Journal of Information Technology and Cyber Security 4(1) January 2026: 26-40 EISSN 2987-386X | PISSN 2987-3878
DOI:10.30996/jitcs.133023

Classification of Volcanic Status Events Using Autocorrelation and Support Vector Machine Methods

Fridy Mandita [1,*], Muhammad Arif Fajriyansah [2]

[1] Department of Robotics and Artificial Intelligence, Universitas 17 Agustus 1945 Surabaya, Indonesia
[2] Department of Informatics Engineering, Universitas 17 Agustus 1945 Surabaya, Indonesia
* Corresponding author: fridymandita@gmail.com

Received: 06 January 2026
Accepted: 28 January 2026
Revised: 28 January 2026
Available online: 02 February 2026

To cite this article: Mandita, F., & Fajriyansah, M. A. (2026). Classification of Volcanic Status Events Using Autocorrelation and Support Vector Machine Methods. *Journal of Information Technology and Cyber Security, 4*(1), 26-40. https://doi.org/10.30996/jitcs.133023

Abstract
Volcanic eruption disasters occur frequently in Indonesia due to the high density of active volcanoes, posing persistent risks to surrounding communities and infrastructure. Effective mitigation of these hazards is challenged by limitations in monitoring systems, particularly related to instrumentation coverage and the availability of expert human resources. One critical aspect of volcanic monitoring is the accurate classification of seismic activity, which reflects subsurface volcanic processes and supports timely hazard assessment. This study addresses the challenge of reliably classifying volcanic seismic events by proposing an integrated framework that combines autocorrelation-based signal characterization with Support Vector Machine (SVM)–based multi-class classification, supported by Z-score normalization during data preprocessing. The framework is designed to enhance feature consistency and robustness against noise commonly present in volcanic seismic signals. To evaluate its effectiveness, three SVM kernel functions—linear, polynomial, and radial basis function (RBF)—are systematically assessed under identical experimental conditions. The results demonstrate that the polynomial SVM kernel with a degree of two provides the most reliable classification performance, achieving an accuracy of 0.9605. In addition, the application of Z-score normalization substantially improves model stability and overall performance across all kernel configurations, indicating that feature scaling plays a critical role in SVM-based seismic classification. Performance variations among kernels suggest that non-linear feature representations are better suited to capture the complex characteristics of volcanic seismic signals, while classification errors are primarily influenced by class imbalance in underrepresented event types. These findings indicate that the proposed framework effectively supports automated volcanic seismic signal analysis and has the potential to enhance the reliability of seismic-based volcanic activity monitoring.
Keywords: autocorrelation, seismic signal classification, support vector machine, volcano.

1. Introduction
Volcanic eruptions are natural hazards that frequently occur in Indonesia due to the large number of active volcanoes distributed across the region. These events pose significant risks to surrounding communities and infrastructure, making continuous monitoring and early identification of volcanic activity essential (Tempola, Muhammad, & Khairan, 2018; Marzocchi, Selva, & Jordan, 2021). Limitations in monitoring systems, particularly in terms of instrumentation availability and human resources, further increase the risk of delayed hazard assessment and mitigation (Ririh, Laili, Wicaksono, & Tsurayya, 2020; Thelen, Matoza, & Hotovec-Ellis, 2022). Seismic signals represent one of the most important parameters in volcano monitoring, as they reflect subsurface processes such as magma movement, rock fracturing, and material collapse. Different types of volcanic seismic events exhibit distinct temporal and spectral characteristics; however, waveform similarity, low signal-to-noise ratio (SNR), and complex environmental noise often complicate manual identification (McNutt, 2025; Chouet & Matoza, 2013). These challenges motivate the adoption of automated approaches for seismic signal analysis.

© 2026 The Author(s). Published by Department of Information Systems and Technology, Universitas 17 Agustus 1945 Surabaya, Indonesia. This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial, and no modifications or adaptations are made.

Recent advances in artificial intelligence, particularly machine learning (ML), have enabled the development of data-driven methods capable of learning complex patterns from high-dimensional data (Alzubi, Nayyar, & Kumar, 2018; Bergen, Johnson, Hoop, & Beroza, 2019; Mousavi, Ellsworth, Zhu, Chuang, & Beroza, 2020). ML techniques have been widely applied to volcanic seismic event classification and have demonstrated promising performance in improving accuracy and efficiency (Anggian, Hidayat, & Furqon, 2020; Tempola, Muhammad, & Khairan, 2018; Ross, Meier, Hauksson, & Heaton, 2018). Among various ML algorithms, Support Vector Machine (SVM) has shown strong performance in handling non-linear data distributions and constructing optimal decision boundaries in high-dimensional feature spaces (Handayanto, Latifa, Saputro, & Waliyansyah, 2019; Rahutomo, Saputra, & Fidyawan, 2018). In the context of seismology, SVM has been successfully applied to classify seismic events and distinguish different types of volcanic signals, particularly when training data are limited (Tang, Zhang, & Wen, 2020; Manley, et al., 2022). Furthermore, recent studies have shown that machine learning approaches remain effective under noisy conditions when combined with appropriate preprocessing and feature extraction strategies (Meier, et al., 2019).

Data preprocessing plays a crucial role in improving classification performance. Autocorrelation remains a reliable technique for seismic event detection due to its ability to emphasize coherent and repetitive signal patterns (Perdana, Fatichah, & Purwitasari, 2015; Titos, Bueno, García, Benítez, & Ibañez, 2019; Gibbons & Ringdal, 2006). In addition, data normalization techniques such as Z-score normalization are widely adopted to reduce feature scale disparities, improve model convergence, and enhance classification robustness (Ambarwari, Adrian, & Herdiyeni, 2020; Karo & Hendriyana, 2022; Singh & Singh, 2020). Based on these considerations, this study proposes a volcanic seismic activity classification framework that integrates autocorrelation-based event detection, Z-score normalization for data preprocessing, and SVM for multi-class classification. The proposed approach aims to improve classification accuracy and robustness against noise, thereby supporting more reliable volcanic activity monitoring.

