# Recognition of Geothermal Surface Manifestations: A Comparison of Machine Learning and Deep Learning

**Yongzhu Xiong** ( xiongyz@jyu.edu.cn )
School of Geography and Tourism, Jiaying University
**Mingyong Zhu**
School of Geography and Tourism, Jiaying University
**Yongyi Li**
Institute of Deep Earth Sciences and Green Energy, College of Civil and Transportation Engineering, Shenzhen University
**Kekun Huang**
School of Mathematics, Jiaying University
**Yankui Chen**
School of Geography and Tourism, Jiaying University
**Jingqing Liao**
The eighth Geologic Survey, Guangdong Geological Bureau

**Research Article**

**Keywords:** geothermal manifestation, geothermal energy, deep learning (DL), support vector machine (SVM), decision tree (DT), k-nearest neighbor (KNN), photograph

**Posted Date:** March 25th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1377072/v2



**Version of Record:** A version of this preprint was published at Energies on April 15th, 2022. See the published version at https://doi.org/10.3390/en15082913.

# Abstract

Geothermal surface manifestations (GSMs) are direct clues towards hydrothermal activities of a geothermal system in the subsurface and significant indications for geothermal resource exploration. It is essential to recognize various GSMs for potential geothermal energy exploration. However, there is a lack of work to fulfill this task using Deep Learning (DL), which has achieved unprecedented successes in computer vision and image interpretation. This study aims to explore the feasibility of using a DL model to fulfill the recognition of GSMs with photographs. A new image dataset was created for the GSM recognition by preprocessing and visual interpretation with expert knowledge and a high-quality check after downloading images from the Internet. The dataset consists of seven GSM types, i.e., warm spring, hot spring, geyser, fumarole, mud pot, hydrothermal alteration, crater lake, and one type of none GSM, including 500 images of different photographs for each type. The recognition results of the GoogLeNet model were compared with those of three Machine Learning (ML) algorithms, i.e., Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbor (KNN), by using the assessment metrics of Overall Accuracy (OA), Overall $F_1$ score (OF) and Computational Time (CT) for training and testing the models via cross-validation. The results show that the retrained GoogLeNet model using transfer learning has significant advantages of accuracies and performances over the three ML classifiers with the highest OA and the biggest OF for both validation and test, and the fastest CT for both validation and test. On the contrary, the three selected ML classifiers perform poorly for this task due to their low OA, small OF, and long CT. It suggests that transfer learning with a pretrained network be a feasible method to fulfill the recognition of the GSMs. Hopefully, this study provides a reference paradigm to help promote further research on the application of state-of-the-art DL in the geothermics domain.


# 1. Introduction

The world is confronting three significant challenges, namely overpopulation, resources depletion, and environmental deterioration, which intertwine with each other and influence the world's sustainable development significantly. Geothermal energy, as an alternative resource for the 21st Century [1], is a green, clean, efficient, renewable, and non-carbon-based new energy. Together with other new energies such as solar, wind, and biomass, it can play an essential role in energy-saving and emission reduction, coping with energy shortages, environmental and climate change, and realizing green and sustainable development that the world is facing [2-5]. However, the development of geothermal resources has been slowly progressing, although the proven long-term sustainability of geothermal energy has remained one of the attractions for further exploration and exploitation [6]. Geothermal resource exploration is the basis for its development and utilization, which need first make geothermal potential mapping and evaluation. The surface manifestations of a geothermal system in a volcanic-geothermal area are generally the features that first stimulate mapping and exploration [1]. These manifestations, such as fumaroles, warm and hot springs, and mud pods, can give important hints to the availability and abundance of geothermal resources [7] and reveal the exploitation potential of geothermal resources [8]. Geothermal Surface

Manifestations (GSMs) are crucial for preliminary surveys of geothermal potential. Therefore, recognizing these features and manifestations is essential for geothermal potential mapping and evaluation.

Many GSMs, especially warm and hot springs, geysers, and fumaroles, have drawn plenty of attention from the geothermal community in recent decades. Many researchers investigated geothermal anomalies related to GSMs such as hot springs using geophysics [9-11], geochemistry [12-14], remote sensing [15-19], Geographic Information System (GIS) [20], statistical modeling [21] and conventional Machine Learning (ML) [6, 22, 23]. For example, Gentana *et al*. (2019) demonstrated that the fault system is correlated with the appearances of the GSMs in the Indonesia volcanic zone [24]; Freski *et al*. (2021) tested the effects of alteration degree, moisture, and temperature on laser return intensity for the GSMs. Most of these works revealed the distribution and formation of geothermal resources by integrating multi-source data with traditional approaches [25]. Compared to the previous ML models, Artificial Neural Network (ANN) is getting more and more attention nowadays [6]. Dramsch (2020) made an overview of the development of ML in geoscience in the recent 70 years with an emphasis on technical explanations of some popularly used ML models [26]. Muther *et al*. (2022) outlined Artificial Intelligence (AI) technology integration with geothermal reservoir characterization and management, discussed its potentials, limitations, and ways forward, compared different statistical, numerical, and AI/ML methods, and put forward a concept of Geothermal 4.0 stemmed from the concept of Industry 4.0 [6].

