Geoscience Letters

RESEARCH LETTER

Open Access

A machine learning-based approach for constructing a 3D apparent geological model using multi-resistivity data

Jordi Mahardika Puntu[1], Ping-Yu Chang[1,2,3*], Haiyina Hasbia Amania[1], Ding-Jiun Lin[1], M. Syahdan Akbar Suryantara[1], Jui-Pin Tsai[4], Hwa-Lung Yu[4], Liang-Cheng Chang[5], Jun-Ru Zeng[1] and Lingerew Nebere Kassie[1,6]

Abstract

This study presents a comprehensive approach for constructing a 3D Apparent Geological Model (AGM) by integrating multi-resistivity data using statistical methods, supervised machine learning (SML), and Python-based modeling techniques. Demonstrated through a case study in the Choushui River Alluvial Fan (CRAF) in Taiwan, the methodology enhances data coverage significantly, from 62 to 386 points, by incorporating resistivity data sets from Vertical Electrical Sounding (VES), Transient Electromagnetic (TEM), and borehole information. A key contribution of this work is the rigorous harmonization of these data sets, ensuring consistent resistivity values across different methods before constructing the 3D resistivity model, addressing a gap in previous studies that typically handled these data sets separately, either building models individually or comparing results side-by-side without fully integrating the data. Furthermore, python-based modeling and radial basis function interpolation were employed to construct the 3D resistivity model for greater flexibility and effectiveness than conventional software. Subsequently, this model was transformed into a 3D AGM using the SML technique. Four algorithms, namely, random forest (RF), decision tree (DT), support vector machine (SVM), and extreme gradient boosting (XGBoost) were implemented. Following evaluation via confusion matrix analysis, evaluation metrics, and examination of receiver operating characteristics curve, it emerged that the RF algorithm exhibits superior performance when applied to our multi-resistivity data set. The results from the 3D AGM unveil distinct resistivity anomalies correlated with sediment types. The clay layer exhibited low resistivity ($\leq 59.98\ \Omega m$), while the sand layer displayed medium resistivity ($59.98 < \rho < 136.14\ \Omega m$), and the gravel layer is characterized by high resistivity ($\geq 136.14\ \Omega m$). Notably, in the proximal fan, gravel layers predominate, whereas the middle fan primarily consists of sandy clay layers. Conversely, the distal fan, located in the western coastal area, predominantly comprises clayey sand. To conclude, the findings of this study provide valuable insights for researchers to construct the 3D AGM from the resistivity data, applicable not only to the CRAF but also to other target areas.

Keywords 3D apparent geological model, 3D resistivity model, Supervised machine learning, Python-based modeling, Choushui river alluvial fan

*Correspondence:
Ping-Yu Chang
pingyuc@ncu.edu.tw
Full list of author information is available at the end of the article

Introduction

A comprehensive understanding of subsurface lithology distribution holds significant importance for geoscientists, as it serves as the foundation for diverse applications, spanning from resource exploration to environmental protection. Traditionally, this distribution is derived from borehole data; however, the high costs associated with borehole acquisition pose challenges for covering extensive areas like basins or alluvial fans, limiting detailed spatial lithology distribution of the area. To address this challenge, our approach integrates multi-resistivity data obtained from geophysical measurements. This data is then utilized to construct a comprehensive three-dimensional (3D) resistivity model. Subsequently, machine learning (ML) techniques are employed to transform this model into a 3D apparent geological model (AGM), enabling a thorough lithology distribution analysis. In this case, these approaches are demonstrated in the Choushui river alluvial fan (CRAF) in Taiwan.

This area has been extensively studied, motivated not only by its designation as one of the major groundwater basins in Taiwan but also due to the presence of critical transportation infrastructure, such as the Taiwan high-speed rail (THSR), which crosses the subsidence zone in the CRAF. Land subsidence poses a significant risk to the THSR, as uneven ground settling can lead to infrastructure damage and potential operational disruptions (Chen et al. 2021; Hsu 1998; Huang et al. 2024). Previous studies have examined the role of subsurface conditions in this phenomenon. For instance, Liu et al. (2001) investigated the effect of clay dehydration on land subsidence in the coastal area of the CRAF. Liu et al. (2004) used leveling surveys and groundwater monitoring wells to investigate subsidence, linking layer compression to groundwater extraction, particularly in clayey and sandy layers. Hung et al. (2009) suggests that the land subsidence in the region is linked to the compaction of clay materials at various depths, where 70% of subsidence (> 3 cm/year) occurred along the THSR route, with a peak rate of 8.2 cm/year Lu et al. (2016). Used kriging to analyze subsidence trends, noting a shift from coastal to central areas Chu et al. (2021). Developed a spatial regression model to map subsidence bowls. The study revealed that the subsidence bowl was found in the inland area of Yunlin, which was consistent with the observed subsidence bowl location. Based on these studies, it can be concluded that subsidence is strongly correlated with subsurface compaction, particularly in clay-rich layers, since clays are known to be highly compressible, especially when affected by groundwater extraction and changes in moisture content, which makes them a significant factor in subsidence processes (Lin et al. 2016; Liu et al. 2001). Thus, understanding the distribution of these sediments, especially in a 3D context, is essential for effective research, monitoring, and management.

Despite extensive research, a significant gap remains in developing a comprehensive 3D Apparent Geological Model (AGM) for the subsurface in this area. Current subsurface models rely heavily on 1D resistivity data from borehole records, which are spaced between 1 and 17 km apart, limiting their ability to capture detailed subsurface features, particularly in large areas like the CRAF (Cheng & Hsu 2021; GSMMA 2023). While borehole data provide precise lithological information at specific points, they lack the spatial resolution needed to detect lateral variations, leaving critical gaps in understanding the subsurface. Several attempts have been made to utilize resistivity data in studying this area. For example, Yang and Lee (2002) mapped apparent resistivity using direct current resistivity sounding, and Kassie et al. (2023) used 1D Transient Electromagnetic (TEM) data to explore subsurface structures for hydrogeological purposes. However, these studies fell short of constructing a comprehensive 3D resistivity model, stopping at distribution mapping without fully integrating resistivity data with borehole information to link resistivity values with lithology. In terms of 2D models, the existing 2D hydrogeological models, such as those provided by GSMMA, rely on sparsely distributed borehole data and often use manually drawn cross sections to connect points between profiles. This manual process can lead to misalignments between cross sections, further reducing the accuracy of the models. The limitations of both 1D and 2D approaches highlight the need for a more advanced 3D model that can accurately represent both vertical and lateral subsurface variations (Chiang 1999; GSMMA 2023). By incorporating additional geophysical measurements and developing a comprehensive 3D AGM, we aim to address these challenges and provide a more detailed and consistent representation of lithological distribution across the region.

Constructing the 3D model is crucial, as it offers several advantages compared to one-dimensional (1D) and 2D models. Specifically, it enables the visualization of the geometric distribution of subsurface features, fully capturing the complexity of the geological subsurface. In addition, it greatly facilitates enhanced interpretation, providing a clearer understanding of the 3D spatial relationships among various types of subsurface data. Furthermore, when communicating with non-technical audiences, 3D models can often prove to be more effective than traditional maps or 2D cross sections (Aldiss et al. 2012; Rabeau et al. 2010; Witter and Melosh 2018).

Over time, several studies have utilized geophysical measurements, especially geoelectrical methods to construct 3D models, demonstrating their effectiveness in

different contexts. For instance, Cardarelli and De Donno (2017) employed 1D, 2D, and 3D electrical resistivity methods to estimate bedrock depth. The 3D resistivity model was constructed from four parallel 2D Electrical Resistivity Imaging (ERI). In the same year, Chabaane et al. (2017) combined both Vertical Electrical Sounding (VES) and 2D ERI methods for geothermal groundwater characterization. Osinowo and Falufosi (2019) conducted research aimed at examining the foundations before construction by creating a 3D model through the amalgamation of multiple 2D resistivity profiles. Arowoogun and Osinowo (2021) presented a case study on the use of a 3D resistivity model from 1D VES data to assess groundwater potential and aquifer protective capacity. A recent study by Abu Rajab et al. (2023) constructed a 3D resistivity model by interpolating 1D Transient Electromagnetic (TEM) data and combined it with ERI results to observe the impact of the seawater intrusion.

Despite various attempts outlined above to construct 3D models using geoelectrical methods, persistent challenges demand attention, including limited data coverage for extensive studies, software limitations in managing model boundaries, and the need for a method to transform the 3D resistivity model into a 3D AGM. To tackle the issue of limited data coverage, we initiated the collection of three additional measurements, such as VES, TEM, and Normal Borehole Resistivity (NBR) data from boreholes, along with lithological information, where data harmonization is performed to integrate these data sets. In response to the limitations of traditional software, characterized by its stiffness and tendency to generate square models even in data-sparse regions, we introduced a Python-based modeling and visualization approach to enhance model realism. Furthermore, to transform the 3D resistivity model into a 3D AGM, we shifted our focus from employing a conventional method solely reliant on direct visual comparison of borehole lithology sections with resistivity data to adopting a Supervised Machine Learning (SML) technique. This technique utilizes statistical methods and borehole information as ground truth data, offering a more advanced approach. Although this technique has been successfully applied in various studies using geophysical data sets (Bressan et al. 2020; Dong et al. 2023; Kumar et al. 2022; Marzán et al. 2021; Piegari et al. 2023; Puntu et al. 2023; Puntu et al. 2021; Tilahun & Korus 2023), very few studies have implemented this approach for 3D modeling. Therefore, our study aims to fill this gap by applying it to our 3D data set. In this case, we utilized four SML algorithms, including decision tree (DT), random forest (RF), support vector machine (SVM), and extreme gradient boost (XGBoost), and compared them to identify the most suitable algorithm for our data sets.

In summary, the objective of this study is to build a 3D AGM by integrating multi-resistivity data and utilizing statistical methods, machine learning techniques, and Python-based modeling and visualization tools marking a transition from traditional methodologies to more advanced approaches.

Materials and methods

Background of the study area

The present study is located in the Choushui River Alluvial Fan (CRAF) in Central Taiwan, as depicted in Fig. 1. The study area is demarcated by a bold red line. The CRAF covers approximately 2000 km^2 of the Changhua and Yunlin Counties and is divided into three sections: proximal fan, mid-fan, and distal fan. This area is a Holocene deposit consisting of clay, sand, and gravel (Hung et al. 2009; Liu et al. 2002). Figure 2 shows a conceptual hydrogeological profile (**A–A'**) parallel to the Choushui River, it was constructed from the available borehole data from the Haifeng borehole on the west to the Chukou borehole on the east. This conceptual model was obtained from the Geological Survey and Mining Management Agency (GSMMA) of Taiwan. The western section is near the coastal area, whereas the eastern area with higher altitude is near the Douliu Hill and Bagua Tableland. In general, it shows that the proximal fan is mainly composed of gravel and sand with high permeability, whereas the middle and distal fans are primarily composed of clay and fine sand with low permeability. The CRAF is located at the interface between the mountains and the coastal plain and surrounded by natural geographical boundaries, including the Wu River to the north, the Bagua Tableland and the Douliu Hill to the east, the Beigang River to the South, and the Taiwan Strait to the west (GSMMA 2023).

The geology of the eastern CRAF is complex and includes various formations, such as terrace deposits, lateritic terrace deposits, the Nanchuang Formation (alternation of sandstone and shale), the Cholan Formation (interbeds of sandstone, mudstone, and shale), the Toukoshan Formation (TF), and the Kueichulin Formation (KF). The TF can be divided into four types: TFa–TFd, as shown in Fig. 1. TFa consists of gravel with sandstone lentils intercalated with thick-bedded sandstone with mudstone and gravel lentils, TFb consists of gravel with sandstone lentils, TFc consists of sandstone, mudstone, and shale, and TFd consists of sandstone, siltstone, mudstone, and sandstone interbedded with mudstone. The KF can be divided into three types: KFa–KFc, as shown in Fig. 1. KFa is composed of muddy sandstone with shale, KFb is composed of shale with intercalated sandstone, and KFc is composed of muddy sandstone and alteration