2. Literature Review

2.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm originally developed by Vapnik and colleagues, grounded in the principle of Structural Risk Minimization (SRM). This principle aims to enhance generalization performance by balancing empirical error and model complexity. The fundamental objective of SVM is to construct an optimal separating hyperplane that maximizes the margin between classes within the feature space (Cortes & Vapnik, 1995; Vapnik, 1998; Handayanto, Latifa, Saputro, & Waliyansyah, 2019). SVM supports both linear and non-linear classification. For linearly separable data, a linear hyperplane is sufficient, whereas non-linearly separable data are handled through kernel-based transformations. Kernel functions implicitly map the original input space into a higher-dimensional feature space, enabling linear separation in the transformed domain. Commonly used kernels include linear, polynomial, and radial basis function (RBF) kernels (Bishop, 2006; Schölkopf & Smola, 2001).

a) Architecture and Workflow

From an architectural standpoint, Support Vector Machine (SVM) consists of an input feature space, a set of support vectors, an optimal hyperplane, and a margin that separates classes. Support vectors correspond to training samples located closest to the decision boundary and play a decisive role in determining the hyperplane position. The margin, defined as the distance between the hyperplane and the nearest support vectors, is maximized to improve classification robustness and generalization capability (Vapnik, 1998; Cortes & Vapnik, 1995). The SVM workflow begins with feature vector input, followed by kernel-based mapping when required to handle non-linearly separable data. The optimal hyperplane is then obtained by solving a constrained optimization problem that balances margin maximization and classification error. Classification of unseen data is performed by evaluating the sign of the resulting decision function (Bishop, 2006; Schölkopf & Smola, 2001). The overall architecture and workflow of the SVM classifier are illustrated in Fig. 1.

b) Mathematical Formulation

For linear SVM, the decision function is expressed in Eq. 1:

$$f(x) = w \cdot x + b \tag{1}$$

where w represents the weight vector and b denotes the bias term. This formulation aims to determine an optimal separating hyperplane that maximizes the margin between classes in the feature space (Cortes & Vapnik, 1995; Vapnik, 1998).

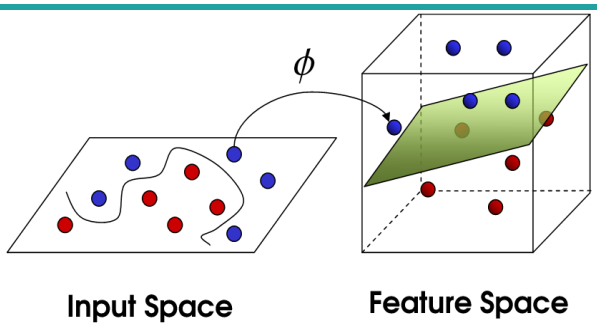


Fig. 1. Architecture and workflow of the Support Vector Machine.

For non-linear SVM, the decision function is defined in Eq. 2:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \quad (2)$$

where α_i are the Lagrange multipliers, y_i are the class labels, x_i denote the support vectors, and $K(\cdot)$ represents the kernel function that implicitly maps the input data into a higher-dimensional feature space (Schölkopf & Smola, 2001).

In this study, three kernel functions: linear, polynomial, and radial basis function (RBF)—are employed to evaluate different levels of feature space complexity in volcanic seismic signal classification. The linear kernel is defined in Eq. 3:

$$K(x_i, x_j) = x_i \cdot x_j \quad (3)$$

The linear kernel assumes linear separability in the original feature space and serves as a baseline model for assessing the discriminative capability of the extracted seismic features (Bishop, 2006). The polynomial kernel is expressed in Eq. 4:

$$K(x_i, x_j) = (\gamma x_i \cdot x_j + r)^d \quad (4)$$

where γ is a scaling parameter, r is a constant term, and d denotes the polynomial degree. The polynomial kernel enables the modeling of non-linear feature interactions, which are commonly observed in seismic signals due to complex subsurface volcanic processes (Schölkopf & Smola, 2001; Bishop, 2006). The RBF kernel is defined in Eq. 5:

$$K(x_i, x_j) = \exp\left(-\gamma \parallel x_i - x_j \parallel^2\right) \quad (5)$$

The RBF kernel is frequently adopted due to its flexibility in modeling complex non-linear relationships and its effectiveness in handling noisy and high-dimensional data (Bishop, 2006; Tang, Zhang, & Wen, 2020). This property makes the RBF kernel particularly suitable for seismic signal classification, where overlapping waveform characteristics and environmental noise are commonly encountered (Malfante, et al., 2018).

The evaluation of these three kernel functions allows a systematic assessment of model robustness and generalizability across different feature space representations, which is essential for multi-class volcanic seismic signal classification.

2.2. Autocorrelation

Autocorrelation is a signal processing technique that quantifies the similarity between a signal and a time-shifted version of itself. By evaluating the degree of self-similarity across different time lags, autocorrelation reveals repeating patterns, periodic structures, and coherent energy that may not be clearly observable in the time domain. In seismic analysis, autocorrelation has been widely adopted for event detection and signal characterization, particularly in environments dominated by noise and low signal-to-noise ratios (Gibbons & Ringdal, 2006).

a) Mathematical Formulation

For a discrete seismic signal $x[n]$ of length N, the autocorrelation function $R_{xx}(k)$ is defined in Eq. 6:

$$R_{xx}(k) = \sum_{n=0}^{N-k-1} x[n]x[n+k] \quad (6)$$

where k denotes the time lag and $x[n+k]$ represents the signal shifted by k samples. This formulation measures the similarity between the original signal and its delayed version.

To facilitate comparison across signals with different amplitudes, a normalized autocorrelation function is commonly employed as defined Eq. 7:

$$\rho_{xx}(k) = \frac{R_{xx(k)}}{R_{xx(0)}} \quad (7)$$

where $R_{xx}(0)$ corresponds to the signal energy at zero lag. Normalization constrains the autocorrelation values to the interval [-1,1], enabling consistent interpretation of correlation strength.

b) Interpretation and Application in Seismic Analysis
In seismic applications, significant peaks in the autocorrelation function indicate the presence of coherent or repeating waveform structures. Such behavior is often associated with repeating earthquakes, volcanic tremor, or resonance processes within volcanic systems. In contrast, incoherent background noise typically produces low and irregular autocorrelation values across time lags. Previous studies have demonstrated that autocorrelation-based methods are effective in enhancing the detection of low-magnitude seismic events that may be missed by conventional amplitude-based techniques (Gibbons & Ringdal, 2006). Furthermore, autocorrelation has been successfully integrated with feature extraction and pattern recognition approaches to improve the interpretability and robustness of seismic signal analysis (Perdana, Fatichah, & Purwitasari, 2015).

c) Application Context in This Study
In this study, autocorrelation is utilized as a signal characterization approach to emphasize intrinsic temporal structures within seismic waveforms. By highlighting coherent energy patterns prior to classification, autocorrelation contributes to improved feature discrimination and supports subsequent machine learning–based analysis.

