Journal of *Marine Science and Engineering*

MDPI

Article

A Machine Learning Approach for the Clustering and Classification of Geothermal Reservoirs in the Ying-Qiong Basin

Yujing Duan [1], Yuan Liang [2,3,4,*], Qingyun Ji [2,4] and Zhong Wang [3]

[1] Education Evaluation and Supervision Division, Chengdu University of Technology, Chengdu 610059, China; duanyj@cdut.edu.cn
[2] School of Mathematical Sciences, Chengdu University of Technology, Chengdu 610059, China; 2022051185@stu.cdut.edu.cn
[3] College of Management Science, Chengdu University of Technology, Chengdu 610059, China; wangzhong@cdut.edu.cn
[4] Geomathematics Key Laboratory of Sichuan Province, Chengdu University of Technology, Chengdu 610059, China
* Correspondence: liangyuan13@cdut.edu.cn

Abstract: The exploration and development of marine geothermal energy is a field with significant potential, but it is also one that presents considerable challenges and costs. The assessment of marine geothermal reservoir potential is currently based on subjective analysis, and this study proposes an innovative clustering-based method to classify marine geothermal reservoirs systematically. The Yingqiong Basin was analysed to develop a machine learning framework to predict the potential of marine geothermal reservoirs (CPPOGR). The study integrated eight key geothermal features into a unified dataset, employing dimensionality reduction techniques (principal component analysis and sparse autoencoder) and SMOTE to balance the sample size. Machine learning classifiers, including XGBoost, BP Neural Networks, Support Vector Machines, K-Nearest Neighbours, and Random Forests, were utilised for prediction. The experimental results demonstrate that XGBoost is the most suitable classifier, achieving an excellent performance of 0.96 precision, 0.9556 recall, 0.9528 F1 score, and 0.9623 accuracy. These results demonstrate the effectiveness of the proposed CPPOGR in accurately classifying marine geothermal reservoirs based on intrinsic features. This study underscores the potential of integrating cluster analysis with machine learning for efficient reservoir characterisation, thereby offering a novel approach for marine geothermal resource assessment.

Keywords: Ying-Qiong basin; oceanic geothermal reservoir; cluster analysis; machine learning

check for updates

Academic Editor: Mohamed Benbouzid

Received: 23 January 2025
Revised: 17 February 2025
Accepted: 20 February 2025
Published: 23 February 2025

Citation: Duan, Y.; Liang, Y.; Ji, Q.; Wang, Z. A Machine Learning Approach for the Clustering and Classification of Geothermal Reservoirs in the Ying-Qiong Basin. *J. Mar. Sci. Eng.* **2025**, *13*, 415. https://doi.org/10.3390/jmse13030415

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Geothermal energy is a significant source of renewable energy, with considerable development potential, minimal environmental pollution, sustainable use, and a low impact on natural climate [1,2]. The rapid development of geothermal energy technology has enabled the reliable provision of heat and electricity for industrial and civil use [3]. The increased demand for clean and low-carbon energy production has led to a surge in research activities in the geothermal energy industry, in line with the general trend in the energy sector [4–6]. Technological advances in geothermal energy development have made it an increasingly important part of the global energy mix.

Marine geothermal exploration faces a unique set of challenges when compared with land-based geothermal exploration. Despite advances in seafloor drilling technology, which have enabled scientists to study seafloor heat sources at greater depths and with greater

precision, there are still many technical barriers to drilling in deep ocean environments. These include drilling depths, equipment pressure tolerance, and the high cost of seafloor operations [7]. These challenges make the assessment and exploitation of ocean geothermal resources complex and expensive [8]. The collection of ocean heat flow data is similarly limited by environmental conditions. Although some information on seafloor heat sources can currently be obtained through seafloor heat flow soundings and seafloor thermal gradient analyses, data access is still relatively scarce [9]. Furthermore, the assessment of marine geothermal exploration relies heavily on geophysical techniques, such as seismic and electromagnetic surveys, which aid researchers in conducting preliminary analyses of seafloor heat source distribution, heat flow characteristics, and thermal storage [10]. However, there are technical barriers to the application of these exploration methods in deep-sea environments, especially in terms of high-precision measurements and multilevel data fusion.

The focus of contemporary research in the domain of marine geothermal exploration is directed towards specific regions, notably tropical and subtropical seas, with a particular emphasis on specific hydrothermal systems in the South China Sea, the Pacific Ocean, and the Atlantic Ocean. These regions have been the focus of particular attention due to the existence of documented evidence of active ocean geothermal activity, thereby enhancing the comprehension of their ocean geothermal resource potential [11]. However, less information is available for marine geothermal than for terrestrial geothermal, which makes it more difficult to determine which geothermal resources should be prioritised for exploitation [12]. Consequently, the development of regional geothermal models is imperative prior to undertaking offshore exploratory endeavours, with the objective of optimising resource utilisation and mitigating exploration risks [13].

Cluster analysis has been demonstrated to be a valuable tool in the assessment and development of geothermal resources. The utilisation of cluster-based analytical models in geothermal resource development and utilisation has been shown to facilitate the identification of potential hotspot areas [14]. Furthermore, the application of cluster analysis to geothermal reservoirs has enabled the identification of high resource density areas in neighbouring regions [15], thereby enabling developers to focus on hotspot areas, enhancing resource utilisation efficiency, and providing a blueprint for the development of geothermal resources. In recent years, clustering algorithms have been extensively applied to address various challenges in the geothermal field. For instance, Unternährer et al. developed a mixed-integer linear programming model to optimise urban energy infrastructure by dividing urban systems into spatial clusters using integer linear programming [16]. Lindsey et al. utilised principal component analysis and clustering identification to evaluate exploration sites and direct subsequent development towards high-potential geothermal clusters [17].

Di Giuseppe et al. demonstrated the application of a k-mean cluster analysis-based technique to the Solfatara-Pisciarelli volcano geothermal system, which is used to characterise buried structures in complex geothermal environments and to reveal local relationships between different physical parameters [18]. Hranić et al. proposed a two-stage clustering method using a temperature similarity and spatial proximity combination This approach was used to optimise the placement of geothermal production facilities, with the objective of achieving efficient production [19]. In a similar vein, Cavur et al. implemented an unsupervised Self-Organising Mapping (SOM) algorithm to analyse the distribution of geothermal zones in the state of Nevada. This study successfully defined areas of subsidence and uplift of the ground surface, as well as the classification of geothermal zones [20]. In the field of geothermal prediction methods, a substantial body of research has been conducted by numerous scholars. Ullah et al. utilised geothermal parameters to identify fault constraints and proposed a method to estimate the thermal structure of

faults [21]. This method possesses the potential for application to many basins worldwide that exhibit similar sedimentary and igneous geology. Addressing the limitations of the mineral classification model in the Sichuan Basin, China, which underutilises logging data, Ullah et al. developed a mineral classification model by integrating three machine learning methods and logging data, achieving an accuracy of over 70% [22].

Xiao et al. identified a lack of a feasible economic assessment model for the closed-loop geothermal system (CGS) and proposed a novel economic assessment model based on a coupled technology–economy model [23]. Furthermore, Xiao et al. established a technology–economy coupling model for a low-grade thermal energy utilisation system during drilling and developed an efficient low-grade thermal energy utilisation method based on insulated drill pipes [24]. Ullah et al. proposed a history matching method based on Bayesian PSO and GRU for optimising geothermal reservoir modelling, which is robust and effective in addressing the issues of non-uniqueness, subsurface uncertainty, and computational intensity [25]. Gudala et al. utilised a fully connected neural network to establish mathematical relationships between reservoir parameters and temperatures [26], which can be employed to study the temperature evolution through numerical simulations. Similarly, Suzuki et al. utilised machine learning to estimate input parameters for geothermal numerical simulations, with the estimated distributions displaying a comparable trend to the actual distributions [27]. Khaled et al. proposed a machine learning model capable of accurately predicting bottomhole circulating temperatures (BHCTs) in real time [28], and this enhancement in active temperature management is anticipated to alleviate the challenges of thermal genesis in geothermal drilling operations.

Notwithstanding the plethora of advances achieved by preceding studies, the present study is confronted with a distinctive set of challenges:

1. The paucity of reference materials for the construction of marine geothermal datasets is predominantly ascribed to the dearth of information and the exorbitant exploration costs [4,7].

2. The identification of the most suitable clustering method and the number of clusters for datasets in this domain necessitates specific problem-oriented considerations [29,30].

3. Information redundancy between ocean geothermal reservoir features is inevitable; without mitigation measures, subsequent machine learning models may exhibit poor accuracy and low robustness.

The main objective of this study is to develop a powerful machine learning model that can effectively classify unknown geothermal resources in the Yingqiong Basin of the South China Sea and neighbouring basins and assess their development potential. While scholars have previously used different machine learning approaches to statistically analyse geothermal resources, and these studies have provided valuable insights for resource exploration [31]; the uniqueness of this study lies in its particular challenges and its innovative solutions.

In order to address the aforementioned issues, the present study initially collated recent drilling data from offshore platforms in the Ying-Qiong Basin. The data were then subjected to a rigorous screening process based on four criteria: the presence of a dynamic heat source, the existence of a thermal conduction channel, the presence of a heat-storage rock mass, and the presence of heat-retaining cap rocks. The application of cluster analysis to the dataset yielded eight features that characterise oceanic geothermal reservoirs. Prior to conducting the cluster analysis, correlation coefficient thermal maps were employed to ascertain the suitability of this approach by identifying low feature correlations. The number of clusters was determined using both the elbow method and the silhouette coefficient. The K-Means++ clustering method was employed for analysis, with visual validation through two-dimensional and three-dimensional scatterplots, followed by an examination of oceanic

geothermal reservoir classification in the Ying-Qiong Basin. In addition, dimensionality reduction techniques such as principal component analysis (PCA) and sparse autoencoder (SAE) were applied to address sample imbalance using Synthetic Minority Oversampling Technique (SMOTE) before inputting processed results into various machine learning models to develop a Classifier for Predicting Oceanic Geothermal Reservoir Potential (CPOGRP).

2. Materials and Methods

2.1. Study Area

The South China Sea is located at the intersection of the Eurasian, Indo-Australian, and Pacific tectonic plates, and is regarded as one of the most extensive marginal seas in the western Pacific Ocean. It is also China's most substantial and profound offshore sea region. The region exhibits relatively intricate geological features and has undergone a complex tectonic evolution process. Specifically, the northern margin has experienced extension, the eastern margin has undergone subduction, the southern margin has experienced extrusion, and the western margin has undergone shearing, resulting from the interactions between the three major plates: Eurasian, Australian, and Pacific plates.

It is widely acknowledged that complex tectonic processes are frequently accompanied by thermal evolution, and it is this process that has resulted in the current geothermal characteristics of the South China Sea Basin. The heat flow in the South China Sea ranges from approximately 9 to 181 mW/m^2. The heat flow in the northern and southern parts of the South China Sea is lower, with an average of around 75 mW/m^2 [32]. A gradual increase in heat flow is observed from the north and south basins towards the central basin. Two centres of high heat flow have been identified in the central region, located in proximity to the Yinggehai Basin and the Qiongdongnan Basin. The Yinggehai Basin and the Qiongdongnan Basin are combined and known as the Ying-Qiong Basin, containing rich hydrothermal activities and diapiric structures, exhibiting a high mean heat flow value of 78.7 mW/m^2 [33,34], and the exact locations are shown in Figure 1.

