, we employed the non-parametric Kruskal–Wallis test to assess whether the data were significantly different [54]. With a p-value of 0.422, greater than the significance level of 0.05, we cannot conclude that the machine learning methods used in our analysis could be considered significantly different. Nor did pairwise comparisons performed using the Dwass–Steel–Critchlow–Fligner test reveal significant differences between methods [55]. The statistical pairwise comparison test results are shown in Table S1.

Even if we cannot prove statistical significant differences existed among the methods overall, the low p-value (very near to the significance level) and each method’s accuracy allow us to revisit our data and consider the characteristics of certain techniques, particularly k-nearest neighbor and random forest. These methods, unlike others, do not rely on assumptions about the underlying data distribution but instead make decisions based on the specific positions of individual data points in the feature space. From this, we can suppose that the precise positioning of data points in the feature space is at least as important as the overall data distribution for this type of classification. In other words, classifying the volcanic state appears to be achievable by focusing on local relationships between data points rather than relying on global statistical distributions or alternatively we should assume that such distributions are geometrically very complex.

It can be observed that we used a very small dataset for each group in our statistical analysis, consisting of only five accuracy values for each ML method. However, increasing the number of accuracy values for each method would have significantly reduced the size of the dataset used to test the machine learning algorithms due to the specific k-fold method we implemented (see paragraph 2.3.1), with consequent issues in terms of accuracy’s reliability and eventually overfitting. In fact, according to [56, 57], the most common train-test splits in the literature are 70:30 and 80:20. These ratios offer a good balance by providing sufficient data for both training and testing and are frequently chosen for their robustness and reliability in various contexts. For this reason, although working with small datasets can present challenges for statistical analysis, we opted to use fewer accuracy values while ensuring greater confidence in their reliability.

This study, conducted on one of the most closely monitored volcanoes in the world, employs supervised learning techniques with time-labeled instances, incorporating field-based data. It paves the way for future research that could

leverage semi-supervised or unsupervised learning approaches, or transfer learning methods in the case of volcanoes with similar characteristics.

However, the generalization of this methodology to less-monitored volcanoes is a challenging task, highly dependent on both the nature of the volcano and the availability of observational data. Our work represents an initial step aimed at evaluating the feasibility and potential of machine learning in an ideal monitoring context. Future developments will focus on adapting these techniques to more data-scarce environments, possibly by integrating alternative data sources, employing domain adaptation strategies, or using models trained on well-monitored systems to inform those with limited ground-based data.

5 Permission to use third-party material

All of the material is owned by the authors and no permissions are required.

Acknowledgements Christian Bignami and Cristiano Tolomei are acknowledged for providing access to the InSAR products by MASE (see Data availability statement). We are also thankful to the project Pianeta Dinamico—SAFARI—code CUP D53J19000170001—funded by Italian Ministry MIUR (“Fondo Finalizzato al rilancio degli investimenti delle amministrazioni centrali dello Stato e allo sviluppo del Paese”, legge 145/2018), VT-SAFARI 2023-25, as it has partially funded this work.

Author contributions All the authors collectively participated in the development and composition of all sections of the paper. The manuscript reflects ongoing collaboration, exchanges of ideas, and active discussions among the team. More specifically: C.P, A.P. and G.R made substantial contributions to the implementation, application, and evaluation of machine learning techniques. G.G. Conceptualization, coordination, funding. G.B. Data acquisition, validation. M.D. processed MODIS thermal data and contributed to the writing of the part related to sulphur dioxide and the supplementary material. E.T. Data acquisition, validation, writing. F.Z. classification of the Volcanic States.

Funding This work has been partially funded by the Pianeta Dinamico—SAFARI—code CUP D53J19000170001—funded by Italian Ministry MIUR (“Fondo Finalizzato al rilancio degli investimenti delle amministrazioni centrali dello Stato e allo sviluppo del Paese”, legge 145/2018), VT-SAFARI 2023–25.

Data availability InSAR products are freely available at the geo-data portal of the MASE (Italian Ministry of Environment and Energy Security), <https://gn.mase.gov.it/portale/prodotti-interferometrici>, under the National Remote Sensing Plan project (PST). SEVIRI data are freely available from EUMETSAT data store.