2.3. Related Work

Numerous studies have explored the use of machine learning techniques for seismic signal classification in both tectonic and volcanic contexts. Tang, Zhang, & Wen (2020) applied Support Vector Machine (SVM) to classify seismic events in the Tianshan orogenic belt using spectral features derived from P-wave and S-wave characteristics, achieving high classification accuracy. Their findings highlight the effectiveness of SVM in handling complex seismic datasets. Lara-Cueva, Benítez, Paillacho, Villalva, & Rojo-Álvarez (2018) investigated multi-class SVM for classifying volcanic seismic signals recorded at Cotopaxi volcano, with a focus on long-period (LP) and volcano-tectonic (VT) events. By integrating an initial detection stage with SVM-based classification, their study demonstrated promising performance while also revealing challenges related to waveform similarity across event types.

Beyond SVM-based approaches, Malfante, et al. (2018) conducted a comprehensive study on machine learning applications for volcanic seismic signals and emphasized that model performance is strongly influenced by preprocessing strategies and feature representation. Their work underscores the importance of addressing noise complexity and non-stationary signal characteristics. Autocorrelation-based detection methods have also been widely reported in seismic research. Gibbons & Ringdal (2006) demonstrated that autocorrelation significantly improves the detection of low-magnitude seismic events that are difficult to identify using conventional techniques. Similar approaches have been adopted in various seismic monitoring studies to enhance detection sensitivity under low signal-to-noise ratio conditions.

Despite these advances, most previous studies focus primarily on direct classification of seismic signals without explicitly integrating time-series based event detection and systematic data normalization. In addition, the impact of feature normalization, particularly Z-score normalization, on SVM performance in multi-class volcanic seismic classification remains insufficiently explored. These limitations motivate the development of an integrated framework that combines autocorrelation-based event detection, data normalization, and comprehensive kernel evaluation to improve the robustness and reliability of volcanic seismic signal classification.

3. Methods

In general, this study consists of three main stages: data preprocessing, data modeling, and model evaluation. An overview of the research stages is presented in Fig. 2. Fig. 2 illustrates the overall research workflow adopted in this study, beginning with problem identification to define the scope and objectives of the investigation. This initial stage focuses on recognizing key challenges in volcanic seismic data analysis and formulating research questions to be addressed, ensuring that subsequent methodological steps remain aligned with the study aims. Following this stage, data collection is conducted by acquiring volcanic seismic signal recordings relevant to the case study. These recordings constitute the primary dataset and represent various types of volcanic seismic activity, providing a solid foundation for further processing and analysis. The workflow then proceeds to data preprocessing, which is essential for ensuring data quality and suitability for modeling. During this phase, raw seismic signals are subjected to event detection and normalization procedures to reduce noise effects, standardize feature scales, and preserve important signal

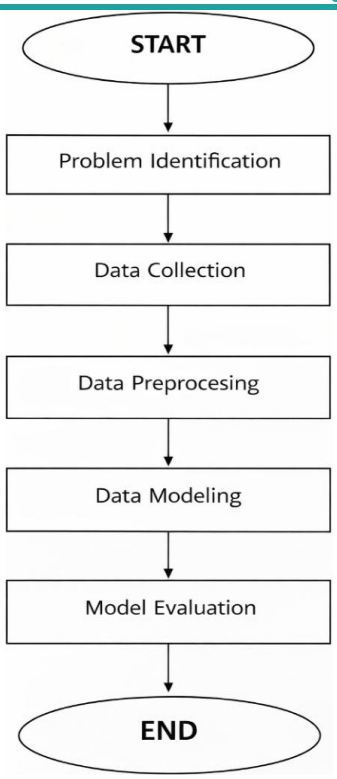


Fig. 2. Research workflow.

characteristics required for reliable analysis. After preprocessing, data modeling is performed as a single, clearly defined stage. In this phase, machine learning techniques are applied to develop a classification model capable of learning discriminative patterns from the processed seismic data and distinguishing among different seismic event classes. Finally, the developed model is evaluated to assess overall performance and reliability of the results. Although the workflow stages are presented using general labels, each stage reflects specific processes described in the Methods section, including autocorrelation-based event detection, data normalization, and machine learning–based classification using Support Vector Machine.

3.1. Dataset Collection

The data used in this study consist of volcanic seismic signal recordings collected from monitoring stations surrounding Mount Merapi, Indonesia, during the period 2019 - 2021. The seismic signals were recorded within a frequency range of 0.5 - 50 Hz, which is appropriate for capturing a wide range of volcanic seismic activities. The dataset comprises 5,000 seismic signal samples, which were labeled by seismic analysts according to established volcanic seismic classifications.

The seismic signals were categorized into eight classes: AP (we use a localized term *Awan Panas*), associated with pyroclastic density currents or hot cloud events; DG (Deep Volcanic Earthquake), representing seismic activity originating from deeper volcanic structures related to magma movement; Low Frequency events, characterized by dominant low-frequency components and commonly linked to fluid or gas movement within the volcanic system; Tremor, referring to continuous or semi-continuous seismic vibrations typically associated with sustained magma or gas flow; Multiple-phase events, which contain more than one identifiable seismic phase within a single signal; Rockfall events, generated by surface material collapse or gravitational mass movement; VT-A, representing shallow volcano-tectonic earthquakes caused by brittle rock failure; and VT-B, corresponding to deeper volcano-tectonic earthquakes associated with stress changes within the volcanic structure.

For the classification process, the Support Vector Machine (SVM) model is trained using feature representations extracted from seismic signals after event detection and normalization. These features capture essential temporal and frequency-related characteristics of the waveforms, enabling effective discrimination among different seismic event classes. Autocorrelation is applied directly to the raw seismic signals prior to normalization to preserve their intrinsic temporal structure, while time–frequency analysis is used to retain both temporal patterns and spectral content. Details of the preprocessing and feature extraction procedures are provided in the Methods section.

Table 1
Example of the dataset.

No	X1	X2	X3	Y
1	3.734	1.437	6.262	1
2	8.820	−1.354	1.196	3
3	−4.266	1.122	−1.818	4
4	2.205	−9.590	8.223	3
5	1.296	−1.614	5.618	7

Table 2
Example of Z-score normalized data.

No	X1	X2	X3	Y
1	−0.224	−0.190	−0.166	1
2	1.001	−0.299	−0.262	3
3	−0.199	1.122	−0.170	4
4	0.114	−0.179	0.901	3
5	−0.302	0.561	−0.236	7

To evaluate the proposed classification model, the dataset is divided into training and testing subsets using a 90:10 ratio, ensuring sufficient data for model learning while retaining an independent test set for objective performance assessment in a multi-class setting. The training set consists of labelled seismic signals used for model learning, whereas the testing set contains unseen samples for evaluating model generalization. An example of the dataset structure is presented in Table 1.