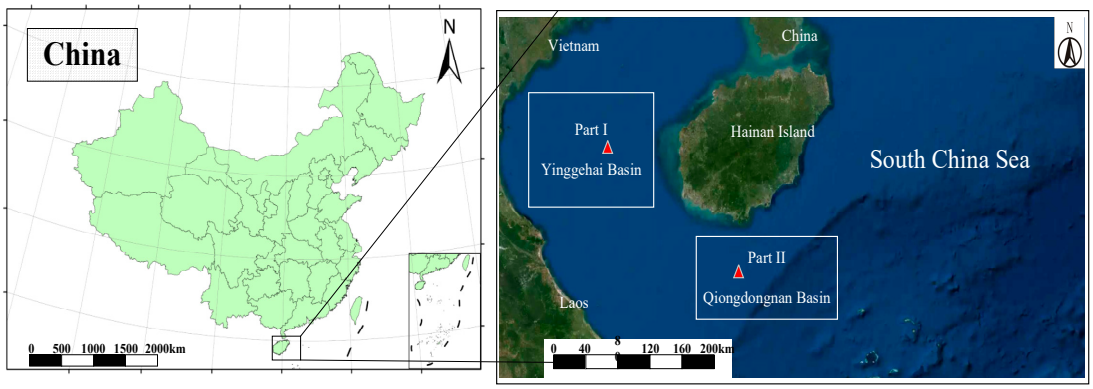


Figure 1. Schematic diagram of the Ying-Qiong Basin research area. (**a**) Regional map of China; (**b**) Ying-Qiong Basin study area, Part I of which is Yinggehai Basin and Part II of which is Qiongdongnan Basin.

The Ying-Qiong Basin is located in the coastal region where the Tethys and Pacific tectonic domains converge. This encompasses the Yinggehai Basin, which covers an area of approximately 7.7 × 104 km^2, and the Qiongdongnan Basin, which covers an area of about 8.9 × 104 km^2. The total distribution area of the Ying-Qiong Basin is more than 16 × 104 km^2 [35]. The current geothermal field in the Ying-Qiong Basin is principally influenced by the area's geology and deep crustal structure. Geophysical data indicates that

the Moho surface in the Ying-Qiong Basin is shallow, and crustal thickness is comparatively thin. This extensional thinning of the crust has resulted in the upwelling and partial melting of soft fluvial materials. Consequently, these molten materials were transferred to the lower and middle crust by convection along the deep ruptures and subsequently accumulated in the basin to form a magma chamber. This process has been identified as a primary factor in the generation of volcanism within the basin. Consequently, the contemporary geothermal field of the Ying-Qiong Basin is distinguished by elevated heat flow and a pronounced geothermal gradient [36], as shown in Figure 2.

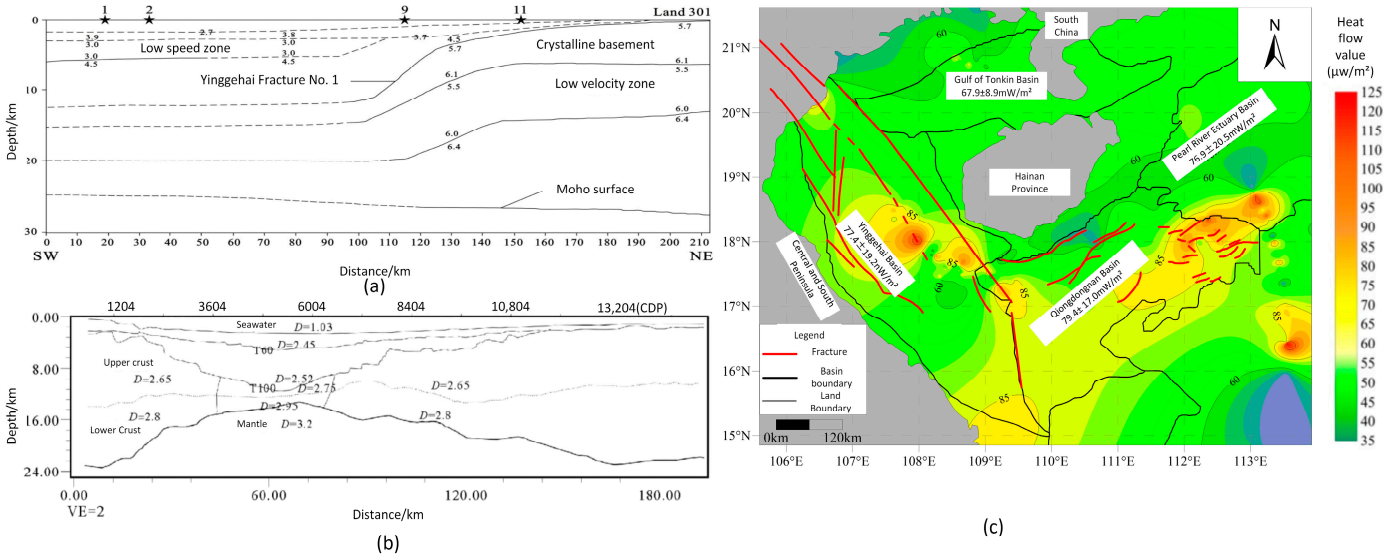


Figure 2. Map of crustal structure and geothermal flow in the depression of Yingqiong Basin [37]. (**a**) Crustal structure of the central depression in the Yinggehai Basin; (**b**) crustal structure of Changchang Depression, Qiongdongnan Basin; (**c**) geothermal heat flow map of Yingqiong Basin.

In the South China Sea region, from the northern land margin to the southern land margin, the crustal thickness shows a gradual thinning from both sides towards the centre of the basin. Furthermore, the trend of crustal heat flow is consistent with the crustal thickness. Consequently, the lithospheric thermal structure of the entire South China Sea region is expressed as a 'cold crust and hot mantle'. The overall distribution of heat flow in the Qiongdongnan Basin of the northern land margin exhibits the trend of 'low in the north and high in the south, low in the west and high in the east', and the fracture structure in this region is highly developed, with some fractures remaining active, thereby affecting the contemporary distribution of local heat flow, as shown in Figure 3.

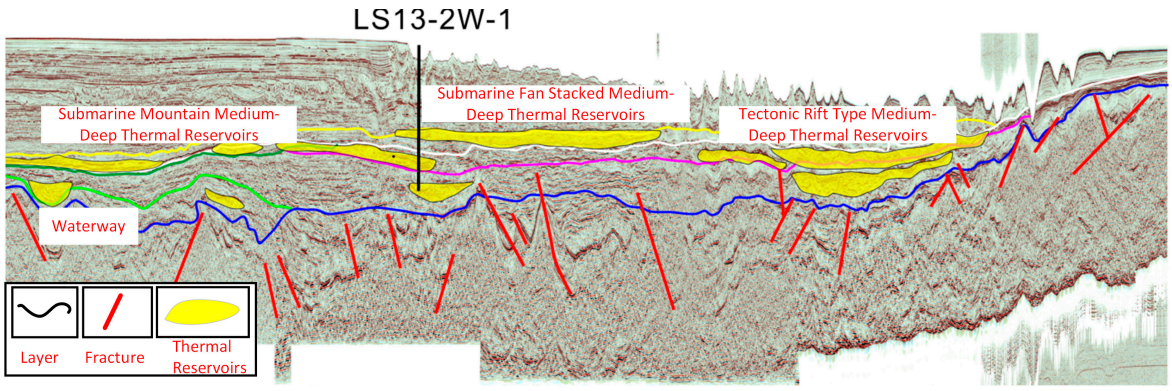


Figure 3. Geological profile of the Yingqiong Basin.

The Yinggehai Basin, a high-heat basin under the influence of the Red River Fracture, has developed a series of bottom-breaking tectonics as a result of slip fracture zones in the centre of the basin. These tectonics facilitate the upward transmission of deep heat, contributing to the high heat flux characteristics of the region. The western part of the southern margin of the South China Sea is characterised by a tensional environment, resulting from the sliding and torsion of the boundary faults. The upwelling of the deep mantle heat source is evidenced by high heat flow.

The Yinggehai Basin is currently in the late stages of geothermal development. In this paper, 26 sandstone thermal reservoirs of the Huangliu Formation in the central depression of the Yinggehai Basin were analysed using logging, seismic interpretation, and development data, and the depths and thicknesses of the top and bottom of each thermal reservoir were determined. The results show that the depths of the top and bottom of Section 1 of the Huangliu Formation are 462.5 m–4700.0 m and 567.7 m–4805.1 m, with average depths of 2576.8 m and 2723.6 m, respectively, and the average thickness of the reservoirs is 50.7 m. For Section 2 of the Huangliu Formation, the depths of the reservoirs range from the top of the section, 844.9 m, to 4477. For Section 2 of the Huangliu Formation, the depths of the reservoirs range from the top of the section, 844.9 m, to 4477.3 m, and the average thickness of the reservoir is 50.7 m. For Section 3 of the Huangliu Formation, the depths of the reservoirs range from the top of the section, 844.9 m, to 4477.3 m, and the average depth of the reservoir is 3135.5 m; the average thickness of the reservoir is 3920.1 m, respectively.

Therefore, the temperature variations in the Yinggehai Basin from −1 km to −5 km depth (distal to the seafloor) were investigated. Figure 4 shows the dynamic relationship between temperature variation and depth in the Yinggehai Basin.

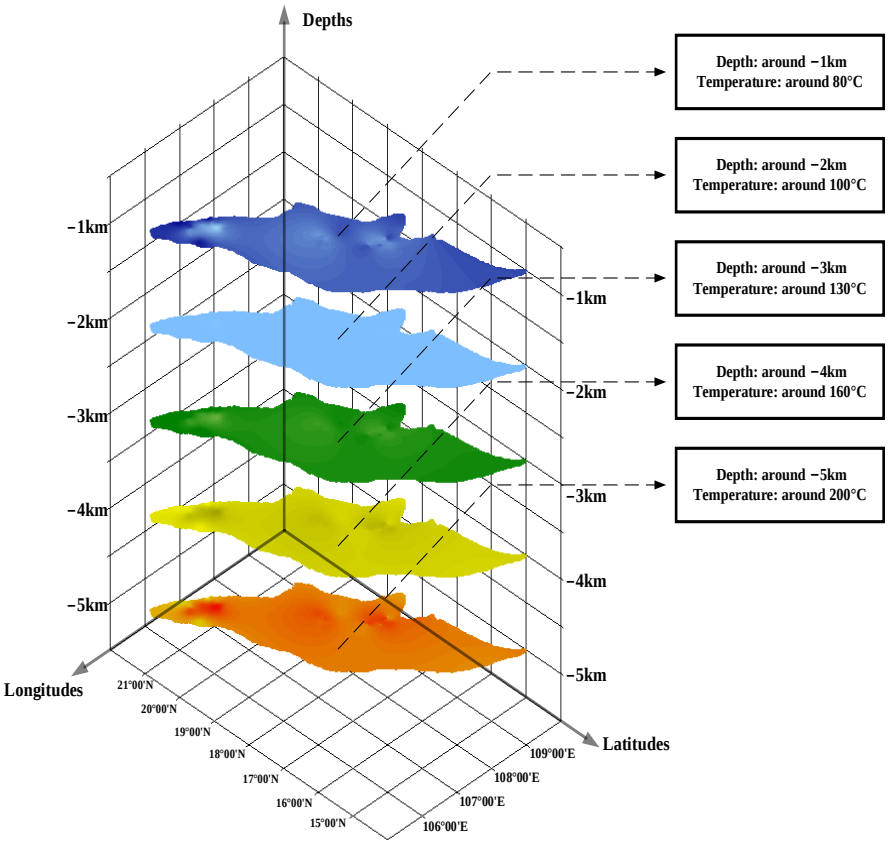


Figure 4. Dynamics of temperature and depth in the Yinggehai Basin.

