https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

The impact of data preprocessing in machine-learning models for spatial continuity of fault detection using borehole data and Triangulated Irregular Networks

Adam Lewiński[1], Michał Michalak[1], Paulina Leonowicz[2], Agnieszka Kulawik[3]

[1]Faculty of Geology, Geophysics and Environmental Protection, AGH University of Science and Technology, Mickiewicza 30, 30-059 Cracow, Poland

[2]Faculty of Geology, University of Warsaw, Żwirki i Wigury 93, 02-089 Warsaw, Poland

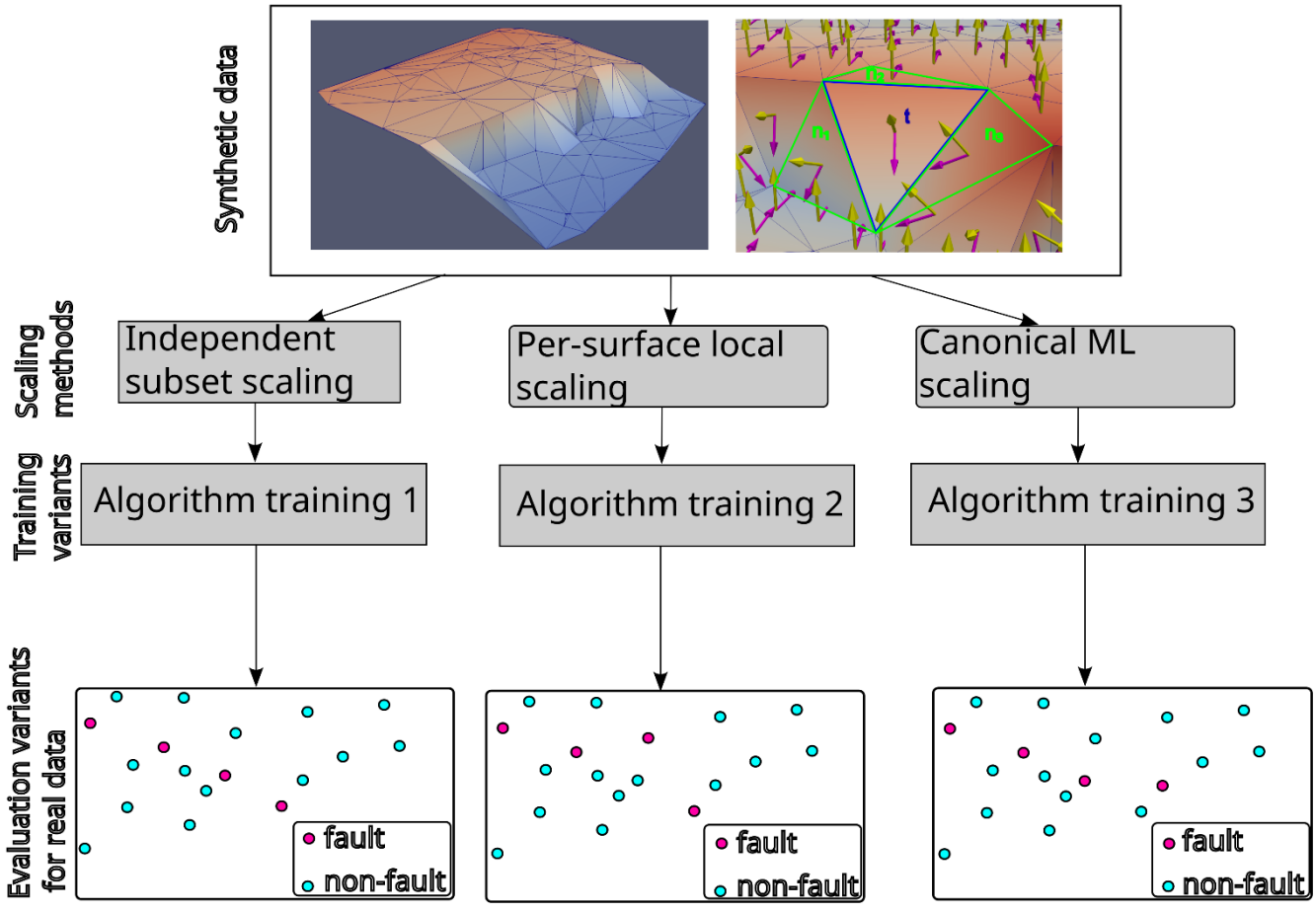
[3]University of Silesia in Katowice, Faculty of Science and Technology, ul. Bankowa 14, 40-007 Katowice, Poland

Correspondence to: Michał P. Michalak (michalm@agh.edu.pl)

Abstract.

Accurate fault detection is crucial for 3D subsurface structural modeling, yet the success of the method fundamentally depends on a critical property: the spatial continuity of fault-traces. In this study, we illustrated that spatial continuity strongly depends on spatial additivity of machine-learning-based classification. Spatial additivity dictates whether a machine-learning (ML) model yields consistent predictions when evaluating an entire structural domain versus its localized spatial subsets. While supervised ML can be used to map tectonic discontinuities, the impact of data preprocessing on the important properties (spatial additivity, spatial continuity) remains unaddressed. This study compares three data-scaling workflows influencing a Support Vector Machine (SVM) classifier: independent subset scaling, per-surface local scaling, and canonical ML scaling. Validated on structural data from the Kraków-Silesian Homocline, the results demonstrate that local and independent scaling frameworks violate spatial additivity, introducing potential artifacts. Conversely, the canonical global architecture satisfies spatial additivity, ensuring stable, continuous fault networks. Crucially, maintaining spatial additivity is vital for the incremental updating of geological models, enabling the seamless integration of newly acquired borehole data without triggering complete recalculations of the global feature space.

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.



30

1 Introduction

Faults represent critical structural discontinuities that partition the rock mass into discrete fragments, thereby playing a
35 fundamental role in 3D geological modelling, where omitting these structures from a model (false negatives) often leads to oversimplifications of the subsurface. With the gradual advent of machine learning (ML), automated geological modeling has increasingly adopted supervised algorithms to detect complex geological objects (Mousavi and Beroza, 2022). In geophysics, for instance, seismic methods frequently utilize 2D slices to reduce the immense computational costs associated with 3D data volumes (Dou et al., 2022).

40 Despite the rapid development of these techniques, significant research gaps remain. Specifically, insufficient attention has been paid to the data preprocessing stage. In the broader ML community, it is standard practice to rigorously justify the

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

separation of training and testing sets to ensure they do not interact (leakage), which is crucial for the reliability of the results (Kapoor and Narayanan, 2023). In machine learning, scaling represents one of the preprocessing methods, which is a common requirement for solving classification tasks. This procedure is required to mitigate different orders of magnitude when geometric dissimilarity is taken into account. For example, angular distance can have values between 0 and 90 (if acute angle is considered) and Euclidean distance can have values between 0 and 2 if unit vectors are used (points on the unit sphere). Leaving the data unscaled, would pose a risk that certain variables would dominate the objective function and other potentially important features would bring little value to the learning process. As such, data scaling is a particularly vital preprocessing step for algorithms that are not scale-invariant or, in other words, sensitive to different orders of magnitude, such as the K-nearest neighbors or Support Vector Machine (SVM) (Pedregosa et al., 2011). The choice of scaling architecture can fundamentally alter the performance of the model and its ability to generalize. We note that there may be many scaling approaches. In this study, we focus on the StandardScaler method which is considered a common scaling method and which was used in the referenced study on detecting faults using triangulated models of geological surfaces (Michalak et al., 2025). According to the documentation of the StandardScaler tool (StandardScaler, 2025; Pedregosa et al., 2011), scaling is performed by subtracting mean and dividing by standard deviation for every feature from the training set (Apicella et al., 2023). Therefore, mean and variance (or standard deviation) can be regarded as transformation parameters which are used for later sets such as test and production tests (Pedregosa et al., 2011). In this context, the role of data scaling becomes particularly relevant for geological applications, yet it remains insufficiently explored.

To investigate this problem, we used borehole data which document a subsurface stratigraphic boundary which is regarded as a homoclinal surface in regional geological models (Hermański, 1971). To emulate the homoclinal scenario, we utilize a series of synthetic structural models represented as Triangulated Irregular Networks (TIN) using Delaunay triangulation (De Berg et al., 2008). These models are organized into specific datasets where individual triangulated surfaces can be aggregated in various ways - either as independent surfaces or grouped sets of interfaces - to simulate different data-handling scenarios. By evaluating these methods on real data, we seek to establish a framework for preprocessing sparse borehole-derived data in 3D structural modelling which preserves spatial continuity of fault traces. In the field of machine learning, a fundamental distinction is made between inter-subject approaches (the all-subject schema), where a model operates on data derived from multiple entities, and intra-subject approaches (the single-subject schema) (Apicella et al., 2023), where processes occur within the boundaries of a specific, individual unit. In this study, a “subject” is defined as a specific geological entity, such as a single triangulated surface. We compare three distinct data-scaling architectures and investigate their impact on the spatial continuity of fault classification results for sparse borehole data and Triangulated Irregular Networks (Fig. 1).

- Method 1 (Independent subset scaling method) uses three independent scalers for train, test and real data (an inter-subject approach).
- Method 2 (Per-surface local scaling) uses an independent scaler for every triangulated model of the surface (an intra-subject approach).

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

- Method 3 (Canonical ML scaling) uses only one scaler for train data and the computed parameters are used as transformation parameters for test and real data (an inter-subject approach).

The comparison of these approaches is particularly relevant in the context of geological modelling, where preserving the spatial continuity of fault structures is essential for reliable interpretation. We will illustrate that certain scaling architectures may result in the lack of spatial additivity of classification: the classification result of one function performed on the entire study area can differ from the joined classification result of two classification functions for two spatial subsets of the study area. This effect should be discussed in the context of spatial continuity of the fault traces and re-integration of the fault network with the broader structural context.