Deep Learning (DL), which has achieved unprecedented successes in computer vision and image interpretation in the latest decade, has witnessed its emerging uses in the geothermics domain. Gangwani *et al*. (2021) provided an approach for predicting geothermal energy production using a long-short term-memory sequence to sequence encoder-decoder neural network architecture [27]. Shahdi *et al*. (2021) explored the applicability of four ML models (i.e., deep neural network, ridge regression, extreme gradient boosting, and random forest) in predicting subsurface temperatures in the northeastern United States using bottom-hole temperature data and geological information from 20,750 wells [28]. Yang *et al*. (2022) proposed an innovative method for identifying the formation temperature field based on a deep belief network and successfully applied the technique to identify the formation temperature field for the southern Songliao Basin, northeast China [29]. Besides the application of DL in the geothermal domain, DL has also found its way to the geophysical domain and more. For example, Petrov *et al*. (2022) investigated shape carving methods of geologic body interpretation from seismic data based on DL. They found that the dilated fully convolutional network was suitable for handling the task of seismic data interpretation. However, it is unclear whether DL can be feasible for the recognition of GSMs. Inspired by this latest progress on the domains of geothermics and geophysics, we hypothesized that using DL technology could realize the recognition of GSMs.

Two key challenges are obstructing fulfilling the recognition task of GSMs: (i) lack of a suitable GSM dataset for the task and (ii) how to select an optimal DL model from a great number of deep neural network architectures such as Convolutional Neural Networks (CNN), deep belief networks [29], recurrent neural networks, generative adversarial networks to train and test on this dataset for obtaining a suitable DL model for the task. Therefore, it is necessary to create a GSM dataset at first and investigate the

application of a selected DL model in recognition of GSMs to verify our hypothesis. In this study, we attempted to compare the accuracy and performance metrics of one DL model, GoogLeNet, with those of three traditional ML algorithms, i.e., Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbor (KNN). The aim of this study is to explore the feasibility of these four models and find the best model by this comparison. More specifically, we further compared different training strategies of the DL model, GoogLeNet, for obtaining an optimal one to fulfill the GSM recognition task with better performance. It is desired to provide reference information to help promote further research on the application of state-of-the-art DL in the geothermics domain.

The main contributions of the present study consist of the following four aspects.

1. A novel dataset for recognizing the GSMs, called JiaYing University Geothermal Surface Manifestation (JYU-GSM) dataset, was manually created by visual interpretation with expert knowledge and a high-quality check.

2. It is the first attempt to compare the applications of DL and ML models in recognition of GSMs in the geothermics and AI domains.

3. A retrained DL model, namely the GoogLeNet deep neural Network model (GSM-Net), for recognizing the GSMs is obtained by using transfer learning and finetuning of GoogLeNet.

4. It is found that there is high feasibility to use a pretrained GoogLeNet model to fulfill the task of recognizing GSMs.

The rest of this paper is organized as follows. The materials and methods used are introduced in Section 2. In Section 3, the results are presented and analyzed. In Section 4, several influencing factors of DL accuracy are discussed, followed by the limitations and future work analyses. The paper concludes with a summary in Section 5.

The overall workflow chart for the present study is shown in Figure 1, indicating our main research ideas and framework.

# 2. Materials And Methods

This section introduces the concept and its classification of GSMs briefly at first. Then the workflow of data preparation, preprocessing, and dataset creation, three ML (SVM, DT, and KNN) models and one DL (GoogLeNet) model used, assessment metrics applied, and implementation details in the present experiments are presented successively.

## 2.1. Geothermal Surface Manifestation

GSMs, also known as geothermal leakage indications, are thermal activities influenced by abnormal subsurface temperatures and exposed to the Earth's surface. Depending on the reservoir temperatures

and discharge rates, these surface manifestations take the forms of warm springs, hot springs, hot-water rivers/lakes/ponds, boiling springs, seeps, fumaroles, geysers, warm/steaming grounds, mud pots/volcanoes, hydrothermal explosion, phreatic explosion craters, zones of acid alteration, volcanic lakes, and so on [24, 30, 31]. In addition, there are some deposits of silica sinter, travertine, and/or the bedded breccias that surround phreatic craters [31]. The GSMs are crucial for the preliminary survey to determine the geothermal potential in a geothermal field because they can provide important information about the thermal propagation from the subsurface of the Earth.