2.2. Data Sources

The Yingqiong Basin is located in the northern part of the South China Sea and has been affected by several phases of tectonic movements, resulting in a complex crustal structure and a rich depositional environment. The stratigraphy of the area is well developed and includes sedimentary sequences from the Cenozoic to the Mesozoic, providing a rich source of organic matter for hydrocarbon generation. The Yingqiong Basin has developed a variety of favourable sedimentary systems, such as deep-water channels, shelf-edge deltas, and channelled fans. These depositional systems not only provide good reservoirs but also converge the material sources to the deepwater area of the basin by gravity flow and other means, forming high-quality oil and gas reservoirs. A number of oil and gas wells have been drilled in the area in recent years, covering a wide range of types such as high-temperature, high-pressure, and deepwater wells. Examples include the Lingshui series of gas fields, the Dongfang series of gas fields, the Yacheng series of gas fields, and the Ledong series of gas fields.

The Yinggehai Basin includes the First Member of the Yinggehai Formation, the Second Member of the Yinggehai Formation, the First Member of the Huangliu Formation, the Second Member of the Huangliu Formation, the Meishan Formation, the Sanya Formation, and the Lingshui Formation (see Table 1). The Qiongdongnan Basin, meanwhile, encompasses the Huangliu Formation, Meishan Formation, Sanya Formation, Lingshui Formation, and Yacheng Formation [38], as illustrated in Table 1.

Table 1. Distribution range of the dataset.

Basin Type	Formation Type
Yinggehai Basin	First Member of the Yinggehai Formation Second Member of the Yinggehai Formation First Member of the Huangliu Formation Second Member of the Huangliu Formation Meishan Formation Sanya Formation Lingshui Formation
Qiongdongnan Basin	Huangliu Formation Meishan Formation Sanya Formation Lingshui Formation Yacheng Formation

The data utilised in this study are drawn from a preceding set of drilling records obtained from offshore platforms, encompassing a total of 316 temperature readings from 37 wells in the Ying-Qiong Basin, of which 178 temperature readings are derived from 21 wells in the Yinggehai Basin and 138 temperature readings are sourced from 16 wells in the Qiongdongnan Basin. Subsequently, the dataset of oceanic geothermal reservoirs in the Ying-Qiong Basin, with a capacity of 102 records, was organised and calculated. The dataset comprises eight data features: heat flow, geothermal gradient, fracturing structure, reservoir depth, reservoir temperature, porosity, resource abundance, and caprock thickness. These features are encompassed by the four dimensions of dynamic heat source, thermal conduction channel, heat-storage rock mass, and heat retaining cap rocks in the evaluation of geothermal reservoirs. In the subsequent classification algorithms, we split the dataset into a training set and a test set, with 70% for the training set and 30% for the test set.

2.3. Research Framework

As demonstrated in Figure 5, this study employs a five-step approach to cluster analysis and classification predictive modelling of geothermal reservoirs in the Ying-Qiong Basin. In the initial step, data features are selected and feature scaling is performed to normalise the data features. In the subsequent step, the K-means++ method is employed to cluster the data. The number of clusters is determined by the elbow method and the silhouette coefficient. The ensuing step involves the visualisation of the clustering effect. The third step involves the use of principal component analysis (PCA) and sparse autoencoder (SAE) to downscale the data features and remove redundant information from the original features. In the fourth step, the Synthetic Minority Oversampling Technique (SMOTE) is employed to balance the sample data. In the fifth step, the K-means++ clustering results are used as data labels in machine learning and the two groups of dimensionally reduced features are used as inputs. The training effect was then compared using various machine learning methods combined with K-fold cross-validation with grid search. Ultimately, eXtreme Gradient Boosting (XGBoost) was selected as the automatic classifier model in this study.

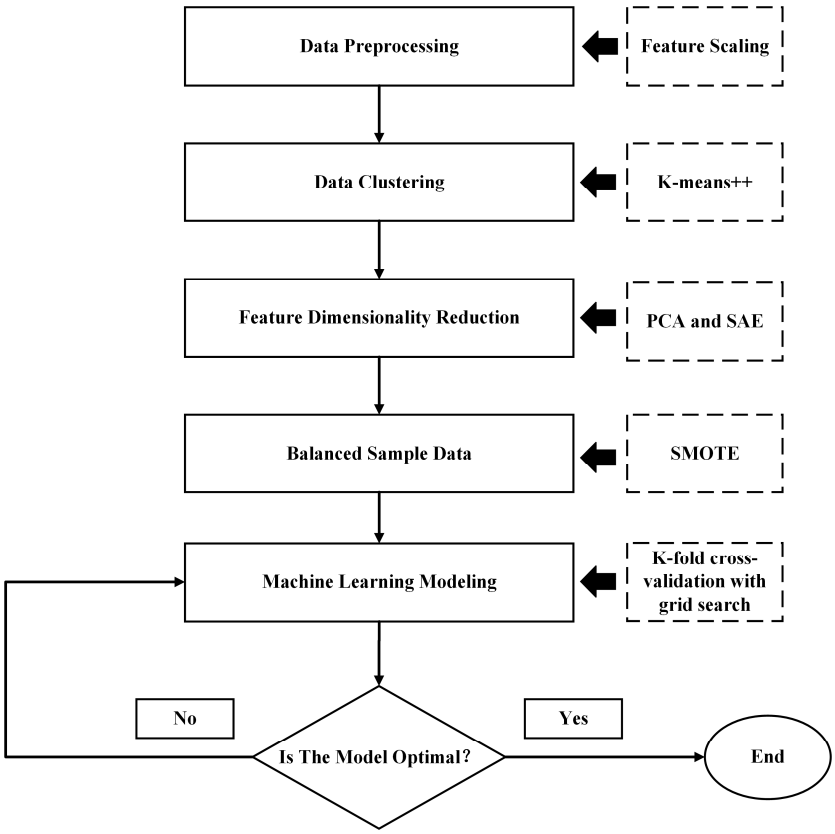


Figure 5. Workflow diagram of the study.

2.4. Method

2.4.1. Feature Scaling

Cluster analysis and machine learning are more sensitive to the distribution of the data. If the influence of the scale is not eliminated, then excessive training of the model will be observed for variables with larger values, whilst inadequate training will be observed for variables with smaller values. Consequently, prior to the input of a substantial dataset into the network, it is imperative to implement feature scaling, a process that concomitantly accelerates network training and enhances the efficacy of model training. A range of commonly used dataset feature scaling methods exist, including Min-Max Normalisation and Z-Score Normalisation [39]. In this study, Min-Max Normalisation is utilised to

transform the data into the range [0, 1] by leveraging the maximum difference in the data, as depicted in Equation (1).

$$x_{new} = \frac{x - \min(x)}{max(x) - min(x)}, \quad (1)$$

2.4.2. Cluster Analysis

Cluster analysis is an unsupervised learning technique that reveals the intrinsic structure of data by algorithmically dividing N items into K similar groups. It was first used in recommender systems to group users for personalised recommendations. In recent years, algorithms such as K-means++ (based on partitioning theory), DBSCAN (based on density theory) and Gaussian mixture model (based on probabilistic model) have been widely used. This study compares and tests these three clustering algorithms, citing the relevant literature [40–42].

The objective of the K-means++ clustering algorithm is to partition the data into k clusters. The probability of a sample point being selected as another centre of clustering is dependent on its distance from the centre of the other clusters. Each sample belongs to exactly one of the clusters whose distance from the centre of clustering is the shortest. The centre point corresponding to each sample data is also determined. The calculation of the Euclidean distance from each object to each centre of clustering is outlined in Equation (2).

$$\mathrm{dis}(X_i, C_j) = \sqrt{\sum_{t=1}^{m} (X_{it} - C_{jt})^2} \quad (2)$$

where X_i represents the ith object $(1 \leq i \leq n)$, C_j represents the jth centre of clustering $(1 \leq j \leq k)$, X_{it} represents the tth feature of the ith object $1 \leq t \leq m$, and C_{jt} represents the tth feature of the jth centre of clustering.

Specifically, K-mean++ is based on the following three steps through iteration:

1. Choose a sample point at random to serve as the initial centre of clustering C_1.

2. Calculate the shortest distance between each sample and the current centre of clustering. Express this as $D(x)$. The larger the $D(x)$, the greater the probability that the sample is chosen as the centre of clustering. Finally, select the sample point that corresponds to the maximum probability value as the next centre of clustering.

3. Repeat step 2 until the k centre of clustering is selected.

The k clusters generated by K-means++ provide the basis for the classification of oceanic geothermal reservoirs in the Ying-Qiong Basin, which is a prerequisite for further classification predictions.

2.4.3. PCA and SAE

(1) PCA

The specific algorithm flow of PCA is shown as follows. First, the sample set $X = \{x_1, x_2, \ldots, x_m\}$ was inputted, where X is an $n \times m$ sample matrix, and the goal is to downscale the n-dimensional space to d-dimensions. The samples were then zero-centred as shown in Equation (3).

$$x_i \leftarrow x_i - \frac{1}{m} \sum_{i=1}^{m} x_i, \quad (3)$$

The covariance matrix XX^T is computed for all samples, followed by eigenvalue decomposition of XX^T. Subsequently, m eigenvalues are obtained and sorted in descending order: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. Then, the unit eigenvector corresponding to the first d

eigenvalues is determined. The unit eigenvector corresponding to the ith eigenvalue is shown in Equation (4).

$$\omega_i = (a_{1i}, a_{2i}, \ldots, a_{mi})^T, i = 1, 2, \ldots, d, \tag{4}$$

From the eigenvectors $\omega_1, \omega_2, \ldots, \omega_{\mathrm{d}}$ corresponding to the largest d eigenvalues, the matrix $\mathrm{W} = (\omega_1, \omega_2, \ldots, \omega_{\mathrm{d}})$ is constructed, where W is an n × d orthogonal matrix, and the principal component of the ith sample is shown in Equation (5).

$$\hat{x}_i = a_i^T x,\ i = 1, 2, \ldots, d, \tag{5}$$

By mapping the n-dimensional space to d-dimensional, the sample matrix of d × m is obtained. Where the principal component matrix $\hat{\mathrm{X}} = \mathrm{W}^{\mathrm{T}}\mathrm{X}$, where the output projection matrix is W and the sample principal component matrix is $\hat{\mathrm{X}}$.

(2) SAE

SAE is a deep learning model for learning a low-dimensional representation of data. It consists of two parts, an encoder and a decoder, and learns a compact representation of the input data by minimising the reconstruction error. The goal of SAE is to learn sparsity, i.e., to make the neuron activity during encoding as sparse as possible, thus inducing the model to selectively retain key features of the input data.

The SAE is mainly divided into 4 parts,

1. Encoder stage: first input sample set $\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_{\mathrm{m}}\}$, through the encoder network to obtain the low dimensional representation h; the encoder consists of fully connected layers, and the encoder's computational process can be expressed as shown in Equation (6).

$$h = f_e(X; \theta_e), \tag{6}$$

where $\mathrm{f_e}$ denotes the encoder network and θ_{e} denotes the encoder parameters.

2. Decoder stage: The low-dimensional representation h is reconstructed from the original data by a decoder network $\hat{\mathrm{X}}$. The goal of the decoder network is to restore the input data as much as possible. The computational procedure of the decoder is shown in Equation (7).

$$\hat{X} = f_d(h; \theta_d), \tag{7}$$

where $\mathrm{f_d}$ denotes the decoder network and θ_{d} denotes the decoder parameters.

3. Loss function: the sparse autoencoder trains the model parameters by the sum of minimum reconstruction error and regularisation for sparsity, the loss function is shown in Equation (8).

$$Loss = \|X - \hat{X}\|^2 + \lambda \|h\|_1, \tag{8}$$

where λ is a hyperparameter that controls the weight of the regularisation for sparsity.