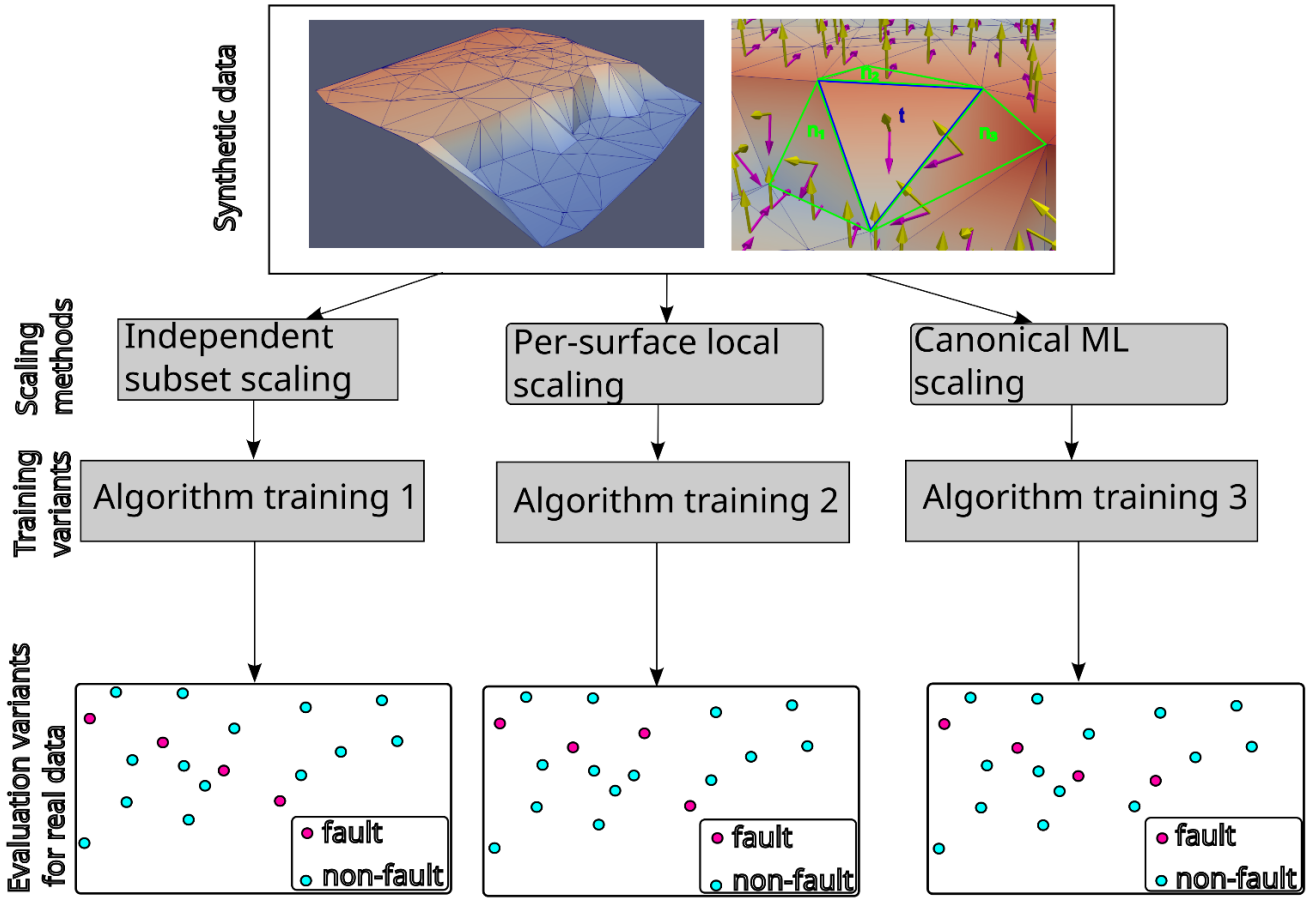


Fig. 1 Workflow applied in this study. Synthetic triangulated models of subsurface homoclinal horizons are input data for the supervised fault detection algorithm. Three distinct pre-processing methods are applied for the data. The first scaling method (Method 1: "Independent subset scaling") uses three independent scalers for train, test and real data. The second scaling method (Method 2: "Per-surface local scaling") uses an independent scaler for every triangulated model of the surface. The third scaling method (Method 3: "Canonical ML scaling") uses only one scaler for train data and the computed parameters are used as transformation parameters for test and train data. The SVM algorithm is then trained for three distinct data sets. After algorithm training, we apply the classification model to real data. Portions of this figure are modified from (Michalak et al., 2025) which also described the process of generating synthetic data.

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

2 Background

2.1 Classification tasks in 3D geological modeling

Machine learning is a widely adopted tool to detect geological features, such as salt (Muller et al., 2022), channels (Gao et al., 2021) or faults (Kaur et al., 2023), for seismic data. For example, the fault identification problem can be posed as a binary classification problem in which fault pixels are represented as one and non-fault pixels as zero (Lin et al., 2024). One of the limitations of using supervised methods in geological or geophysical modeling is that they require large amounts of labeled data (Mousavi and Beroza, 2022), which represents a challenge due to time-consuming labelling by experts. In geological modeling, synthetic data (Oakley et al., 2025; Wang et al., 2024; Wu et al., 2020) with automatically inserted faults are often used to train the algorithm. This strategy constitutes slightly less than 1/3 of the supervised learning studies to detect geological features (Lin et al., 2024).

Recent advances in applying machine learning models to subsurface borehole data relate mainly to lithological predictions (Zhang et al., 2023) or 3D subsurface soil stratigraphy modeling (Hu et al., 2024). Borehole data are sometimes combined with geophysical data in the clustering procedure to better visualize fault zone (Doyoro et al., 2025). In 3D geological modeling, triangulated irregular network (TIN) is considered a valid tool for geometric modeling of geological surfaces (Ji et al., 2023), and in particular sparse borehole data (Wu et al., 2005).

An attempt has also been made to detect faults using supervised methods for subsurface homoclinal data using individual triangles and their geometric attributes such as local orientation and distances with neighbors (Michalak et al., 2025). The underlying assumption is that relative displacement of stratigraphy on either side of a fault allows the discontinuity to be detected (Lin et al., 2024), provided that the surface of interest exhibits regional trend. We note that assumption of homoclinal structure may seem narrow, however a homoclinal model which is not relevant at a larger scale may be relevant at a local scale. In that approach, training and tests sets were comprised of many triangulated models of synthetic homoclinal surfaces with slightly different parameters controlled by a user such as regional trend, fault throw or noise. After fine-tuning of the algorithm on synthetic data, the classification tool attempted to detect faults for real data, assuming that training data represented a useful representation of the reality (Michalak et al., 2025). We note that without using machine-learning methods, faults can also be detected for DEM raster data using the assumption of abrupt elevation changes and edge-detection algorithms which are helpful in this context (Gayrin et al., 2026).

2.2 Data leakage

According to the definition, data leakage occurs when information that would not be available at prediction time is used when building the model (Common pitfalls and recommended practices, 2025; Pedregosa et al., 2011). There are several categories proposed with respect to data leakage. The most evident category corresponds to the lack of clean separation of training and test dataset. In this context, there can be following examples of data leakage: 1) using the same dataset for training and testing the model; 2) using the entire dataset including train and test data for scaling; 3) using test set to select best features during model training; and 4) duplicates in train and test data sets. (Kapoor and Narayanan, 2023). This lack of separation can result

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

in overoptimistic results (Kapoor and Narayanan, 2023) because data used for purely evaluative purposes such as test data should not be used for creating the model. We note that the judgement of whether a machine-learning design constitutes a leakage may not be straightforward, as the discussion often comes down to domain-specific expertise (Kapoor and Narayanan, 2023).

3 Methods

In this section, we explain three distinct data-scaling methods used in the pre-processing stage of a supervised machine learning workflow. We note that the factual description of the methods is presented in Methods section, while discussion about specific advantages or disadvantages is reserved for Discussion.

In all scaling methods, a user of the method had to supply ranges of parameters of the triangulated models including geometric attributes such as dip angle, dip direction and fault throw, among many. Then, several triangulated models were generated using a randomly selected value from the supplied ranges, the local (specific to an individual triangle) and relational (distances with neighbours) geometric attributes were calculated and stored in corresponding N files. We note that the values of dip angle and dip direction were not directly used as attributes but first converted to three-dimensional representation of normal and dip vectors. We adhere to a reasonable generalization assumption that synthetic data should share similar attributes with target field (real) data (Choi et al., 2025).

3.1 Method 1: Independent subset scaling

Method 1: “Independent subset scaling” was used in the supervised approach of detecting faults using geometric attributes of triangulated models representing subsurface interfaces (Michalak et al., 2025). An aggregated data frame was generated using the N files. After data aggregation, a random split into training and test sets was performed. In order to reduce the impact of different ranges of various relational geometric attributes such as angular distances (values 0-90) or Euclidean distance (values 0-2), we performed an independent scaling to the training and test sets. After algorithm fine-tuning on synthetic data, the algorithm was applied for real data with an independent scaler. Therefore, in this approach, three independent scalers were used: two for synthetic data and one for real data (Fig. 2).

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

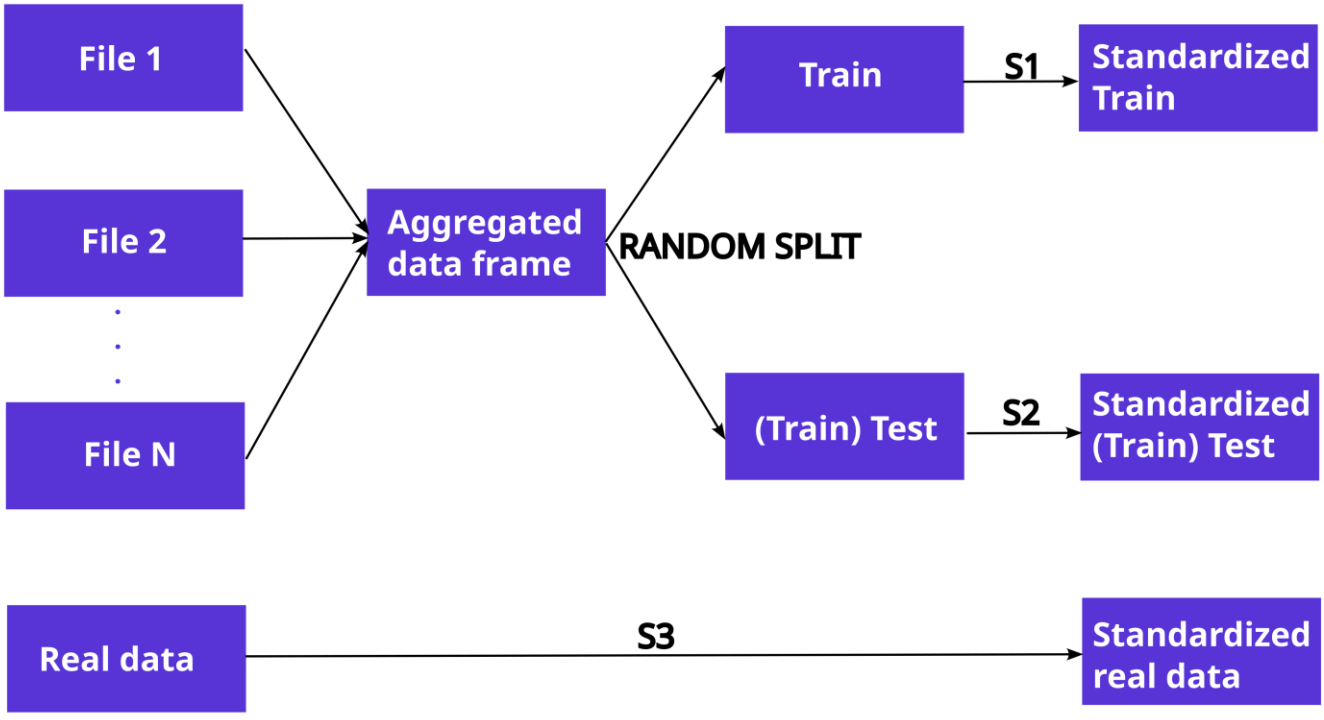


Fig. 2 Method 1: "Independent subset scaling" applies three scalers and it starts with creating several files (numbered from **1** to ***N***) representing triangulated models of homoclinal interfaces. Then, the files are aggregated into a single data frame. In the next step, random splitting is performed into a train and test set which are scaled independently into standardized sets. After fine tuning, the algorithm can be applied to real data which are scaled independently using the third scaler.

3.2 Method 2: Per-surface local scaling

Method 2: "Per-surface local scaling", similarly to Method 1: "Independent subset scaling", creates N triangulated models of subsurface interfaces, as described in the first paragraph of the Methods section. However, as opposed to the first method, scaling is applied before data aggregation and scaling is independently applied to every created model (file - Fig. 3). Then, standardized data frames are aggregated into one data frame. The second difference with the first method is that we apply a non-random split instead of a random split to obtain train and test sets. The non-random split is a chronological split in which the first surfaces of the aggregated data frame constitute the train set, while its last surfaces correspond to the test set. The application of non-random (chronological) split corresponds to a request that a researcher needs to justify why the test set does not interact with training data (Kapoor and Narayanan, 2023). In this case, this non-random split guarantees that entire surfaces

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

are either in the train or the test set. If this was not the case, one could argue about a mild "leakage" because geometric attributes of the same surface could be found in both train and test set. At first and without domain knowledge, the idea of a non-random split of the surfaces could be viewed controversial. However, we would like to point out that the randomness was already applied in the program based on the CGAL library (Michalak et al., 2025; Michalak, 2025) during creation of the surfaces – the parameters were drawn from a range specified by a user using a uniform distribution and the user has no control over the surface parameters. After algorithm fine-tuning on synthetic data, the algorithm was applied for real data with an independent scaler. Therefore, in this approach, $N + 1$ scalers were used: N for synthetic data and one for real data.