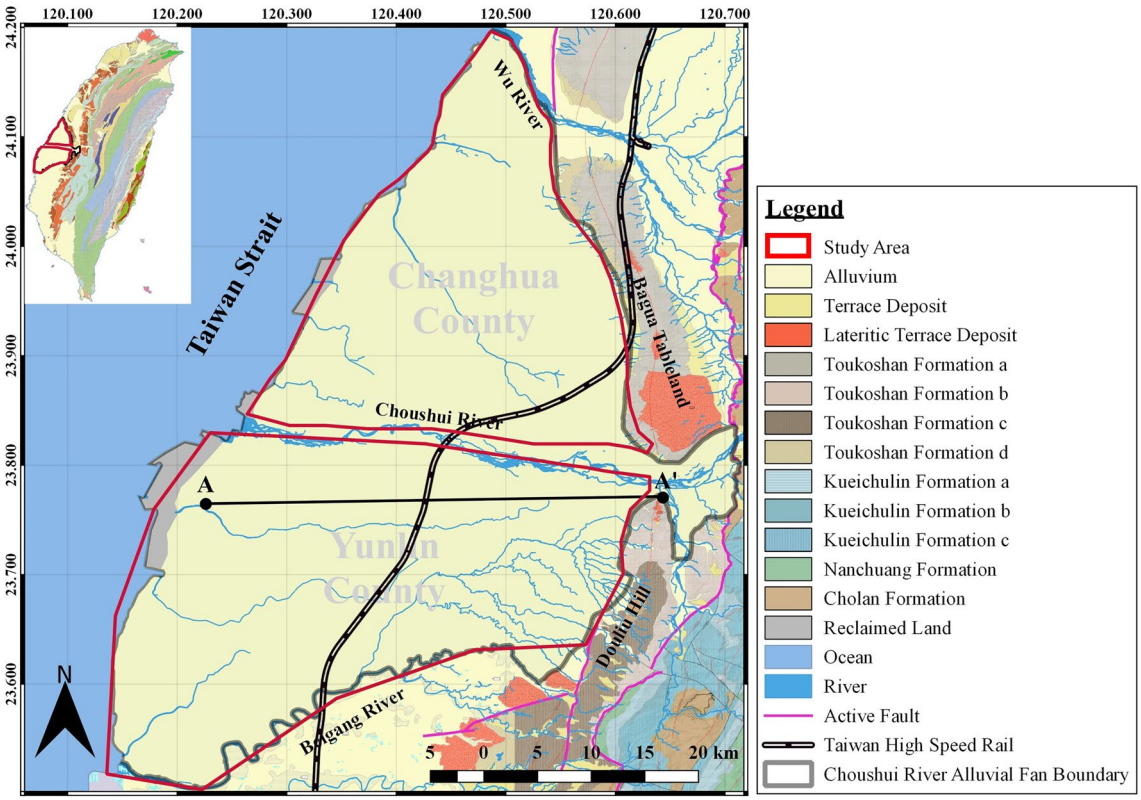

Fig. 1 Study area is situated in the Choushui River Alluvial Fan in central Taiwan, outlined by a bold red line. The CRAF encompasses around 2000 km2 within Changhua and Yunlin Counties (Modified from GSMMA (2023)

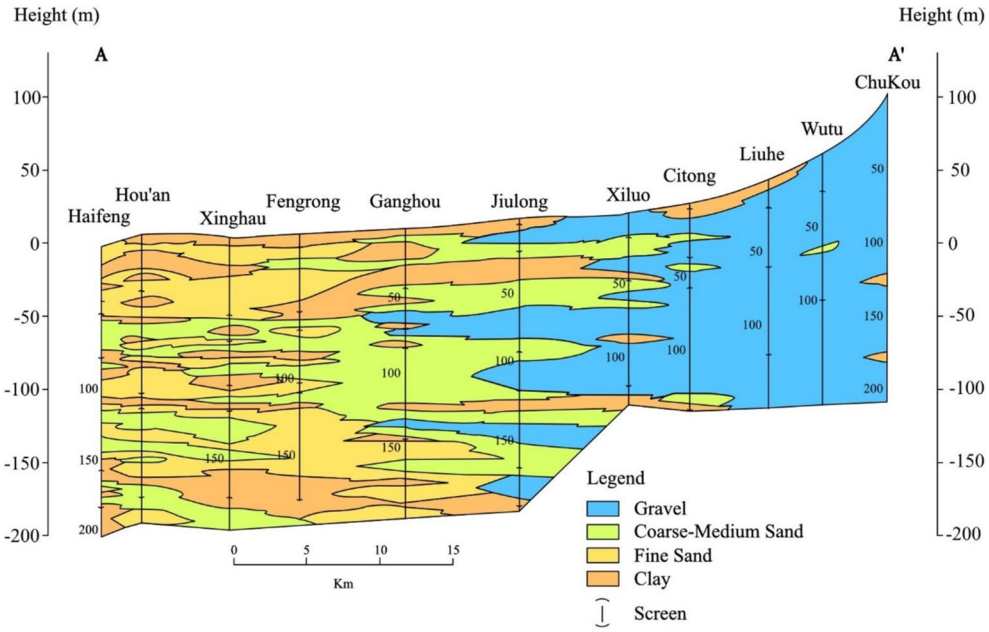

Fig. 2 Conceptual hydrogeological profile (**A**–**A'**), *redrawn from* GSMMA (2023)

of sandstone and shale (Chu et al., 2021; GSMMA 2023; Hung et al. 2009; Liu et al. 2002; Lu et al. 2020).

The geoelectrical data

Previous geological and hydrogeological surveys in the CRAF have primarily relied on individual data sets, such as borehole information, TEM or VES (Kassie et al. 2023; Tsai et al. 2019; Yang and Lee 2002), without integrating these data sets comprehensively. This fragmented approach limits spatial coverage and reduces the ability to accurately characterize subsurface heterogeneity. Specifically, without joint comparison, these surveys fail to resolve inconsistencies between different data sources, such as variations in resistivity measurements across methods. In this study, we utilized three geophysical methods: VES, TEM, and NBR. Without integration, significant geological features, such as lithological transitions, could be misinterpreted or overlooked. Joint interpretation allows for better calibration and validation of resistivity data against borehole lithological information, leading to more accurate and cohesive 3D Resistivity and Apparent Geological Models (AGMs). This shortcoming highlights the need for a multi-resistivity data integration approach, as proposed in this study, to provide a more accurate representation of subsurface conditions.

Figure 3 illustrates the distribution of geoelectrical measurements, where the study area is indicated by a red line, the THSR route by a bold black and white line, the township by a black pentagon, and the borehole (NBR), VES, and TEM by an orange circle, green triangle, and blue rectangle, respectively.

Normal borehole resistivity (NBR)

We obtained a total of 62 borehole data located in the southern region of the CRAF, which were retrieved from the GSMMA database (GSMMA 2023). From this data set, we exclusively selected 54 borehole records within the study area boundary. Subsequently, the Normal Borehole Resistivity (NBR) data was extracted and inverted to derive the true resistivity distribution, which served as the basis for constructing a 3D Apparent Geological Model (AGM). These borehole data sets were serving as the ground truth by providing both resistivity log data and sediment type information from the core samples.

Vertical electrical sounding (VES)

A total of 221 VES data points was derived from the GSMMA report within the study area (Dong et al. 1996; Tsai et al. 2019). To enhance the inversion result, we conducted a re-analysis of the VES raw data in python using

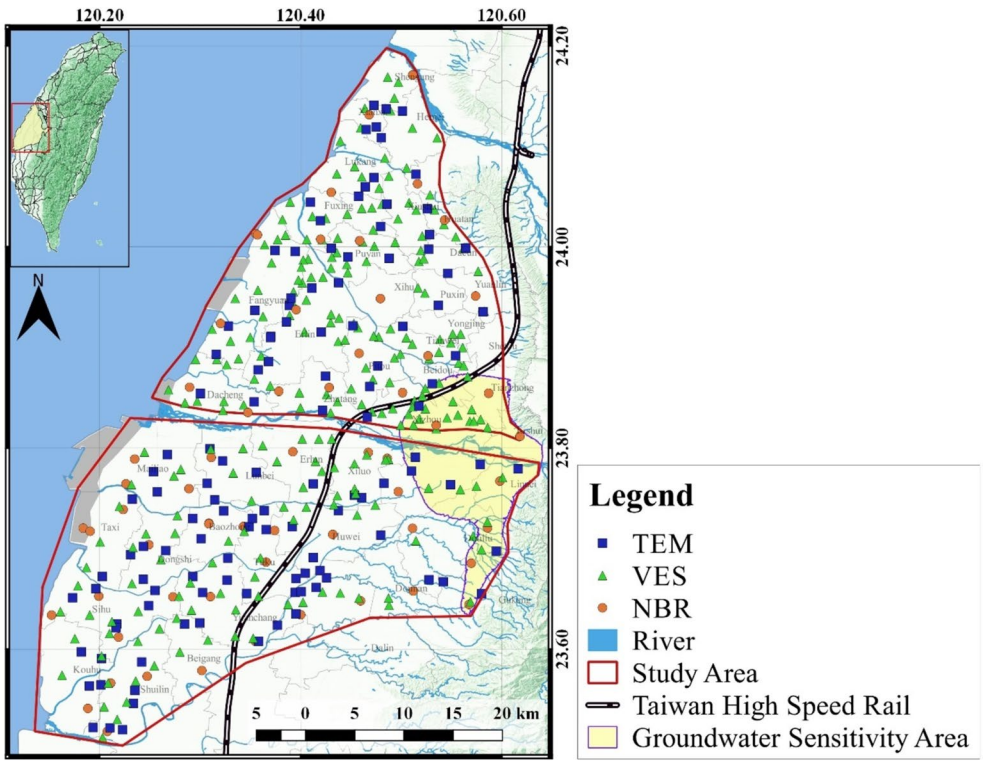

Fig. 3 Distribution of geoelectrical data within the study area, with boreholes, VES, and TEM represented by orange circles, green triangles, and blue rectangles, respectively

simPEG, an open source python package designed for geophysical applications that provides simulation and gradient-based parameter estimation. There are three main steps in simPEG, including inputs (Data, Uncertainty estimates, governing equation, and prior knowledge), inversion implementation (Forward simulation and inversion components), and evaluation (evaluate and assess results). The inverse analysis of VES data inherently involves uncertainties stemming from measurement noise, data quality, and the non-uniqueness of geophysical inversion results. In this study, we addressed these uncertainties by conducting a re-analysis of the VES raw data using the simPEG package in Python. SimPEG allows for uncertainty estimates to be explicitly incorporated during the inversion process. Specifically, the inputs for simPEG include the raw data, estimates of uncertainty, governing physical equations, and prior geological knowledge, which together help constrain the inversion and reduce uncertainty. For further details on SimPEG, please refer to Cockett et al. (2015) and Heagy et al. (2017).

Transient electromagnetic (TEM)

We deployed 111 TEM sounding points in the study area using the FASTNAP system with a transmitter loop size of 50 m × 50 m and a receiver loop size of 3 m × 3 m. The system continuously operates at low, medium, and high modes by injecting a current of 0.34–30A. Several pulses are transmitted, and response signals are enhanced by averaging the records at each time slot. The different mode records are stitched together, producing the voltage transient decay curve. Resistivity and thickness are calculated based on gradient-based inversion (Kassie et al. 2023). To build the 1D model, we utilized prior information, i.e., resistivity log and vertical electrical sounding data, as a guide and used MODEL 3.1 software to invert the data (Sharlov 2015).

The resistivity data harmonization

Each geoelectrical method (VES, NBR, and TEM) generates a 1D resistivity–depth model (Fig. 4). Traditionally, these data sets are directly compared and combined to construct a 3D model, with borehole information used for direct interpretation. However, due to inherent differences in measurement techniques, each method operates on a different scale despite measuring resistivity. To enable meaningful comparisons and integrated analysis, a crucial preprocessing step known as data harmonization was implemented prior to lithology analysis and 3D model construction. In general, data harmonization refers to the process of combining data from different sources or heterogeneous data into a cohesive data set by adjusting parameters such as measurement units, scales, maximum and minimum boundaries, or data formats (Cheng et al. 2024; Kumar et al. 2021; Nan et al. 2022). In this case, it involved adjusting the scale of resistivity data from each method to a common reference scale using the feature scaling method, ensuring compatibility and consistency across data sets. Harmonizing the resistivity data facilitated effective combination and integration of disparate measurements, providing a comprehensive understanding of subsurface properties and enhancing the accuracy of subsequent geological interpretations. To achieve this, several steps are required.

Initially, we adjusted the sampling rate of each one-dimensional data set, spacing them at 1-m intervals employing the Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) method, and restricted our usage to data reaching a depth of 200 m (Chapra 2012; Rabbath and Corriveau 2019). Following this, feature scaling was performed using the normalization technique known as Min–Max Re-scaling. This technique linearly transforms the original data range to maintain relationships within a predetermined boundary (Morrow 2020; Patro and Sahu 2015), expressed as

$$\rho_i' = a + \frac{(\rho_i - \min(\rho_i))}{\max(\rho_i) - \min(\rho_i)}(b - a) \quad (1)$$

Here i represents the specific data set (VES, NBR, or TEM) for which the scaling is being applied. ρ_i' represents the scaled value of resistivity from the data set i. ρ_i denotes the original resistivity data point from the data set i. $\min(\rho_i)$ and $\max(\rho_i)$ are the minimum and maximum absolute value of resistivity in the data set i. a and b are the minimum and maximum absolute value of resistivity in the VES data set. All resistivity data sets were transformed based on the VES resistivity range within the study area, a process we termed data retrieval. This retrieval normalized resistivity data to a range of 0–1, which was then rescaled to match the VES resistivity values using the sci-analysis package in Python (Morrow 2019). This efficient Python package facilitates rapid exploratory data analysis (EDA) by abstracting the underlying SciPy, NumPy, and Matplotlib commands. Furthermore, we utilized seaborn, a Python-based data visualization library to visualize the results (Waskom 2021). Figure 5 explains the workflow of this study, streamlining the methodology process.