4. Optimisation update: The backpropagation algorithm facilitates the calculation of the gradient of the loss function with respect to both the encoder and the decoder. Subsequently, an optimisation algorithm is employed to update the parameters and minimise the loss function.

2.4.4. Over-Sampling Treats the Classification of Unbalanced Samples

The sample labels derived from the cluster analysis frequently exhibit an imbalance in the sample classification labels. The failure to balance the samples can result in machine learning algorithms tending to favour the dominant class. This can further lead to a deterioration in the model's performance. To address this issue, this study proposes the

use of the Synthetic Minority Oversampling Technique (SMOTE) as a means of balancing the proportion of samples in each category [43].

SMOTE is based on the feature space of the samples and synthesises the samples by interpolating the minority class samples. The following steps are involved:

1. For each minority class sample, the distance from all other minority class samples is calculated and the K nearest neighbours are identified (K = 3 is set in this study);

2. A randomly selected sample from the K nearest neighbours is then calculated, and the difference between that sample and the current sample is determined;

3. A new synthetic sample is then generated based on the proportion of difference, which is located on the connecting line between the two samples;

4. The aforementioned steps should be repeated to generate the requisite number of synthetic samples.

The SMOTE oversampling operation enables the proportion of samples in the training set of the oceanic geothermal reservoir constructed in this study to be balanced, thereby facilitating effective improvement in the accuracy of the prediction for the test set. This demonstrates that SMOTE has a superior balancing effect on the imbalanced classification dataset of the oceanic geothermal reservoir in this study.

2.4.5. XGBoost

XGBoost is a machine learning algorithm based on gradient boosting, proposed by Dr Tianqi Chen of the University of Washington, and based on gradient boosted regression trees (GBRT) and classification and regression trees (CART) [44,45]. The algorithm achieves integrated learning of multiple trees by integrating multiple decision trees (CART), with efficient computational power and strong predictive performance as well as good interpretability. It is found that XGBoost performs well in Kaggle competitions and is widely used in scientific prediction and is able to improve prediction accuracy with less computational resources. Its loss function consists of two parts: the first part is the loss function, which is used to measure the error between the predicted value and the true value; the second part is the regularisation term, which is used to control the model complexity to prevent overfitting.

2.5. Modelling Evaluation

2.5.1. Evaluation Indicators

This paper evaluates the performance of several machine learning algorithms using metrics like precision, recall, F1-score, and accuracy within the specified dataset. In classification tasks, the model categorises samples into two groups: positive and negative cases.

Precision measures the proportion of correctly predicted positive cases (true positives) among all predicted positives. A higher precision indicates better ability to identify negative cases. The calculation process is delineated in Equation (9).

$$Precision = \frac{TP}{TP + FP}, \tag{9}$$

Recall assesses how many positive cases were correctly identified. A higher recall reflects better performance in distinguishing positive cases. The calculation process is delineated in Equation (10).

$$Recall = \frac{TP}{TP + FN}, \tag{10}$$

The F1-score combines precision and recall, ranging from 0 to 1. A higher value suggests better balance between precision and recall, especially useful in imbalanced datasets. The calculation process is delineated in Equation (11).

$$F1 - score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)}, \quad (11)$$

Accuracy is the overall percentage of correct predictions. The calculation process is shown in Equation (12).

$$ccuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (12)$$

2.5.2. ROC Curve

The ROC curve is a graphical tool employed in the evaluation of the performance of a classification model. The ROC curve delineates the relationship between the true positive rate (TPR) and the false positive rate (FPR) of a classifier at varying classification thresholds. The ideal ROC curve is situated as close as possible to the upper left corner, indicating that the model exhibits a high true positive rate while maintaining a low false positive rate. Conversely, the closer the ROC curve is to the diagonal, the poorer the performance of the classifier is, as it is unable to effectively distinguish between positive and negative classes [46].

3. Results

3.1. Correlation Analysis

In order to assess the independence between the seven selected geothermal features and their interrelationships, a correlation coefficient analysis is required and a heat map of the correlation coefficients is drawn [47]. This heat map can visualise the degree of correlation between the geothermal features and thus help to analyse the correlation between them.

If the correlation between these geothermal features is too high, it may lead to a poorer interpretation of the model and the model may be overfitted in subsequent analyses, thus affecting the prediction accuracy of the test set. For example, as shown in Figure 6, the correlation coefficient between geothermal gradient and reservoir temperature is 0.56, while the correlation coefficient between reservoir depth and reservoir temperature is 0.88. This suggests that, in the subsequent classification analyses, a dimensionality reduction approach is necessary to reduce the feature dimensions.

3.2. Oceanic Geothermal Reservoir Clustering Based on Geothermal Features

As demonstrated in Figure 7a, a distinct 'inflexion point' at k = 5 is evident, signifying that the decline in SSE attributable to an increase in the number of clusters beyond this threshold becomes negligible. Conversely, Figure 7b reveals that the maximum value (~0.223) is attained at k = 5, suggesting that at this point, the data exhibits maximum density within the clusters, thereby ensuring optimal separation between distinct clusters. The findings from these two complementary evaluation metrics indicate that k = 5 not only significantly reduces the clustering error but also achieves an optimal balance between model complexity and clustering quality, thus avoiding overfitting problems while ensuring the interpretability and usefulness of the clustering results.

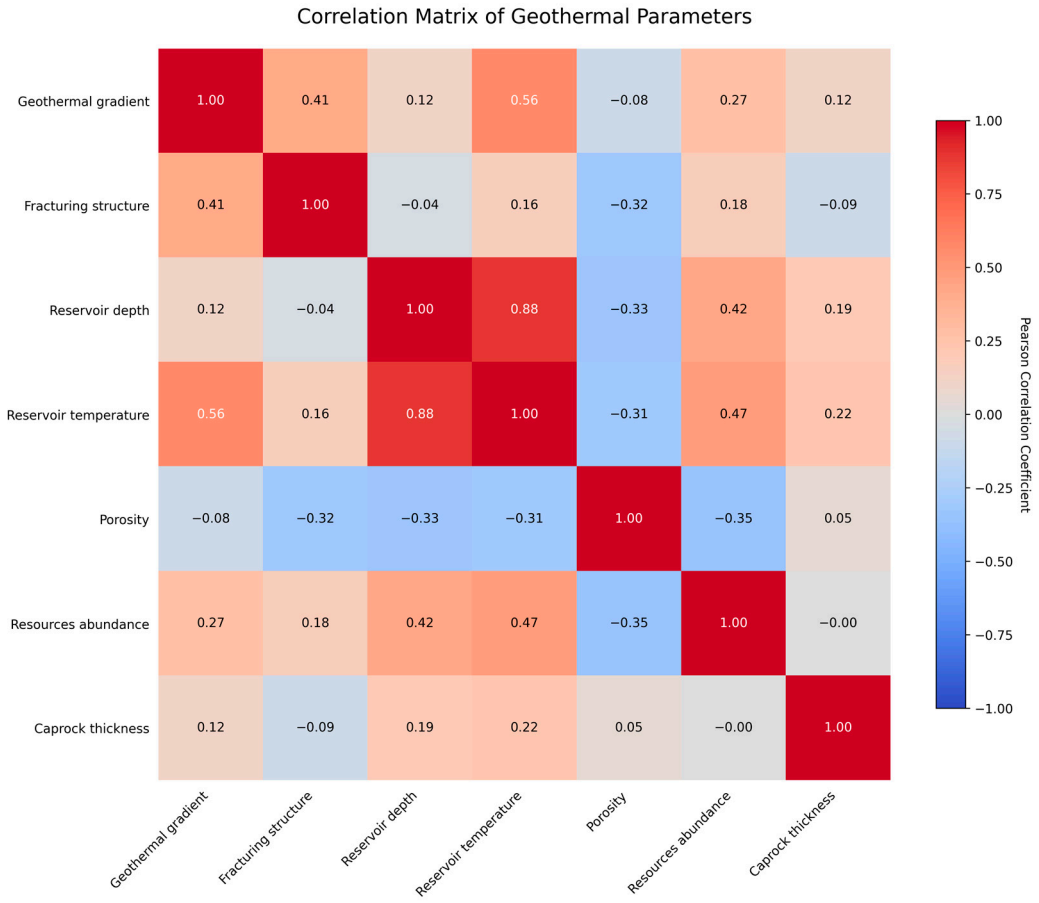


Figure 6. Correlation heatmap.

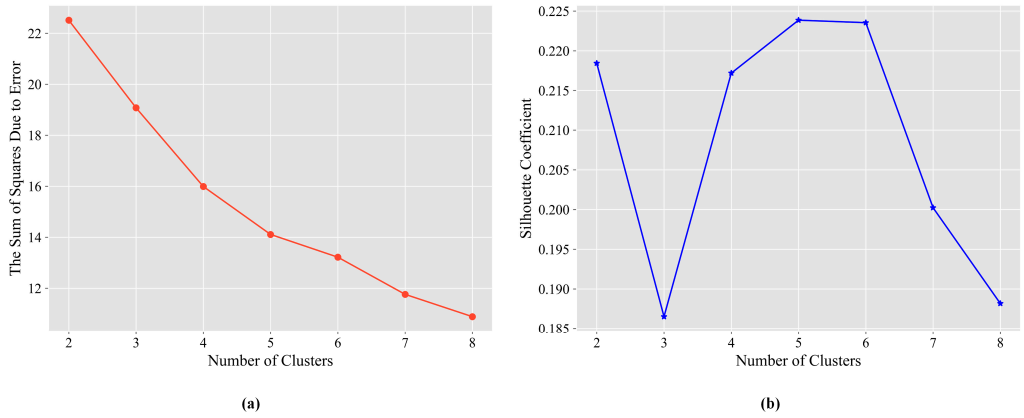


Figure 7. The process of obtaining the optimal value of K. (**a**) Elbow Method. (**b**) Silhouette Coefficient.

In order to establish a classification system for oceanic geothermal reservoirs in the Ying-Qiong Basin, the K-means++ clustering algorithm was utilised for the effective classification of these reservoirs into five distinct classes: Class A, Class B, Class C, Class D, and Class E. Subsequent to the implementation of the clustering process, additional categorical labels were appended to the initial oceanic geothermal reservoir dataset in the Ying-Qiong Basin. The classification labels represent the five clustering results derived from the cluster analysis. However, it should be noted that the interpretation of these results remains contingent upon statistical analysis and expression. Consequently, a statistical analysis of the clustering results is undertaken in conjunction with geothermal features. Utilising the geothermal features in four dimensions of dynamic heat source, thermal conduction channel, heat-storage rock mass, and heat retaining cap rocks, the average values of geothermal

features of various types of oceanic geothermal reservoirs can be calculated, and the values of the centroids of the five clusters are shown in Table 2.

Table 2. Mean Geothermal Features of Various Types of Oceanic Geothermal Reservoirs: Minimum and Maximum Values Across Five Types.