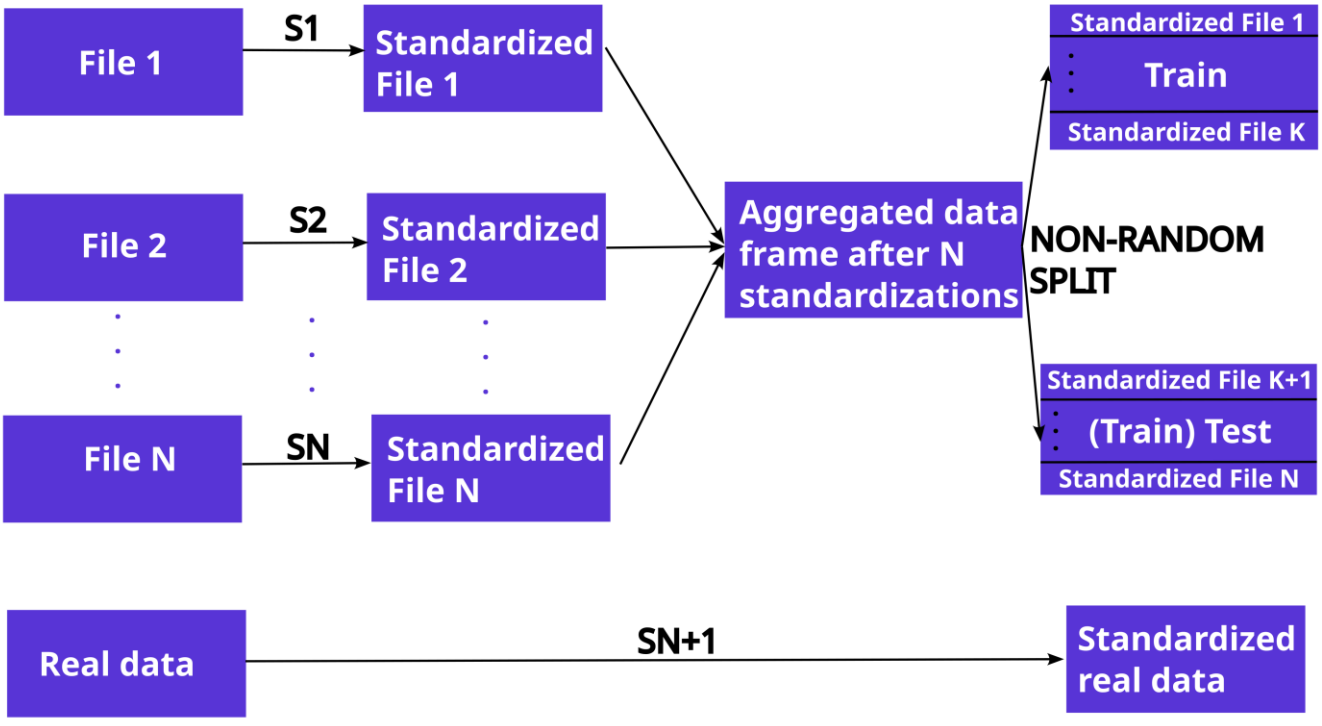


Fig. 3 Method 2: "Per-surface local scaling" with application of $\boldsymbol{N + 1}$ scalers. In this variant, $\boldsymbol{N}$ triangulated models of homoclinal interfaces receive their own scaler. As a result, $\boldsymbol{N}$ standardized data frames are aggregated into one data frame and split in a chronological order with first $\boldsymbol{k}$ standardized data frames in the train set and the last $\boldsymbol{N - k}$ standardized files in the test set. After fine tuning, the algorithm can be applied to real data which are scaled independently using the last $\boldsymbol{N + 1}$ scaler.

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

3.3 Method 3: Canonical ML scaling

Method 3: "Canonical ML scaling" begins with the creation of *N* data frames representing *N* triangulated models of subsurface interfaces, as described in the first paragraphs of the Methods section. An aggregated data frame was then created from the *N* data frames, and a random split was performed, as described in Section 3.1.

After performing the random split, we calculate the transformation parameters and transform the train set (Fig. 4). However, as opposed to the Method 1: "Independent subset scaling", we do not apply any independent scaler for the next sets (test and real data). Instead, we use the transformation parameters calculated for the train set to transform the test and real data sets. Therefore, in this approach, only one scaler was used.

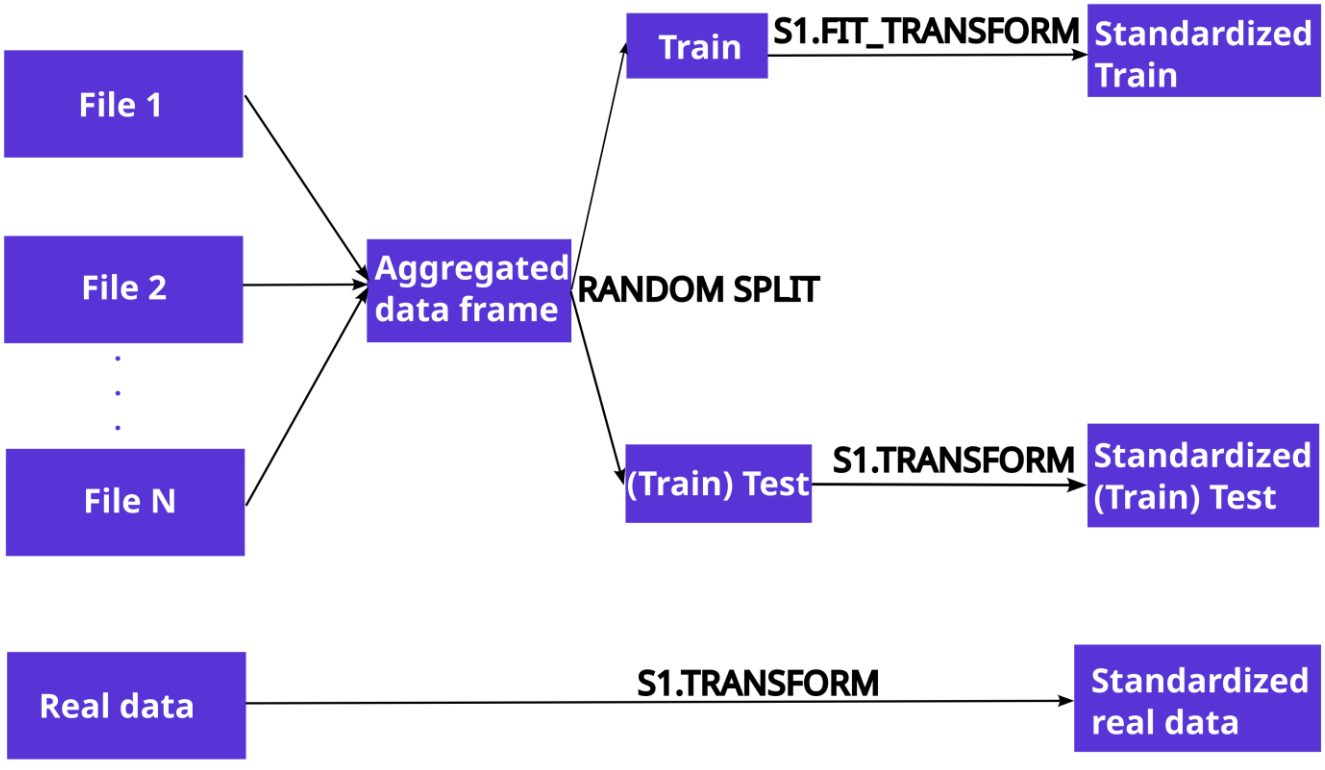


Fig. 4 Method 3: "Canonical ML scaling" with only one scaler. In this variant, the procedure starts with creating several files (numbered from **1** to ***N***) representing triangulated models of homoclinal interfaces. Then, the files are aggregated into a single data frame. In the next step, random splitting is performed into a train and test set. For the train set, we calculate the transformation parameters and transform the set (fit_transform function) using the only scaler (denoted as S1). For the test set, we apply the transformation using the parameters calculated for the train set. After fine tuning, the algorithm can be applied to real data which are transformed using the parameters obtained for train data.

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

3.4 SVM algorithm

Binary classification was performed using the Support Vector Machine (SVM) algorithm, a robust two-class classifier (Bishop, 2006; Vapnik, 2000), implemented through the scikit-learn library (Pedregosa et al., 2011). The SVM framework operates as an optimization problem that identifies a separating hyperplane with the maximum margin, defined by the shortest distance between the training samples and the decision boundary. This maximization of the margin ensures that the model remains stable even under minor perturbations of the data instances (Shalev-Shwartz and Ben-David, 2013). Formally, the optimization task is formulated as follows (Bishop, 2006):

$$arg \max_{w,b}\{ \min_{n} \left[t_n \left(\frac{w^T f(x_n)+b}{||w||} \right) \right] \}, \quad (5)$$

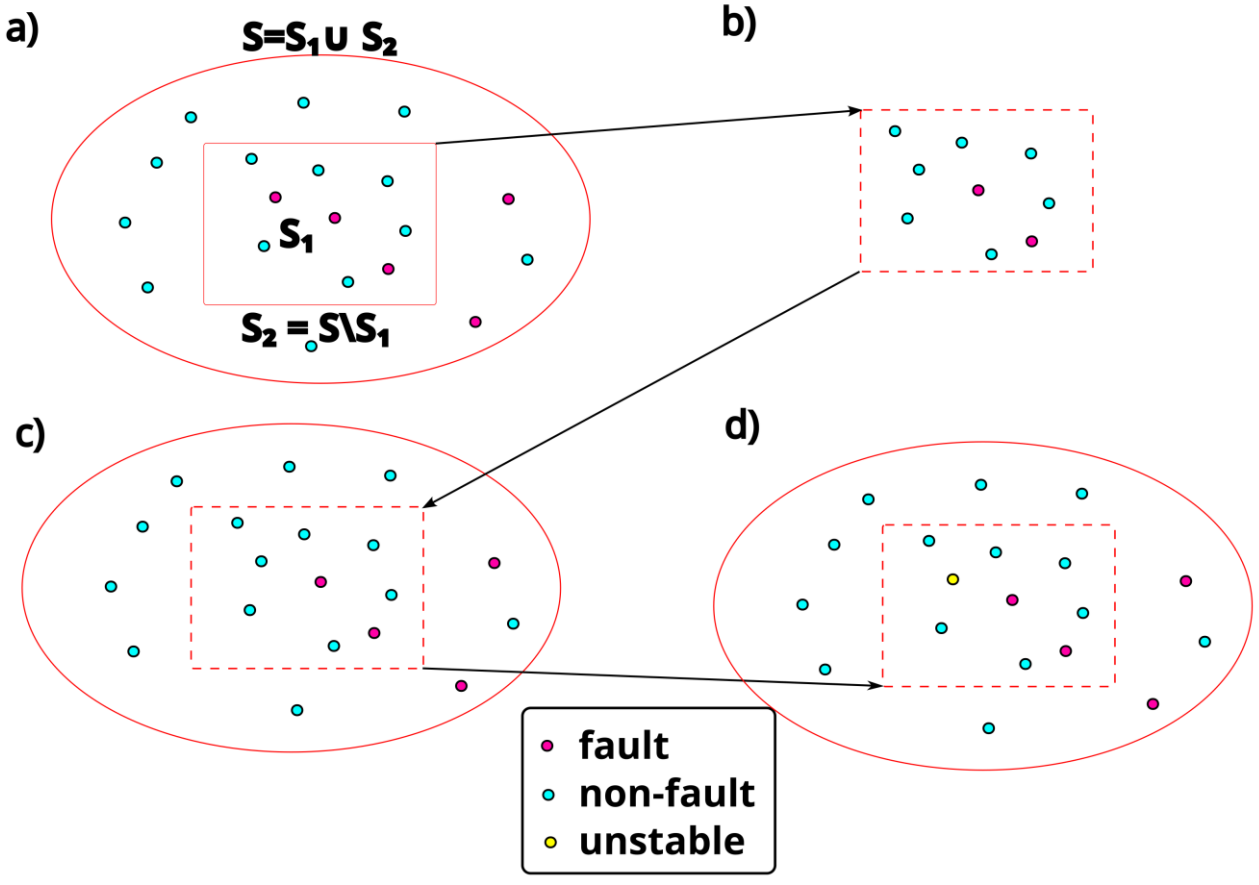
where $t_n \in \{-1,1\}$ denotes the target labels and $f(x)$ represents a fixed feature-space transformation (kernel) intended to facilitate the separation of data that is not linearly separable in its original dimensions. The orientation of the decision surface is determined by the weight vector w while b serves as the bias parameter. The condition $t_n(w^T f(x) + b)$ ensures that the optimization identifies solutions where all data points are correctly categorized.