According to the concept stated above, seven GSM types, namely warm spring, hot spring, geyser, fumarole, mud pot, hydrothermal alteration, and crater lake, as well as one type of none GSM, were determined as the task's classification categories on the base of expert knowledge and the visible form and identifiability of the features of GSM photographs. Many photographs were taken in Yellowstone National Park in the USA, Fuji Mountain in Japan, Tengchong, Changbaishan, Tibet volcanic-geothermal areas in China, and other famous geothermal areas. Tens of thousands of photographs were downloaded and processed manually as samples for the recognition of GSMs. Figure 2 shows the example images for each type of GSMs with different sizes. The main features of these seven types of GSMs and one type of none GSM used in the present study are characterized as follows.

1. Warm Spring (WS): It refers to the geothermal water outcrop whose temperature at the spring mouth is significantly higher than the local annual average temperature and less than or equal to 45℃., The temperature of springs cannot be directly observed from a photograph. Consequently, the warm spring type is regarded as seeps, warm-water pools/ponds/lakes/rivers usually without emitting steam from the water surface viewing in a photograph.

2. Hot Spring (HS): Theoretically, a hot spring refers to the geothermal water outcrop whose temperature at the spring mouth is higher than 45℃ and lower than the boiling point of the local surface water. Hot springs are the most visible manifestation of hot-water geothermal systems that transfer heat to the ground surface, from which the reservoir type can be predicted hypothetically [32]. Hot springs include boiling springs and hydrothermal explosions in the present study. Whether the water surface emits steam or not is regarded as the sign of distinguishing hot springs from warm springs. A hot spring usually has a phenomenon emitting steam from the water surface, while a warm spring often has no such phenomenon. In reality, a warm spring will emit steam in the cold season, confusing the ML models.

3. Geyser (GE): A geyser is generally a hole within a cone on the Earth's surface from which hot water and steam are forced out, usually at irregular intervals. The geyser is an obvious indicator of the water domination reservoir. This type usually refers to the geyser spraying water and steam like a fountain in the present study. Otherwise, it is regarded as a mud pod type or hydrothermal alteration type depending on whether there is mud in the pod.

4. Fumarole (FU): A fumarole refers to an opening in or near a volcano or ground surface through which hot gases escape. It is an evident sign of the high-temperature geothermal field. This type includes

holes, craters, or grounds spraying gases or smokes but no steam or water. Sometimes, it is difficult to distinguish between gas and steam visually from a photograph, which will make a fumarole and a geyser easy to confuse and difficult to classify manually.

5. Mud Pod (MP): A mud pod depicts mud with pop bubbles because of captured gases like carbon dioxide ($CO_2$). This type of GSMs includes mud pots, pools, and volcanoes.

6. Hydrothermal Alteration (HA): The alteration rock, zone, or deposit is the surface manifestation by contact between rocks and geothermal fluid. Hydrothermal alteration rocks, phreatic explosion craters, zones of acid alteration, and deposits of silica sinter, travertine, and bedded breccias are included in this type.

7. Crater Lake (CL): A crater lake is a lake caused by a volcano after its eruption for a long time in a volcanic area, which is different from the other lakes. Its temperature is below the local annual average, different from warm lakes/ponds. The photographs of many famous crate lakes are downloaded and processed as samples of this type.

8. None GSM (NG): This type includes negative samples for recognizing the GSMs, which can improve the performance of robustness of a classifier model. It contains a lot of kinds of photographs other than the GSMs, such as ordinary lakes, mountains, rivers, animals, plants, fountains, clouds, sky, smoke from thermal power plants. Some of them may be similar to a certain GSM which would increase the uncertainty of recognizing the GSMs but improve the robustness of a model.

### 2.2. Data Preparation, Preprocessing, and Dataset Creation

It is well known that a high-quality dataset is a prerequisite to achieving a good performance in the DL domain. Hence, we created a novel high-quality image dataset elaborately for this GSM recognition task by hand with expert knowledge and visual interpretation. Figure 3 shows the workflow chart of the data acquirement and preprocessing. The GSM-related keywords (i.e., warm spring, hot spring, geyser, fumarole, mud pot, hydrothermal alteration, crater lake, geothermal, and geothermal surface manifestation) were adopted to search the relevant types of photographs on the Internet. More than 10,000 photograph images were downloaded from the Baidu Image Search Engine (https://image.baidu.com), the Microsoft Bing Search Engine (https://en.bing.com), and a great deal of tourism service websites and tourists' blogs.

The duplicated photographs and photographs that do not belong to any type of GSMs (false images) were then manually removed one by one for many rounds. The persons in some photographs (most belong to the warm spring type) were manually masked as much as possible. To avoid the uneven effect of samples of different types, 500 images of each type of GSMs were retained, totally containing 4,000 photographs in the dataset. Afterward, the 4,000 images were converted from different formats (e.g., PNG, TIFF, JFIF, GIF, WEBP) to JPEG, the most used image file format. Then they were resized to no larger than 448 pixels in both width and height while keeping their ratio unchanged. If the size of a photograph

is larger than 448 pixels, it will be reduced to 448 pixels in either width or height; otherwise, it will be enlarged to be 448 pixels. This reduction may help fit the input size 224-224-3 of the GoogLeNet DL model. As a result, the image sizes of the eight GSM types distribute unevenly for spanning a lot in width and height, as shown in Table 1.