Geothermal Reservoir Classification	Dynamic Heat Source		Thermal Conduction Channel	Heat-Storage Rock Mass				Heat-Retaining Cap Rocks	Number
	Heat Flow (mW/m^2)	Geothermal Gradient (°C/km)	Fracturing Structure (pcs)	Reservoir Depth (m)	Reservoir Temperature (°C)	Porosity (%)	Resources Abundance (J/m^2)	Caprock Thickness (m)	
Class A	76.36	39.24 (minimum)	1.00 (minimum)	2575.44 (minimum)	127.37 (minimum)	16.35 (maximum)	1.77×10^{10} (minimum)	171.01	33
Class B	69.48	41.15	2.00	6003.82 (maximum)	271.00	7.09	1.60×10^{11} (maximum)	201.47	15
Class C	68.30 (minimum)	40.30	3.91	3426.56	163.75	6.34	4.20×10^{10}	120.12 (minimum)	35
Class D	71.95	39.67	2.87	4055.66	185.19	6.16	4.64×10^{10}	314.80 (maximum)	15
Class E	84.41 (maximum)	73.15 (maximum)	17.50 (maximum)	4118.09	324.03 (maximum)	4.54 (minimum)	1.43×10^{11}	181.00	4

3.3. Feature Extraction Based on PCA and Sparse Autoencoder

In addressing the challenge of oceanic geothermal reservoir classification, we initially observe the presence of eight original features. However, these features inevitably contain redundant information. To address this challenge, we propose the utilisation of dimensionality reduction methods to streamline the features. This approach enables the retention of effective information while eliminating redundant information that holds minimal significance for the prediction of the classification of oceanic geothermal reservoirs. Despite the limited number of features in the dataset under consideration, low feature counts tend to yield superior results on small datasets. Consequently, the following two dimensionality reduction methods were selected for the purpose of experimental comparison.

In this study, two different ways of reducing the dimensionality of the data features are used to downscale the data, including principal component analysis (PCA) and sparse autoencoder (SAE). PCA is a method of dimensionality reduction based on linear transformation. This method projects the data into a new coordinate system by finding the principal components in the original data. The principal components are defined as linear combinations that exhibit the highest degree of variability within the original data. This process enables the retention of the most significant information within the data.

The optimal number of features extracted by dimensionality reduction is then tested, and the performance of machine learning is compared for different numbers of dimensionality reduction features input. The three possible scenarios were tested when the number of dimensionality reduction features were 3, 4, and 5, and the dimensionality reduction features were input into the XGBoost model to test the effect, respectively. The optimal setting for the PCA dimensionality reduction features was determined to be 4, and the SAE dimensionality reduction features were set to 4 to achieve the same effect. Consequently, the total number of dimensionality reduction features was established as 4, as illustrated in Table 3.

After the dimensionality reduction processing, the prompting of the next step of the machine learning effect can be achieved more effectively; the detailed dimensionality reduction processing flow is shown in Figure 8.

Table 3. Performance of dimensionally reduced data on XGBoost model.

Model	Number of Features	Precision	Recall	F1-Score	Accuracy	Is It Optimal?
PCA-XGBoost	3	0.8861	0.8369	0.8344	0.8302	
PCA-XGBoost	4	0.9222	0.9106	0.9128	0.9245	√
PCA-XGBoost	5	0.9081	0.8546	0.8717	0.8679	
SAE-XGBoost	3	0.9029	0.8825	0.8719	0.8868	
SAE-XGBoost	4	0.9600	0.9556	0.9528	0.9623	√
SAE-XGBoost	5	0.9271	0.9213	0.9204	0.9245	

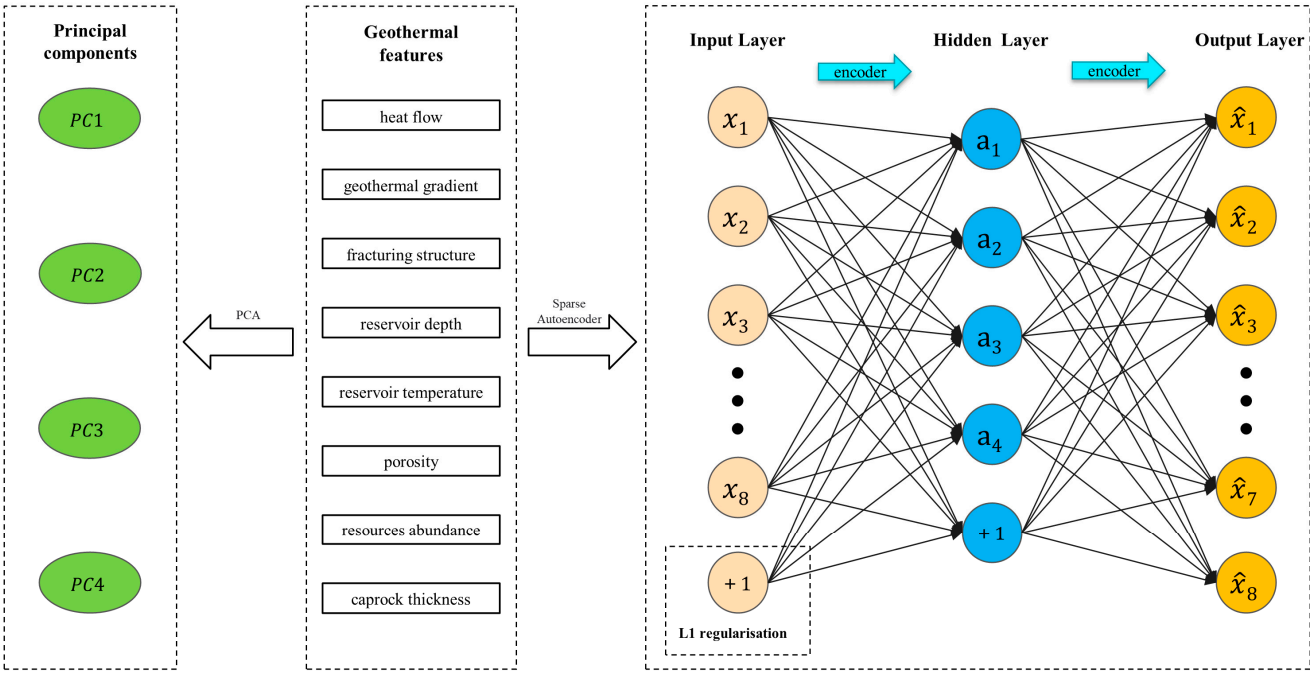


Figure 8. Example of dimensionality reduction in an oceanic geothermal reservoir feature.

The objective of this study is to ascertain the most efficacious dimensionality reduction method for the purpose of feature fusion extraction. To this end, the effectiveness of the features extracted by dimensionality reduction via PCA is compared with those extracted by SAE in order to determine the most suitable dimensionality reduction method for the task of classifying oceanic geothermal reservoirs. The efficacy of the dimensionality reduction method is determined by comparing the prediction evaluation metrics of the machine learning model on the test set, in which XGBoost is selected for the machine learning model. The XGBoost prediction results without dimensionality reduction, with PCA dimensionality reduction, and with SAE dimensionality reduction are shown in Table 4.

Table 4. Comparison of the effects of dimensionality reduction methods.

Model	Precision	Recall	F1-Score	Accuracy
XGBoost	0.9444	0.9328	0.9375	0.9434
PCA-XGBoost	0.9222	0.9106	0.9128	0.9245
SAE-XGBoost	0.9600	0.9556	0.9528	0.9623

It is evident that the impact of the features processed by the SAE input to the machine learning model has undergone a substantial enhancement, as evidenced by the precision of 0.9600, in comparison to the impact of the undimensioned features prior to enhancement of 0.0556; recall being 0. 9556, compared with the effect of the undimensioned features before the enhancement of 0.0228; F1-score being 0.9528, compared with the effect of the undimensioned features before the enhancement of 0.0147; and accuracy being 0.9623, compared with the effect of the undimensioned features before the enhancement of 0.0189. This demonstrates that the sparse self-coder can achieve superior dimensionality reduction in the Ying-Qiong Basin oceanic geothermal reservoir dataset constructed in this study. The sparse self-coder effectively combines features, retains effective information in the machine learning classification task, and eliminates redundant information. Furthermore, the Loss curve and Sparsity Loss curve of the SAE during the iteration of 100 Epochs should be observed.

As demonstrated in Figure 9, the Loss curve demonstrates a decline from 0.21 to 0.0185 after 100 Epochs, reaching a state of stability, while the Sparsity Loss curve exhibits a decrease from 1.28 to 0.178 after 100 Epochs. This suggests that the training of the SAE has converged, the model parameters have stabilised, and the model has attained an optimal state. Consequently, it can be employed for the subsequent feature reduction task.

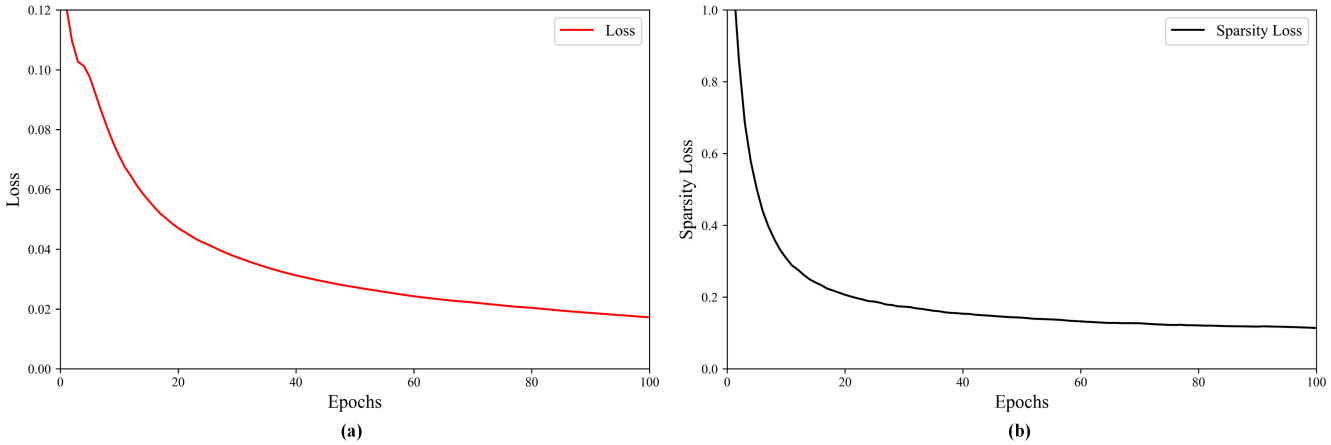


Figure 9. Loss curves and Sparsity Loss curves for sparse autoencoder. (**a**) Loss curves. (**b**) Sparsity Loss curves.

It is also evident that the application of PCA to the output of the machine learning model results in a reduction in predictive ability. This phenomenon can be attributed to the inherent nature of PCA as a linear dimensionality reduction method. In the event of a robust nonlinear relationship between the original features, PCA is unable to encapsulate these latent structures, which ultimately results in suboptimal dimensionality reduction.

Consequently, an analysis of various dimensionality reduction techniques was conducted, and the sparse autoencoder was identified as the optimal technique for processing the oceanic geothermal reservoir data features. This approach involves transforming the original eight features in the dataset into four reduced and fused features, thereby enhancing the effectiveness of the machine learning model.

3.4. Machine Learning for Geothermal Reservoir Classification Modelling

In order to efficiently categorise undiscovered and unexplored geothermal heat reservoirs within oceans, it is proposed that a machine learning approach be utilised for the classification of said reservoirs. The clustering results described above have resulted in the allocation of category labels to 102 data points on ocean geothermal energy. These data points were then modelled using the XGBoost classifier. In the subsequent classification

algorithms, we split the dataset into a training set and a test set, with 70% for the training set and 30% for the test set.

It is important to note the significant imbalance of samples in the given class labels. For instance, Class E of the study's dataset contains only four samples. This imbalance could potentially result in overfitting of the majority of samples in the classes when directly inputting the data into the XGBoost model without oversampling. To address this imbalance, an optimisation method based on the SMOTE for the classification of imbalanced samples is employed in this paper to enhance XGBoost. In instances where the number of samples belonging to certain categories is limited, over-sampling can be employed to effectively augment the number of minority class samples and enhance their representation in the training process. This, in turn, leads to an improvement in the model's performance, particularly in terms of the accuracy and recall of the classifier.