To account for potential outliers, a soft-margin approach is employed, allowing for controlled misclassifications to improve overall model resilience (Shalev-Shwartz and Ben-David, 2013). In this study, Radial Basis Function (RBF) kernel was utilized. Its performance is governed by the penalty parameter C, which regulates the balance between training accuracy and the simplicity of the decision boundary, and the gamma parameter, which determines the influence radius of individual training points (Pedregosa et al., 2011). More specifically, to keep focus on the scaling methods, we used only one configuration of hyperparameters – the optimal combination of hyperparameters confirmed in the original classification study (Michalak et al., 2025): C=10, gamma=0.01 with radial basis function as the kernel function.

3.5 Investigating spatial continuity of classification results

To evaluate the practical applicability and spatial continuity of the classification, a multi-stage prediction workflow was implemented. First, the classification is performed using data from the whole study area (Fig. 5a). Then, classification is performed using data from the spatial subset of the study area (Fig. 5b). In the next step (Fig. 5c), labels of classification results corresponding to the spatial subset of the study area computed for the whole study area (Fig. 5a) are replaced with the labels computed for the spatial subset (Fig. 5b). In the final step (Fig. 5d), discrepancies between panels (a) and (c) are marked using yellow colour. By analyzing the changes after inserting the subset, we quantify the impact of different scaling architectures on the spatial continuity of detected faults.

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.



225

Fig. 5 A multi-stage prediction workflow to investigate spatial continuity of classification results. (a) the classification is performed using data from the whole study area; (b) the classification is performed using data from the spatial subset of the study area; (c) labels of classification results corresponding to the spatial subset of the study area computed for the whole study area (a) are replaced with the labels computed for the spatial subset (b); (d) discrepancies between panels (a) and (c) are marked using yellow colour. From a set-theoretic
230 viewpoint, the study area can be denoted as $\boldsymbol{S}$, the rectangular spatial subset can be denoted as $\boldsymbol{S_1}$ and the remaining observations (between the rectangle and the ellipsoid-like boundary) as $\boldsymbol{S_2} = \boldsymbol{S} \backslash \boldsymbol{S_1}$. Then, $\boldsymbol{S} = \boldsymbol{S_1} \cup \boldsymbol{S_2}$.

4 Study area: Geology and hydrogeology

4.1 Lithostratigraphy and structural framework

We used borehole data (810 boreholes – see Data availability section) documenting a subsurface Jurassic horizon, situated
235 between the older sandstones of the Kościeliska beds (Kościeliska sandstones) and the younger clays and mudstones of the Ore-Bearing Częstochowa Clay Formation, known also as ore-bearing clays (Dayczak-Calikowska et al., 1997; Kopik, 1998; Fig. 6C). This sequence is located within the Kraków-Silesian Homocline (Fig. 6A, B), a geological unit characterized by a subtle regional dip toward the NE (Deczkowski, 1977; Matyszkiewicz et al., 2015; Michalak et al., 2019; Znosko, 1960). The Jurassic strata lie discordantly over folded Paleozoic basement rocks (Górecka, 1993).

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

4.2 Hydrogeological context and mining history

From a hydrogeological perspective, the younger ore-bearing clays act as a confining layer for the aquifer hosted within the more permeable Kościeliska sandstones. The groundwater flow follows the regional dip of the strata toward the NE (Hermański, 1971; Więckowski, 1973; Pich and Pokora, 1982). Historically, the exploitation of iron ore within the lower parts of the ore-bearing clays (the siderite horizon; Fig. 6C) encountered significant hydrogeological hazards. The confined aquifer posed a risk (Wang and Park, 2003) of flooding mine faces excavated in the upper Kościeliska sandstones (Górniak and Sałaciński, 1981; Figs. 6 and 7). Consequently, intensive pumping was required to lower the water table below the mining levels (Fig. 7).

4.3 The role of fault zones

In the study area, the structural pattern of Kraków-Silesian Homocline is dominated by a dense grid of normal and strike-slip faults. These features are predominantly oriented in NW-SE and NE-SW (primarily) directions, reflecting the complex polyphase tectonic evolution of the Jurassic strata (Krokowski, 1984; Hermański, 1993). These fault zones significantly altered local hydrogeological conditions. During mining, increased water inflow was frequently observed when intersecting these zones, with pressurized water migrating along faults from the downthrown blocks of the Kościeliska sandstones (Górniak, 1969; Fig. 7). Notably, these faults remained impermeable in their upper sections (ore-bearing clays) because the fractures were sealed with clay material, preventing downward migration from Quaternary aquifers (Górniak, 1974; Fig. 7).

4.4 Current engineering context

Archived documentation from the mining era provides critical context for contemporary infrastructure projects, such as the E75, 46 and 91 roads (Fig. 6). During the construction of the E75, geophysical surveys identified various anomalies. These were attributed to lithological contrasts as well as anthropogenic features, including both horizontal and vertical historical mine shafts (GDDKIA ODDZIAŁ W KATOWICACH, 2019).

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

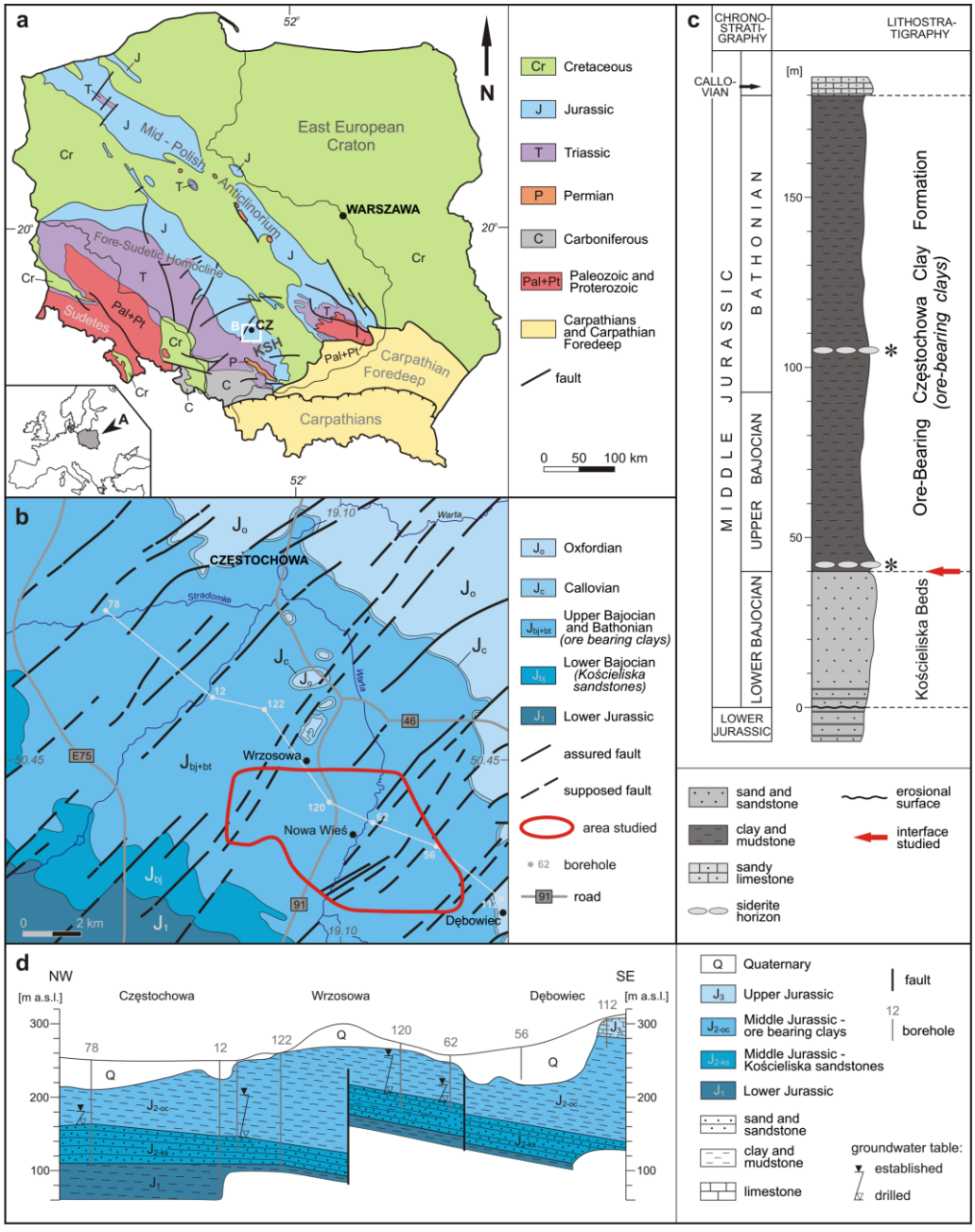


Fig. 6 . A - A simplified geological map of Poland without Cenozoic formations (after Karnkowski, 2008; Osika et al., 1972, modified), and location of the area studied. CZ - Częstochowa, KSH - Kraków-Silesian Homocline. B - Geological map of the studied part of the Kraków-Silesian Homocline (after Bardziński et al., 1986, modified) and the location of area studied. C – Generalised lithological log, lithostratigraphy and generalised chronostratigraphy of the Middle Jurassic deposits of the Kraków-Silesian Homocline from the Częstochowa region (after Matyja and Wierzbowski, 2000, modified). The bottom and the top ore levels are indicated by asterisk. D – Geological cross-section across the Kraków-Silesian Homocline in the Częstochowa region (after Kieńć et al., 2008, modified). Location shown on B. The faults were identified during mining activity due to exploitation of ore-bearing clays in the region (Hermański, 1993).

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

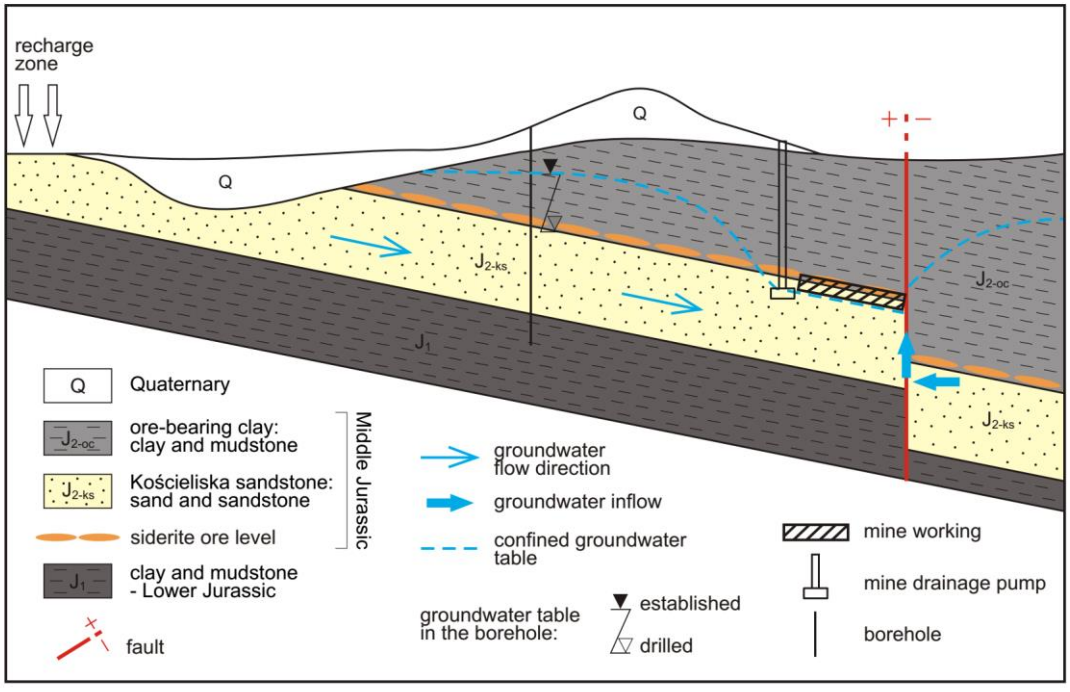


Fig. 7 A schematic diagram of the hydrogeological conditions of the Middle Jurassic deposits illustrating the recharge zone of the confined Kościeliska sandstone aquifer and the formation of a cone of depression due to mine drainage. Note the groundwater inflow from the downthrown block along the fault zone towards the mine workings, driven by the hydraulic head (after Górniak, 1969; Górniak and Sałaciński, 1981; Hermański, 1971, modified).

5 Results

In the results section, we present only analyses for real data and three different scaling methods described in the Methods section. We also provide confusion matrices and summary data of the classification task for synthetic data in Appendix A (Tables A1-A6).