[Table 1 is in the supplementary files section.]

Lastly, the preprocessed images were labeled manually into eight types of GSMs based on expert knowledge with a high-quality check. Hence, a novel GSM image dataset, called JiaYing University Geothermal Surface Manifestation photographs dataset, namely the JYU-GSM dataset, was established at the end. To evaluate the accuracy and performance of the DL model obtained by training on GoogLeNet, the JYU-GSM dataset was divided into three subsets, i.e., training, testing, and validation subsets, according to the ratios of 0.8:0.1:0.1, 0.6:0.2:0.2, and 0.4:0.3:0.3, which is commonly used in the data preprocessing setup of DL. These three ratios were applied to analyze the effect of data division on the performance of the GoogLeNet model. The ratio of 0.8:0.1:0.1 was mainly applied to split the dataset for the accuracy and performance comparison of the DL and ML models used.

### 2.3. Machine Learning Models

ML is the science of getting computers to act without being explicitly programmed. In recent decades, ML has proven to be a powerful tool for deriving insights from data [33, 34]. It has been applied successfully in self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome thanks to many practical algorithms developed. Traditional ML models mainly include KNN, DT, SVM, ANN, random forest, extreme gradient boosting, and naïve Bayesian network. ML is so pervasive that it is believed to be the best way to make progress towards human-level AI.

More recently, the development of DL has yielded further performance improvements thanks to its capacity to extract a variety of features from large datasets. As a new component of ML, DL has become the most promising direction of ML for its excellent performances in many challenging tasks such as image recognition and detection, text classification, and natural language processing. It should be noted that the detailed explanations of theories and algorithms of ML and DL, which have been expounded in the literature of computer vision and pattern recognition,  are beyond the purpose and scope of this paper. We empirically selected three ML models, i.e., KNN, DT, and SVM, briefly introduced below as a comparative study to one DL model, GoogLeNet.

Feature extraction is crucial for conventional ML algorithms except for DL models as a critical feature engineering method. The Histogram of Oriented Gradient (HOG) is one commonly used feature descriptor for object recognition and detection in computer vision and image processing. It forms the feature by

calculating and counting the gradient direction histogram of the local area of the image. In the present study, the HOG feature combined with the SVM, DT, and KNN classifiers was used to recognize the GSMs, as shown in Figure 1.

2.4.1. Support Vector Machine

The SVM is a generalized linear classifier that classifies data according to supervised learning. The SVM uses the hinge loss function to calculate empirical risk and adds a regularization term to the solution system to optimize structural risk. It is a sparse and robust classifier. SVM can carry out nonlinear classification through the kernel method, one of the common kernel learning methods. The SVM has been applied in pattern recognition problems in various fields, including image classification [35-37], building detection [38], surface-wave separation [39], face recognition [40], cancer recognition [41], etc.

The Standard SVM is an algorithm based on the binary classification problem, which cannot directly deal with the multiple-classification problem. Based on the calculation process of the standard SVM, multiple decision boundaries are orderly constructed to realize multiple classifications of samples. The usual implementation is one-versus-many and one-versus-one. The one-versus-many SVM establishes $m$ decision boundaries for $m$ classifications, and each decision boundary determines the attribution of one type to all other categories. It can calculate all decision boundaries in one iteration by modifying the optimization problem of the standard SVM. The one-versus-one SVM is a voting method to establish decision boundaries for any two $m$ classifications; there are a total of $m(m\text{-}1)/2$ decision boundaries. The sample category is selected according to the category with the highest score in the discrimination results of all decision boundaries. Since a detailed analysis of the theory of SVM is beyond the scope of this paper, we refer the reader to [40, 42] for more detail on SVM. In the present study, the one-versus-one SVM was used as an ML model to recognize eight types of GSMs.

2.4.2. Decision Tree

The DT is a decision analysis method based on the known probability of occurrence of various situations. It is a graphical method of intuitively using probability analysis to calculate the probability that the expected value of net present value is greater than or equal to zero, evaluate the project risk, and judge its feasibility. Because this decision-making branch is drawn as a graph, which is similar to the branches of a tree, it is called a DT. The DT is a prediction model representing a mapping relationship between object attributes and values. Entropy, messy degree of the system, is used in ID3, C4.5, and C5.0 algorithms. This measure is based on the concept of entropy in informatics theory.