In addition, in order to enhance the efficacy of the XGBoost model's predictions, a range of hyperparameters were adjusted, encompassing nine pivotal hyperparameters (namely, n_estimators, max_depth, learning_rate, gamma, reg_alpha, reg_lambda, min_child_weight, subsample and colsample_bytree) [48]. To ascertain the optimal combination of hyperparameters, a grid search and K-fold cross-validation for hyperparameter selection were employed [49,50]. Through a comprehensive evaluation and comparative analysis of various hyperparameter configurations, it was ascertained that the XGBoost model attains optimal performance when the hyperparameters are set to the values delineated in Table 5. This set of hyperparameter settings can inform machine learning models when confronted with other oceanic geothermal reservoir datasets.

Table 5. Hyperparametric setting.

Hyperparameter	Value
n_estimators	100
max_depth	8
learning_rate	0.05
gamma	0.1
reg_alpha	0.001
reg_lambda	0.001
min_child_weight	2
subsample	0.8
colsample_bytree	1

As demonstrated in Table 6, the classification efficacy of XGBoost for oceanic geothermal reservoirs on the test set indicates that precision, recall, F1-score, and accuracy are all more satisfactory. To facilitate comparative analysis, four machine learning models with distinct prediction principles were selected for consideration, including the BP Neural Network, SVM, KNN, and Random Forest. XGBoost demonstrated superior performance in all four evaluation metrics, outperforming the other models. Among the machine learning models employed for comparison, Random Forest exhibited the most favourable results. This may be attributed to the fact that both XGBoost and Random Forest are decision tree-based models. In the context of the classification problem of oceanic geothermal reservoirs in this study, it can be hypothesised that a decision tree-based model may be more effective in achieving accurate predictions than other models.

Table 6. Evaluation metrics for various machine learning model classifiers.

Model	Precision	Recall	F1-Score	Accuracy
SAE-XGBoost	0.9600	0.9556	0.9528	0.9623
SAE-BP Neural Network	0.9306	0.9306	0.9294	0.9434
SAE-SVM	0.9017	0.9071	0.8978	0.9057
SAE-KNN	0.9229	0.9179	0.9095	0.9057
SAE-Random Forest	0.9355	0.9357	0.9325	0.9434

The efficacy of machine learning performance was also assessed through the use of ROC curves, which are fed into XGBoost and other machine learning models for comparison. The resulting micro-average ROC curves are illustrated in Figure 10. The ROC curves for XGBoost demonstrate the greatest distance from the diagonal and the closest proximity to the upper left corner. This finding suggests that the XGBoost model has the potential to attain a higher true positive rate (TPR) while maintaining a lower false positive rate (FPR). The area under the ROC curve (AUC) is estimated to be one.

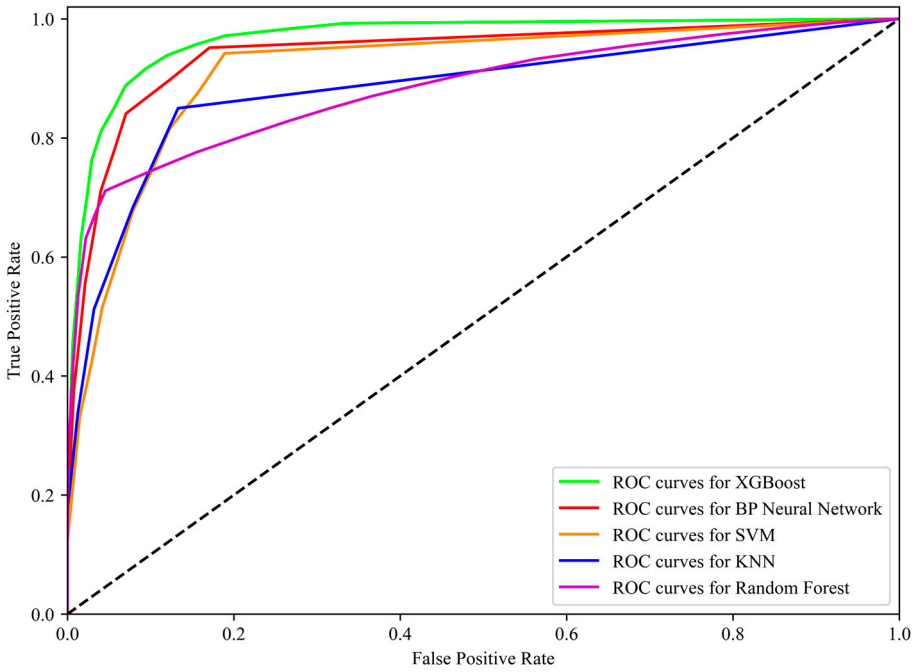


Figure 10. Comparison of ROC curves for five machine models.

In summary, the oceanic geothermal reservoir dataset constructed in this study was subjected to SAE dimensionality reduction and SMOTE processing, and it was confirmed that this dataset is effective when input into various types of machine learning models. Among these models, XGBoost was found to have the best effect. The machine learning classification evaluation metrics and ROC curves demonstrate that XGBoost is more effective than other machine learning models. To further investigate the predictive performance of XGBoost on diverse oceanic geothermal reservoirs, the output evaluation metrics are presented in Table 7.

Table 7. Predictive performance of XGBoost classifier on various types of oceanic geothermal reservoirs.

Classifications	Precision	Recall	F1-Score
A	Approximate to 1	0.7778	0.8750
B	Approximate to 1	Approximate to 1	Approximate to 1
C	0.8000	Approximate to 1	0.8889
D	Approximate to 1	Approximate to 1	Approximate to 1
E	Approximate to 1	Approximate to 1	Approximate to 1

The prediction performance of various types of oceanic geothermal reservoirs indicates that the optimal XGBoost classifier among different machine learning models exhibits shortcomings in predicting Classes A and C oceanic geothermal reservoirs while achieving evaluation metrics approximate to 1 for other types of oceanic geothermal reservoirs. This discrepancy may be attributed to the potential overfitting risk associated with SMOTE treatment on a limited sample size. Notwithstanding this potential limitation, the oceanic geothermal reservoir classification predictor developed in this study exhibits a commendable degree of predictive capability within the Ying-Qiong Basin. It facilitates the efficient automatic prediction of the class to which an oceanic geothermal reservoir belongs. Subsequent research should prioritise data collection for oceanic geothermal reservoirs in neighbouring basins to validate the model's robustness.

4. Discussion

4.1. Analysis of Clustering Results for Marine Geothermal Reservoirs

The geothermal gradient, fracturing structure, reservoir depth, reservoir temperature, and resources abundance of Class A oceanic geothermal thermal reserves are all at their lowest in Class A–E, and the porosity is at its maximum in class A-E, with the remaining geothermal indexes at a medium level. Class A is characterised by low-temperature, shallow geothermal energy with low geothermal temperature gradients, weak thermal conductivity channels, and large porosity, with low reserves. In contrast, class B oceanic geothermal thermal reserves demonstrate the largest reservoir depth and resource abundance in class A–E, with the remaining geothermal indicators exhibiting median values. Class B, on the other hand, is characterised by deep geothermal energy with abundant resources.

The heat flow and caprock thickness of Class C oceanic geothermal heat storage are the smallest in class A–E, and the rest of the geothermal indexes are in the middle degree. Class C is positioned in the middle layer of geothermal energy, characterised by a dynamic heat source and suboptimal heat preservation. The caprock thickness of Class D oceanic geothermal reservoirs is the largest in Classes A–E, and the rest of the geothermal indicators are in the middle range. Class D is a mid-level geothermal energy with good heat retention. The heat flow, geothermal gradient, fracturing structure, and reservoir temperature of Class E oceanic geothermal energy storage are the largest in Classes A–E, and the porosity is the smallest in Classes A–E. The remaining geothermal indicators are found in the median range. Class E is positioned in the median layer of geothermal energy, characterised by a substantial heat source, effective heat conduction channels, and elevated heat storage temperatures.

The objective of this study is to assess the validity of the K-means++ clustering outcomes for oceanic geothermal reservoirs. The principal component analysis (PCA) dimensionality reduction technique was employed to reduce the dimensionality of eight oceanic geothermal reservoir features, and the clustering results were assigned to each sample. Two and three principal components were extracted for the oceanic geothermal reservoir features for the two-dimensional and three-dimensional scatterplots, respectively.

Following the assessment of the clustering effect, it was determined that Classes A–E can all be regarded as both representative and distinguishable. The scatter plot reveals that Class A–E can be distinguished from each other by a specific division law, as illustrated in Figure 11.

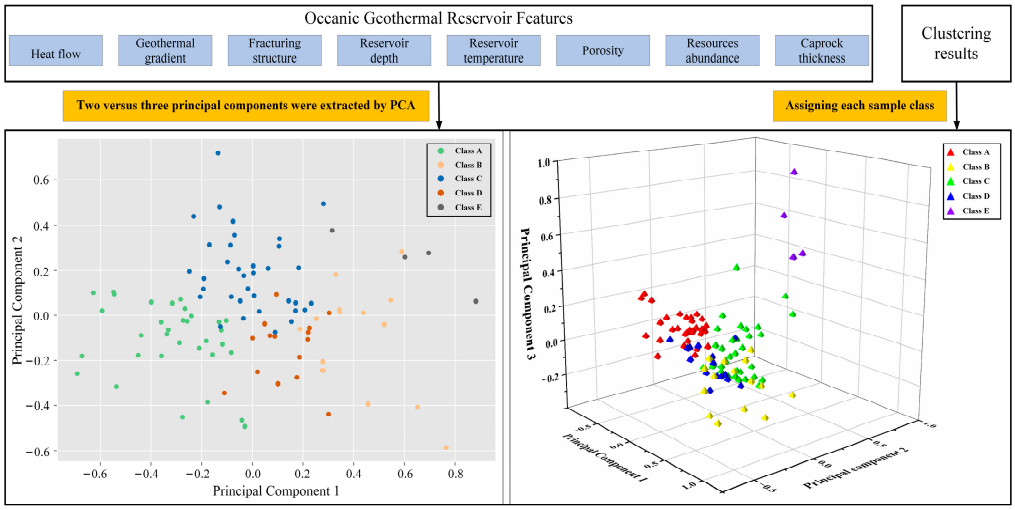


Figure 11. Two-dimensional and three-dimensional scatterplots to test clustering effects.(**a**) 2D clustering diagram; (**b**) 3D clustering diagram.

4.2. Analysis of Marine Geothermal Reservoir Clusters

The final clustering result is illustrated in Figure 12; the *X*-axis is the five categories and the *Y*-axis is the number of individuals. Different colours represent different formations. The 33 samples from the Yinggehai Basin are classified as belonging to Class A, with 5 samples originating from the First Member of the Yinggehai Formation, 9 from the Second Member of the Yinggehai Formation, 14 from the First Member of the Huangliu Formation, 3 from the Second Member of the Huangliu Formation, and 2 from the Lingshui Formation. The Class A category is notable for its substantial sample size, and it is therefore recommended as a shallow low-temperature geothermal potential group. The composition of Class B consists of nine samples from the Yinggehai Basin and six samples from the Qiongdongnan Basin. Of the samples from the Yinggehai Basin, two samples are from the First Member of the Huangliu Formation, one sample is from the Meishan Formation, four samples are from the Sanya Formation, two samples are from the Lingshui Formation, and one sample is from a combination of the Meishan Formation and the Sanya Formation. Of the samples obtained from the Qiongdongnan Basin, one originates from the Meishan Formation, one from the Sanya Formation, and four from the Lingshui Formation. Class B is a sample group of intermediate quantitative level, which we posit as a group of deep high-reserve geothermal potential. Class C consists of 9 samples from the Yinggehai Basin and 26 samples from the Qiongdongnan Basin. Of the samples from the Yinggehai Basin, two are from the First Member of the Yinggehai Formation, four are from the Second Member of the Huangliu Formation, one is from the Meishan Formation, one is from the Lingshui Formation, four are from the Huangliu Formation, six are from the Meishan Formation, and one is from the Lingshui Formation.