Code availability There is no code to share.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Ganci G, Cappello A, Bilotta G, Del Negro C. How the variety of satellite remote sensing data over volcanoes can assist hazard monitoring efforts: the 2011 eruption of Nabro volcano. *Remote Sens Environ.* 2020;236:111426. <https://doi.org/10.1016/j.rse.2019.111426>.
2. Kumari S, Agarwal S, Agrawal NK, Agarwal A, Garg MC. A comprehensive review of remote sensing technologies for improved geological disaster management. *Geol J.* 2025;60(1):223–35. <https://doi.org/10.1002/gj.5072>.

3. Andries A, Morse S, Murphy RJ, Lynch J, Woolliams ER. Using data from earth observation to support sustainable development indicators: an analysis of the literature and challenges for the future. *Sustainability*. 2022. <https://doi.org/10.3390/su14031191>.
4. Valade S, Ley A, Massimetti F, D'Hondt O, Laiolo M, Coppola D, et al. Towards global volcano monitoring using multisensor sentinel missions and artificial intelligence: the MOUNTS monitoring system. *Remote Sens*. 2019. <https://doi.org/10.3390/rs11131528>.
5. Han W, Zhang X, Wang Y, Wang L, Huang X, Li J, et al. A survey of machine learning and deep learning in remote sensing of geological environment: challenges, advances, and opportunities. *ISPRS J Photogramm Remote Sens*. 2023;202:87–113. <https://doi.org/10.1016/j.isprsjprs.2023.05.032>.
6. Schmitt M, Ahmadi SA, Xu Y, Taşkın G, Verma U, Sica F, et al. There are no data like more data: datasets for deep learning in earth observation. *IEEE Geosci Remote Sens Mag*. 2023;11(3):63–97. <https://doi.org/10.1109/MGRS.2023.3293459>.
7. Corradino C, Ganci G, Cappello A, Bilotta G, Calvari S, Del Negro C. Recognizing eruptions of Mount Etna through machine learning using multiperspective infrared images. *Remote Sens*. 2020. <https://doi.org/10.3390/rs12060970>.
8. Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm. *Pattern Recognit*. 2003;36(2):451–61.
9. Pandey D, Niwaria K, Chourasia B. Machine learning algorithms: a review. *Mach Learn*. 2019;6(02): 02.
10. Mahesh B. Machine Learning Algorithms -A Review, 9. 2019. <https://doi.org/10.21275/ART20203995>.
11. Petrelli M, Anantrasirichai N, Bean CJ, Biggs J, Malfante M, Wilding JD, et al. Methodological advances in volcanology: the role of artificial intelligence in volcano monitoring, modelling, and hazard assessment –part 1: tectonics and plumbing systems. *Encycl Volcanoes*. 2025. <https://doi.org/10.3122/X5043X>.
12. Behncke B, Neri M. Cycles and trends in the recent eruptive behaviour of Mount Etna (Italy). *Can J Earth Sci*. 2003;40(10):1405–11. <https://doi.org/10.1139/e03-052>.
13. Proietti C, De Beni E, Cantarero M, Ricci T, Ganci G. Rapid provision of maps and volcanological parameters: quantification of the 2021 Etna volcano lava flows through the integration of multiple remote sensing techniques. *Bull Volcanol*. 2023;85(10):58. <https://doi.org/10.1007/s00445-023-01673-w>.
14. Rosi M, Acoella V, Cioni R, Bianco F, Costa A, De Martino P, et al. Defining the pre-eruptive states of active volcanoes for improving eruption forecasting. *Front Earth Sci*. 2022;10:795700. <https://doi.org/10.3389/feart.2022.795700>.
15. Branca S, Carlo PD. Types of eruptions of Etna volcano AD 1670–2003: implications for short-term eruptive behaviour. *Bull Volcanol*. 2005;67(8):732–42. <https://doi.org/10.1007/s00445-005-0412-z>.
16. Cappello A, Bilotta G, Zuccarello F, Proietti C, Ganci G. An improved methodology for lava flow hazard mapping at Etna volcano. *Ann Geophys*. 2025. <https://doi.org/10.4401/ag-9157>.
17. National Institute of Geophysics and Vulcanology (INGV). Bollettini settimanali multidisciplinari. [Online]. Available from: <https://www.ct.ingv.it/index.php/monitoraggio-e-sorveglianza/prodotti-del-monitoraggio/bollettini-settimanali-multidisciplinari?filter%5Bsearch%5D=Etna>