The training dataset comprises 5,000 seismic signal samples and is arranged to maintain a relatively balanced distribution across the eight seismic event classes, thereby reducing potential class bias and supporting fair performance evaluation. Prior to classification, seismic event detection is performed using autocorrelation-based analysis with a normalized threshold of 0.5, which effectively distinguishes coherent seismic events from background noise.

During model optimization, kernel-specific parameter ranges are selected according to the characteristics of each SVM kernel. For the polynomial kernel, higher values of the regularization parameter C are evaluated to avoid underfitting after Z-score normalization and to capture non-linear feature relationships. In contrast, the RBF kernel is tested starting from moderate C values, as very small values tend to over-smooth the decision boundary and reduce class separability. These parameter selections follow established SVM tuning practices and enable a fair and meaningful performance comparison.

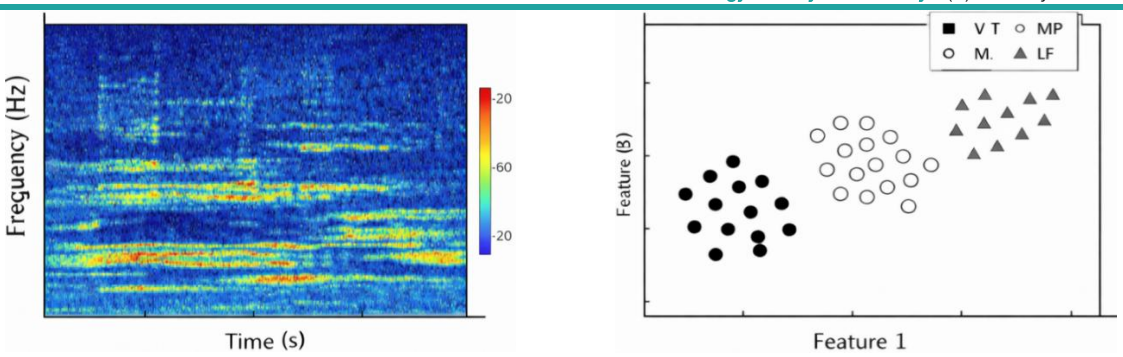
The dataset consists of feature vectors X_n that represent seismic signal attributes extracted from waveform data, including amplitude, frequency, and temporal characteristics. These features are used as input variables for the machine learning model. The target variable Y denotes the seismic event class, where 0 corresponds to AP, 1 to DG, 2 to Low Frequency events, 3 to Tremor, 4 to Multiple-phase events, 5 to Rockfall, 6 to VT-A, and 7 to VT-B.

The seismic signals are modeled using feature vectors that encode key waveform characteristics. Feature X_1 represents amplitude-based attributes associated with signal energy, X_2 describes frequency-related properties derived from spectral analysis, and X_3 captures temporal or morphological characteristics of the waveform. These features collectively form a compact representation of seismic signal attributes that facilitates the classification of seismic event types, as indicated by the output variable Y.

3.2. Data Preprocessing

Data preprocessing is a crucial step in optimizing the SVM model to improve classification accuracy. The seismic signals used in this study contain a significant number of low-frequency components; therefore, signal features and events must be analyzed prior to modelling. The preprocessing steps applied in this study are described as follows.

a) Autocorrelation
   In this stage, seismic signals are processed for event detection by transforming the raw waveforms into autocorrelation representations using the np.correlate() function from the Python NumPy library. Event detection is performed by comparing the maximum amplitude of the autocorrelation output with a predefined threshold value of 0.5. Signals whose autocorrelation amplitudes exceed this threshold are classified as containing seismic events. The detected signals are subsequently converted into spectrogram representations and then assigned corresponding labels for further analysis.
b) Data Normalization



(a) (b)

Fig. 3. (a) Before and (b) after data modeling results.

Before model training, data normalization is performed using the Z-score method. Zero-mean normalization is applied by using the mean and standard deviation of the data, resulting in standardized features with a mean value of 0 and a standard deviation of 1 (Han, Kamber, & Pei, 2012; Singh & Singh, 2020). Z-score normalization is employed to reduce the effect of unbalanced value ranges across features. An example of Z-score normalized data is shown in Table 2.

3.3. Data Modelling

After completing the data preprocessing stage, a Support Vector Machine (SVM) model is constructed to classify seismic signals into multiple classes. The preprocessed seismic features serve as the input data before modeling, where feature consistency is ensured through autocorrelation-based characterization and Z-score normalization. During the modeling process, the SVM maps the input features into an appropriate feature space and learns decision boundaries to achieve class separation.

Three kernel functions: linear, polynomial, and radial basis function (RBF) are evaluated to assess their capability in modeling the data. For the linear kernel, different values of the regularization parameter *C* are tested to balance margin maximization and classification error. The polynomial kernel is examined by varying both the polynomial degree and *C* to capture non-linear patterns, while the RBF kernel is analyzed using different combinations of *C* and *γ* to model complex decision boundaries. The resulting feature space reflects the data representation after modeling, where class separation is achieved based on the learned decision functions.

The transformation of data representations before and after modeling is illustrated in Fig. 3. Prior to modeling, the seismic signals are transformed into representative features through autocorrelation-based characterization and normalized using the Z-score method. After modeling, the Support Vector Machine (SVM) maps the input features into a discriminative feature space and learns decision boundaries to separate different types of volcanic seismic events. This illustration highlights the transformation of volcanic seismic data during the modeling process.

3.4. Model Evaluation

Model evaluation in this study is conducted using a confusion matrix to assess the performance of the classification model. The confusion matrix compares predicted class labels with the true class labels, providing information on both correctly classified and misclassified samples for each class. This representation allows the evaluation of overall classification accuracy while also highlighting the distribution of classification errors across different seismic event categories. Furthermore, the confusion matrix enables the calculation of class-specific performance metrics, including precision, recall, and F1-score, which are particularly important for multi-class seismic signal classification with potential class imbalance.