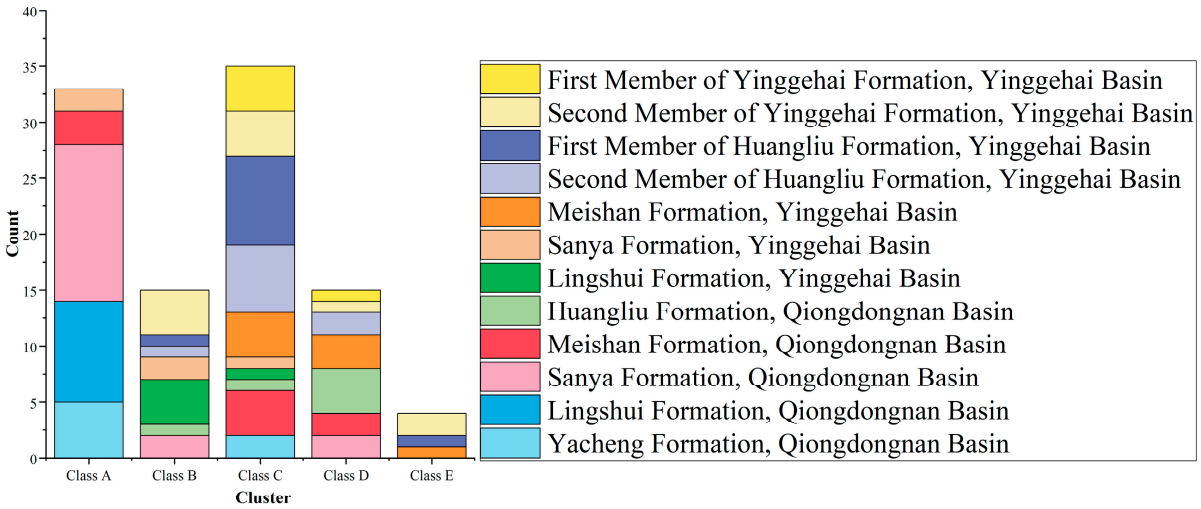


Figure 12. The data points are represented separately by clusters.

The samples from the Qiongdongnan Basin are as follows: four from the Huangliu Formation, six from the Meishan Formation, eight from the Sanya Formation, and four from the Lingshui Formation. In addition, four samples were obtained from the Huangliu Formation, six from the Meishan Formation, eight from the Sanya Formation, four from the Lingshui Formation, and four from the Yacheng Formation. The most numerous sample group is Class C, and it is suggested that this group may be a geothermal potential group, with weak heat retention from low heat sources in the middle layer. Class D consists of eight samples from the Yinggehai Basin and seven samples from the Qiongdongnan Basin. Of the samples from the Yinggehai Basin, two samples are from the First Member of the Huangliu Formation, two samples are from the Second Member of the Huangliu Formation, and four samples are from the Meishan Formation. In turn, the samples from the Qiongdongnan Basin comprise three from the Huangliu Formation, two from the Meishan Formation, one from the Lingshui Formation, and one from the Yacheng Formation. Class D is a sample group with intermediate numbers, which we recommend as a mid-level high heat-retention geothermal potential group. Class E, which comprises four samples from the Qiongdongnan Basin (one from the Huangliu Formation, one from the Sanya Formation, and two from the Lingshui Formation), is a notable example. Class E is one of the least numerous sample groups, and it is recommended as the middle high-heat source high-temperature geothermal potential group.

The findings of this study indicate that Class E is the most viable geothermal heat storage option in the Ying-Qiong Basin, despite having the fewest samples within the clusters. This is due to its distinct advantages over other classes. These advantages include higher temperatures, favourable dynamic heat source characteristics, and excellent thermal conductivity. The following section presents a sample of sand bodies within the Class E cluster containing Example Geothermal Reservoir No. 1, Example Geothermal Reservoir No. 2, Example Geothermal Reservoir No. 3, and Example Geothermal Reservoir No. 4, as shown in Table 8.

Following a comparative analysis of the key exploratory work in the Ying-Qiong Basin, it has been determined that the samples in Class E should be prioritised as targets for exploitation. It is evident that the samples in category E exhibit analogous geothermal characteristics, characterised by a robust dynamic heat source, effective thermal conduction channels, an excellent heat-storage rock mass, and moderate heat-retaining cap rocks. The E samples thus provide a reference point for the subsequent search for excellent oceanic geothermal reservoirs in the Ying-Qiong Basin and analogous basins.

Table 8. Sample data for specific sand bodies within Class E clusters.

Thermal Storage	Heat Flow (mW/m^2)	Geothermal Gradient (°C/km)	Fracturing Structure (pcs)	Reservoir Depth (m)	Reservoir Temperature (°C)	Porosity (%)	Resources Abundance (J/m^2)	Caprock Thickness (m)
Example Geothermal Reservoir No. 1	70.46	83.3	25.00	3202.61	291.78	9.46	1.65×10^{11}	263.00
Example Geothermal Reservoir No. 2	94.69	63.0	20.00	3337.21	235.24	4.70	5.70×10^{10}	140.00
Example Geothermal Reservoir No. 3	99.00	63.0	15.00	5333.93	361.04	1.80	1.58×10^{11}	94.00
Example Geothermal Reservoir No. 4	73.50	83.3	10.00	4598.60	408.06	2.20	1.91×10^{11}	227.00

4.3. Machine Learning Classification Prediction Effectiveness Analysis

The objective of this study is to make a comparison of the performance of stacked autocoder (SAE)-based hybrid models in the task of geothermal reservoir classification. Geothermal reservoir classification is a highly complex task involving the identification of nonlinear relationships and the extraction of multi-scale features. The study analysed five models: SAE-XGBoost, SAE-BP neural network, SAE-Random Forest, SAE-KNN, and SAE-SVM. The experimental results show that the SAE-XGBoost model exhibits the best performance (Precision = 0.9600, Recall = 0.9556, F1-score = 0.9528, Accuracy = 0.9623). The superior performance of the SAE-XGBoost model is attributed to the autocoder's capacity to model complex nonlinearities and interactions, in conjunction with XGBoost's proficiency in handling geothermal geological parameters.

In contrast, SAE-BP neural network and SAE-Random Forest have balanced performances under specific geological conditions, and each has its own advantages in dealing with geothermal fluid dynamics and fault features, respectively, but more suitable models may need to be selected according to the geological background in practical applications. SAE-KNN and SAE-SVM perform poorly due to their shortcomings which are attributed to limitations inherent in the SAE-KNN and SAE-SVM, such as spatial heterogeneity and their limited capacity to address multi-class problems. Conversely, the superior performance of the SAE-BP neural network and SAE-Random Forest is attributed to their capacity to model complex nonlinearities and interactions and the ability of XGBoost to effectively classify geothermal geological parameters. It is noteworthy that the SAE-BP neural network and SAE-Random Forest demonstrate particularly poor performance in the spatial distribution characteristics of geothermal reservoirs.

It is evident that these models demonstrate varying degrees of adaptability in addressing the multifaceted and varied characteristics inherent in geothermal data. Geothermal data are typically sourced from diverse origins, including geophysical exploration, drilling operations, geochemical analyses, and other methods. Self-coders have proven effective in the effective integration of these heterogeneous data sources, while different classifiers possess distinct capabilities in the interpretation of these integrated features. This is especially crucial for the detection and evaluation of deep geothermal resources, where high model accuracy is paramount to reduce exploration risks and optimise investment decisions.

Future research could further explore the applicability of these models in different geological contexts, such as the effectiveness of classifying different types of geothermal resources (e.g., volcanic, basin or geothermal enhancement system (EGS)) and the incorporation of time-series data (e.g., long-term temperature monitoring) into the models to improve predictive capability. Simultaneously, the integration of these sophisticated classification models with geostatistical methodologies, encompassing the quantification of geological uncertainties and risk evaluation, has the potential to furnish more comprehensive technical assistance for geothermal resource evaluation.

5. Conclusions

In this paper, we put forward a machine learning method for the automated classification of oceanic geothermal reservoirs in the Yingqiong Basin. Initially, a comparative analysis of several clustering methods was conducted, leading to the selection of K-means for the categorisation of oceanic geothermal reservoirs. The efficacy of this clustering approach was subsequently evaluated. The clustering results were then employed as labels, and eight geothermal features were collectively integrated into the dataset utilised in this study. Prior to the implementation of the machine learning classification algorithm, PCA and SAE dimensionality reduction were performed on the features of the dataset, with SAE being selected as the dimensionality reduction method in this study. This method is effective in the elimination of redundant information from the original features, the retention of important information, and the improvement of the effectiveness of the subsequent machine learning classification model. Following sample balancing by the SMOTE method, several machine learning classification algorithms were applied, with XGBoost being selected. The precision of this model was found to be 0.96, the recall was 0.9556, the F1-score was 0.9528, and the accuracy was 0.9623 on the test set.

This study has successfully demonstrated the feasibility of the model in the classification of oceanic geothermal reservoirs. The model has been demonstrated to facilitate the investigation of unknown oceanic geothermal reservoirs with minimal financial expenditure, while also enabling the classification of oceanic geothermal reservoirs within the Ying-Qiong Basin and the determination of development priorities. In comparison to traditional geothermal reservoir exploration methods, which often rely on extensive drilling, high operational costs, and lengthy timeframes for data collection, the model provides a more cost-effective and time-efficient alternative. The classification and evaluation of oceanic geothermal reservoirs represents a pivotal aspect within the broader context of geothermal development. The Classifier for Predicting the Potential of Oceanic Geothermal Reservoirs (CPPOGR) has been demonstrated to facilitate the advancement of geothermal development.