18. Allard P, Behncke B, D'Amico S, Neri M, Gambino S. Mount Etna 1993–2005: anatomy of an evolving eruptive cycle. *Earth-Sci Rev*. 2006;78(1):85–114. <https://doi.org/10.1016/j.earscirev.2006.04.002>.
19. Corsaro RA, Miraglia L. Dynamics of 2004–2005 Mt. Etna effusive eruption as inferred from petrologic monitoring. *Geophys Res Lett*. 2005. <https://doi.org/10.1029/2005GL022347>.
20. Palano M, Viccaro M, Zuccarello F, Gresta S. Magma transport and storage at Mt. Etna (Italy): a review of geodetic and petrological data for the 2002–03, 2004 and 2006 eruptions. *J Volcanol Geotherm Res*. 2017;347:149–64. <https://doi.org/10.1016/j.jvolgeores.2017.09.009>.
21. Norini G, De Beni E, Andronico D, Polacci M, Burton M, Zucca F. The 16 November 2006 flank collapse of the south-east crater at Mount Etna, Italy: study of the deposit and hazard assessment. *J Geophys Res Solid Earth*. 2009. <https://doi.org/10.1029/2008JB005779>.
22. Andronico D, Scollo S, Lo Castro MD, Cristaldi A, Lodato L, Taddeucci J. Eruption dynamics and tephra dispersal from the 24 November 2006 paroxysm at South-East Crater, Mt Etna, Italy. *J Volcanol Geotherm Res*. 2014;274:78–91. <https://doi.org/10.1016/j.jvolgeores.2014.01.009>.
23. Andronico D, Cristaldi A, Scollo S. The 4–5 september 2007 lava fountain at South-East Crater of Mt Etna, Italy. *J Volcanol Geotherm Res*. 2008;173(3):325–8. <https://doi.org/10.1016/j.jvolgeores.2008.02.004>.
24. Bonaccorso A, Bonforte A, Calvari S, Del Negro C, Di Grazia G, Ganci G, et al. The initial phases of the 2008–2009 Mount Etna eruption: a multidisciplinary approach for hazard assessment. *J Geophys Res Solid Earth*. 2011. <https://doi.org/10.1029/2010JB007906>.
25. Ferretti A, Prati C, Rocca F. Permanent scatterers in SAR interferometry. *IEEE Trans Geosci Remote Sens*. 2001;39(1):8–20. <https://doi.org/10.1109/36.898661>.
26. Ganci G, Bilotta G, Cappello A, Hérault A, Del Negro C. HOTSAT: a multiplatform system for the thermal monitoring of volcanic activity using satellite data. *Geol Soc Lond Spec Publ*. 2016;426(1):207–21. <https://doi.org/10.1144/SP426.21>.
27. Corradini S, Guerrieri L, Stelitano D, Salerno G, Scollo S, Merucci L, et al. Near real-time monitoring of the Christmas 2018 Etna eruption using SEVIRI and products validation. *Remote Sens*. 2020;12:1336. <https://doi.org/10.3390/rs12081336>.
28. Girona T, Realmuto V, Lundgren P. Large-scale thermal unrest of volcanoes for years prior to eruption. *Nat Geosci*. 2021;14(4):238–41. <https://doi.org/10.1038/s41561-021-00705-4>.
29. Levelt PF, van den Oord GHJ, Dobber MR, Malkki A, Visser H, de Vries J, et al. The ozone monitoring instrument. *IEEE Trans Geosci Remote Sens*. 2006;44(5):1093–101. <https://doi.org/10.1109/TGRS.2006.872333>.
30. Stoiber R, Malinconico LL, Williams SN. Use of the correlation spectrometer at volcanoes. *Forecast. Volcan. Events*. 1983;425–444.
31. Williams-Jones G, Stix J, Hickson C. The COSPEC cookbook: making SO₂ measurements at active volcanoes. Geneva: IAVCEI; 2008. <https://doi.org/10.13140/RG.2.2.13728.99845>.
32. Merucci L, Burton M, Corradini S, Salerno G. Reconstruction of SO₂ flux emission chronology from space-based measurements. *J Volcanol Geotherm Res*. 2011;206:80–7. <https://doi.org/10.1016/j.jvolgeores.2011.07.002>.
33. Krueger A, Schnetzler C, Walter L. The December 1981 eruption of Nyamuragira Volcano (Zaire), and the origin of the “mystery cloud” of early 1982. *J Geophys Res*. 1996;101:15191–6. <https://doi.org/10.1029/96JD00221>.
34. Chen J, Cazenave A, Dahle C, Llovel W, Panet I, Pfeffer J, et al. Applications and challenges of GRACE and GRACE follow-on satellite gravimetry. *Surv Geophys*. 2022;43(1):305–45. <https://doi.org/10.1007/s10712-021-09685-x>.