Precision is defined as the proportion of correctly predicted positive samples among all predicted positive samples and is calculated as in Eq. 8:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall, as defined in Eq. 9, represents the proportion of correctly predicted positive samples among all actual positive samples:

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

The F1-score, as formulated in Eq. 10, provides a balanced measure between precision and recall and is computed as the harmonic mean of these two metrics:

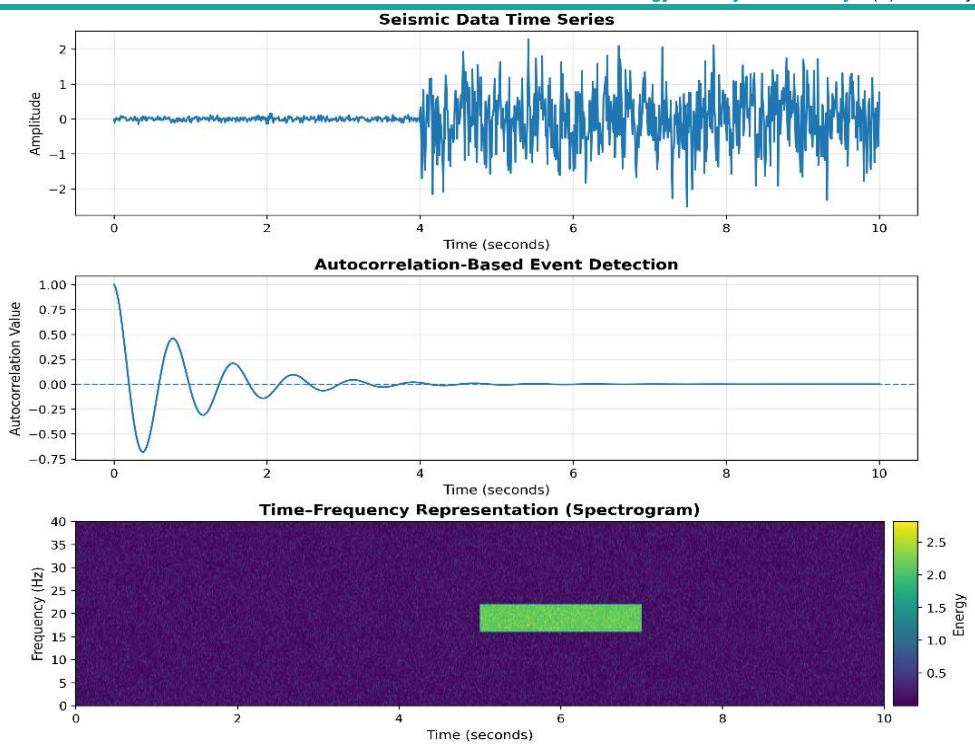


Fig. 4. Autocorrelation-based event detection results.

$$F1 - score = \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

where TP denotes true positives, FP false positives, and FN false negatives. By examining these metrics, the strengths and limitations of the model in distinguishing between different types of seismic events can be more clearly identified, thereby providing a comprehensive assessment of model performance (Bishop, 2006; Han, Kamber, & Pei, 2012). This study adopts a comparative machine learning performance evaluation approach, focusing on empirical differences among SVM kernel configurations under identical experimental settings. Therefore, inferential statistical analyses such as p-values, confidence intervals, or effect size measurements are not applied, as the objective is model performance comparison rather than hypothesis testing (Schölkopf & Smola, 2001; Vapnik, 1998).

4. Results and Discussion

In this study, the optimal parameters for each SVM kernel are evaluated. The best-performing kernel configuration is then selected for model evaluation. In addition, this study compares the performance of the SVM model using Z-score normalized data and non-normalized data. The results of the experiments conducted in this study are presented as follows.

4.1. Autocorrelation Results

At this stage, seismic event analysis is conducted using autocorrelation applied to the dataset, and the results are visualized through spectrograms. Fig. 4 presents the seismic signal analysis using time-domain, autocorrelation, and time–frequency representations. The time-series signal shows a noticeable increase in amplitude after approximately 4 seconds, indicating the onset of a seismic event. The autocorrelation function displays oscillatory patterns that gradually decay over time, reflecting the temporal structure of the signal as well as the growing influence of background noise. In the time–frequency domain, the spectrogram highlights a localized high-energy region, represented by yellow–green colors, within the mid-frequency range of approximately 15–20 Hz between 5 and 7 seconds, while darker purple regions indicate lower energy levels. These representations together support the identification and characterization of the observed seismic event.

4.2. SVM Kernel Results

In this experiment, three different SVM kernel functions: linear, radial basis function (RBF), and polynomial are evaluated to determine the best classification accuracy. The results of the SVM kernel evaluations are described below.

4.2.1 Linear SVM results

Table 3
Linear SVM kernel parameter testing.

No	Cost Parameter (C)	Testing Accuracy
1.	0.001	0.8817
2.	0.01	0.9014
3.	0.1	0.9261
4.	1	0.9310
5.	10	0.9261
6.	100	0.9261

Table 4
Polynomial kernel parameter evaluation.

No	Cost Parameter (C)	Accuracy		
		Degree=1	Degree=2	Degree=3
1.	100	0.9162	0.9605	-
2.	200	0.9261	0.9605	-
3.	300	0.9261	0.9605	-
4.	400	0.9162	0.9605	-
5.	500	0.9162	0.9605	-

During the optimization of the linear SVM model, the regularization parameter C was systematically evaluated to balance the trade-off between model complexity and generalization performance. The tested values of C (0.001, 0.01, 0.1, 1, 10, and 100) were selected to span a wide range on a logarithmic scale, which is commonly adopted in SVM parameter tuning to examine different regularization strengths. Smaller values of C impose stronger regularization, which may lead to underfitting, whereas larger values allow the model to fit the training data more closely and may increase the risk of overfitting. This range is particularly relevant for the seismic signal classification task addressed in this study, as the extracted feature vectors exhibit variability and may contain noise, requiring an appropriate balance between smooth and flexible decision boundaries.

As shown in Table 3, the testing accuracy increases as the value of C rises from 0.001 to 1, indicating improved model flexibility and more effective decision boundary formation. Although higher values of C (10 and 100) also produce relatively high accuracy, no further improvement is observed, suggesting that increasing model complexity beyond this point does not enhance generalization performance. Therefore, since this study emphasizes testing accuracy as the primary performance criterion, $C = 1$ is selected as the optimal parameter for the linear SVM model.

4.2.2 Polynomial SVM results

The polynomial kernel is a non-linear kernel that is particularly suitable when the training dataset has undergone normalization. It should be noted that a polynomial kernel with degree $d = 1$ is mathematically equivalent to a linear kernel. Therefore, in this experiment, the evaluation of the polynomial kernel focuses on higher-degree configurations in order to better capture non-linear patterns in the seismic data. Optimization is performed on the regularization parameter C and the polynomial degree d. The values of C tested are 100, 200, 300, 400, and 500, while the polynomial degrees evaluated are $d = 2$ and $d = 3$. The results of the polynomial kernel evaluation are presented in Table 4.