Author Contributions: Conceptualization, Y.D.; methodology, Q.J. and Z.W.; software, Y.L.; validation, Y.D. and Y.L.; formal analysis, Y.L.; investigation, Z.W.; resources, Z.W.; data curation, Y.L.; writing—original draft preparation, Q.J. and Y.L.; writing—review and editing, Y.D., Q.J., Y.L. and Z.W.; visualisation, Y.L.; supervision, Y.D.; project administration, Y.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China under Grant No. 52192622 and 2024 Natural Science Foundation of Sichuan Province (2024NSFSC0091) and the Philosophy and Social Science Research Fund of Chengdu University of Technology in 2022, grant number: YJ2022-ZD012.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Barbier, E. Geothermal energy technology and current status: An overview. *Renew. Sustain. Energy Rev.* **2002**, *6*, 3–65. [CrossRef]
2. Gallup, D.L. Production engineering in geothermal technology: A review. *Geothermics* **2009**, *38*, 326–334. [CrossRef]
3. Fridleifsson, I.B. Geothermal energy for the benefit of the people. *Renew. Sustain. Energy Rev.* **2001**, *5*, 299–312. [CrossRef]
4. Lund, J.W.; Boyd, T.L. Direct utilization of geothermal energy 2015 worldwide review. *Geothermics* **2016**, *60*, 66–93. [CrossRef]

5. Lund, J.W.; Toth, A.N. Direct utilization of geothermal energy 2020 worldwide review. *Geothermics* **2021**, *90*, 101915. [CrossRef]
6. Tester, J.W.; Anderson, B.J.; Batchelor, A.S.; Blackwell, D.D.; DiPippo, R.; Drake, E.M.; Garnish, J.; Livesay, B.; Moore, M.C.; Nichols, K. The future of geothermal energy. *Mass. Inst. Technol.* **2006**, *358*, 1–3. [CrossRef]
7. Compernolle, T.; Welkenhuysen, K.; Petitclerc, E.; Maes, D.; Piessens, K. The impact of policy measures on profitability and risk in geothermal energy investments. *Energy Econ.* **2019**, *84*, 104524. [CrossRef]
8. Jouanne, A.V.; Brekken, T.K.A. Ocean and Geothermal Energy Systems. *Proc. IEEE* **2007**, *105*, 2147–2165. [CrossRef]
9. Suárez-Arriaga, M.-C.; Bundschuh, J.; Samaniego, F. Assessment of submarine geothermal resources and development of tools to quantify their energy potentials for environmentally sustainable development. *J. Clean. Prod.* **2014**, *83*, 21–32. [CrossRef]
10. Kana, J.D.; Djongyang, N.; Raïdandi, D.; Nouck, P.N.; Dadjé, A. A review of geophysical methods for geothermal exploration, Renewable and Sustainable. *Energy Rev.* **2015**, *44*, 87–95. [CrossRef]
11. Wang, Y.; Liu, Y.; Dou, J.; Li, M.; Zeng, M. Geothermal energy in China: Status, challenges, and policy recommendations. *Util. Policy* **2020**, *64*, 101020. [CrossRef]
12. Zheng, H.; Luo, J. Progress in research on the exploration and evaluation of deep geothermal resources in the Fujian-Guangdong-Hainan region, China. *Energy Geosci.* **2024**, *5*, 100232. [CrossRef]
13. Yuan, H.; Jing, T.; Yin, Y.; Wei, S.; Gan, H.; Zhao, W.; Zhang, J. Geothermal model and development area of a fault-controlled geothermal zone along the Fujian coastal area of southeastern China. *Nat. Gas Ind. B* **2024**, *11*, 28–41. [CrossRef]
14. Kanzari, E.; Fazio, G.; Fricano, S. Analysing the energy landscape in Africa using cluster analysis: Drivers of renewable energy development. *Energy Policy* **2024**, *195*, 114366. [CrossRef]
15. Ali, M.; Zhu, P.; Jiang, R.; Huolin, M.; Ehsan, M.; Hussain, W.; Zhang, H.; Ashraf, U.; Ullaah, J. Reservoir characterization through comprehensive modeling of elastic logs prediction in heterogeneous rocks using unsupervised clustering and class-based ensemble machine learning. *Appl. Soft Comput.* **2023**, *148*, 110843. [CrossRef]
16. Unternährer, J.; Moret, S.; Joost, S.; Maréchal, F. Spatial clustering for district heating integration in urban energy systems: Application to geothermal energy. *Appl. Energy* **2017**, *190*, 749–763. [CrossRef]
17. Lindsey, C.R.; Neupane, G.; Spycher, N.; Fairley, J.P.; Dobson, P.; Wood, T.; McLing, T.; Conrad, M. Cluster analysis as a tool for evaluating the exploration potential of Known Geothermal Resource Areas. *Geothermics* **2018**, *72*, 358–370. [CrossRef]
18. Di Giuseppe, M.G.; Troiano, A.; Patella, D.; Piochi, M.; Carlino, S. A geophysical k-means cluster analysis of the Solfatara-Pisciarelli volcano-geothermal system, Campi Flegrei (Naples, Italy). *J. Appl. Geophys.* **2018**, *156*, 44–54. [CrossRef]
19. Hranić, J.; Raos, S.; Leoutre, E.; Rajšl, I. Two-Stage Geothermal Well Clustering for Oil-to-Water Conversion on Mature Oil Fields. *Geosciences* **2021**, *11*, 470. [CrossRef]
20. Cavur, M.; Moraga, J.; Duzgun, H.S.; Soydan, H.; Jin, G. Displacement Analysis of Geothermal Field Based on PSInSAR and SOM Clustering Algorithms a Case Study of Brady Field, Nevada—USA. *Remote Sens.* **2021**, *13*, 349. [CrossRef]
21. Ullah, J.; Luo, M.; Ashraf, U.; Pan, H.; Anees, A.; Li, D.; Ali, M.; Ali, J. Evaluation of the geothermal parameters to decipher the thermal structure of the upper crust of the Longmenshan fault zone derived from borehole data. *Geothermics* **2022**, *98*, 102268. [CrossRef]
22. Ullah, J.; Li, H.; Ashraf, U.; Heping, P.; Ali, M.; Ehsan, M.; Asad, M.; Anees, A.; Ren, T. Knowledge-based machine learning for mineral classification in a complex tectonic regime of Yingxiu-Beichuan fault zone, Sichuan basin. *Geoenergy Sci. Eng.* **2023**, *229*, 212077. [CrossRef]
23. Xiao, D.; Liu, M.; Li, L.; Cai, X.; Qin, S.; Gao, R.; Liu, J.; Liu, X.; Tang, H.; Li, G. Model for economic evaluation of closed-loop geothermal systems based on net present value. *Appl. Therm. Eng.* **2023**, *231*, 121008. [CrossRef]
24. Xiao, D.; Xiao, H.; Song, W.; Li, G.; Zhang, J.; Deng, H.; Guo, B.; Tang, G.; Duan, M.; Tang, H. Utilization method of low-grade thermal energy during drilling based on insulated Drill pipe. *Renew. Energy* **2024**, *225*, 120363. [CrossRef]
25. Ullah, J.; Li, H.; Soupios, P.; Ehsan, M. Optimizing geothermal reservoir modeling: A unified bayesian PSO and BiGRU approach for precise history matching under uncertainty. *Geothermics* **2024**, *119*, 102958. [CrossRef]
26. Gudala, M.; Tariq, Z.; Govindarajan, S.K.; Yan, B.; Sun, S. Fractured Geothermal Reservoir Using CO_2 as Geofluid: Numerical Analysis and Machine Learning Modeling. *ACS Omega* **2024**, *9*, 7746–7769. [CrossRef] [PubMed]
27. Suzuki, A.; Shi, S.; Sakai, T.; Fukui, K.-i.; Onodera, S.; Ishizaki, J.; Hashida, T. Automated parameter estimation for geothermal reservoir modeling using machine learning. *Renew. Energy* **2024**, *224*, 120243. [CrossRef]
28. Khaled, M.S.; Wang, N.; Ashok, P.; van Oort, E.; Wisian, K. Real-time prediction of bottom-hole circulating temperature in geothermal wells using machine learning models. *Geoenergy Sci. Eng.* **2024**, *238*, 212891. [CrossRef]
29. Smith, C.; Faulds, J.; Brown, S.; Coolbaugh, M.; Lindsey, C.; Treitel, S.; Ayling, B.; Fehler, M.; Gu, C.; Mlawsky, E. Characterizing Signatures of Geothermal Exploration Data with Machine Learning Techniques: An Application to the Nevada Play Fairway Analysis. In Proceedings of the 46th Workshop on Geothermal Reservoir Engineering, Stanford, CA, USA, 15–17 February 2021.
30. Vesselinov, V.V.; Ahmmed, B.; Mudunuru, M.K.; Pepin, J.D.; Burns, E.R.; Siler, D.L.; Karra, S.; Middleton, R.S. Discovering hidden geothermal signatures using non-negative matrix factorization with customized k-means clustering. *Geothermics* **2022**, *106*, 102576. [CrossRef]

31. Al-Fakih, A.; Abdulraheem, A.; Kaka, S. Application of machine learning and deep learning in geothermal resource development: Trends and perspectives. *Deep. Undergr. Sci. Eng.* **2024**, *3*, 286–301. [CrossRef]
32. Wan, Z.; Xu, X.; Wang, X.; Xia, B.; Sun, Y. Geothermal analysis of boreholes in the Shenhu gas hydrate drilling area, northern South China Sea: Influence of mud diapirs on hydrate occurrence. *J. Pet. Sci. Eng.* **2017**, *158*, 424–432. [CrossRef]
33. He, L.; Xiong, L.; Wang, J.; Yang, J.; Dong, W. Tectono-thermal modeling of the Yinggehai Basin, South China Sea. *Sci. China Ser. D Earth Sci.* **2001**, *44*, 7–13. [CrossRef]
34. Xie, Y.; Li, X.; Fan, C.; Tan, J.; Liu, K.; Lu, Y.; Hu, W.; Li, H.; Wu, J. The axial channel provenance system and natural gas accumulation of the Upper Miocene Huangliu Formation in Qiongdongnan Basin, South China Sea. *Pet. Explor. Dev.* **2016**, *43*, 570–578. [CrossRef]
35. Zhang, Q.; Hao, F. Evolution and petroleum systems of the Ying-Qiong Basin. *Sci. China Ser. D Earth Sci.* **1997**, *40*, 553–560. [CrossRef]
36. Qiu, N.; Wang, Z.W.; Wang, Z.F.; Sun, Z.P.; Sun, Z.; Zhou, D. Stratigraphic structure and crustal extension characteristics of the Qiongdongnan Basin in the northern South China Sea. *Chin. J. Geophys.* **2014**, *57*, 3189–3207. [CrossRef]
37. Chen, H.; Zheng, F.; Song, R.; Zhang, C.; Dong, B.; Zhang, J.; Zhang, Y.; Wu, T. Geothermal Resource Assessment and Development Recommendations for the Huangliu Formation in the Central Depress. *Yinggehai Basin. Sustain.* **2024**, *16*, 7104. [CrossRef]
38. Tang, X.; Zhang, G.; Liang, J.; Yang, S.; Rao, S.; Hu, S. Modelling of Thermal Effects of Igneous Intrusions on the Temperature Field and Organic Maturity in the Changchang Sag, Qiongdongnan Basin, South China Sea. *Chin. J. Geophys.* **2014**, *57*, 219–229. [CrossRef]
39. Patro, S.G.K.; Sahu, K.K. Normalization: A preprocessing stage. *arXiv* **2016**, arXiv:1503.06462. [CrossRef]
40. Arthur, D.; Vassilvitskii, S. k-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA, 7 June 2006. [CrossRef]
41. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2007.
42. Reynolds, D. Gaussian Mixture Models. In *Encyclopedia of Biometrics*; Li, S.Z., Jain, A., Eds.; Springer: Cham, Switzerland, 2009; pp. 659–663. [CrossRef]
43. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *arXiv* **2002**, arXiv:1106.1813. [CrossRef]
44. Zhang, Z.; Li, Y.; Jin, S.; Zhang, Z.; Wang, H.; Qi, L.; Zhou, R. Modulation Signal Recognition Based on Information Entropy and Ensemble Learning. *Entropy* **2018**, *20*, 198. [CrossRef] [PubMed]
45. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
46. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
47. Haffou, F.Z.; Ouzzaouit, L.A.; Qarbous, A.; Boudad, L. Geothermal reservoir temperature prediction using hydrogeochemical data of Northern Morocco: A machine learning approach. *Geothermics* **2025**, *127*, 103259. [CrossRef]
48. Jiang, Y.; Tong, G.; Yin, H.; Xiong, N. A Pedestrian Detection Method Based on Genetic Algorithm for Optimize XGBoost Training Parameters. *IEEE Access* **2019**, *7*, 118310–118321. [CrossRef]
49. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; Morgan Kaufman Publishing: Montreal, QC, Canada, 1995; Volume 2.
50. Wong, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit.* **2015**, *48*, 2839–2846. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.