The establishment of a 3D resistivity model

Previous studies predominantly utilized conventional software for this task, facing limitations that resulted in a fixed structure. These constraints stemmed from the software's inability to handle model boundaries effectively. For instance, despite data distribution extending beyond square-shaped areas, the outcomes from the 3D

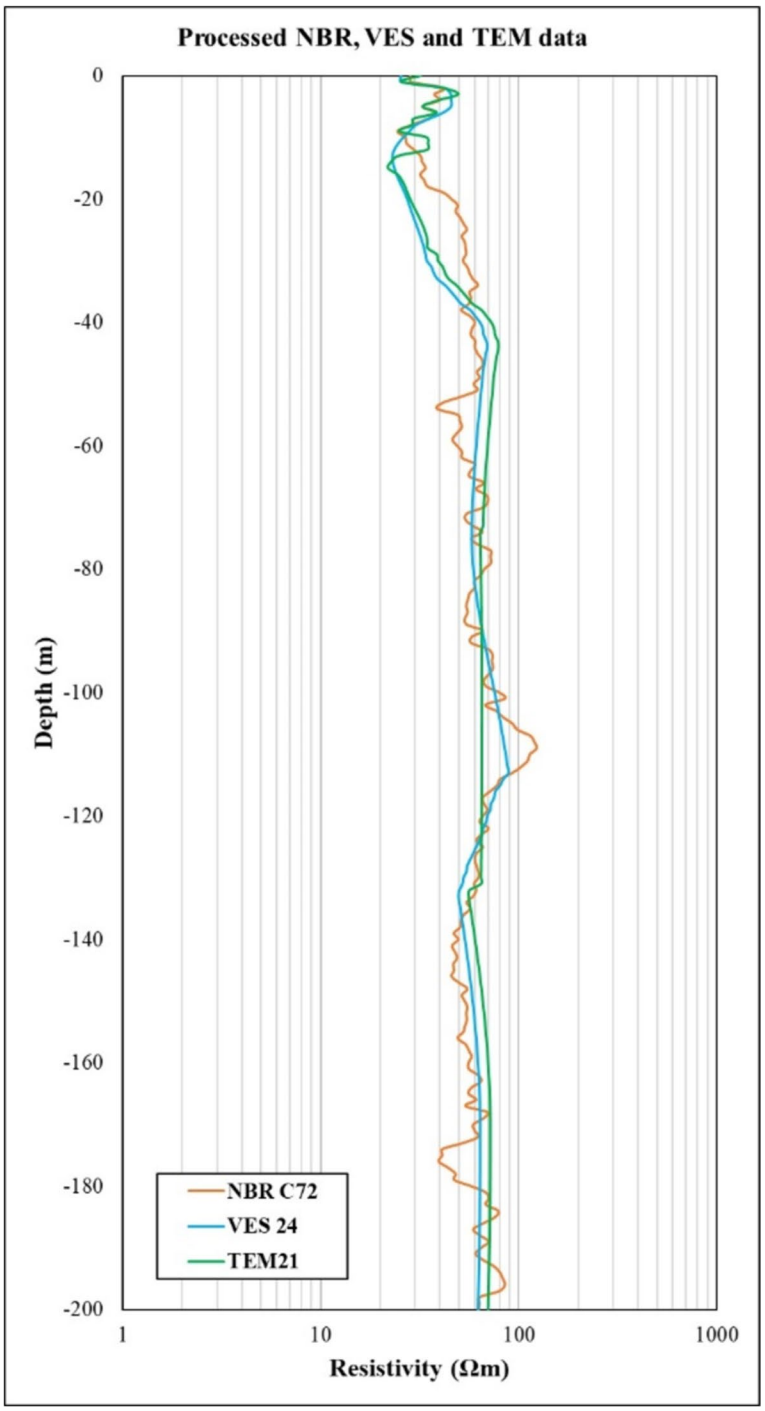

Fig. 4 Processed 1D NBR, VES, and TEM data

model typically conform to a square shape. Therefore, we employed PyVista, a Python-based visualization to develop the 3D resistivity model (Sullivan and Kaszynski 2019). PyVista is a pythonic framework that provides a high-level API to the Visualization Toolkit (VTK), including mesh data structures and spatial data set filtering algorithms. Its 3D plotting capabilities are intended to handle huge and complicated data geometries, simplifying the visualization process. Rather than relying on traditional VTK interfaces, PyVista utilizes NumPy and direct array access to interface with VTK. This approach facilitates rapid prototyping, analysis, and integration of

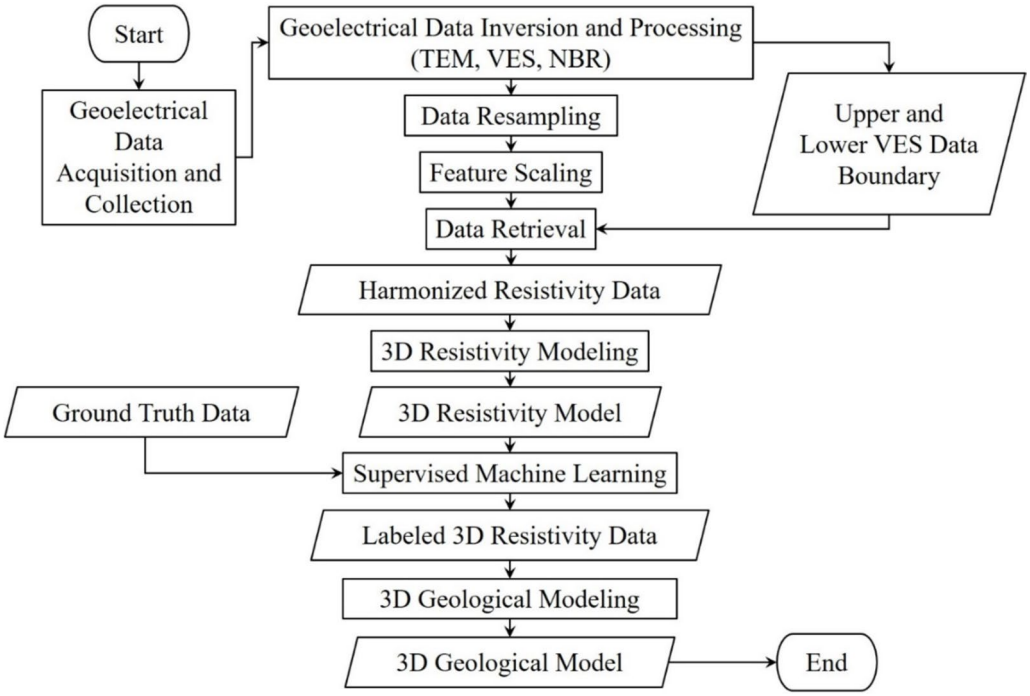

Fig. 5 Workflow of this study, streamlining the methodology process

spatially referenced data sets through a well-documented interface. Figure 6 depicts the methodology employed for the construction of a three-dimensional (3D) model.

In this study, we utilized PyVista to automate the mesh generation procedure for our computational model. Figure 7 depicts the mesh dimensions for the Changhua and Yunlin models, which were established according to the parameters outlined in Table 1. The boundary of the model was determined using a polygon derived from the coverage data, encompassing an area of 852 km^2, and 1174 km^2 for the Changhua and Yunlin models, respectively. These polygons served as an effective constraint

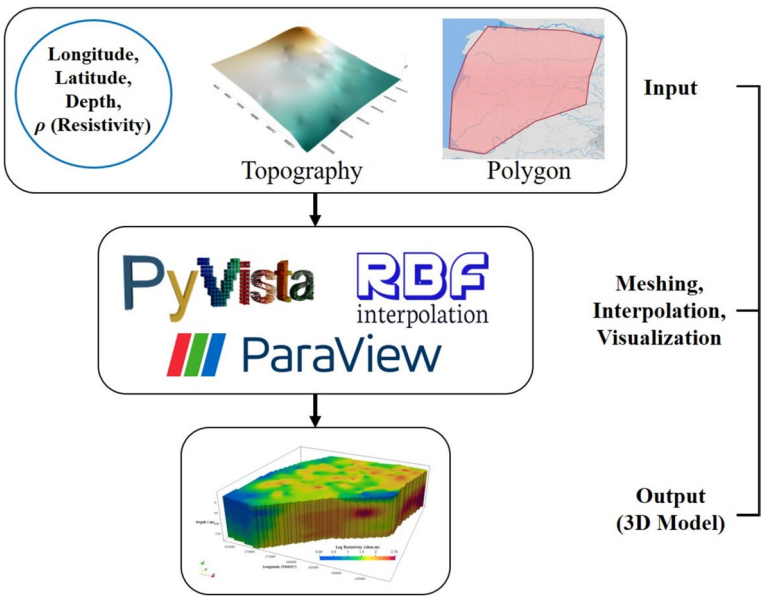

Fig. 6 Methodology employed for the construction of a three-dimensional (3D) model

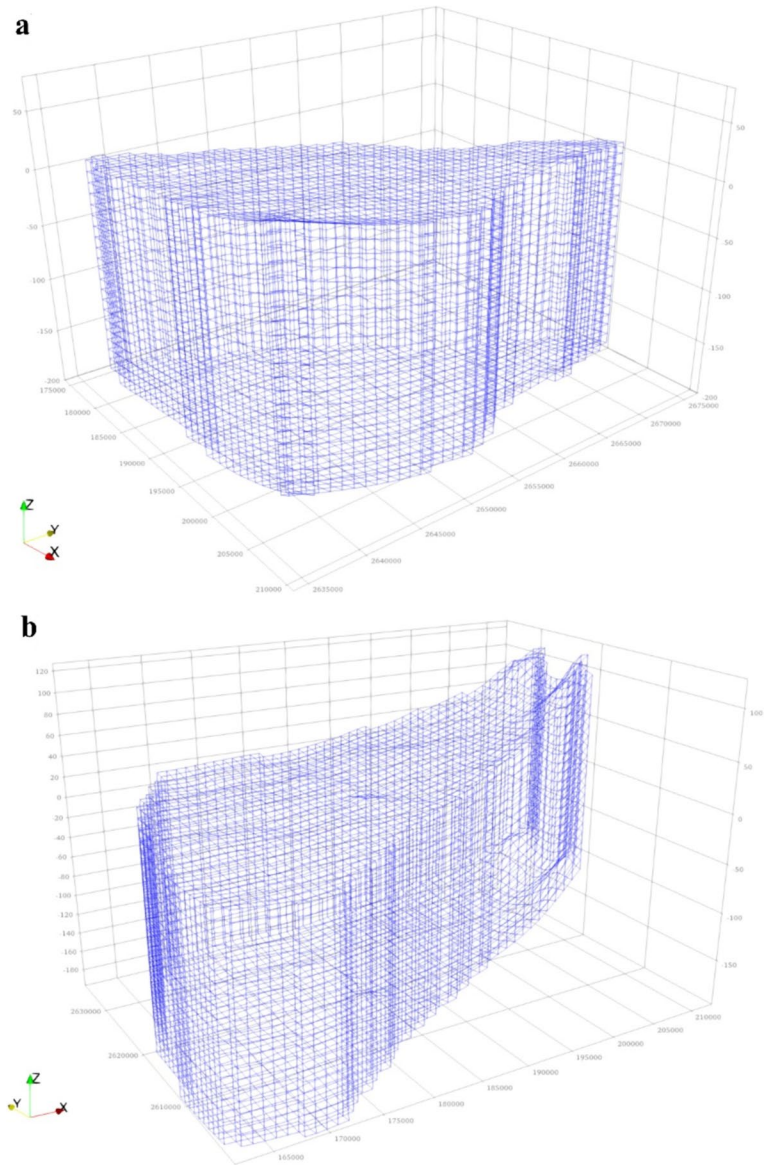

Fig. 7 Mesh dimensions for the Changhua and Yunlin models

Table 1 Parameter of the 3D model for the **(a)** Changhua and **(b)** Yunlin area

Parameter	Changhua	Yunlin
d_x (Horizontal X)	1000 m	1000 m
d_y (Horizontal y)	1000 m	1000 m
d_z (Vertical)	10 m	10 m
Depth	200 m	200 m
Area	852 km^2	1174 km^2
Perimeter	132 km	146 km
Number of cells without boundary: before	36,800	42,120
Number of cells with boundary: after	17,140	23,480

to prevent model exaggeration and over-interpolation in regions without data. It restricted the 3D shapes within the coverage area while removing any meshes outside of it from the analysis. For the spatial interpolation of resistivity data, we employed the Radial Basis Function Interpolation (RBFI) technique, specifically using the linear basis function. This method was chosen for its versatility in handling irregularly spaced data and its ability to generate smooth, continuous interpolation surfaces, making it ideal for subsurface resistivity modeling. A key hyperparameter in this process is the smoothing factor, set to 500 in this study. The smoothing factor controls the trade-off between accurately fitting the data points

and ensuring a smooth interpolation surface. By using a higher smoothing value, we mitigated the risk of overfitting to noisy or sparse data sets, leading to a more generalized and robust model. The use of RBFI with these strategies allows for the generation of smooth and continuous 3D resistivity models that reflect both vertical and horizontal variations in the subsurface. By transforming the resistivity model into a 3D Apparent Geological Model (AGM) using Supervised Machine Learning techniques, we achieved a comprehensive representation of the subsurface, providing insights that would be difficult to detect using 1D or 2D models. For a comprehensive description of the interpolation procedure with RBFI, we refer readers to Carr et al. (2001), Shi and Wang (2021), and Chen et al. (2024). Afterwards, the Paraview software was employed to generate visual representations of the 3D model by using the output VTK file. It is an open-source multiple-platform application for interactive scientific visualization (Ahrens et al. 2005).