The results in Table 4 demonstrate that the polynomial kernel with degree $d = 2$ consistently yields the highest classification accuracy, achieving an average accuracy of 0.9605 across all tested values of the regularization parameter C. Compared to a polynomial degree of $d = 1$, which is equivalent to a linear model, the $d = 2$ configuration provides superior performance, indicating its effectiveness in capturing non-linear patterns in the seismic data. The identical accuracy values obtained for different C settings further indicate that the model performance is relatively insensitive to the choice of C within the tested range. Increasing the polynomial degree to $d = 3$ does not result in additional performance gains and may introduce unnecessary model complexity. Therefore, the polynomial kernel with degree $d = 2$ is selected as the optimal configuration for this study.

For the linear SVM model, comparable accuracy values are observed for $C = 0.1$, 10, and 100. However, the regularization parameter $C = 1$ achieves the highest testing accuracy of 0.931, suggesting an optimal balance between bias and variance. Increasing C beyond this value does not lead to further performance improvement, indicating that excessive relaxation of regularization does not significantly enhance generalization. Consequently, $C = 1$ is selected as the optimal regularization parameter for the linear SVM model.

4.2.3 RBF results

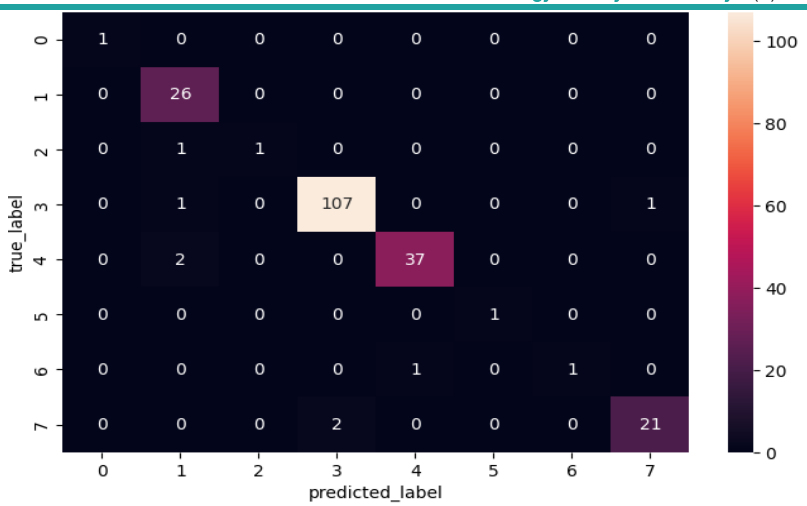


Fig. 5. Confusion matrix of the best-performing model.

Table 5
RBF kernel parameter evaluation.

No	Cost Parameter (C)	Accuracy				
		Gamma				
		1	2	3	4	5
1.	1	0.802	0.817	0.822	0.832	0.827
2.	10	0.857	0.852	0.857	0.857	0.866
3.	50	0.866	0.857	0.866	0.866	0.866
4.	100	0.862	0.857	0.866	0.866	0.866

Table 6
Model accuracy without normalization.

No	Kernal	Accuracy
1.	Linear	0.8325
2.	Polynomial	0.5369
3.	RBF	0.6551

Table 7
Model accuracy with z-score normalization.

No	Kernal	Accuracy
1.	Linear	0.9310
2.	Polynomial	0.9605
3.	RBF	0.8669

The third approach evaluated in this study employs SVM with a radial basis function (RBF) kernel, which is commonly used for data that are not linearly separable. Optimization is performed on the regularization parameter C and the kernel parameter γ. The tested values for C include 1, 10, 50, 100, and 500, while the values of γ range from 1 to 5. The evaluation results for the RBF kernel are presented in Table 5.

Based on the results in Table 5, the parameter $\gamma = 2$ produces the lowest accuracy among the tested values. The best performance for the RBF kernel is achieved with $C = 10$, $C = 50$, and $C = 100$, all yielding an accuracy of 0.866.

4.3. Analysis of Z-Score Normalization Accuracy Results

This section presents a comparative analysis of model performance using normalized and non-normalized data. Tables 6 and 7 summarize the SVM classification accuracy obtained without normalization and with Z-score normalization, respectively, using the best-performing parameters for each kernel.

As shown in Tables 6 and 7, there is a significant improvement in classification accuracy when Z-score normalization is applied. Models trained on normalized data consistently outperform those trained on non-normalized data. This improvement can be attributed to differences in feature value ranges across classes, which make it difficult for the model to learn meaningful patterns without normalization. Without normalization, the model tends to produce biased predictions by focusing on features with larger value ranges.

Table 8
Classification report of the proposed model.

Class	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	1
1	0.87	1.00	0.93	26
2	1.00	0.50	0.67	2
3	0.98	0.98	0.98	109
4	0.97	0.95	0.96	39
5	1.00	1.00	1.00	1
6	1.00	0.50	0.67	2
7	0.95	0.91	0.93	23
Accuracy			**0.96**	**203**
Macro Average	**0.97**	**0.86**	**0.89**	**203**
Weighted Average	**0.96**	**0.96**	**0.96**	**203**

Table 9
Comparison of the proposed method with previous studies.

Study/Method	Feature Extraction	Classifier	No. of Classes	Accuracy
Tempola et al. (2018)	Time-domain features	KNN	4	0.89
Lara-Cueva et al. (2017)	Spectral features	SVM	5	0.92
Tang et al. (2020)	Frequency features	SVM	6	0.93
Proposed Method	Autocorrelation-based features	Polynomial SVM, degree = 2	8	0.9605

4.4. Confusion Matrix

The results of the confusion matrix evaluation are presented in Fig. 5. As shown in Fig. 5, the confusion matrix indicates that the proposed model achieves strong classification performance across eight classes, with most samples correctly classified along the main diagonal. Class 3 exhibits the highest accuracy, with 107 true positives and minimal misclassification, followed by Classes 1, 4, and 7, which also show high correct prediction rates. Misclassifications are mainly observed in classes with limited samples, particularly Classes 2 and 6, suggesting that class imbalance affects predictive reliability for underrepresented classes. Minor confusion between certain class pairs indicates partial feature overlap. Overall, the confusion matrix confirms that the model effectively discriminates among the majority of classes, consistent with the reported accuracy and F1-score results.

As shown in Table 8, the proposed model achieves an overall accuracy of 0.96 on 203 samples, indicating strong classification performance. Classes with sufficient data, particularly Class 3, demonstrate robust results with precision and recall values of 0.98, while Classes 4 and 7 also exhibit high F1-scores of 0.96 and 0.93, respectively. Perfect scores observed in Classes 0 and 5 should be interpreted cautiously due to their limited sample sizes. In contrast, reduced recall values in Classes 2 and 6 highlight the impact of class imbalance. The macro-average F1-score of 0.89 reflects inter-class performance variability, whereas the weighted F1-score of 0.96 confirms reliable overall model performance.