The DT is a tree structure in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category. It is a supervised learning method used widely in remote sensing classification [43-45]. The so-called supervised learning is that given a pile of samples, each sample has a set of attributes and a category. These categories are determined in advance, then a classifier can be obtained through learning, and this classifier can predict correct classification to new objects.

### 2.4.3. K-Nearest Neighbor

The KNN is a mature statistics-based method in theory and one of the simplest supervised learning algorithms. The idea of this method can be expressed that in the feature space, if most of the *k* nearest samples near a sample belong to a certain category, the sample also belongs to this category. As the most basic classifier in ML, The KNN can be used for both binary and multiple classifications. It can be used not only for classification but also for regression. By finding the *k* nearest neighbors of a sample, the average value of the attributes of these neighbors is assigned to the sample as the predicted value. The algorithm involves three main factors: the training set, the measurement of distance and similarity, and the size of *k*. The second factor, i.e., distance and similarity, is the primary consideration when using KNN.

Although the KNN has been popularly used in remote sensing classification [35] and more, there are two disadvantages of KNN: (i) too much computation, time-consuming and memory consuming; (ii) If the sample is unbalanced, for example, there are too many labels, the prediction of the labels to be tested will be greatly affected during voting, and the error rate will increase.

## 2.4. Deep Learning Model: GoogLeNet

DL allows computational models composed of multiple processing layers to learn data representations with multiple levels of abstraction [46]. As a subset of ML and the core of AI, DL methods have dramatically improved the state-of-the-art performances in speech recognition, visual object recognition, object detection, and many other domains such as drug discovery and genomics [46]. For this reason, a DL model, GoogLeNet, is empirically selected in the present study and introduced separately. We evaluated the model's performance for the GSM recognition task as a comparison with three DL algorithms, SVM, DT, and KNN, as shown in Figure 1. Beyond elaborating the principle of CNN, GoogLeNet is solely described below for brevity.

**Table 2.** The network architecture of GoogLeNet [47].

| Type | Patch size/stride | Output size | Depth | #1×1 | #3×3* reduce | #3×3 | #5×5 reduce | #5×5 | Pool proj | Params | ops |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Convolution | 7×7/2 | 112×112×64 | 1 | | | | | | | 2.7 K | 34 M |
| Max pool | 3×3/2 | 56×56×64 | 0 | | | | | | | | |
| Convolution | 3×3/1 | 56×56×192 | 2 | | 64 | 192 | | | | 112 K | 360 |
| Max pool | 3×3/2 | 28×28×192 | 0 | | | | | | | | |
| Inception(3a) | | 56×56×256 | 2 | 64 | 96 | 128 | 16 | 32 | 32 | 159 K | 128 M |
| Inception(3b) | | 56×56×480 | 2 | 128 | 128 | 192 | 32 | 96 | 64 | 380 K | 304 M |
| Max pool | 3×3/2 | 14×14×480 | 0 | | | | | | | | |
| Inception(4a) | | 14×14×512 | 2 | 192 | 96 | 208 | 16 | 48 | 64 | 364 K | 73 M |
| Inception(4b) | | 14×14×512 | 2 | 160 | 112 | 224 | 24 | 64 | 64 | 437 K | 88 M |
| Inception(4c) | | 14×14×512 | 2 | 128 | 128 | 256 | 24 | 64 | 64 | 463 K | 100 M |
| Inception(4d) | | 14×14×528 | 2 | 112 | 144 | 288 | 32 | 64 | 64 | 580 K | 119 M |
| Inception(4e) | | 14×14×832 | 2 | 256 | 160 | 320 | 32 | 128 | 128 | 840 K | 170 M |
| Max pool | 3×3/2 | 7×7×832 | 0 | | | | | | | | |
| Inception(3a) | | 7×7×832 | 2 | 256 | 160 | 320 | 32 | 128 | 128 | 1072 K | 54 M |
| Inception(3a) | | 7×7×1024 | 2 | 384 | 192 | 384 | 48 | 128 | 128 | 1388 K | 71 M |
| Avg pool | 7×7/1 | 1×1×1024 | 0 | | | | | | | | |
| Dropout (40%) | | 1×1×1024 | 0 | | | | | | | | |
| Linear | | 1×1×1000 | 1 | | | | | | | 1000 K | 1 M |
| Softmax | | 1×1×1000 | 0 | | | | | | | | |

* The "#3x3 reduce" and "#5x5 reduce" in the table indicate the number of 1x1 convolutions used before 3x3 and 5x5 convolution operations, respectively.