The establishment of a 3D apparent geological model with supervised machine learning

After obtaining the 3D resistivity model, Supervised Machine Learning (SML) techniques were used to convert this model into a 3D Apparent Geological Model (AGM). SML is a computational approach that involves classifying or predicting data based on prior information, enabling the identification of overarching patterns and hypotheses through the training data to predict the characteristics of test data (Singh et al. 2016; Sotomayor et al., 2023). Several steps are involved in establishing the 3D AGM, as shown in Fig. 8.

First, both ground truth data and geophysical data sets were prepared. The ground truth data, derived from 54 borehole records within the study area, includes sediment type and resistivity log data. The resistivity data from these boreholes were subjected to various procedures such as inversion, resampling, harmonization, and labeling. During the labeling phase, sediment type labels were assigned to the resistivity data, which acted as the target for SML process. This labeling was informed by details from the resistivity log and borehole sediment type, facilitating direct comparison and correlation across three categories: clay, sand, and gravel to the resistivity value. Following this, Exploratory Data Analysis (EDA) was conducted to detect any missing data or outliers. Subsequently, the data was divided into train-test partitions, with 70% allocated for training and 30% for testing.

According to Archie (1942), the relationship of the in-situ resistivity of saturated sedimentary rock, its porosity, and pore-water resistivity can be expressed as follows:

$$\rho_b = a \cdot \rho_w \cdot \emptyset^{-m} \tag{2}$$

where ρ_b denotes bulk resistivity, a represents the tortuosity factor, ρ_w is pore-water resistivity, $\varnothing$ signifies porosity, and m is the cementation exponent specific to the rock.

This equation highlights that the measured bulk resistivity is influenced by the resistivity of pore water. Consequently, if pore water resistivities vary widely, the bulk

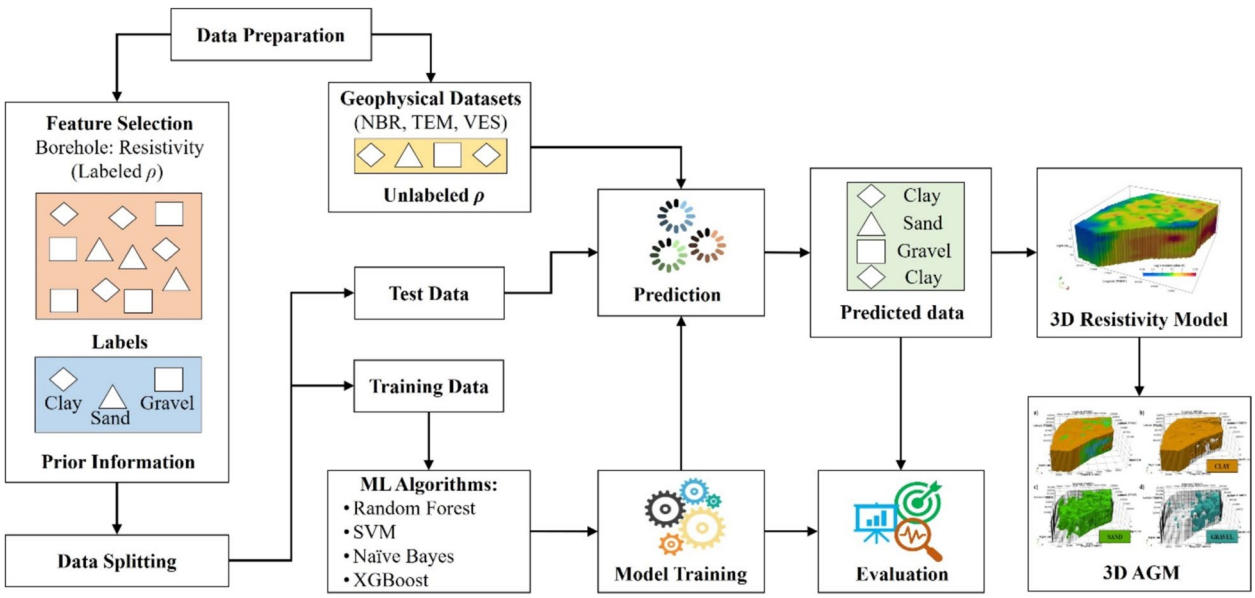

Fig. 8 Supervised machine learning (SML) procedure to convert the resistivity model into an apparent geological model

resistivity of the same sediment type can fluctuate significantly, regardless of the sediment type. Conversely, when the resistivity of groundwater is similar, the resistivity distribution within sediment tends to exhibit a consistent pattern, reflecting the different sediment types present.

To assess the impact of groundwater resistivity on our analysis, we present in Fig. 9, depicting the distribution of groundwater resistivity gathered from observation wells in the CRAF. The figure indicates that gravel (cyan), sand (green), and clay (orange) sediments can be roughly differentiated based on their measured resistivity. This differentiation is facilitated by the fact that groundwater resistivities (blue) primarily fall within the range of 5–50 Ωm, akin to the resistivity of mud sediments, and do not exert a significant influence on the classification analysis. Thus, the influence of groundwater resistivity on our study is deemed negligible. Similar trends were observed in the earlier study by Chang et al. (2024), which was conducted in a different basin area in Taiwan, specifically the Yilan basin, where the groundwater resistivity ≤ 50 Ωm.

Second, four SML algorithms, specifically decision tree (DT), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGBoost) were employed to predict the sediment type of the test data. The selection of these algorithms was based on their suitability for handling geological and geophysical data, as well as their varying strengths in classification tasks. Ensemble methods like random forest and XGBoost were chosen for their ability to manage complex data sets and reduce overfitting through the combination of multiple decision trees, which leads to more robust predictions. Decision Tree was included for its simplicity and ease of interpretation, making it useful for understanding the decision process. Support Vector Machine was selected for its potential to handle non-linear relationships in the data. These models were compared to identify the most appropriate algorithm for the resistivity data set. Random forest ultimately performed best due to its accuracy and reliability across multiple performance metrics (Breiman 2001; Chen & Guestrin 2016; Quinlan 1986). A brief description of these algorithms is described below:

a. Decision tree (DT)

The Decision Tree (DT) algorithm is a method in SML that operates akin to an inverted tree structure. It is a non-parametric approach used for classification and regression tasks. As implied by its name, the algorithm employs a tree-like flowchart structure to display predictions generated by a sequence of splits based on features. It commences at a root node and concludes with decisions at the leaves. The hierarchical structure comprises three key components: the root node which is the starting point of the DT, branches also known as sub-trees, delineate specific decision paths and outcomes within the tree, and leaf

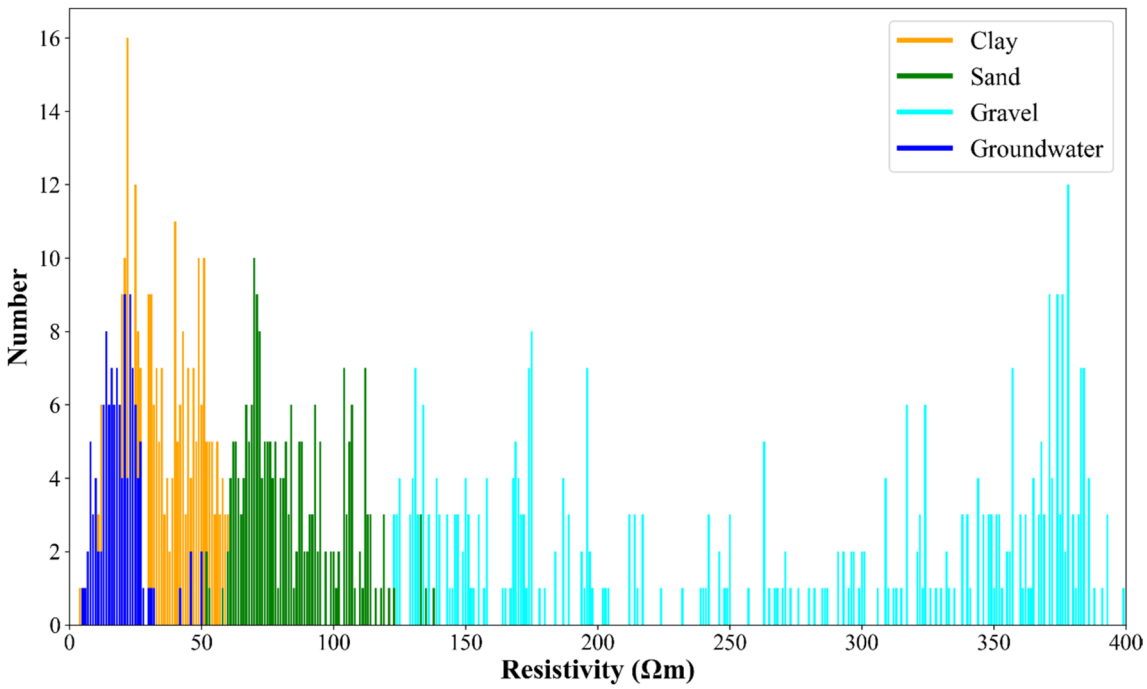

Fig. 9 Statistics for measured bulk resistivity of various sediment types from borehole logging and groundwater resistivity measurements

nodes which is the points within the tree where further division ceases, typically signifying the ultimate classification or outcome. In essence, a DT algorithm can be understood as a sequence of IF-ELSE statements constructing the entire tree from the root node to the leaf node (Kumar et al. 2022; Quinlan 2014; Rokach & Maimon 2005).

b. Random forest (RF)

The random forest (RF) algorithm is a type of ensemble learning method that utilizes multiple decision trees as predictors. Initially, this algorithm developed by Ho (1995) that later modified by Breiman (1996) based on the bagging or bootstrap aggregation approach. In this procedure, a significant number of trees are generated, and each tree contributes to the final prediction by voting for the most commonly predicted class. Through the implementation of the bagging method in RF, it enhances overall accuracy and mitigates overfitting by utilizing the average of predictions obtained from multiple decision trees. (Breiman 2001).

c. Support vector machine (SVM)

The support vector machine (SVM) is a useful technique in SML, extensively applied in both classification and regression tasks. It employs a hyperplane to effectively divide the attribute space, enabling optimal separation between instances of different classes or class values. The data points lying closer to the hyperplane on either side are referred to as support vectors, and the distance between these vectors is termed as the margin or street. A hyperplane with a large margin or street is deemed to be a good classification, while a small margin indicates a poorer classification that requires further parameter tuning. In short, this approach aims to maximize the minimum distance or margin from the hyperplane to the nearest data point (Lorena et al. 2011; Singh et al. 2016; Vapnik 1999).

d. Extreme gradient (XGBoost)

The Extreme Gradient Boosting (XGBoost) technique is an effective machine learning method employed in SML tasks such as regression and classification. It belongs to the ensemble learning family, especially to gradient boosting frameworks. XGBoost builds a predictive model by iteratively combining the predictions of numerous individual models, typically decision trees. This iterative process involves sequentially adding weak learners to the ensemble, with each new learner aiming at correcting the errors made by the previous ones. To reduce overfitting and complexity, a regularization term is included alongside the loss function (Chen and Guestrin 2016; Kumar et al. 2022).

Third, the trained models were used to predict the geophysical data sets, and their performance was evaluated using appropriate metrics. By comparing the results of several models, we identified the most suitable algorithm for prediction. To further optimize performance, we fine-tuned the models using a systematic grid search approach, which tested a predefined range of hyperparameters for each algorithm to find the best combination. Grid search cross-validation (CV) was employed to prevent bias from specific train-test splits by evaluating the model over multiple training and validation sets (Berrar 2019; Bressan et al. 2020; Refaeilzadeh et al. 2009). The average performance across these splits was used to identify the optimal model parameters, which are the hyperparameters that result in the best balance between accuracy and generalization on unseen data, minimizing overfitting. For the Decision Tree model, we fine-tuned parameters such as the criterion (entropy), which determines how to split the data at each node, the maximum depth, which controls how deep the tree can grow, the minimum samples per leaf, which ensures that leaf nodes are not created with too few data points, and the minimum samples per split, which specifies the minimum number of samples required to split a node. These parameters helped balance model complexity and prevent overfitting (Kumar et al. 2022; Quinlan 1986; Rokach and Maimon 2005). In the random forest model, we optimized the number of estimators (trees), which improves accuracy but increases computation time, and the maximum depth, which controls tree complexity and prevents overfitting (Breiman 2001). For the Support Vector Machine (SVM), we adjusted the kernel (RBF) to transform the input data for non-linear separability, gamma, which controls how far the influence of a single training example reaches, the regularization parameter (C), which manages the trade-off between maximizing the margin and minimizing classification error, and the maximum iterations, which controls how many iterations the optimization algorithm is allowed to perform. These adjustments ensured the SVM model could find the optimal decision boundary without overfitting (Lorena et al. 2011; Pedregosa et al. 2011; Singh et al. 2016). Finally, in XGBoost, we fine-tuned the number of estimators, which determines the number of boosting rounds, the learning rate (eta), which controls how much the model is updated with each boosting round, and the maximum depth, which limits the depth of each tree to prevent overfitting

Table 2 Hyper-parameters considered for DT, RF, SVM and XGBoost algorithms

No	SML algorithm	Tuned hyperparameters
1.	Decision tree (DT)	• Criterion: Entropy • mim_sample_leaf: 2 • Min_sample_split: 5 • Max_depth: 10
2.	Random forest (RF)	• N_estimators: 100 • Max_depth: 10
3.	Support vector machine (SVM)	• Kernel: 'rbf' • Gamma: 0.10 • C: 10 • Max_Iter: 100
4.	XGBoost	• N_estimators: 100 • η: 0.3 • Max_Depth: 6

while still capturing important patterns in the data (Chen and Guestrin 2016; Kumar et al. 2022). By optimizing these parameters, we ensured that each model was both accurate and generalizable, leading to robust predictions in our resistivity data set. Table 2 details the final hyper-parameters selected for each.

a. Confusion matrix

The confusion matrix illustrates the summary of predictions in a matrix format, reveals the accuracy of predictions for each class and identifies any class confusion within the model. This matrix displays the TP, FP, FN, and TN values for the respective classes, highlighting correct and incorrect classifications (Bajaj & Sinha 2022; Tiwari 2022).

b. Evaluation metrics

Various evaluation metrics were also computed in order to measure the model performance, such as accuracy, F1 score, precision, and recall. Accuracy represents the proportion of correctly classified examples, while the F1 score is a weighted harmonic mean of precision and recall. Precision measures the proportion of true positives among instances classified as positive, whereas recall quantifies the proportion of true positives among all positive instances in the data set. The equations utilized for computing the performance metrics are expressed as follows (Bressan et al. 2020; Kumar et al. 2022):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{3}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

c. Receiver operating characteristics (ROC) curve

Receiver Operating Characteristics or ROC curve is graphically depicting a classification model's performance. It is worth noting that the ROC curve provides insights into the trade-off between sensitivity and specificity for different threshold values, helping assess the model's discriminatory power and determining an optimal threshold for classification. This plot showcases two parameters: the true positive rate (TPR) and the false positive rate (FPR) that defined as follows (Bressan et al. 2020; Kumar et al. 2022; Yang & Berdine 2017):

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{8}$$

An ideal classification of the desired class by the ML algorithm is shown by a ROC curve that rapidly ascends from the origin to (0,1) and becomes flat. Points in the upper left corner of the ROC spectrum indicate greater performance, whereas curves closer to the diagonal line indicate poor classifier performance.