5.1 Method 1: "Independent subset scaling" – results for real data

In this section, we present results using Method 1: "Independent subset scaling" which uses three independent scalers for training, test and real data sets (Section 3.1). A remarkable effect using this scaling variant is that the classification results may differ between the entire data set and its any subset. We selected a specific subset (marked using a red rectangle in Fig. 8a) and conducted classification (compare panels Fig. 8a and Fig. 8b). In this experiment, the selected subset was treated as an additional „real data file" and processed in the same way as the original data set, following the given data-scaling method. This implies that an additional independent scaler was used for the selected subset. Then, to investigate spatial continuity of the results, we replaced labels predicted for the entire data set with the labels obtained for the subset (Fig. 8c). Finally, to better

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

illustrate the dependence of the results on the selected area, we plotted panel (d) which shows locations where the predictions differ (labelled as “unstable”) between panels (a) and (c).

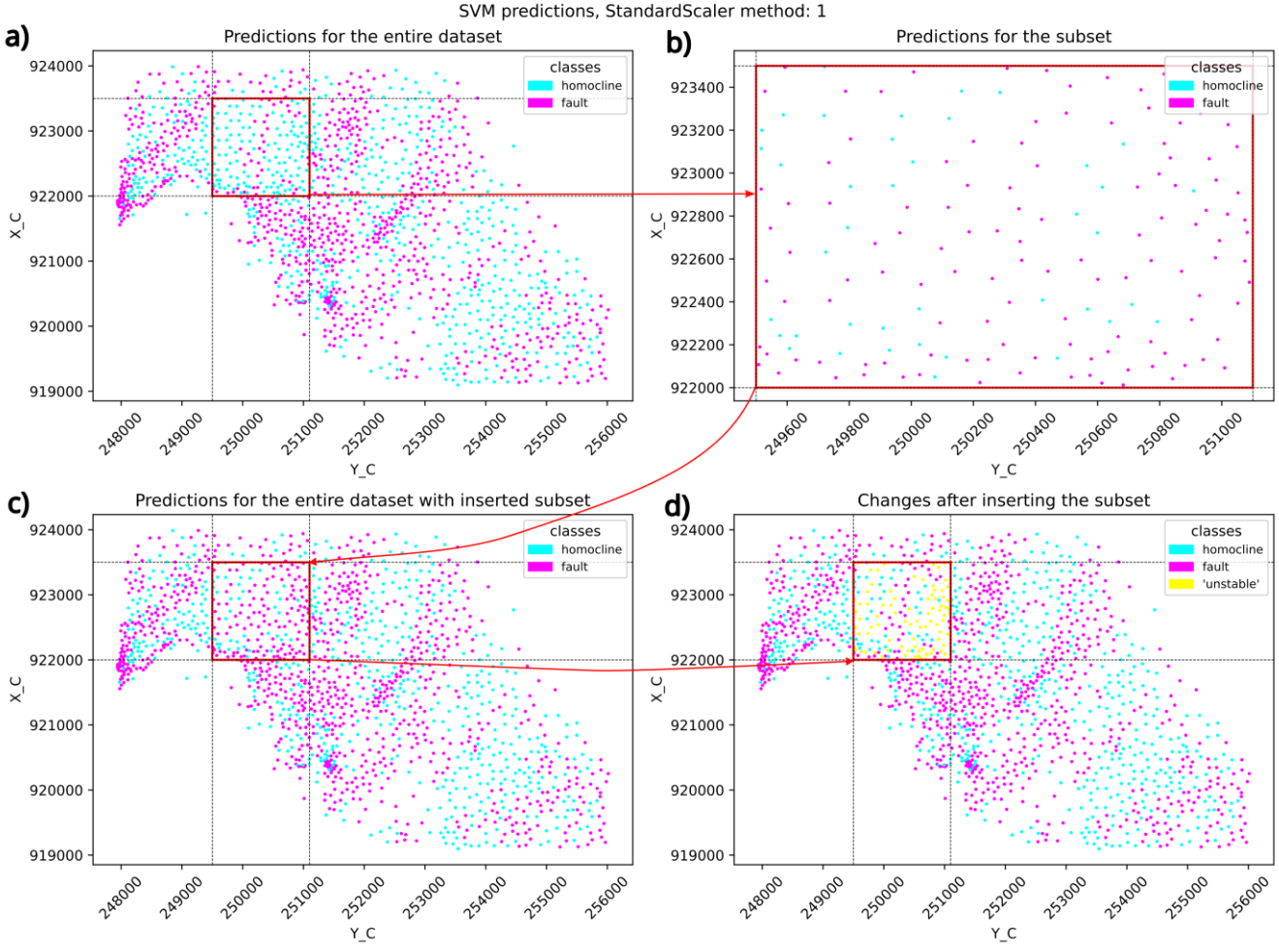


Fig. 8 Results for the classification task for real data using Method 1: “Independent subset scaling” as presented in the Methods section: (a) predictions for the entire dataset; we marked a subset using a red rectangle for further analyses in the next panels. (b) a subset from the entire dataset was selected for the classification task and the classification results are presented accordingly; (c) the results here are equivalent to those presented on panel (a) except the area marked with a red rectangle. In this area, we inserted the results from panel (b). (d) to better observe spatial continuity of the results, we marked discrepancies in the classification results using yellow labels.

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

5.2 Method 2: "Per-surface local scaling" – results for real data and

Here, we present results using Method 2: "Per-surface local scaling" which uses *N* independent scalers for synthetic data and an additional scaler for the real data set (Section 3.2). Similarly to the results in section 5.1, we observe that classification results differ between the entire data set and a selected subset marked using a red rectangle in Fig. 9a. We note that using the entire data set for the classification, the area corresponding to the selected subset (Fig. 9a) is considered mostly homoclinal without faults intersecting the surface. After insertion of the classification results predicted for the subset (Figs. 9b,c – and as described in section 5.1), we obtained a result which has a potential fault in the central part as in Fig. 8a.

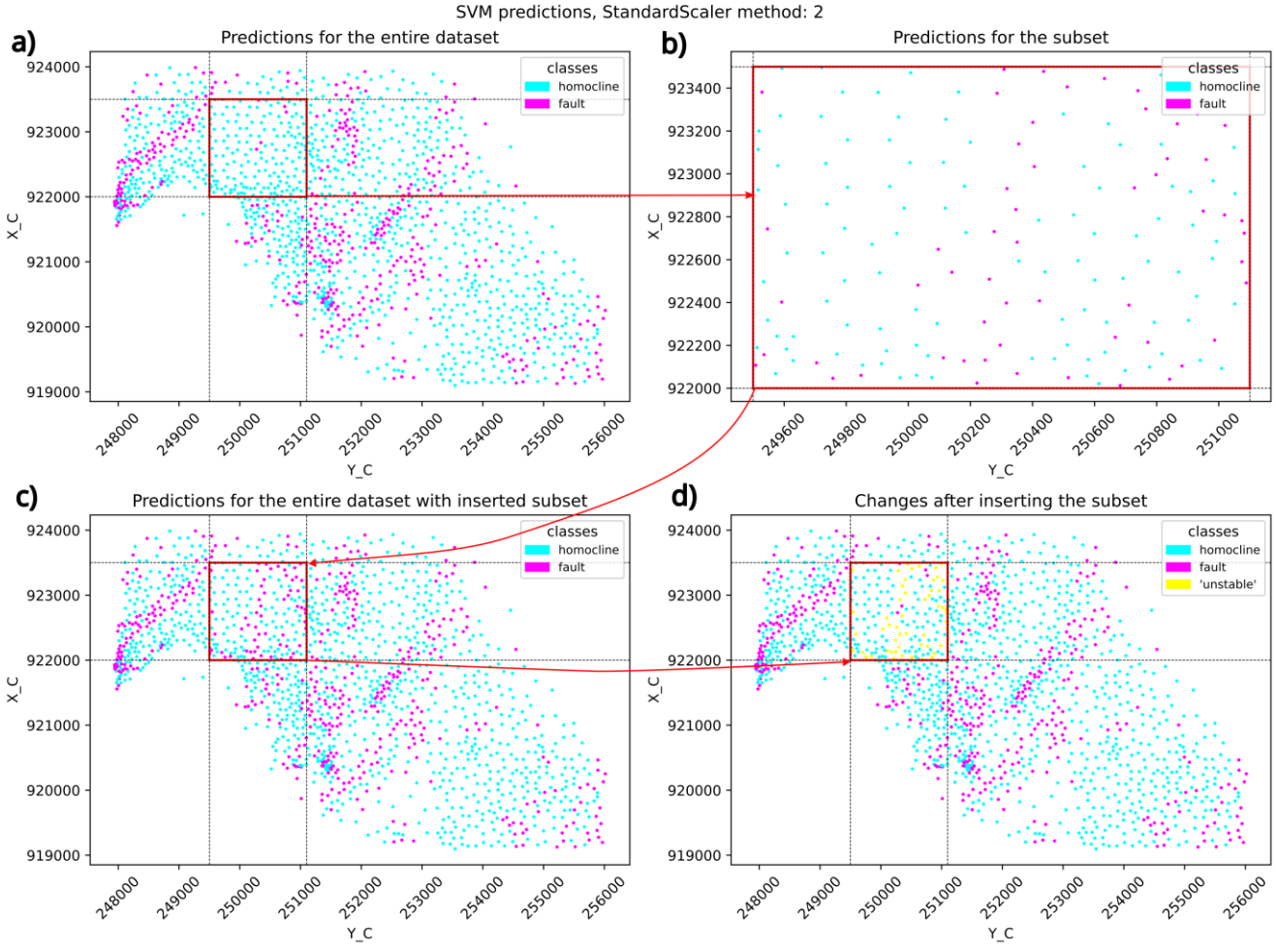


Fig. 9 Results for the classification task for real data using Method 2: "Per-surface local scaling" as presented in the Methods section: (a) predictions for the entire dataset; we marked a subset using a red rectangle for further analyses in the next panels. (b) a subset from the entire dataset was selected for the classification task and the classification results are presented accordingly; (c) the results here are equivalent to those presented on panel (a) except the area marked with a red triangle. In this area, we inserted the results from panel (b). (d) to better observe dependence of the classification results on the area taken for the analysis, we marked discrepancies in the classification results using yellow labels.

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

5.3 Method 3: "Canonical ML scaling" – results for real data

In this section, we present results using Method 3: "Canonical ML scaling" with only one scaler applied to training data (Section 3.3). This method computes the transformation parameters for the training data and shifts them accordingly, without re-computation of the parameters for test and real data. As such, test and real data are shifted using the accessible and already computed parameters. The most notable difference between this approach and previous approaches (sections 5.1 and 5.2) is that the results of classification performed on a subset are equal to results of the classification performed on the entire data set restricted to the selected subset (compare Fig. 10a and Fig. 10b). As such, the insertion (as described in Fig. 5 and section 5.1) does not alter the results corresponding to panel (a) and the set comprising "unstable" observations is consequently an empty set (Fig. 10d).