GoogLeNet, the winner of ILSVRC 2014, is a CNN that is 22 layers deep [47]. Table 2 shows the architecture of GoogLeNet [47]. More specifically, GoogLeNet possesses roughly 6.8 million parameters with nine inception modules, two convolutional layers, one convolutional layer for dimension reduction, two normalization layers, four max-pooling layers, one average pooling, one fully-connected layer, and a linear layer with Softmax activation in the output. Each inception module, in turn, contains two convolutional layers, four convolutional layers for dimension reduction, and one max-pooling layer (Table 2). GoogLeNet also uses dropout regularization in the fully-connected layer and applies the ReLU activation function in all the convolutional layers. To avoid the disappearance of gradient, two Softmax losses in the intermediate layers are added and connected with two auxiliary classifiers in the GoogLeNet architecture, so there are three losses in GoogLeNet. During training, the loss of the whole network is obtained by weighted addition of the three losses, while at inference time, these two middle losses are discarded.

A network pretrained on GoogLeNet was retrained on the JYU-GSM dataset. The network trained on ImageNet1000 classifies images into 1,000 object categories, such as keyboard, mouse, pencil, animal, geyser, fountain, lakeside, and lakeshore, similar to some of the GSM categories semantically but strictly not the same meanings. This ImageNet1000 dataset spans 1,000 object classes and contains 1,281,167 training images, 50,000 validation images and 100,000 test images [48]. The network has learned different feature representations for a wide range of images but no GSM. So, it cannot recognize most of the GSMs correctly. The network has an image input size of 224-by-224 in the RGB format. We retrained the GoogLeNet network pretrained on the ImageNet1000 dataset on the JYU-GSM dataset to perform the recognition task better using transfer learning. The weights of the tenth layers of the pretrained network were frozen for faster training. In the end, a new network fitting the JYU-GSM dataset based on

GoogLeNet was obtained. We name it GSM-Net for short, which could be one of our contributions in the present study for the interdisciplinary fields of Geothermics and DL.

### 2.5. Assessment Metrics

In the AI domain, many assessment metrics are used to evaluate the accuracy and performance of a classifier model. Among these, the confusion matrix, precision), recall, $F_1$ score, overall accuracy, and computational time are the popularly adopted metrics in scientific research areas and industrial applications for their great help in understanding the assessment performance of an AI model. These metrics are usually used for the classification assessment of remote sensing imagery.

#### 2.5.1. Confusion Matrix

Confusion matrix, also known as possibility table or error matrix, is a standard format for accuracy evaluation, expressed in a matrix with n rows and n columns. Each column represents the predicted value, and each row represents the actual category. The name comes from the fact that it can easily indicate whether multiple types are confused (that is, one class is predicted to be another class). It is a specific matrix used to present the visualization effect of algorithm performance, usually for supervised learning.

**Table 3.** Confusion matrix of binary classification of artificial intelligence.

| Confusion matrix | | Predicted label | |
|---|---|---|---|
| | | true | false |
| actual label | positive | TP * | FP |
| | negative | TN | FN |

* True positive (TP): the actual category of the sample is positive, and the result predicted by the model is also positive. True negative (TN): the actual category of the sample is negative, and the model predicts it to be negative. False positive (FP): the actual category of the sample is negative, but the model predicts it to be positive. False negative (FN): the actual category of the sample is positive, but the model predicts it as negative.

Table 3 shows the confusion matrix for a classic example of binary classification. The confusion matrix of multiple classifications is similar to Table 3. The present study is an example of multiple classifications.

According to Table 3, a precision (*P*) and recall (*R*) metric is then defined as Equation 1 and 2, respectively. As shown in Equation 1, the precision denotes the proportion of actual positive samples among all the results predicted as positive samples. As shown in Equation 2, the recall denotes the proportion of the samples predicted as positive examples by the classifier to the actual number of positive examples, also known as sensitivity, describing the classifier's sensitivity to the category of positive examples.

$$P = TP/(TP+FP), \quad (1)$$

$$R = TP/(TP+FN), \quad (2)$$

where, *P* and *R* denotes the precision and recall, respectively; *TP*, *FP*, and *FN* indicate the same meanings as in Table 3.

As for the present multiple-classification task, the overall precision (OP) and overall recall (OR) metrics were able to compare the accuracy performance of the four classifiers used, whose computational method is shown in Equations 3 and 4, respectively.

$$OP = (\sum_{i=1}^{n} P_i)/n, \quad (3)$$

$$OR = (\sum_{i=1}^{n} R_i)/n, \quad (4)$$

where, *OP* means the overall precision and *OR* means the overall recall; $P_i$ and $R_i$ denotes the precision and recall of the *i* type of the GSMs, respectively; *n* is the total classification number equals eight in the present study.

2.5.2. Accuracy and Overall Accuracy

As shown in Equation 5, the accuracy metric denotes the proportion of all correctly predicted samples of a classifier model in the total samples. As for the multiple-classification task, the Overall Accuracy (OA) metric was specifically used to compare the four classifiers' performance, which equals the average accuracy percentage mathematically, as shown in Equation 6.

$$A = (TP+TN)/(TP+TN+FP+FN), \quad (5)$$

$$OA = ((\sum_{i=1}^{n} A_i)/n)*100\%, \quad (6)$$

where, *A* denotes the accuracy; *TP*, *TN*, *FP*, and *FN* indicate the same meanings as in Table 3; *OA* means the overall accuracy; $A_i$ denotes the accuracy of the *i* type of the GSMs; *n* is the total classification number which equals to eight in the present study.