Results

The resistivity data harmonization

Despite conducting all measurements within the same research area and limiting the data to a depth of 200 m, significant discrepancies were observed in the upper and lower bounds of resistivity values, underscoring the necessity for data harmonization. Figure 10 depicts the result of employing min–max scaling for resistivity data harmonization, with Fig. 10a showcasing the data before harmonization and Fig. 10b illustrating the data after harmonization. Before data harmonization, the grand mean and median resistivity values were 48.98 Ωm and 45.71 Ωm. The mean resistivity values for TEM, VES, and NBR measurements were 42.67

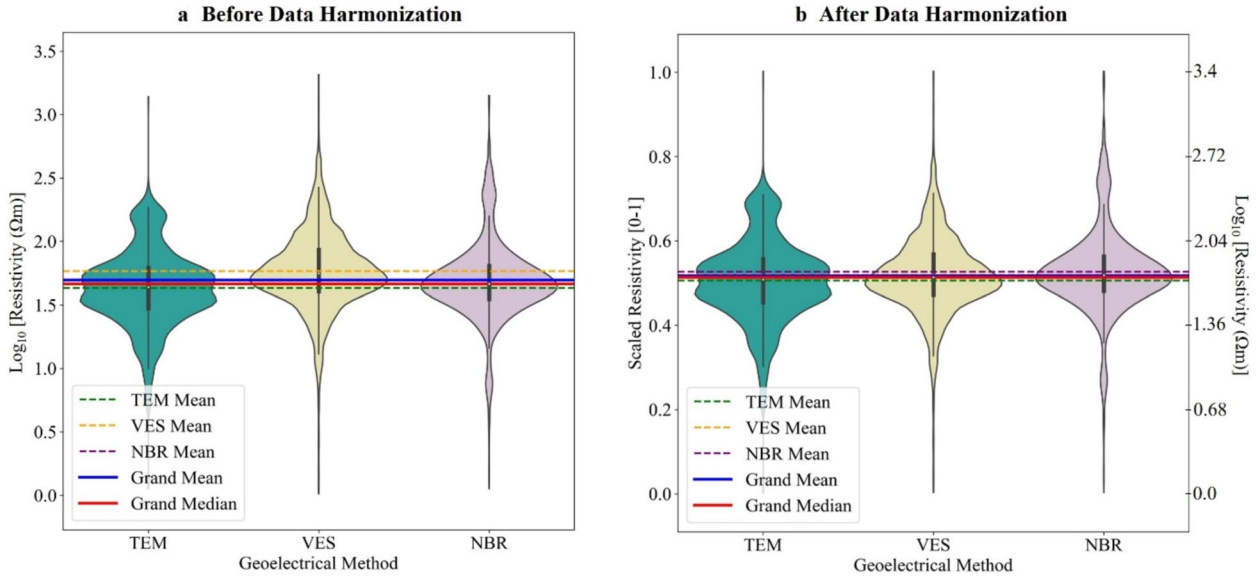

Fig. 10 Results of min–max scaling for resistivity data harmonization: **a** data before harmonization and **b** data after harmonization

Ωm, 57.54 Ωm, and 48.98 Ωm, respectively. After data standardization, the grand mean and median resistivity values shifted to 56.23 Ωm and 55.95 Ωm. The mean resistivity values for TEM, VES, and NBR measurements became 52.48 Ωm, 57.54 Ωm, and 58.25 Ωm, respectively. In this study, we opted to harmonize the resistivity data primarily based on VES measurements for two reasons. First, VES data exhibits both lower and higher boundaries of resistivity values compared to NBR and TEM data, registering at 1.02 Ωm and 2512 Ωm, respectively. Second, VES data offers the most densely distributed measurements among the three methods.

The 3D resistivity model

Figure 11 presents the 3D resistivity model covering the entirety of the Choushui River Alluvial Fan (CRAF), including the Changhua region to the north and the Yunlin region to the south. Figure 11a depicts the 3D compact resistivity model, while Fig. 11b shows the 3D cross-sectional resistivity model. The model reflects the topography of the study area, with higher elevations observed in the eastern portion near the tableland and hills, gradually decreasing towards the western coastal region. This elevation gradient is based on denser field measurements from VES, TEM, and NBR data points. In addition, the model's shape delineates the study area's boundary, which was carefully defined to include areas with sufficient data coverage while excluding regions without data to prevent bias and overextrapolation, as discussed in Sect. "The establishment of a 3D resistivity model". This approach ensures a more realistic representation compared to conventional 3D blocky models typically in rectangular shapes. Overall, the resistivity values range from 1.02 Ωm to 2512 Ωm, with higher resistivity values predominating in the eastern region, particularly in the proximal areas of the CRAF such as the Bagua tableland and Douliu hill. In contrast, resistivity values gradually decrease towards the western coastal areas, where lower values are predominantly observed.

Validation and evaluation results

Confusion matrix analysis, evaluation metrics and the Receiver Operating Characteristic (ROC) curve were employed to assess the final performance of the SML models on the testing data set. The results of these evaluations are outlined below:

a. Confusion matrix result

Figure 12 illustrates the confusion matrices for each Supervised Machine Learning (SML) algorithm, providing insight into their classification performance on the testing data set. In the confusion matrix, columns correspond to predicted values, and rows signify the true values. The main diagonal, represented by green shades, indicates correct predictions, with darker shades reflecting a larger number of correctly classified instances. Random forest (RF), decision trees (DT), and XGBoost exhibit high accuracy, with over 90% of predictions falling on the main diagonal. In

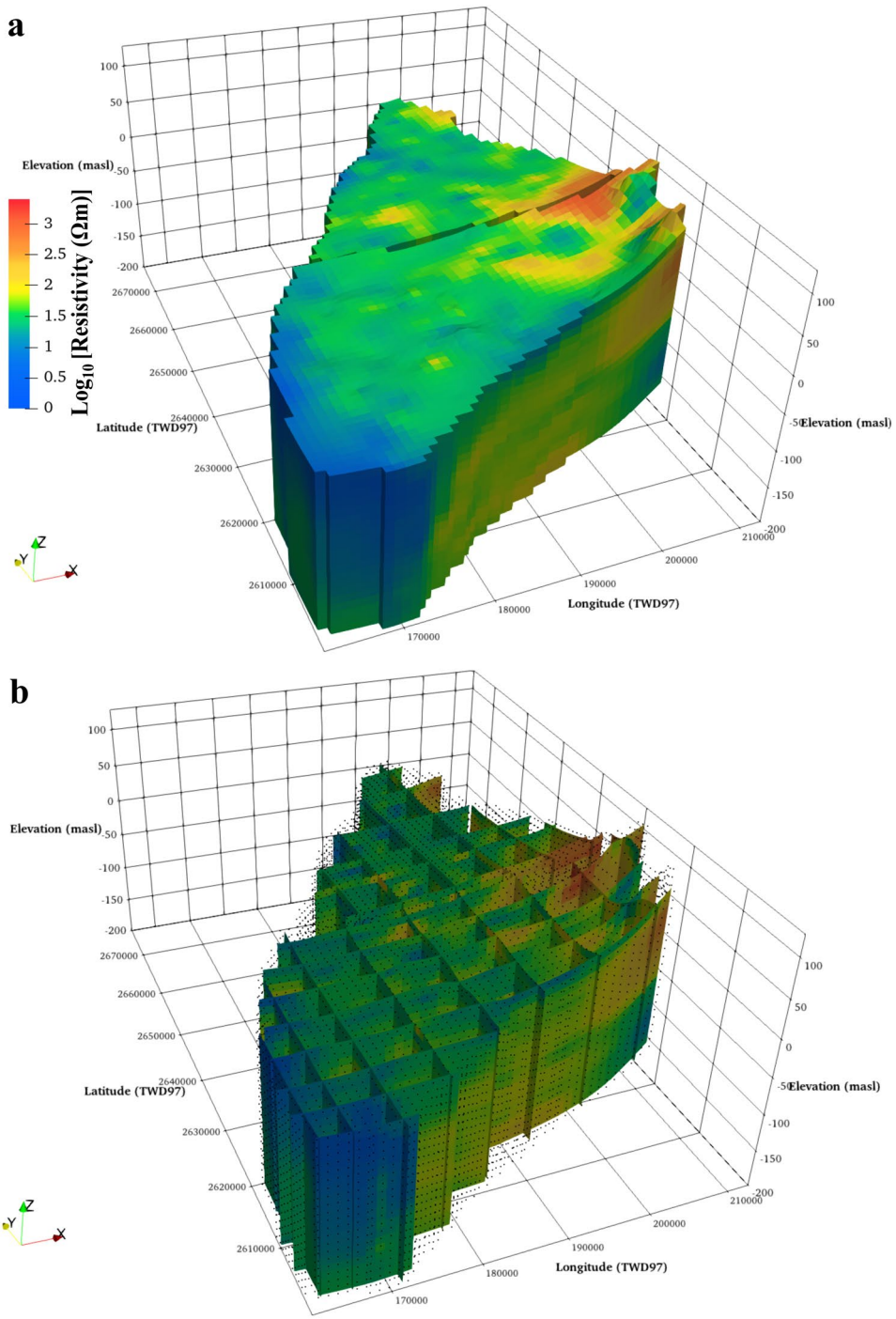

Fig. 11 3D resistivity model of the Choushui River Alluvial Fan (CRAF): **a** 3D compact resistivity model and **b** 3D cross-sectional resistivity model

contrast, the Support Vector Machine (SVM) shows a correct prediction rate of approximately 55.8% for clay, while 44.2% of the clay data is misclassified as sand.

b. Evaluation metrics result

The evaluation metrics for the machine learning models, as shown in Table 3, demonstrate that random forest (RF) achieved the highest overall performance, with an accuracy, F1 score, precision, and recall of 0.952 across all metrics. Decision Tree (DT) followed closely with an accuracy and F1 score of

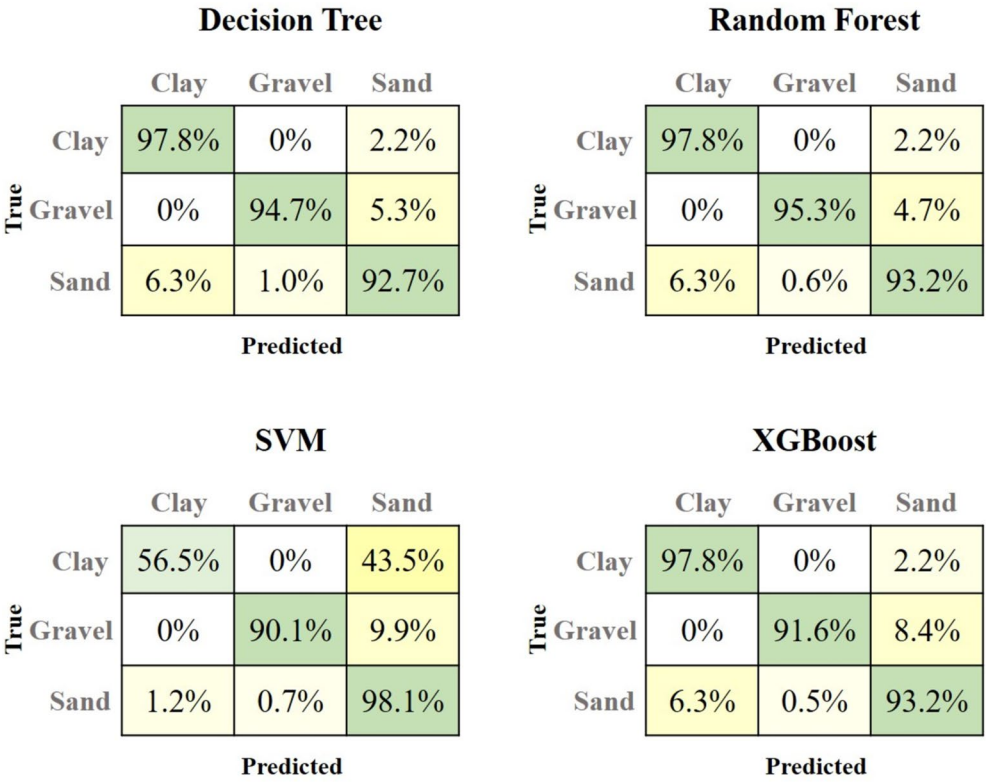

Fig. 12 Confusion matrix of all four ML classifiers used in the study

Table 3 Performance metrics for the machine learning models

SML algorithm	Accuracy	F1	Precision	Recell
Random forest (RF)	0.952	0.952	0.952	0.952
Decision tree (DT)	0.950	0.950	0.950	0.950
XGBoost	0.949	0.949	0.949	0.949
Support vector Machine (SVM)	0.739	0.749	0.843	0.744