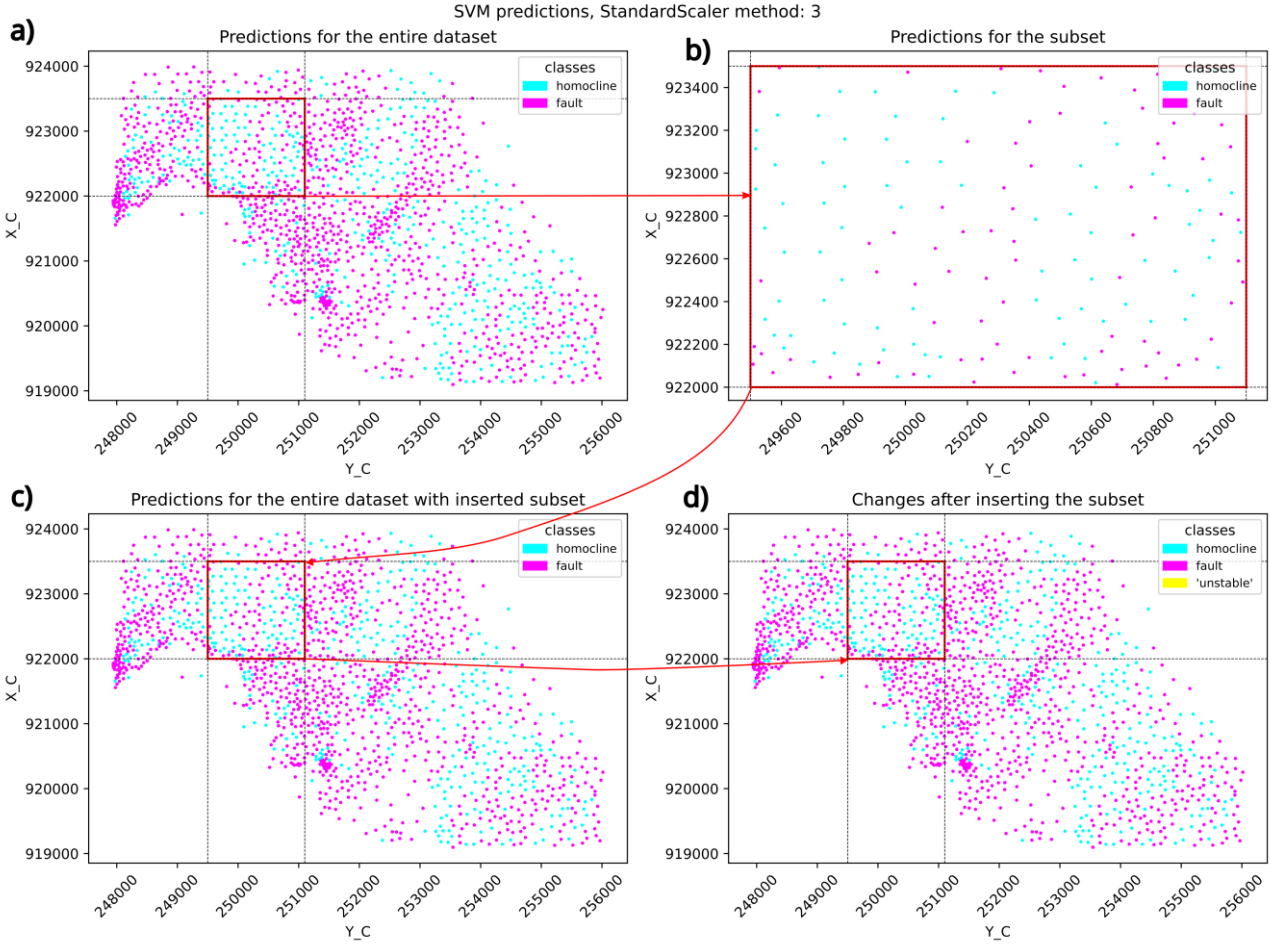


Fig. 10 Results for the classification task for real data using Method 3: "Canonical ML scaling" as presented in the Methods section: (a) predictions for the entire dataset; we marked a subset using a red rectangle for further analyses in the next panels. (b) a subset from the entire dataset was selected for the classification task and the results are presented accordingly; (c) the results here are equivalent to those presented on panel (a). (d) this panel shows that the set corresponding to "unstable" observations is an empty set due to common transformation parameters for training, test and real data.

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

6 Discussion

In 3D geological modelling leveraging implicit interpolation methods, the knowledge about location of faults should be known in advance (Lajaunie et al., 1997; De la Varga et al., 2019). In this study, we presented three scaling approaches as pre-processing steps in the fault detection procedure for subsurface horizons represented by triangulated models of homoclinal interfaces. As such, the analysis can be likened to 2.5D approaches for fault identification in seismic data (Tang et al., 2023; Dou et al., 2022). The conducted experiments revealed significant differences in the behaviour of the three tested data-scaling approaches, both for synthetic and real datasets. Although all methods rely on standardization, their design fundamentally affects how the model generalizes to unseen data.

6.1 Consistency of transformations and data leakage considerations

The main conceptual distinction between the examined approaches lies in whether the scaling transformation depends on the data used for model evaluation (test or real data).

In Method 1: “Independent subset scaling”, all synthetic surfaces are treated as realizations of a single, pooled dataset, and individual surfaces are not considered independent units. Under this assumption, a random train-test split is consistent with the modeling framework. Therefore, in the first method, each dataset (train, test, and real data) is scaled independently and individual surfaces are aggregated into a single data frame before scaling. As such, Method 1: “Independent subset scaling” can be likened to inter-subject approaches (or all-subject schema - (Apicella et al., 2023)) which create a global model which is valid for all synthetic surfaces (Arevalillo-Herráez et al., 2019). However, this design introduces an inconsistency between transformations, which, from the perspective of machine learning best practices, may lead to the lack of spatial continuity. We note that in this approach, the lack of spatial continuity follows purely from the construction of this method, i.e. different transformation parameters for the whole area and its subsets. However, to better summarize the differences in spatial additivity and continuity between three scaling methods, we provided a summary table in which we presented the number of changed labels (Tab. 1)

On the other hand, this version could be preferred if there is a degree of distrust in how the synthetic models represent true data. If this is the case, independent scaling of real data in the process could partially alleviate the exclusive reliance on synthetic models.

Method 2: “Per-surface local scaling” where each surface model is scaled separately before aggregation involves applying a “chronological” data split, where early surface models form the training set and later ones are reserved for testing. As such, this method can be likened to intra-subject approaches (or single-subject schema (Apicella et al., 2023)), where independent models are built for every subject (Arevalillo-Herráez et al., 2019).

Method 3: “Canonical ML scaling” where transformation parameters are derived exclusively from the training set - fully adheres to machine learning standards. This ensures that the model’s predictions are stable and independent of other observations present in the test data. However, as opposed to the first method, the classification tool generated in this approach

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

360 involves a high degree of confidence in how the synthetic models represent the reality. We note that in this approach, an aggregation of data is performed before scaling which shares similarities with inter-subject approaches, as it was the case with the first method (Arevalillo-Herráez et al., 2019).

Table 1. Comparison of three scaling methods in terms of the number of changed labels (Figs. 8d, 9d, 10d).

Method	Changed Labels	Changed Labels%	Stable (spatially additive) ?
Method 1	82	50.93	No
Method 2	53	32.92	No
Method 3	0	0	Yes

365

6.2 Spatial continuity and interpretability of predictions

Empirical results confirm that Method 3: “Canonical ML scaling” produces the most stable predictions. In this case, the scaling transformation is determined solely by the training data, which guarantees identical outcomes whether predictions are made for the entire dataset or only for a subset of points. By contrast, Method 1:”Independent subset scaling” and Method 2: “Per-
370 surface local scaling” produce different results depending on whether the model processes the complete dataset or a partial selection.

Visual inspection of classification maps suggests that Method 2: "Per-surface local scaling" often generates smoother and less noisy spatial patterns (Fig. 9c). This may make it attractive for structural geologists focused on tectonic relationships, as longitudinal and transverse fault systems appear more clearly. However, in several tested cases, this approach failed to detect
375 certain known faults.

6.3 Statistical considerations and model generalization

Statistical analyses and tests provide evidence that transformation parameters such as arithmetic means and variances for variables in train and test sets (after aggregation of the files, Figs. 2 and 4) are statistically equal (see Appendix B and Table B1). This implies that, given a complete dataset, classification results for the following methods should be very similar or even
380 equal for synthetic data: Method 1: "Independent subset scaling" and Method 3: "Canonical ML scaling". However, this high degree of similarity of classification results should only be expected for synthetic data. If real (production) data are considered, Method 1: "Independent subset scaling" requires re-calculation of the transformation parameters for real data and the classification results exhibit slight differences (compare Fig. 8a with Fig. 10a).

We note that from a generalization standpoint, Method 1: "Independent subset scaling" and Method 2: "Per-surface local
385 scaling" may yield inconsistent scaling between synthetic and real datasets, potentially limiting transferability. In contrast, Method 3: "Canonical ML scaling" enforces a single, fixed transformation learned from training data, which supports the

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

concept of generalization – defined as the ability of the model to adapt properly to new, unseen data drawn from the same distribution.

6.4 Geological and practical implications

In geological interpretation, Method 3: "Canonical ML scaling" can be seen as the most reliable for identifying fault structures, since it provides stable predictions even when analyzing local fragments of the study area and ensures spatial continuity of fault traces. Therefore, we recommend this method in projects that need the same results for the whole area and its parts.

Method 2: “Per-surface local scaling”, while potentially advantageous for visualizing regional relationships (Fig. 9c), may produce artifacts if local data characteristics differ substantially from those of the training set. The choice of scaling strategy should therefore depend on the intended use: for exploratory structural analysis, smoother results Method 2: "Per-surface local scaling" might be preferred, whereas for engineering-based predictive models used in real-world applications, Method 3: "Canonical ML scaling" offers better reproducibility and interpretability.

However, we note that conclusions from this study should not be extrapolated to other data sets, in particular to data with different orientation distribution.

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

6.5 Limitations and formal presentation of the effects

Throughout the paper, we used only one fixed set of SVM hyperparameters to compare the three scaling methods to allow more direct comparisons, In theory, different data scaling strategies directly alter the feature distribution, and the optimal hyperparameter settings should vary with the scaling method.

We note that currently our approach does not clearly demonstrate the performance of the proposed methods specifically on real sparse borehole-derived data. Geophysical methods are required to confirm or reject the suggestions proposed by the algorithm, which requires a separate study.

The core issue with Methods 1 and 2 is the lack of spatial additivity for classification: the global model function F applied to the entire study area $S = \bigcup s_n$ is not equivalent to the sum (re-integration) of local model functions f_n applied to individual sub-areas s_n . Formally:

$$F(S) = F\left(\bigcup s_n\right) \neq \bigcup f_n\,(s_n)$$

For example, in our case (Figs. 5 and 6-8), we had two subsets: S_1 and $S_2 = S \backslash S_1$, with the entire study area $S = S_1 \cup S_2$. Then $F(S_1 \cup S_2) \neq f_1(s_1) \cup f_2(s_2)$, as confirmed in Figs. 8d and 9d against Figs. 8a and 9a with the non-empty set of „unstable" measurements.

This discrepancy arises because each local function f_n is derived from a feature space unique to its sub-area, leading to spatial discontinuities upon re-integration. The consequence of this effect can be the lack of spatial continuity of fault traces near the boundaries of the spatial subsets.

6.6 Bridging the research gap

In 3D geological modeling, machine-learning (ML) frameworks are extensively applied to 2D or 3D seismic data, to reconstruct structural architecture (Fernandes et al., 2025; Choi et al., 2025). For seismic images, fault detection is one of the most prominent problems, commonly posed as a binary classification or segmentation task (Wei et al., 2022; Wu et al., 2019b) or, less frequently, as a multiclass problem (Wu et al., 2019a). In these dense-data scenarios, deep Convolutional Neural Networks (CNNs) serve as the primary tools for classification.

In contrast, while adhering to the binary classification concept for fault detection, this study addresses the challenge of data processing in structural modeling from sparse data. borehole networks and 2.5D vector-based Triangulated Irregular Networks (TIN). Existing 2D or 2.5D fault detection methods typically leverage potential field geophysical data, such as gravity and magnetic anomalies (Xu and Green, 2023), without integrating regional geometric information. For example, where borehole data has been previously utilized for ML-driven fault detection, models have relied primarily on physical borehole logging observations-such as identifying fault breccia-rather than analyzing structural elevations within a geometric context (Xu et al., 2024).

Most importantly, this study bridges a critical gap regarding methodological rigor in geoscientific ML applications. To the best of our knowledge, no comprehensive analysis currently exists regarding data-preprocessing pitfalls, such as data leakage

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

or spatial non-additivity, within geomodeling literature. While critical evaluations of data leakage have triggered an active reproducibility crisis debate in medical and biological fields (Kapoor and Narayanan, 2023), the Earth sciences remain largely unexamined. Remarkably, even in notable reviews of seismic ML workflows (Fernandes et al., 2025), data leakage is not recognized as a systemic challenge. This lack of discussion leaves an open question: is the geoscientific field truly immune to these specific machine-learning pitfalls, or is the topic severely underestimated and under-explored? By exposing how standard data-scaling practices can introduce spatial artifacts and disrupt the incremental updating of geological models, this work provides a vital methodological advancement toward more reproducible, stable, and geologically sound workflows.

7 Conclusions

The results of this study demonstrate that automated detection of tectonic discontinuities is strongly dependent on how data preprocessing preserves the spatial continuity of fault traces. In particular, different scaling strategies may lead to substantially different interpretations of fault geometry, especially when integrating local data subsets into regional models.