4.5. Overall Performance Comparison

To evaluate the broader applicability of the proposed framework and to contextualize its performance within existing research, a comparative assessment is conducted using representative results reported in previous studies. This comparison highlights differences in feature extraction techniques, classification models, the number of seismic event categories, and achieved classification accuracy.

As summarized in Table 9, earlier studies have reported competitive performance using different combinations of features and classifiers for seismic signal classification. Tempola, Muhammad, & Khairan (2018) employed time-domain features with a KNN classifier to distinguish four classes, achieving an accuracy of 0.89. Lara-Cueva, Benítez, Paillacho, Villalva, & Rojo-Álvarez (2018) extended the classification task to five event types by incorporating spectral features and SVM, resulting in improved accuracy. A further increase in classification performance was reported by Tang, Zhang, & Wen (2020), who utilized frequency-based features and SVM to classify six seismic event categories.

In contrast, the proposed approach attains the highest accuracy of 0.9605 while simultaneously handling a more challenging classification scenario involving eight volcanic seismic event classes. This improvement reflects the combined effect of autocorrelation-based feature representation, which

emphasizes inherent temporal structures in seismic signals, and Z-score normalization, which enhances model stability by reducing feature scale disparities. Moreover, the polynomial SVM kernel with degree d = 2 provides an effective balance between model complexity and generalization capability, enabling accurate discrimination of non-linear seismic patterns.

It should be noted that this study does not perform a direct quantitative comparison with a manual seismic classification system. Instead, manual analysis is treated as a conceptual baseline representing expert-driven interpretation, which is commonly characterized by subjectivity, high time consumption, and limited scalability under large data volumes and noisy conditions. Therefore, the evaluation in this study focuses on assessing whether the proposed automated framework can deliver stable, reproducible, and high-accuracy classification results that address the operational limitations inherent in manual seismic analysis, rather than benchmarking numerical performance against human interpretation.

Overall, the comparative results indicate that the proposed framework not only achieves higher classification accuracy than previously reported automated methods but also demonstrates improved robustness and generalizability as the number of target classes increases. These findings confirm the effectiveness of the proposed method for complex multi-class volcanic seismic signal classification tasks.

4.6. Discussion

Overall, the experimental findings demonstrate that the proposed framework is capable of capturing salient patterns in seismic signals by integrating autocorrelation-based characterization with SVM classification. Variations in performance across different SVM kernels can be attributed to differences in feature space non-linearity and the stabilizing effect of data normalization. In this context, Z-score normalization substantially enhances classification robustness by reducing feature scale disparities and facilitating more effective kernel optimization. Classification errors are predominantly observed in classes with limited data availability and overlapping signal characteristics, a well-known issue in seismic event classification. Nevertheless, the model maintains reliable performance for dominant classes, as evidenced by the strong diagonal dominance in the confusion matrix and consistently high weighted evaluation metrics. These results establish a clear foundation for further discussion regarding model limitations and potential areas for improvement, which are addressed in the following chapter. Despite the promising results, this study has several limitations that should be acknowledged. One important limitation is class imbalance, which arises from the limited availability of seismic records for certain event types. Although the dataset contains sufficient samples for dominant classes, some classes are underrepresented, leading to uneven class distributions. This imbalance contributes to higher misclassification rates in minority classes, particularly when signal characteristics overlap. Therefore, the observed classification errors are influenced not only by limited data availability but also by the resulting class imbalance, which remains a fundamental challenge in seismic event classification.

5. Conclusions

Based on the findings of this study, the integration of autocorrelation-based signal characterization with Support Vector Machine (SVM) classification is shown to be effective for analyzing volcanic seismic signals. The experiments were conducted using volcanic seismic recordings obtained from authorized monitoring stations, ensuring that the conclusions are supported by real observational data.

Among the evaluated kernel functions, the polynomial SVM kernel with a degree of two delivers the most reliable classification performance, demonstrating its strong capability in capturing non-linear patterns inherent in seismic data. This is supported by its highest achieved classification accuracy of 0.9605, which exceeds that of the linear (0.9310) and radial basis function (RBF) (0.8669) kernels.

Furthermore, the results confirm that Z-score normalization plays a critical role in enhancing model stability and classification reliability by standardizing feature distributions. When normalization is applied, the classification accuracy of the polynomial kernel increases substantially from 0.5369 to 0.9605, while the linear kernel improves from 0.8325 to 0.9310, highlighting the importance of feature scaling in SVM-based seismic classification.

Despite the strong overall performance, reduced classification reliability is observed for several seismic classes with limited training samples. This limitation reflects the impact of class imbalance caused by restricted data availability in minority classes rather than shortcomings of the proposed framework itself, as indicated by recall values as low as 0.50 in these categories. Accordingly, future work will focus on data imbalance handling strategies, including undersampling, oversampling, and synthetic data generation techniques such as the Synthetic Minority Over-sampling Technique (SMOTE), to further enhance the robustness and generalization capability of the proposed approach in operational volcanic seismic

monitoring systems.

6. Declaration of AI and AI assisted technologies in the writing process

During the preparation of this manuscript, the author(s) used ChatGPT only to assist with grammatical review. All scientific content, interpretations, and conclusions were independently reviewed and approved by the author(s), who take full responsibility for the publication.

7. CRediT Authorship Contribution Statement

Fridy Mandita: Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Validation, Visualization, and Writing–original draft. **Muhammad Arif Fajriyansah**: Conceptualization, Formal analysis, and Writing–review & editing.

8. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

9. Acknowledgments

The authors would like to thank the volcanic monitoring institutions, particularly Balai Penyelidikan dan Pengembangan Teknologi Kebencanaan Geologi (BPPTKG), and the seismic analysts involved in data acquisition and labeling for their support. Appreciation is also extended to colleagues and reviewers whose constructive feedback contributed to improving the quality of this manuscript.

10. Data Availability

The volcanic seismic data used in this study were obtained from monitoring stations operated by the Balai Penyelidikan dan Pengembangan Teknologi Kebencanaan Geologi (BPPTKG), Indonesia. The data are not publicly available and require official permission from BPPTKG for access, due to institutional data-sharing regulations. Although the raw data cannot be openly distributed, the methodology, preprocessing steps, and experimental workflow are described in sufficient detail to ensure transparency and enable reproducibility using equivalent seismic datasets.

11. Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

12. Ethical Approval

This study does not involve human participants, personal data, or patient-identifying information. Therefore, ethical approval was not required.