0.950, indicating robust and balanced performance. XGBoost also performed well, with accuracy and F1 score both at 0.949. Support Vector Machine (SVM), however, showed the lowest performance, particularly in terms of accuracy (0.739) and recall (0.744), indicating that it struggled to correctly classify a significant portion of the instances. While SVM achieved a relatively high precision of 0.843, the lower recall resulted in a reduced overall F1 score of 0.749.

c. Receiver operating characteristics (ROC) curve result Figure 13 presents the ROC curves for XGBoost, RF, DT, and SVM, represented by green, orange, purple, and magenta curves, respectively. The Area Under the Curve (AUC) values for the ROC curves of all ensemble methods (XGBoost, RF, and DT) exceed the main diagonal, reflecting strong performance across all sediment types. Specifically, the average AUCs over sediment classes are 0.981 for RF, 0.980 for DT, 0.979 for XGBoost, and 0.733 for SVM. However, the discriminative algorithm (SVM) demonstrates poorer performance, particularly for clay and sand. The ROC curve for SVM on clay touches the main diagonal, while the curve for sand falls below the diagonal up to a false positive rate (FPR) of 0.3, suggesting weaker discrimination power for these classes.

The 3D apparent geological model

Based on the evaluation metrics in Table 3, the random forest algorithm proved to be the most suitable choice for modeling our data. Thus, we utilized the RF outcome to convert our 3D resistivity model into a 3D Apparent Geological Model (AGM), depicted in Fig. 14. In Fig. 14a, which presents a comprehensive overview encompassing all sediment types (clay, sand, and gravel), the orange, green, and cyan colors denote clay, sand, and gravel, respectively. This figure reveals a predominance of clay layers in the western area, while in the eastern region near the tableland and hills, gravel layers dominate. For

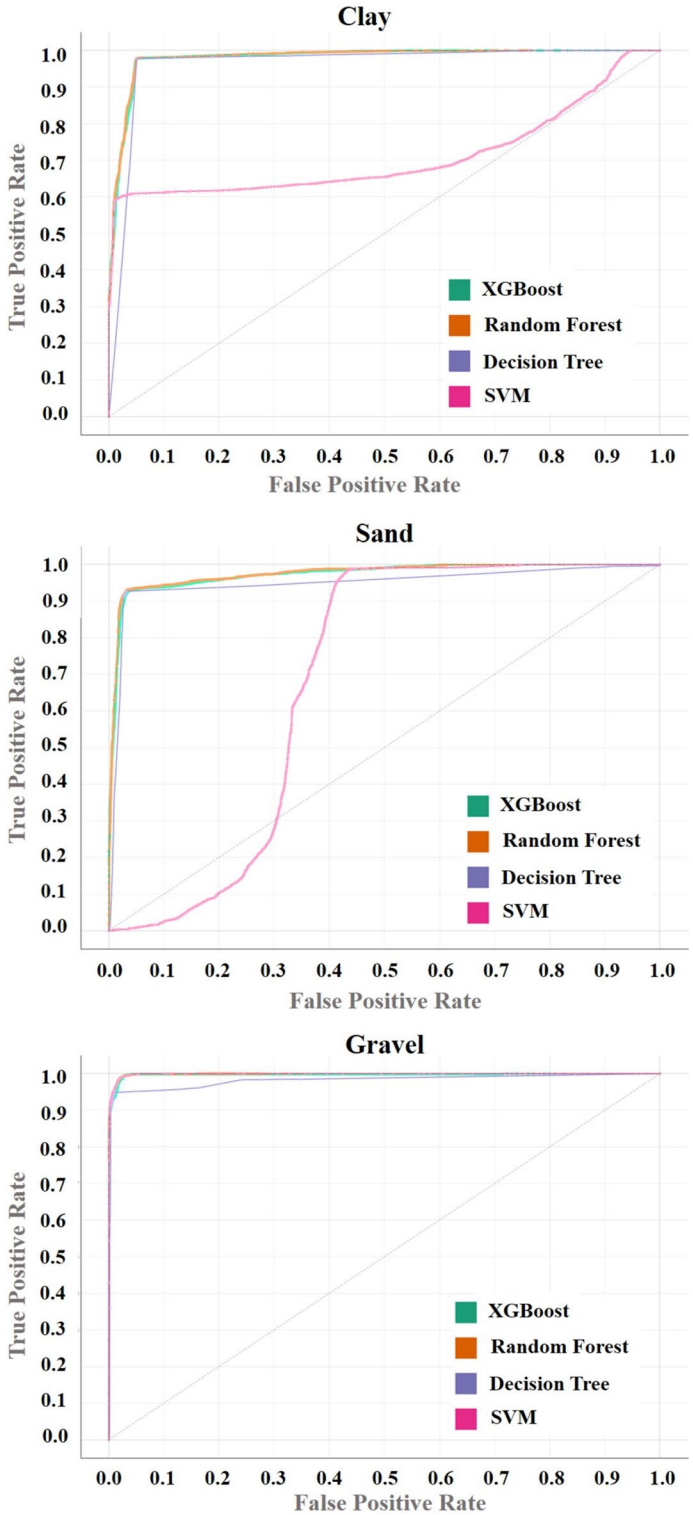

Fig. 13 Receiver operating characteristic (ROC) curves

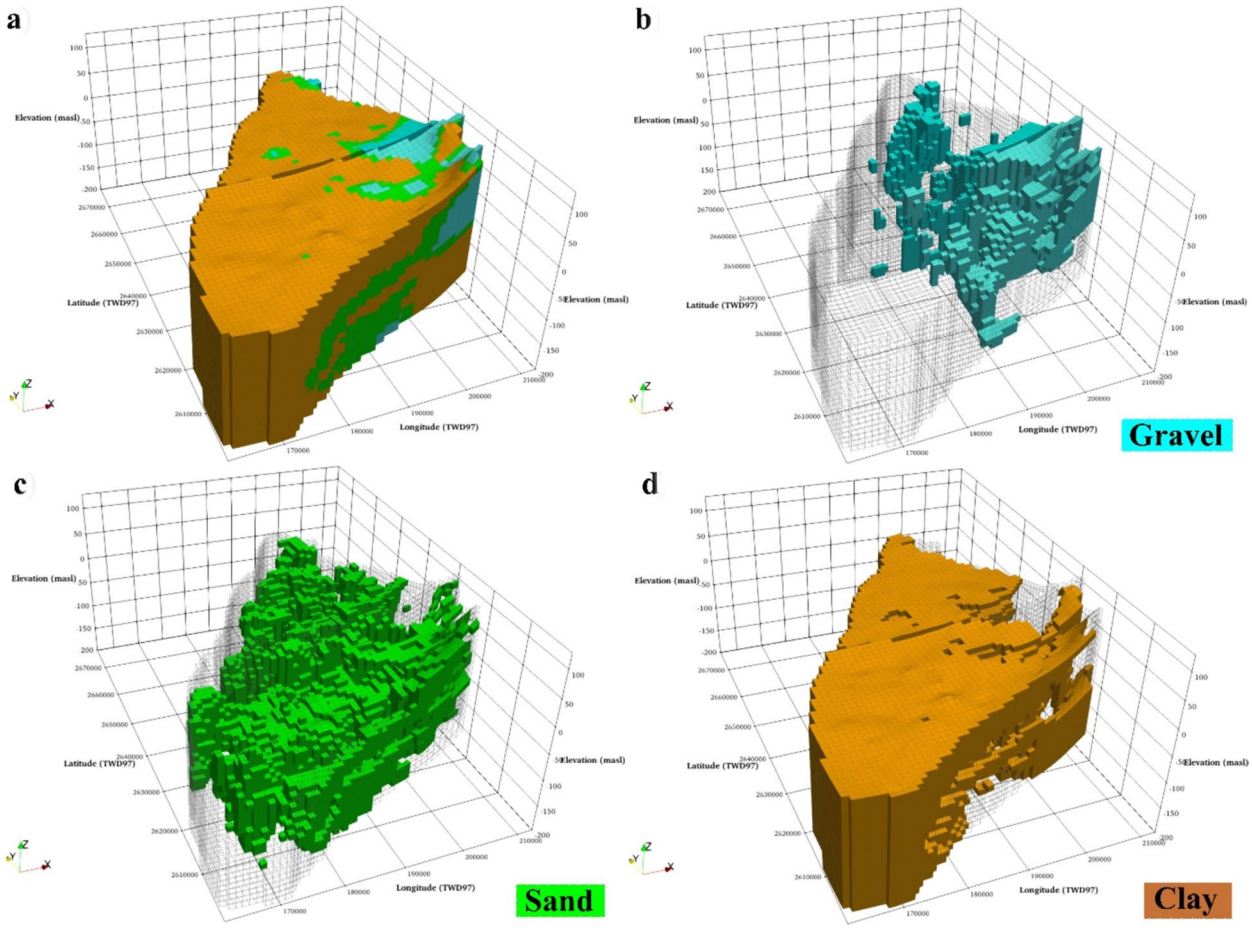

Fig. 14 3D apparent geological model derived from the resistivity model: **a** all sediment types, **b** gravel, **c** sand, and **d** clay

a detailed examination of the distribution of each sediment type, individual representations are provided in Fig. 14b–d. These depictions illustrate that gravel layers primarily accumulate in the eastern area, extending to greater depths for both Changhua and Yunlin models, while clay and sand exhibit more uniform distribution across all areas, with clay notably concentrated near the coastal region on the western side. The resistivity range for each sediment type is approximately ≤ 59.98 Ωm for clay, $59.98 < \rho < 136.14$ Ωm for sand, and ≥ 136.14 Ωm for gravel.

Discussion

This study has successfully increased the number of data points for 3D modeling from 62 to 386 across the Choushui River Alluvial Fan (CRAF), which spans approximately 2000 km^2. This addition was achieved by integrating geophysical data sets, including Vertical Electrical Sounding (VES), Transient Electromagnetic (TEM), Normal Borehole Resistivity (NBR), and available borehole information, as detailed in Sect. "The geoelectrical data". The significant addition of data points has led to a notable enhancement in our spatial coverage, thereby facilitating a more comprehensive understanding of the subsurface properties within the study area. Previously, with only 62 data points, the coverage for each point averaged approximately 32.26 km^2 area. However, with the inclusion of 386 data points, the coverage per data point has reduced to approximately 5.18 km^2. This decrease signifies a remarkable increase in coverage density by approximately 84.02%. The substantial increase in data density enables us to capture finer details and resolutions in our analysis.

Furthermore, by integrating these data sets through the data harmonization technique, we effectively address two significant issues. First, this process aligns the resistivity range across VES, TEM, and NBR data prior to 3D modeling. Logically, despite employing different methods, one would expect the data sets to exhibit a similar resistivity range within the same geographical area. However, Fig. 10a illustrates varying ranges among the data sets, but after applying this process, the resistivity

range becomes uniform, spanning from 1.02 Ωm to 2512 Ωm (Fig. 10b). It is important to emphasize that data harmonization does not alter the trend of resistivity, as evidenced in Fig. 10. Here, the violin plots for each measurement remain consistent, with their modals remaining similar both before and after this process. Second, this procedure proves beneficial for machine learning (ML) applications. Harmonization plays a pivotal role in mitigating bias and enhancing the performance of the ML algorithm by ensuring equitable contribution from all data sets (see Sect. "The resistivity data harmonization").

Regarding the 3D model, the utilization of Python-based modeling and visualization tools has yielded significant advancements compared to results obtained from commercial software. For instance, consider the 3D resistivity model illustrated in Fig. 11. A notable enhancement is observed in the delineation of the mesh boundary, which was predetermined by creating a polygon outlining the area of interest (refer to red line in Fig. 1, and the meshes in Fig. 7). This polygon acts as a constraint, effectively excluding areas lacking data and thereby preventing over-interpolation in regions with no data points, resulting in a more realistic and precise model. This method proves to be both effective and adaptable. Should we need to extend the model to another target area, it simply entails creating a new polygon for that specific region.