35. Centre National d'Études Spatiales/Groupe de Recherche de Géodésie Spatiale (CNES/GRGS). CNES/GRGS RL05. [Online]. Available from: <https://grace.obs-mip.fr/>
36. Elrahman SMA, Abraham A. A Review of Class Imbalance Problem. 2013.
37. Bjarke Skogstad Larsen. Synthetic Minority Over-sampling Technique (SMOTE). (2025). MATLAB. Accessed: May 06, 2025. [Online]. Available from: https://github.com/dkbsl/matlab_smote/releases/tag/1.0
38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002;16:321–57. <https://doi.org/10.1613/jair.953>.
39. Elreedy D, Atiya AF. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Inf Sci.* 2019;505:32–64. <https://doi.org/10.1016/j.ins.2019.07.070>.
40. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Cost-sensitive learning. In: Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F, editors. *Learning from imbalanced data sets*. Cham: Springer International Publishing; 2018. p. 63–78. https://doi.org/10.1007/978-3-319-98074-4_4.
41. Refaeilzadeh P, Tang L, Liu H. Cross-validation. *Encycl Database Syst.* 2009;532–538:532–8. https://doi.org/10.1007/978-0-387-39940-9_565.
42. Barber D. *Bayesian reasoning and machine learning*. Cambridge: Cambridge University Press; 2012. <https://doi.org/10.1017/CBO9780511804779>.
43. Pignatelli A, Piochi M. Machine learning applied to rock geochemistry for predictive outcomes: the Neapolitan volcanic history case. *J Volcanol Geotherm Res.* 2021. <https://doi.org/10.1016/j.jvolgeores.2021.107254>.
44. Lantz B. *Machine learning with R*. Birmingham: Packt publishing Ltd; 2013.
45. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. *Inf Fusion.* 2022;81:84–90. <https://doi.org/10.1016/j.inffus.2021.11.011>.
46. Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G. Deep neural networks and tabular data: a survey. *IEEE Trans Neural Netw Learn Syst.* 2024;35(6):7499–519. <https://doi.org/10.1109/TNNLS.2022.3229161>.
47. Zhu W, Beroza GC. PhaseNet: a deep-neural-network-based seismic arrival-time picking method. *Geophys J Int.* 2019;216(1):261–73. <https://doi.org/10.1093/gji/ggy423>.
48. Massey FJ. The kolmogorov-smirnov test for goodness of fit. *J Am Stat Assoc.* 1951;46(253):68–78. <https://doi.org/10.2307/2280095>.
49. Anderson TW, Darling DA. A test of goodness of fit. *J Am Stat Assoc.* 1954;49(268):765–9. <https://doi.org/10.1080/01621459.1954.10501232>.
50. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika.* 1965;52(3/4):591–611. <https://doi.org/10.2307/2333709>.
51. Mohd Razali N, Yap B. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *J Stat Model Anal.* 2011;2:21–33.
52. Bartlett MS. Properties of sufficiency and statistical tests. *Proc R Soc Lond Ser Math Phys Sci.* 1937;160(901):268–82.
53. Levene H. Robust tests for equality of variances. *Contrib Probab Stat Essays Honor Harold Hotell.* 1960;2:278–92.
54. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc.* 1952;47(260):583–621. <https://doi.org/10.1080/01621459.1952.10483441>.
55. Critchlow D, Fligner M. On distribution-free multiple comparison in the one-way analysis of variance. *Commun Stat - Theory Methods.* 1991;20:127–39. <https://doi.org/10.1080/03610929108830487>.
56. Vrigazova B. The proportion for splitting data into training and test set for the bootstrap in classification problems. *Bus Syst Res J.* 2021;12(1):228–42. <https://doi.org/10.2478/bsrj-2021-0015>.
57. Tan J, Yang J, Wu S, Chen G, Zhao J. A critical look at the current train/test split in machine learning. *arXiv.* 2021. <https://doi.org/10.48550/arXiv.2106.04525>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.