13. References

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series, 1142*. doi:https://doi.org/10.1088/1742-6596/1142/1/012012

Ambarwari, A., Adrian, Q. J., & Herdiyeni, Y. (2020). Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 4*(1), 117-122. doi:https://doi.org/10.29207/resti.v4i1.1517

Anggian, F. C., Hidayat, N., & Furqon, M. T. (2020). Implementasi Metode Modified K-Nearest Neighbor untuk Klasifikasi Status Gunung Berapi. *JPTIIK (Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer), 3*(12), 11027-11033. Retrieved January 29, 2026, from https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/6843

Bergen, K. J., Johnson, P. A., Hoop, M. V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science, 363*(6433). doi:https://doi.org/10.1126/science.aau0323

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer.

Chouet, B. A., & Matoza, R. S. (2013). A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption. *Journal of Volcanology and Geothermal Research, 252*, 108-175. doi:https://doi.org/10.1016/j.jvolgeores.2012.11.013

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*(20), 273–297. doi:https://doi.org/10.1007/BF00994018

Gibbons, S. J., & Ringdal, F. (2006). The detection of low magnitude seismic events using array-based waveform correlation. *Geophysical Journal International, 165*(1), 149–1667. doi:https://doi.org/10.1111/j.1365-246X.2006.02865.x

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques.* Morgan Kaufmann.

Handayanto, A., Latifa, K., Saputro, N. D., & Waliyansyah, R. R. (2019). Analisis dan Penerapan Algoritma Support Vector Machine (SVM) dalam Data Mining untuk Menunjang Strategi Promosi. *JUITA: Jurnal Informatika, 7*(2), 71–79. doi:https://doi.org/10.30595/juita.v7i2.4378

Karo, I. M., & Hendriyana, H. (2022). Klasifikasi Penderita Diabetes menggunakan Algoritma Machine Learning dan Z-Score. *Jurnal Teknologi Terpadu, 8*(2), 94–99. doi:https://doi.org/10.54914/jtt.v8i2.564

Lara-Cueva, R., Benítez, D. S., Paillacho, V., Villalva, M., & Rojo-Álvarez, J. L. (2018). On the use of multi-class support vector machines for classification of seismic signals at Cotopaxi volcano. *2017 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC).* Ixtapa, Mexico: IEEE. doi:https://doi.org/10.1109/ROPEC.2017.8261613

Malfante, M., Mura, M. D., Mars, J. I., Métaxian, J.-P., Macedo, O., & Inza, A. (2018). Automatic Classification of Volcano Seismic Signatures. *Journal of Geophysical Research: Solid Earth, 123*(12), 10,645-10,658. doi:https://doi.org/10.1029/2018JB015470

Manley, G. F., Mather, T. A., Pyle, D. M., Clifton, D. A., Rodgers, M., Thompson, G., & Londo, J. M. (2022). A Deep Active Learning Approach to the Automatic Classification of Volcano-Seismic Events. *Frontiers in Earth Science, 10*. doi: https://doi.org/10.3389/feart.2022.807926

Marzocchi, W., Selva, J., & Jordan, T. H. (2021). A unified probabilistic framework for volcanic hazard and eruption forecasting. *Natural Hazards and Earth System Sciences, 21*(11), 3509–3517. doi:https://doi.org/10.5194/nhess-21-3509-2021

McNutt, S. R. (2025). Volcanic seismology. *Annual Review of Earth and Planetary Sciences, 33*, 461-491. doi:https://doi.org/10.1146/annurev.earth.33.092203.122459

Meier, M.-A., Ross, Z. E., Ramachandran, A., Balakrishna, A., Nair, S., Kundzicz, P., . . . Yue, Y. (2019). Reliable Real-Time Seismic Signal/Noise Discrimination With Machine Learning. *Journal of Geophysical Research: Solid Earth, 124*(1), 788-800. doi:https://doi.org/10.1029/2018JB016661

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020). Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications, 11*. doi:https://doi.org/10.1038/s41467-020-17591-w

Perdana, R. S., Fatichah, C., & Purwitasari, D. (2015). Pemilihan Kata Kunciuntuk Deteksi Kejadian Trivial pada Dokumen Twitter Menggunakan Autocorrelation Wavelet Coefficients. *JUTI: Jurnal Ilmiah Teknologi Informasi, 13*(2), 152–159. doi:https://doi.org/10.12962/j24068535.v13i2.a484

Rahutomo, F., Saputra, P. Y., & Fidyawan, M. A. (2018). Implementasi Twitter Sentiment Analysis untuk Review Film Menggunakan Algoritma Support Vector Machine. *Jurnal Informatika Polinema, 4*(2), 93-100. doi:https://doi.org/10.33795/jip.v4i2.152

Ririh, K. R., Laili, N., Wicaksono, A., & Tsurayya, S. (2020). Studi Komparasi dan Analisis Swot pada Implementasi Kecerdasan Buatan (Artificial Intelligence) di Indonesia. *J@ti Undip: Jurnal Teknik Industri, 15*(2), 122-133. Retrieved January 28, 2026, from https://ejournal.undip.ac.id/index.php/jgti/article/view/29183

Ross, Z. E., Meier, M.-A., Hauksson, E., & Heaton, T. H. (2018). Generalized Seismic Phase Detection with Deep Learning. *Bulletin of the Seismological Society of America, 108*(5A)), 2894–2901. doi:https://doi.org/10.1785/0120180080

Schölkopf, B., & Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing, 97*. doi:https://doi.org/10.1016/j.asoc.2019.105524

Tang, L., Zhang, M., & Wen, L. (2020). Support Vector Machine Classification of Seismic Events in the Tianshan Orogenic Belt. *Journal of Geophysical Research: Solid Earth, 125*(1). doi:https://doi.org/10.1029/2019JB018132

Tempola, F., Muhammad, M., & Khairan, A. (2018). Comparison of Classification between KNN and Naive Bayes at the Determination of the Volcanic Status with k-Kold Cross Validation. *Jurnal Teknologi Informasi dan Ilmu Komputer, 5*(5), 577-584. doi:https://doi.org/10.25126/jtiik.201855983

Thelen, W. A., Matoza, R. S., & Hotovec-Ellis, A. J. (2022). Trends in volcano seismology: 2010 to 2020 and beyond. *Bulletin of Volcanology, 84*. doi:https://doi.org/10.1007/s00445-022-01530-2

Titos, M., Bueno, A., García, L., Benítez, M. C., & Ibañez, J. (2019). Detection and Classification of Continuous Volcano-Seismic Signals With Recurrent Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing, 57*(4), 1936-1948. doi:https://doi.org/10.1109/TGRS.2018.2870202

Vapnik, V. N. (1998). *Statistical Learning Theory.* John Wiley & Sons.