Figure 11 illustrates that the majority of high resistivity anomalies (depicted by yellow to red shades) are concentrated in the elevated regions of the eastern area, particularly within the proximal fan of the CRAF, which is southeast of Changhua and northeast of Yunlin. Interestingly, high resistivity anomalies are observed near the upstream regions of all the rivers, including the Wu River in the north, the Choushui River in the central, and the Beigang River in the southern area. Upon further examination using available borehole data from the Geological Survey and Mining Management Agency (GSMMA) Taiwan, it reveals that these high resistivity anomalies closely correspond to recharge areas by gravel layer. This finding is consistent with previous studies on groundwater sensitivity areas (see Fig. 3) conducted by GSMMA, which identified regions with high potential for groundwater aquifers that mostly cover the proximal fan of the CRAF (GSMMA 2023). In contrast, low resistivity anomalies (depicted by green to blue shades) are primarily situated in the western area, particularly within the distal fan of the CRAF, near the Taiwan Strait, where clay–sand layers predominate.

The 3D resistivity model solely depicts the distribution of resistivity across the study area. Therefore, to derive a more comprehensive understanding, the ML algorithm is employed to transform this valuable data into a 3D Apparent Geological Model (AGM). The term "AGM" is chosen deliberately to highlight that this model is an approximation or representation of the subsurface geology, rather than an exact replica. This acknowledges that the model relies on available data and interpretations, which may not capture all the complexities of the natural system. However, due to significant improvements in data coverage and enhancements in the 3D modeling process, the findings of this study provide a substantial amount of subsurface knowledge that researchers can utilize for further investigation. This includes topics like land subsidence, groundwater resources, protection measures, and more.

After evaluating the performance of supervised machine learning (SML) techniques, it is evident that ensemble learning method such as random forest (RF) and extreme gradient boosting (XGBoost), as well as standalone decision tree learning, provide better outcomes compared to discriminative learning method like Support Vector Machine (SVM). This is indicated by their high accuracy, F1 score, precision, and recall values, as presented in Table 3. Further analysis of the confusion matrix in Fig. 12 reveals that while the SVM model demonstrates high accuracy in predicting gravel and sand classes, achieving rates exceeding 90% for both, its performance in predicting clay appears comparatively lower. In the top-left cell, 55.8% of samples that are truly clay were correctly predicted as clay (true positive), while in the top-right cell, 44.2% of samples that are truly clay were incorrectly predicted as sand (false positive).

Upon examining the Receiver Operating Characteristics (ROC) curve depicted in Fig. 13, it is evident that the performance of SVM models for both clay and sand is comparatively poorer when compared to other algorithms. This is indicated by the fact that both curves touch the diagonal line. Ideally, a desirable ROC curve would closely hug the upper-left corner of the plot, reflecting high sensitivity (true positive rate) and low false positive rate across various threshold values. However, in this case, the ROC curve for clay touches the diagonal line around a false positive rate greater than 0.7 (x-axis) towards the end, indicating a high false positive rate for clay classification, particularly at higher thresholds. This suggests misclassification of a substantial number of negative instances as positive (clay). On the other hand, the ROC curve for sand dips below the diagonal line, particularly noticeable in the initial segment of the curve, roughly within the false positive rate range of 0–0.3. This observation indicates that the SVM model's performance in classifying sand is notably inferior to random guessing at these specific thresholds. It suggests that the SVM ability to distinguish between sand and non-sand samples is considerably poorer than random guessing, particularly at lower threshold levels. Thus, the outcomes derived

from the SVM will not be utilized. In addition, the ROC curve illustrates that the DT curve, represented by purple, is slightly inferior to the RF and XGBoost (orange and green curves), with the RF curve performing the best. This finding aligns well with the outcomes of cross-validation, where the RF exhibited the highest performance among the algorithms evaluated. Hence, we have selected the results obtained from RF algorithm to transform our 3D resistivity model into 3D AGM.

Overall, our observation from the 3D Apparent Geological Model (AGM) depicted in Fig. 14 indicates a correlation between resistivity anomalies and sediment types. Specifically, we found that low resistivity anomalies are associated with the clay layer (Fig. 14d), medium resistivity corresponds to the sand layer (Fig. 14c), and high resistivity anomalies are highly correlated with the gravel layer (Fig. 14b). These correlations were determined based on the RF algorithm, with respective resistivity ranges approximately ≤ 59.98 Ωm, $59.98 < \rho < 136.14$ Ωm, and ≥ 136.14 Ωm. Certainly, this outcome is consistent with the 2D conceptual profile provided by GSMMA, illustrated in Fig. 2 (GSMMA 2023). It is apparent that the eastern region, proximal to the CRAF from Chukou to Jiulong boreholes and situated upstream of the Choushui River, is predominantly characterized by gravel layers. Conversely, as expected, the coastal area on the western side, representing the distal fan of the CRAF from Haifeng to Fengrong boreholes and located downstream, primarily consists of clayey sand. Furthermore, the middle fan surrounding the Ganghou and Jiulong boreholes predominantly features a sandy clay layer. It is important to note that the 2D model in Fig. 2, which relies on widely spaced borehole data and manual interpretation, only captures sediment distribution in the x and z directions, providing a simplified and coarse view of the subsurface. In contrast, our 3D model, built with denser data, offers a more detailed representation of sediment layers by capturing both lateral and vertical variability that is difficult to detect in 1D or 2D models. The 3D model also enables the identification of key features that would otherwise remain hidden. For instance, as shown in Fig. 15, the 3D model clearly reveals a low-resistivity anomaly strongly correlated with clay layers in the center of the subsidence area. This anomaly, difficult to visualize with 1D or 2D models, is well defined in 3D, allowing for a clearer understanding of its boundaries and spatial extent. This detailed view is essential for

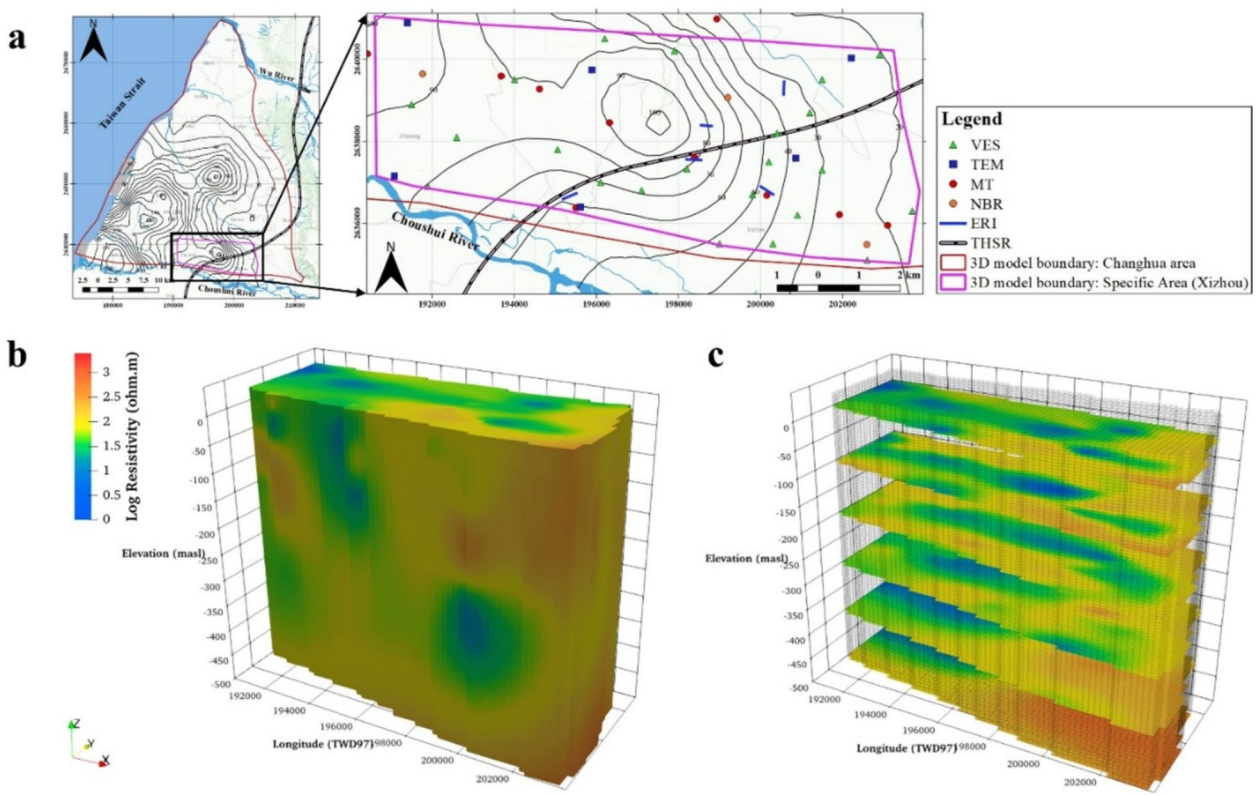

Fig. 15 High-resolution 3D resistivity model: **a** model boundary in magenta, **b** 3D resistivity model, and **c** cross section. The model features a resolution of 200 m × 200 m horizontally and 2 m vertically, with a depth of 500 m

further investigation and provides a level of precision that 1D and 2D approaches cannot achieve.

We aware that our study faces certain limitations, particularly in the 3D modeling process due to the extensive area and large data set involved. Our computer system, equipped with 64.0 GB of RAM and an 11th Gen Intel Core i7-11,700 @ 2.50 GHz processor, encountered memory errors when we initially attempted to assign a finer mesh size of 250 m for horizontal (d_x and d_y) and 2 m for vertical (d_z). This was the finest resolution our system could handle without causing memory allocation issues. Any attempt to reduce the mesh size further resulted in errors. To address this, we divided the CRAF model into two regions (Changhua and Yunlin) and adjusted both the vertical and horizontal mesh resolution to maintain computational feasibility, as described in Sect. "The establishment of a 3D resistivity model".

Despite the unavoidable trade-off in model resolution, we can effectively address this challenge through two main approaches. First, employing high-performance computing resources, such as cluster computer systems, becomes essential for model execution. Second, by utilizing the inherent flexibility and adaptability of our proposed methodology, we can customize the model to cover smaller targeted regions while enhancing both vertical and horizontal resolutions. For instance, we have constructed a finer 3D model within the Critical Area for Land Subsidence in the CRAF, particularly in the Taiwan high-speed rail area (Chen et al. 2021; Hsu 1998). The model dimensions extend to d_x and d_y of 200 m each, with a d_z of 2 m. By integrating additional Magnetotelluric (MT) and surface Electrical Resistivity Imaging (ERI) data specific to the study area, the model depth has been expanded to 500 m below the subsurface, as depicted in Fig. 15. The delineation of the 3D model's boundary within the specified region is highlighted in magenta on the map (Fig. 15a), accompanied by corresponding visualizations of the 3D resistivity model and its cross section in Fig. 15b, c. The enhancements applied to this model have effectively addressed the previously outlined challenge. In addition, it is worth noting that creating multiple models with higher resolutions and combining them for large-scale studies is also feasible.

Another concern of this study is the inherent stratigraphic uncertainty and spatial variability associated with deterministic boundaries. The model assumes fixed boundaries, which may not fully capture subsurface variability due to data sparsity and geological heterogeneity. Stratigraphic uncertainty arises from factors like resistivity variability, geophysical method limitations, and uneven data distribution. Resistivity measurements are affected by environmental noise and electrode limitations, leading to potential errors. In addition, interpolation strategies like Radial Basis Function (RBF) smooth the data but may generalize subsurface features, further contributing to uncertainty. To address this, we incorporated multiple data sets (VES, TEM, borehole data) to reduce uncertainty and better constrain the model. However, variability remains in data-sparse regions. Future studies could integrate more direct measurements, such as new borehole data, to improve boundary accuracy.

Conclusions

This study effectively outlines the procedure for building a 3D Apparent Geological Model (AGM) through the integration of multi-resistivity data, particularly in the Choushui River Alluvial Fan (CRAF) area. It employs statistical methods, machine learning techniques, and Python-based modeling and visualization tools, signifying a shift from conventional methodologies to more advanced approaches. By integrating multiple geophysical data sets and available borehole data, we increased the spatial coverage from 62 to 386 points across the study area, achieving an 84.02% increase in coverage density. Our analysis harmonized all these data sets and established a consistent range of resistivity values of log 0.01 Ωm to log 3.4 Ωm, prior to 3D modeling.

Furthermore, the assessment of supervised machine learning (SML) through confusion matrix analysis, evaluation metrics, and receiver characteristic curves (ROC) underscored the effectiveness of ensemble learning methods, particularly highlighting the random forest algorithm as the top performer among the various algorithms evaluated to transform the 3D resistivity model into 3D AGM. The 3D AGM unveiled distinct resistivity anomalies correlated with sediment types, with low resistivity anomalies associated with clay layers ($\leq$59.98 Ωm), medium resistivity corresponding to sand layers (59.98 $< \rho <$ 136.14 Ωm), and high resistivity anomalies highly correlated with gravel layers ($\geq$136.14 Ωm). Gravel layers were predominantly found in the proximal fan, situated in the eastern area of the CRAF. Conversely, the distal fan, located in the western coastal area, mostly consisted of clayey sand. In addition, the middle fan primarily comprised sandy clay layers. Despite encountering challenges associated with 3D modeling resolution due to memory constraints, our study proposes effective solutions to mitigate this issue. These strategies involve utilizing high-performance computing resources and adjusting models to focus on smaller target areas while increasing the vertical and horizontal resolutions. By combining multiple smaller models, we were able to cover the larger-scale study area.