- In the context of the study, it was found that a specific classification tool preserves spatial continuity of detected faults only if it preserves spatial additivity, that is if the global model function F applied to the entire study area $S = \cup s_n$ is equivalent to the sum (re-integration) of local model functions f_n applied to individual sub-areas s_n.
- It was found that only the canonical scaling approach (Method 3) ensures the structural integrity of the fault network when re-integrating localized data segments into the regional model because the transformation (scaling) parameters do not change after training. This is particularly important for the incremental updating of geological models, as it enables the seamless integration of new borehole data without requiring a complete recalculation of the global feature space. This is because transformation parameters were calculated at the training stage and are not re-calculated at later stages. While local re-triangulation is still necessary to maintain geometric accuracy, the use of a fixed canonical scale prevents the emergence of boundary artifacts, such as artificial offsets or misinterpretation of fault terminations.
- We note that while we used only SVM algorithm with specific hyperparameters, the conducted analysis may be applicable for a variety of algorithms that are not scale-invariant (e.g. K-nearest neighbors).
- The practical significance of this approach is evident in applications involving high hydrogeological pressure. In mining environments, especially those located above confined aquifers, accurate representation of fault continuity is critical. In settings characterized by a homoclinal structure, where groundwater flow is strongly controlled by the regional dip of layers, reliable identification of faults plays a key role in assessing hydraulic connectivity. In this context, differences arising from alternative preprocessing strategies may influence the interpretation of fault continuity and, consequently, affect the evaluation of hydrogeological conditions, including the potential risk of water inrush.

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

- Despite these advances, the current approach is limited by the use of a local neighborhood of three adjacent faces within the TIN structure, which may not fully capture the multi-scale complexity of tectonic deformation. Future work should therefore focus on incorporating higher-order spatial relationships to improve the robustness of fault detection in complex geological settings.

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

Appendices

Appendix A

Results for synthetic data

In this section, we present more granular results of the classification method. The following metrics were used to evaluate the model: $precision = \frac{true\ positive}{true\ positive + false\ positive}$ and $recall = \frac{true\ positive}{true\ positive + false\ negatives}$.

Therefore, precision is maximized if there are zero false positives and the recall is maximized when there are zero false negatives. Based on these definitions the harmonic mean of both can be defined as follows:

$$F_1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}} = 2 * \frac{precision * recall}{precision + recall}.$$

Method 1: " Independent subset scaling"– results for synthetic data

Table A1. Confusion matrix for the classification task

2982 (true negatives)	163 (false positives)
134 (false negatives)	2927 (true positives)

Table A2. Results for the classification of test data (unseen surfaces) for arbitrarily selected hyperparameters

Class	Precision	Recall	F1-score
Non-fault	0.96	0.95	0.95
Fault	0.95	0.96	0.95

Method 2: "Per-surface local scaling"– results for synthetic data

Table A3. Confusion matrix for the classification task

2946 (true negatives)	59 (false positives)
109 (false negatives)	2996 (true positives)

Table A4. Results for the classification of test data (unseen terrains) for arbitrarily selected hyperparameters

Class	Precision	Recall	F1-score
Non-fault	0.96	0.98	0.97
Fault	0.98	0.96	0.97

Method 3: "Canonical ML scaling" – results for synthetic data

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

Table A5. Confusion matrix for the classification task

2993 (true negatives)	170 (false positives)
123 (false negatives)	2920 (true positives)

Table A6. Results for the classification of test data (unseen surfaces) for arbitrarily selected hyperparameters

Class	**Precision**	**Recall**	**F1-score**
Non-fault	0.96	0.95	0.95
Fault	0.94	0.96	0.95

Appendix B

Statistical analysis of transformation parameters

Statistical analyses were conducted for 24 independent pairs (train and test data) of samples, resulting in a total of 24 comparisons. Each pair consisted of independent samples of sizes $n_1 = 6205$ and $n_2 = 18615$. Differences in means between the two groups were assessed using Welch's t-test, which does not assume equal variances. Given the very large sample sizes, the sampling distribution of the difference in means can be approximated by a normal distribution according to the Central Limit Theorem; therefore, the test statistic was interpreted as a z-score compared to the standard normal distribution (Dowdy et al., 2004). No formal tests of normality were conducted, as they are unnecessary for datasets of this size.

The results of the 24 independent comparisons are presented in Table B1, which reports the p-values obtained from each Welch's t-test (the penultimate column). For each comparison, the null hypothesis stated that the population means of the two samples were equal. In all cases, the resulting p-values exceeded the 0.05 significance level, indicating no statistically significant differences between the compared samples. As an additional diagnostic step, equality of variances was examined for each of the 24 comparisons using the Brown–Forsythe test; none of the tests indicated statistically significant differences in variances (all p-values > 0.05, last column in Table B1).

Table B1. Results of the comparisons of transformation parameters.

STATISTICAL VARIABLE	X_test		X_train		p-value for the Welch's t-test	p-value for the Brown–Forsythe test
	sample mean	sample standard deviation	**sample mean**	sample standard deviation		
Angle_D_Min	**9.103**	12.101	**9.061**	11.902	**0.8115836**	0.6719061
Angle_D_Max	**38.121**	26.921	**37.973**	27.000	**0.7087739**	0.6117569
Angle_D_Intermediate	**20.440**	19.547	**20.526**	19.493	**0.7638799**	0.8848307

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

Angle_N_Min	**0.782**	1.392	**0.750**	1.303	**0.1153381**	0.0685081
Angle_N_Max	**2.954**	5.245	**2.878**	4.918	**0.3164987**	0.2815999
Angle_N_Intermediate	**1.507**	2.158	**1.486**	2.044	**0.5187470**	0.3407576
Cosine_D_Min	**0.051**	0.137	**0.049**	0.135	**0.4974963**	0.4853488
Cosine_D_Max	**0.523**	0.652	**0.521**	0.654	**0.8823433**	0.9769524
Cosine_D_Intermediate	**0.182**	0.335	**0.181**	0.328	**0.8470950**	0.7852282
Cosine_N_Min	**0.000**	0.002	**0.000**	0.002	**0.1448928**	0.1494271
Cosine_N_Max	**0.005**	0.036	**0.005**	0.034	**0.3045782**	0.288059
Cosine_N_Intermediate	**0.001**	0.005	**0.001**	0.005	**0.2836416**	0.2518479
Euclidean_D_Min	**0.183**	0.261	**0.181**	0.257	**0.7266988**	0.6014828
Euclidean_D_Max	**0.798**	0.639	**0.795**	0.640	**0.7504548**	0.8709219
Euclidean_D_Intermediate	**0.414**	0.438	**0.415**	0.434	**0.7700155**	0.9582994
Euclidean_N_Min	**0.014**	0.024	**0.013**	0.023	**0.1151902**	0.0683088
Euclidean_N_Max	**0.051**	0.089	**0.050**	0.083	**0.3159479**	0.2792964
Euclidean_N_Intermediate	**0.026**	0.038	**0.026**	0.036	**0.5202675**	0.3415199
X_D	**0.430**	0.563	**0.428**	0.561	**0.8029139**	0.7754447
X_N	**0.015**	0.038	**0.014**	0.036	**0.3283184**	0.384081
Y_D	**0.446**	0.545	**0.437**	0.555	**0.2553933**	0.1729
Y_N	**0.015**	0.035	**0.015**	0.035	**0.2588993**	0.7661097
Z_D	**-0.041**	0.037	**-0.041**	0.036	**0.4657109**	0.3102891
Z_N	**0.998**	0.005	**0.999**	0.004	**0.3243195**	0.2350856

Code availability

Name of code: SubsurfaceBreaks_DataScalingMethods. **License**: GNU General Public License v3.0. **Developer**: Adam Lewiński. **Contact address**: AGH University of Krakow, Poland. E-mail: lewinski@student.agh.edu.pl. **Year first available**: 2026. **Hardware required**: the computer code was run on a PC with Intel(R) Core(TM) i9-14900KF (3.20 GHz), 64 GB RAM. **Software required**: CGAL library (v. 4.8) **Program language**: C++, Python. **Program size**: 2,11 MB. **How to access the source code**: https://github.com/lewi9/SubsurfaceBreaks_DataScalingMethods/blob/main/README.md (last access: 08

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

January 2026). A reader is also advised to visit a Zenodo repository with information about generating synthetic models (Michalak, 2025).

Data availability

Datasets for this research (input and processed data) are available in these intext data citation references (Michalak, 2024).

Author contribution

Conceptualization: AL (mainly), MM, Data curation: MM, AL., Formal analysis: AL, MM, AK, Investigation: AL, MM, PL, Methodology: AL (mainly), MM, Project administration: MM, Resources: MM, Software: AL (mainly), MM, Supervision: MM, Validation: AL., MM, Visualization: AL., MM, PL, Writing - original draft: AL., MM, AK, Writing – review and editing: AL., MM, AK, PL.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

We are grateful to Museum of Częstochowa (Archive of the History Department) for sharing the geological documentation. We used AI methods (chatGPT, Gemini) to improve readability of the manuscript.

References

Apicella, A., Isgrò, F., Pollastro, A., and Prevete, R.: On the effects of data normalization for domain adaptation on EEG data, Eng. Appl. Artif. Intell., 123, https://doi.org/10.1016/j.engappai.2023.106205, 2023.

Arevalillo-Herráez, M., Cobos, M., Roger, S., and García-Pineda, M.: Combining inter-subject modeling with a subject-based data transformation to improve affect recognition from EEG signals, Sensors (Switzerland), 19, https://doi.org/10.3390/s19132999, 2019.

Bardziński, W., Lewandowski, J., Więckowski, R., and Zieliński, T.: Objaśnienia do Szczegółowej Mapy Geologicznej Polski w skali 1:50000, ark. Częstochowa (845), Wydawnictwa Geologiczne, Warszawa, Poland, 72 pp., 1986.

De Berg, M., Cheong, O., Van Kreveld, M., and Overmars, M.: Computational Geometry: Algorithms and Applications, 3rd Ed., Springer, 364 pp., https://doi.org/10.2307/3620533, 2008.

Bishop, C. M.: Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, Berlin,

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

Heidelberg, 2006.

Choi, W., Pyun, S., and Jou, H.-T.: Synthetic Training Data Optimization for Enhanced Fault Detection in Seismic Images, Lithosphere, 2025, https://doi.org/10.2113/2025/lithosphere_2024_240, 2025.

Dayczak-Calikowska, K., Kopik, J., Maliszewska, A., and Marcinkiewicz, T.: Middle Jurassic. In: Marek, S., Pajchlowa M. (Eds), The Epicontinental Permian and Mesozoic in Poland., Pr. Państwowego Inst. Geol., 153, 236–282, 1997.

Deczkowski, Z.: Budowa geologiczna pokrywy permsko-mezozoicznej i jej podłoża we wschodniej części monokliny przedsudeckiej (obszar kalisko-częstochowski), Pr. Inst. Geol., 82, 1–63, 1977.

Dou, Y., Li, K., Zhu, J., Li, X., and Xi, Y.: Attention-Based 3-D Seismic Fault Segmentation Training by a Few 2-D Slice Labels, IEEE Trans. Geosci. Remote Sens., 60, https://doi.org/10.1109/TGRS.2021.3113676, 2022.

Dowdy, S., Weardon, S., and Chilko, D.: Statistics for research: Third Edition, Third Edit., John Wiley & Sons, Hoboken, New Jersey, 1-633 pp., https://doi.org/10.1002/0471477435, 2004.

Doyoro, Y. G., Gelena, S. K., and Lin, C. P.: Improving subsurface structural interpretation in complex geological settings through geophysical imaging and machine learning, Eng. Geol., 344, https://doi.org/10.1016/j.enggeo.2024.107839, 2025.

Fernandes, G. L., Figueiredo, F., Hatushika, R. S., Leão, M. L., Mariano, B. A., Monteiro, B. A. A., de Cerqueira Oliveira, F. T., Panoutsos, T., Pires, J. P., Poppe, T. M., and Zavam, F.: A systematic review of deep learning for structural geological interpretation, Data Min. Knowl. Discov., 39, 1–56, https://doi.org/10.1007/s10618-024-01079-y, 2025.