2.5.3. $F_1$ Score and Overall $F_1$ Score

The $F_1$ score is the harmonic mean of precision and recall taking both metrics into account in Equation 7. Overall $F_1$ score (OF) is used for the GSM recognition task to evaluate the accuracy performance of the different classifier models in general, which equals the average $F_1$ scores mathematically, as shown in Equation 8.

$$F_1 = 2*P*R/(P+R), \quad (7)$$

$$OF = (\sum_{i=1}^{n} F_{1i})/n, \quad (8)$$

where, $F_1$ denotes the $F_1$ score; $P$ and $R$ denotes the precision and recall, respectively; $OF$ means the overall $F_1$ score; $F_{1i}$ denotes the $F_1$ score of the $i$ type of the GSMs; $n$ is the total classification number which equals eight in the present study.

2.5.4. ROC and AUC

ROC is short for receiver operating characteristic. ROC curve is the main analysis tool drawn on a two-dimensional plane (e.g., Figure 9), also called the sensitivity curve. The $x$-coordinate of the plane is the false positive rate, and the $y$-coordinate is the true positive rate. It can visually represent the performance of a classifier algorithm. The steeper the curve, the better is the performance of the algorithm.

AUC is short for the area under the curve, which is a comprehensive number simply implying the performance of a classifier algorithm. The closer the AUC to 1, the better the algorithm's performance, vice versa. An AUC value of 0.5 shows that the model has a random distribution and an AUC value of 1 indicates that the model is entirely consistent with the actual situation. In general, AUC values between 0.7 and 0.9 indicate good accuracy and authenticity, whereas AUC values greater than 0.9 indicate high accuracy. The ROC, together with the AUC, OA, and OF, was adopted to discriminate and compare the performance of the retrained network models from different aspects.

2.5.5. Computational Time

The computational time (CT) is a key performance indicator for DL and ML. It may denote the training, testing, or validation time, even the sum of the two or three processes. The CT is mainly influenced by the hardware, the architecture and size of a model, the input size of an image for the model, the total number of images (size of dataset) for training, testing, and validation, and training strategy and parameter options, and so forth. When using the same number of images to train, test, and validate under the same setting environment, the shorter the computational time, the better is the classifier model. The CT was automatically recorded when carrying on the model training, testing, and validation in MATLAB. The CT can be converted to the frames per second, an indicator showing image recognition speed.

The confusion matrix was calculated and plotted using MATLAB, where the precision, recall, accuracy, and $F_1$ score were computed simultaneously. Based on the confusion matrix, the OA, OP, OR, and OF metrics was calculated as Equation 3, 4, 6, and 8, respectively, the ROC curve was plotted, and the AUC was then calculated. The OA was generally converted to a percentage as computed in Equation 6. These accuracy metrics reflect the accuracy of image classification from different aspects. All the assessment

metrics vary from 0 to 1 except the OA metric from 0% to 100%. The closer the metric to 1, the better is the model, vice versa.

The precision, recall, accuracy, and $F_1$ score metrics were used to evaluate the performance of a classifier model to predict every single type of the GSMs. Meanwhile, the overall accuracy, overall $F_1$ score, and computational time were used to compare the performance of different classifier models in general. In addition, the ROC curve and AUC were also used to discriminate the performance of the models used visually.

## 2.6. Implementation Details

### 2.6.1. Computational Environment

The experiment hardware and software environment for computation is shown below, equipped with a good GPU to accelerate computation. The same computational environment is for training the GoogLeNet, SVM, DT, and KNN models.

- Operating system: Microsoft Windows 10 education version
- CPU: Intel(R) Core (TM) i7-7700K, four cores
- RAM: Kingston 16*3 Gb
- GPU: NVIDIA GeForce RTX 2080 Ti, 11 Gb GDDR 6
- Software: MathWorks MATLAB® 2021a (9.10)

### 2.6.2. Setup of GoogLeNet

The main experiment parameters for the DL model (GoogLeNet) training are listed below, which are the finetuned results.