In conclusion, the results of this study provide a valuable 3D model of subsurface conditions, which can serve as a useful resource for researchers undertaking further investigations that necessitate a comprehensive understanding of the subsurface environment.

Acknowledgements
This study has been funded by the Geological Survey and Mining Management (GSMMA) under the Ministry of Economics, Taiwan. We appreciate the assistance with fieldwork provided by the near-surface geophysics research group at the Department of Earth Sciences, National Central University. We also extend our gratitude to the anonymous reviewers for their constructive suggestions and comments.

Author contributions
JP: Conceptualization, methodology, data processing, formal analysis, draft manuscript, review and modify, data curation, visualization, and validation. PC: Conceptualization, methodology, review and modify, and validation. HA, DL: Resources, and data curation. MS: Resources and visualization. JT, HY, LC: Resources and validation. JC, LK: data curation and TEM data processing.

Funding
This research was funded by the Geological Survey and Mining Management Agency of Taiwan (GSMMA), under project numbers 110-5726901000-05-01, 111-5726901000-05-01, and 112-5726901000-05-01, spanning 3 years.

Availability of data and materials
No datasets were generated or analysed during the current study.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Earth Sciences, National Central University, Taoyuan, Taiwan. [2]Earthquake-Disaster, Risk Evaluation, and Management Centre, National Central University, Taoyuan, Taiwan. [3]Center for Astronautical Physics and Engineering (CAPE), National Central University, Taoyuan, Taiwan. [4]Department of Bioenvironmental Systems Engineering, National Taiwan University, Taipei, Taiwan. [5]Department of Civil Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. [6]Earth System Science, Taiwan International Graduate Program (TIGP), Academia Sinica, Taipei, Taiwan.

Received: 14 July 2024 Accepted: 6 November 2024
Published online: 25 November 2024

References

Abu Rajab J et al (2023) Multiscale geoelectrical characteristics of seawater intrusion along the eastern coast of the Gulf of Aqaba, Jordan. J Appl Geophys. https://doi.org/10.1016/j.jappgeo.2022.104868
Ahrens J et al (2005) 36-paraview: an end-user tool for large-data visualization. Visualization Handbook 717:50038–50031
Aldiss D, et al. (2012). Benefits of a 3D geological model for major tunnelling works: an example from Farringdon, east–central London, UK
Archie GE (1942) The electrical resistivity log as an aid in determining some reservoir characteristics. Trans AIME 146(01):54–62. https://doi.org/10.2118/942054-G
Arowoogun KI, Osinowo OO (2021) 3D resistivity model of 1D vertical electrical sounding (VES) data for groundwater potential and aquifer protective capacity assessment: a case study. Mod Earth Syst Environ 8(2):2615–2626. https://doi.org/10.1007/s40808-021-01254-w
Bajaj V, Sinha G (2022) Artificial intelligence-based brain-computer Interface: Academic Press
Berrar D (2019) Cross-validation. In book: Reference Module in Life Sciences. https://doi.org/10.1016/B978-0-12-809633-8.20349-X
Breiman L (1996) Bagging Predictors. Mach Learn 24:123–140
Breiman L (2001) Random Forests. Mach Learn 45:5–32
Bressan TS et al (2020) Evaluation of machine learning methods for lithology classification using geophysical data. Comput Geosci 139:104475
Cardarelli E, De Donno G (2017) Multidimensional electrical resistivity survey for bedrock detection at the Rieti Plain (Central Italy). J Appl Geophys 141:77–87. https://doi.org/10.1016/j.jappgeo.2017.04.012
Carr JC, et al. (2001) Reconstruction and representation of 3D objects with radial basis functions. In: Paper presented at the Proceedings of the 28th annual conference on Computer graphics and interactive techniques
Chabaane A et al (2017) Combined application of vertical electrical sounding and 2D electrical resistivity imaging for geothermal groundwater characterization: hammam Sayala hot spring case study (NW Tunisia). J Afr Earth Sc 134:292–298. https://doi.org/10.1016/j.jafrearsci.2017.07.003
Chang P-Y et al (2024) Application of machine learning and resistivity measurements for 3D apparent geological modeling in the Yilan plain, Taiwan, at the SW Tip of the Okinawa trough. Geosci Lett 11(1):25
Chapra SC (2012) Applied numerical methods with MATLAB® for engineers and scientists, 3rd edn. McGraw-Hill Education, New York
Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining
Chen Y-A et al (2021) Space-Time Evolutions of Land Subsidence in the Choushui River Alluvial Fan (Taiwan) from Multiple-Sensor Observations. Remote Sensing. https://doi.org/10.3390/rs13122281
Chen W et al (2024) Geotechnical correlation field-informed and data-driven prediction of spatially varying geotechnical properties. Comput Geotech 171:106407
Cheng SY, Hsu KC (2021) Bayesian integration using resistivity and lithology for improving estimation of hydraulic conductivity. Water Resour Res 57(3):e2020WR027346
Cheng C et al (2024) A general primer for data harmonization. Sci Data 11(1):152
Chiang CJ (1999) General report on hydrogeological investigation and study of the Chuoshuixi alluvial fan of the first phase of the groundwater observation network in Taiwan. Retrieved from Taipei.
Chu H-J et al (2021) Development of spatially varying groundwater-drawdown functions for land subsidence estimation. J Hydrol Reg Stud. https://doi.org/10.1016/j.ejrh.2021.100808
Cockett R et al (2015) SimPEG: An open source framework for simulation and gradient based parameter estimation in geophysical applications. Comput Geosci 85:142–154
Dong L-D, et al (1996) Taiwan Groundwater Observation Network Phase I Project: Geophysical Exploration and Stratigraphy Correlation of the Zhoushui River Alluvial Fan. Retrieved from Central Geological Survey of Taiwan: Zonghe New Taipei City
Dong Y et al (2023) 3D pseudo-lithologic modeling via iterative weighted k-means++ algorithm from Tengger Desert cover area. China Front Earth Sci 11:1235468
GSMMA (2023) Hydrogeological Database Combined Query Platform. Retrieved June 13, 2020, from Geological Survey and Mining Management Agency of Taiwan, Taipei. https://hydro.geologycloud.tw/map
Heagy LJ et al (2017) A framework for simulation and inversion in electromagnetics. Comput Geosci 107:1–19
Ho TK (1995) Random decision forests. Paper presented at the Proceedings of 3rd international conference on document analysis and recognition.
Hsu S-K (1998) Plan for a groundwater monitoring network in Taiwan. Hydrogeol J 6(October 1998):405–415. https://doi.org/10.1007/s100400050163
Huang C-W et al (2024) Identifying private pumping wells in a land subsidence area in Taiwan using deep learning technology and street view images. J Hydrol Reg Stud 51:101636
Hung W-C et al (2009) Monitoring severe aquifer-system compaction and land subsidence in Taiwan using multiple sensors: Yunlin, the southern Choushui River Alluvial Fan. Environ Earth Sci 59(7):1535–1548. https://doi.org/10.1007/s12665-009-0139-9
Kassie LN et al (2023) Mapping hydrogeological structures using transient electromagnetic method: a case study of the Choushui River Alluvial Fan in Yunlin. Taiwan Water 15(9):1703
Kumar G et al (2021) Data harmonization for heterogeneous datasets: a systematic literature review. Appl Sci 11(17):8275

Kumar T et al (2022) Lithology prediction from well log data using machine learning techniques: a case study from Talcher coalfield, Eastern India. J Appl Geophys 199:104605
Lin C-W, et al (2016) Land Subsidence Caused by Groundwater Exploitation in Yunlin, Taiwan. Paper presented at the Proceedings of the 12 th International Conference on Hydroscience and Engineering Hydro-Science and Engineering for Environmental Resilience, Tainan, Taiwan
Liu C-W et al (2001) The effect of clay dehydration on land subsidence in the Yun-Lin coastal area. Taiwan Environ Geol 40:518–527
Liu C-W et al (2002) Three-dimensional spatial variability of hydraulic conductivity in the Choushui River alluvial fan. Taiwan Environ Geol 43(1–2):48–56. https://doi.org/10.1007/s00254-002-0648-2
Liu C-H et al (2004) Characterization of land subsidence in the Choshui River alluvial fan. Taiwan Environ Geol 45(8):1154–1166. https://doi.org/10.1007/s00254-004-0983-6
Lorena AC et al (2011) Comparing machine learning classifiers in potential distribution modelling. Expert Syst Appl 38(5):5268–5275
Lu C-H et al (2016) Geostatistical Data Fusion of Multiple Type Observations to Improve Land Subsidence Monitoring Resolution in the Choushui River Fluvial Plain, Taiwan. Terr Atmos Ocean Sci 27(4):505–520. https://doi.org/10.3319/TAO.2016.01.29.02(ISRS)
Lu C-Y et al (2020) The relationship between surface displacement and groundwater level change and its hydrogeological implications in an Alluvial Fan: case study of the Choshui River, Taiwan. Remote Sensing. https://doi.org/10.3390/rs12203315
Marzán I et al (2021) Joint interpretation of geophysical data: applying machine learning to the modeling of an evaporitic sequence in Villar de Cañas (Spain). Eng Geol 288:106126
Morrow C (2019) sci_analysis Documentation, Release 2.2.0
Morrow C (2020) Normalization Techniques in Python Using NumPy. Normalizing datasets with Python and NumPy for analysis and modeling. Retrieved from https://towardsdatascience.com/normalization-techniques-in-python-using-numpy-b998aa81d754
Nan Y et al (2022) Data harmonisation for information fusion in digital healthcare: a state-of-the-art systematic review, meta-analysis and future research directions. Inf Fusion 82:99–122
Osinowo OO, Falufosi MO (2019) 3D Electrical Resistivity Imaging (ERI) for subsurface evaluation in pre-engineering construction site investigation. NRIAG J Astron Geophys 7(2):309–317. https://doi.org/10.1016/j.nrjag.2018.07.001
Patro S, Sahu KK (2015) Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462
Pedregosa F et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
Piegari E et al (2023) A machine learning-based approach for mapping leachate contamination using geoelectrical methods. Waste Manage 157:121–129
Puntu JM et al (2021) A comprehensive evaluation for the tunnel conditions with ground penetrating radar measurements. Remote Sensing. https://doi.org/10.3390/rs13214250
Puntu JM et al (2023) Groundwater monitoring and specific yield estimation using time-lapse electrical resistivity imaging and machine learning. Front Environ Sci 11:1197888
Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106
Quinlan JR (2014) C4. 5: programs for machine learning: Elsevier
Rabbath CA, Corriveau D (2019) A comparison of piecewise cubic Hermite interpolating polynomials, cubic splines and piecewise linear functions for the approximation of projectile aerodynamics. Defence Technol 15(5):741–757. https://doi.org/10.1016/j.dt.2019.07.016
Rabeau O et al (2010) Gold potential of a hidden Archean fault zone: the case of the Cadillac-Larder Lake Fault. Explor Min Geol 19(3–4):99–116
Refaeilzadeh P et al (2009) Encyclopedia of Database Systems. Cross-Validation 5:532–538
Rokach L, Maimon O (2005) Decision trees. Data mining and knowledge discovery handbook, pp. 165–192
Sharlov MV (2015) FastSnap Digital Electroprospecting System Version 3.0. Rusia: Sigma LLC
Shi C, Wang Y (2021) Non-parametric machine learning methods for interpolation of spatially varying non-stationary and non-Gaussian geotechnical properties. Geosci Front 12(1):339–350
Singh A, et al (2016) A review of supervised machine learning algorithms. Paper presented at the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)
Sotomayor LN et al (2023) Supervised machine learning for predicting and interpreting dynamic drivers of plantation forest productivity in northern Tasmania Australia. Comput Electr Agricult. https://doi.org/10.1016/j.compag.2023.107804
Sullivan C, Kaszynski A (2019) PyVista: 3D plotting and mesh analysis through a streamlined interface for the Visualization Toolkit (VTK). J Open Source Software 4(37):1450
Tilahun T, Korus J (2023) 3D hydrostratigraphic and hydraulic conductivity modelling using supervised machine learning. Appl Comput Geosci 19:100122
Tiwari A (2022) Supervised learning: From theory to applications. In: Artificial intelligence and machine learning for EDGE computing (pp. 23–32): Elsevier
Tsai JP et al (2019) Constructing the apparent geological model by fusing surface resistivity survey and borehole records. Groundwater 57(4):590–601
Vapnik V (1999) The nature of statistical learning theory: Springer science & business media
Waskom ML (2021) Seaborn: statistical data visualization. J Open Source Software 6(60):3021
Witter JB, Melosh G (2018) The value and limitations of 3D models for geothermal exploration. Paper presented at the 43rd Workshop on Geothermal Reservoir Engineering, no. article SGP-TR-213, 2018Standford Univeristy, Standford, California
Yang S, Berdine G (2017) The receiver operating characteristic (ROC) curve. Southwest Respir Crit Care Chron 5(19):34–36
Yang CH, Lee WF (2002) Using direct current resistivity sounding and geostatistics to aid in hydrogeological studies in the Choshuichi Alluvial Fan, Taiwan. Groundwater 40(2):165–173

Publisher's Note