Gao, H., Wu, X., and Liu, G.: ChannelSeg3D: Channel simulation and deep learning for channel interpretation in 3D seismic images, Geophysics, 86, IM73-IM83, https://doi.org/10.1190/geo2020-0572.1, 2021.

Gayrin, P., Wrona, T., Brune, S., Neuharth, D., Molnar, N., La Rosa, A., and Naliboff, J.: Fatbox: the fault analysis toolbox, Solid Earth, 17, 555–572, 2026.

GDDKIA ODDZIAŁ W KATOWICACH: Album dotyczy budowy autostrady A1 na odocinkach G, H oraz I, Katowice, 2019.

Górecka, E.: Geological setting of the Silesian-Cracow Zn-Pb deposits, Geol. Q., 37, 127–146, 1993.

Górniak, E.: Analiza wyników pomiarów dopływów wody do kopalni za 1969 r. (Kopalnia Szczekaczka). [Analysis of the results of measurements of water inflows to the mine for 1969 (Szczekaczka Mine)-Unpublished - Museum of Częstochowa (Archive of the History Department)], Częstochowa, Poland, 1969.

Górniak, E.: Dokumentacja zagrożeń wodnych kopalni “Żarki” [Documentation of water hazards in the “Żarki” mine - Unpublished - Museum of Częstochowa (Archive of the History Department)], Osiny, Poland, 1974.

Górniak, E. and Sałaciński, T.: Przekrój szybu głównego Kopalni “Jerzy-Malice” 1:500 [Cross-section of the main shaft of the mine “Jerzy-Malice” - Unpublished - Museum of Częstochowa (Archive of the History Department)], Częstochowa, Poland, 1981.

Hermański, S.: Wpływ prac odwadniających kopalnictwa rud żelaza na kształtowanie warunków hydrogeologicznych w rejonie częstochowsko-kłobuckim., Rudy Żelaza, 9–10, 13–16, 1971.

Hermański, S.: Mapa stropu i miąższości warstw kościeliskich. Rejon Żarki-Wieluń. Skala 1:100000. Centralne Archiwum Geologiczne., Warszawa, Poland, 1993.

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

Hu, Y., Wang, Z. Z., Guo, X., Kek, H. Y., Ku, T., Goh, S. H., Leung, C. F., Tan, E., and Zhang, Y.: Three-dimensional reconstruction of subsurface stratigraphy using machine learning with neighborhood aggregation, Eng. Geol., 337, https://doi.org/10.1016/j.enggeo.2024.107588, 2024.

Ji, G., Wang, Q., Zhou, X., Cai, Z., Zhu, J., and Lu, Y.: An automated method to build 3D multi-scale geological models for engineering sedimentary layers with stratum lenses, Eng. Geol., 317, 107077, https://doi.org/10.1016/j.enggeo.2023.107077, 2023.

Kapoor, S. and Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science, Patterns, 4, https://doi.org/10.1016/j.patter.2023.100804, 2023.

Karnkowski, P. H.: Regionalizacja tektoniczna Polski - Niż Polski, Prz. Geol., 56, 895–903, 2008.

Kaur, H., Zhang, Q., Witte, P., Liang, L., Wu, L., and Fomel, S.: Deep-learning-based 3D fault detection for carbon capture and storage, Geophysics, 88, IM101--IM112, https://doi.org/10.1190/geo2022-0755.1, 2023.

Kieńć, D., Wąsik, M., Śliwka, R., Mżyk, S., Firlit, G., Gawron, M., Kuczer, M., and Ruszkiewicz, P.: Dokumentacja określająca warunki hydrogeologiczne dla ustanowienia obszarów ochronnych zbiornika wód podziemnych Częstochowa /W/ - GZWP nr 325, pow. wieluński, oleski, kłobucki, częstochowski, myszkowski, woj. łódzkie, opolskie, śląskie, Wrocław, Poland, 923 pp., 2008.

Kopik, J.: Lower and Middle Jurassic of the north-eastern margin of the Upper Silesian Coal Basin (in Polish with English summary), Biul. Państwowego Inst. Geol., 378, 67–129, 1998.

Krokowski, J.: Mezoskopowe studia strukturalne w osadach permsko-mezozoicznych południowo-wschodniej części Wyżyny Śląsko-Krakowskiej, Ann. Soc. Geol. Pol., 54, 79–121, 1984.

De la Varga, M., Schaaf, A., and Wellmann, F.: GemPy 1.0: open-source stochastic geological modeling and inversion, Geosci. Model Dev., 12, 1–32, https://doi.org/10.5194/gmd-12-1-2019, 2019.

Lajaunie, C., Courrioux, G., and Manuel, L.: Foliation fields and 3D cartography in geology: Principles of a method based on potential interpolation, Math. Geol., 29, 571–584, https://doi.org/10.1007/bf02775087, 1997.

Lin, L., Zhong, Z., Li, C., Gorman, A., Wei, H., Kuang, Y., Wen, S., Cai, Z., and Hao, F.: Machine learning for subsurface geological feature identification from seismic data: Methods, datasets, challenges, and opportunities, Earth-Science Rev., 257, https://doi.org/10.1016/j.earscirev.2024.104887, 2024.

Matyja, B. and Wierzbowski, A.: Ammonites and stratigraphy of the uppermost Bajocian and Lower Bathonian between Częstochowa and Wieluń, Central Poland, Acta Geol. Pol., 50, 191–209, 2000.

Matyszkiewicz, J., Kochman, A., Rzepa, G., Gołębiowska, B., Krajewski, M., Gaidzik, K., and Żaba, J.: Epigenetic silicification of the Upper Oxfordian limestones in the Sokole Hills (Kraków-Częstochowa Upland): Relationship to facies development and tectonics, Acta Geol. Pol., 65, 181–203, https://doi.org/10.1515/agp-2015-0007, 2015.

Michalak, M.: michalmichalak997/SubsurfaceBreaks: Updated Manuscript Integration with Zenodo (1.0.1)., https://doi.org/10.5281/zenodo.14660007, 2025.

Michalak, M. P.: SubsurfaceBreaks v. 1.0: A supervised detection of fault-related structures on triangulated models of

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

subsurface homoclinal interfaces: Input and Processed Data, https://doi.org/10.5281/zenodo.14589469, 2024.

Michalak, M. P., Bardziński, W., Teper, L., and Małolepszy, Z.: Using Delaunay triangulation and cluster analysis to determine the orientation of a sub-horizontal and noise including contact in Kraków-Silesian Homocline, Poland, Comput. Geosci., 133, 104322, https://doi.org/10.1016/j.cageo.2019.104322, 2019.

Michalak, M. P., Gerhards, C., and Menzel, P.: SubsurfaceBreaks v. 1.0: a supervised detection of fault-related structures on triangulated models of subsurface homoclinal interfaces, Geosci. Model Dev., 18, 4469–4481, https://doi.org/10.5194/gmd-18-4469-2025, 2025.

Mousavi, S. M. and Beroza, G. C.: Deep-learning seismology, Science (80-.)., 377, https://doi.org/10.1126/science.abm4470, 2022.

Muller, A. P. O., Costa, J. C., Bom, C. R., Faria, E. L., Klatt, M., Teixeira, G., De Albuquerque, M. P., and De Albuquerque, M. P.: Complete identification of complex salt geometries from inaccurate migrated subsurface offset gathers using deep learning, Geophysics, 87, R453–R463, https://doi.org/10.1190/geo2021-0586.1, 2022.

Oakley, D., Loiselet, C., Coowar, T., Labbe, V., and Callot, J. P.: GEOMAPLEARN 1.2: Detecting structures from geological maps with machine learning - The case of geological folds, Geosci. Model Dev., 18, 939–960, https://doi.org/10.5194/gmd-18-939-2025, 2025.

Osika, R., Pożaryski, W., Rühle, E., and Znosko, J.: Geological map of Poland without Cainozoic formations, 1:500 000, Wydawnictwa Geologiczne, Warszawa, Poland, 1972.

Pedregosa, F., Varoquaux, G., Gramfort, A., Vincent, M., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, https://doi.org/10.5555/1953048.2078195, 2011.

Pich, J. and Pokora, M.: Depressional cone in area of the Częstochowa-Kłobuck iron mines, Przegląd Geol., 30, 132–141, 1982.

Common pitfalls and recommended practices: https://scikit-learn.org/stable/common_pitfalls.html, last access: 15 December 2025.

StandardScaler: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html#sklearn.preprocessing.StandardScaler, last access: 15 December 2025.

Shalev-Shwartz, S. and Ben-David, S.: Understanding machine learning: From theory to algorithms, Cambridge University Press, 397 pp., https://doi.org/10.1017/CBO9781107298019, 2013.

Tang, Z., Wu, B., Wu, W., and Ma, D.: Fault Detection via 2.5D Transformer U-Net with Seismic Data Pre-Processing, Remote Sens., 15, https://doi.org/10.3390/rs15041039, 2023.

Vapnik, V. N.: The nature of statistical learning theory., Springer-Verlag, New York, 2000.

Wang, J.-A. and Park, H. D.: Coal mining above a confined aquifer, Int. J. Rock Mech. Min. Sci., 40, 537–551, 2003.

Wang, S., Si, X., Cai, Z., Sun, L., Wang, W., and Jiang, Z.: Fast Global Self-Attention for Seismic Image Fault Identification,

https://doi.org/10.5194/egusphere-2026-3591
Preprint. Discussion started: 7 July 2026
© Author(s) 2026. CC BY 4.0 License.

IEEE Trans. Geosci. Remote Sens., 62, https://doi.org/10.1109/TGRS.2024.3436066, 2024.

Wei, X. L., Zhang, C. X., Kim, S. W., Jing, K. L., Wang, Y. J., Xu, S., and Xie, Z. Z.: Seismic fault detection using convolutional neural networks with focal loss, Comput. Geosci., 158, https://doi.org/10.1016/j.cageo.2021.104968, 2022.

Więckowski, R.: Zanieczyszczenia wód wgłębnych Częstochowskiego Okręgu Przemysłowego. [Pollution of deep water in the Częstochowa Industrial District - Unpublished - Museum of Częstochowa (Archive of the History Department)], Częstochowa, Poland, 1973.

Wu, Q., Xu, H., and Zou, X.: An effective method for 3D geological modeling with multi-source data integration, Comput. Geosci., 31, 35–43, https://doi.org/10.1016/j.cageo.2004.09.005, 2005.

Wu, X., Shi, Y., Fomel, S., Liang, L., Zhang, Q., and Yusifov, A. Z.: FaultNet3D: Predicting Fault Probabilities, Strikes, and Dips with a Single Convolutional Neural Network, IEEE Trans. Geosci. Remote Sens., 57, 9138–9155, https://doi.org/10.1109/TGRS.2019.2925003, 2019a.

Wu, X., Liang, L., Shi, Y., and Fomel, S.: FaultSeg3D: Using synthetic data sets to train an end-to-end convolutional neural network for 3D seismic fault segmentation, Geophysics, 84, IM35-IM45, https://doi.org/10.1190/geo2018-0646.1, 2019b.

Wu, X., Geng, Z., Shi, Y., Pham, N., Fomel, S., and Caumon, G.: Building realistic structure models to train convolutional neural networks for seismic structural interpretation, Geophysics, https://doi.org/10.1190/geo2019-0375.1, 2020.

Xu, L. and Green, E. C. R.: Inferring geological structural features from geophysical and geological mapping data using machine learning algorithms, Geophys. Prospect., 71, 1728–1742, https://doi.org/10.1111/1365-2478.13371, 2023.

Xu, L., Green, E. C. R., and Kelly, C.: Inferring fault structures and overburden depth in 3D from geophysical data using machine learning algorithms – A case study on the Fenelon gold deposit, Quebec, Canada, Geophys. Prospect., 72, 3474–3494, https://doi.org/10.1111/1365-2478.13589, 2024.

Zhang, Z., Wang, G., Carranza, E. J. M., Liu, C., Li, J., Fu, C., Liu, X., Chen, C., Fan, J., and Dong, Y.: An integrated machine learning framework with uncertainty quantification for three-dimensional lithological modeling from multi-source geophysical data and drilling data, Eng. Geol., 324, https://doi.org/10.1016/j.enggeo.2023.107255, 2023.

Znosko, J.: Tektonika obszaru częstochowskiego, Przegląd Geol., 8, 418–424, 1960.