- Optimizer: sgdm
- MaxEpochs: 6
- Shuffle: every-epoch
- MiniBatchSize: 16
- InitialLearnRate: 5e-4
- ValidationFrequency: 100
- LearnRateSchedule: piecewise
- LearnRateDropFactor: 0.2
- LearnRateDropPeriod: 2
- ValidationPatience: 3
- L2Regularization: 0.0005
- Momentum: 0.95

- ExecutionEnvironment: GPU

The training dataset for accuracy and strategy comparisons consists of 3,200 images in total, 400 images for each type of GSMs, respectively. The maximum epoch was set to six, shuffled every epoch. It was an early stop epoch when training for 20 epochs. The number of input images for a mini-batch was set to 16, that is, 200 iterations. The initial learning rate was set to 0.0005 while a dynamic mechanism of learning rate was adopted to update learnable parameters (i.e., weights and biases) using stochastic gradient descent with momentum (SGDM) with an L2 regularization and dropout policy to avoid overfitting. This training policy is commonly used to improve a model's performance in the AI domain. A GPU was used to help train a network faster. An example of the training progress for the GoogLeNet model is shown in Figure 4, indicating a little bit of overfitting for the training because the curve of validation accuracy is almost under the training curve. The accuracy of validation is always less than that of the test.

2.6.3. Setup of SVM, DT, and KNN

We used the default setup values for the three traditional ML models (SVM, DT, and KNN) training, which were not optimized in the present study (Table 4). The three models all used Error-Correcting Output Codes (ECOC) method to train based on onevsone coding name. The SVM model used the hinge loss, and the other two models both used the quadratic loss to realize the recognition task of GSMs.

**Table 4.** Parameter setup for training the SVM, DT, and KNN models.

| Model | SVM * | DT | KNN |
|---|---|---|---|
| BinaryLearners | 1×1 Fittemplate | 1×1 Fittemplate | 1×1 Fittemplate |
| CodingName | onevsone | onevsone | onevsone |
| FitPosterior | 0 | 0 | 0 |
| Method | ECOC | ECOC | ECOC |
| Type | classification | classification | classification |
| BinaryLoss | hinge | quadratic | quadratic |
| Surrogate | - | on | - |
| MaxNumSplits | - | 1 | - |
| NumNeighbors | - | - | 5 |
| Standardize | - | - | 1 |

* SVM denotes Support Vector Machine; DT denotes Decision Tree; and KNN denotes K-Nearest Neighbor. ECOC denotes Error-Correcting Output Codes. "–" denotes no such parameter. The division ratio is 0.8:0.1:0.1 for the training (400 images), testing (50 images) and validation (50 images) subsets.

# 3. Results

In this section, we compared the metrics of accuracy and performance of the trained DL and ML models to evaluate their feasibility for the task of recognition and classification of GSMs. First, visual analyses of the test image presentations were made to check the performance intuitively. Second, the accuracy metrics were computed and compared for the DL and ML models and the two DL strategies. Third, the computational time for training, testing, and validation was recorded and analyzed to compare the

performance of models. Last, the ROC curves were made to check the performance further and compare the performance comprehensively.

### 3.1. Visual Comparison and Analysis

A visual comparison was conducted to analyze the four classifier models' performances by presenting testing results from two randomly selected images of each type. Sixteen images, two for each type of eight categories, were randomly selected from the test subset to conduct a visual analysis for testing the four classifier models. Figure 5 shows one of our random multiple test results with similar patterns. It can be clearly seen that 15 images were correctly predicted (except one wrong image, No. 2 predicted from WS to HS), all with very high probabilities using the GoogLeNet model and 16 images. On the contrary, nine images were predicted positively using the SVM model, and too many predicted images were mistaken using the other two models. The test OA for these four models is calculated as 93.75% (15/16), 56.25% (9/16), 25.00% (4/16), and 12.50% (2/16), respectively, indicating a possible advantage of the GoogLeNet model over the three ML models, SVM, DT, and KNN.

In these cases, the No. 2 image was predicted negatively by all four models. It is a photograph showing a warm spring with some people and plenty of heating steam. The characteristic with heating steam or vapor, which belongs mainly to the hot spring type, is probably the reason for this false prediction. Moreover, it is interesting that the GoogLeNet model can positively predict the No. 14 image, which is masked nearly one-third by a stone tablet engraved with the word Tianchi (Heaven Lake). The GoogLeNet can also accurately recognize the geysers with spraying water or vapor and classify the fumaroles with emitting gases or fumes, similar to the vapor to some extent. These indicate the GoogLeNet model has probably learned some significantly different characteristics of various types of GSMs and has a good performance of robustness and generalization.

### 3.2. Comparison of DL and ML Accuracies

The assessment results of accuracies of the trained DL and ML models are shown in Table 5. It can be evidently found that the retrained model based on the GoogLeNet transfer learning (for short, the GoogLeNet model, similarly hereinafter) has a significant advantage of both overall accuracy and overall $F_1$ score over the three ML models, SVM, DT, and KNN whatever for validation or test.

**Table 5.** Assessment results of accuracies of the GoogLeNet and three other ML models.

| Model | Subset | Overall accuracy (%) | Overall $F_1$ score |
|---|---|---|---|
| GoogLeNet | validation | 91.25 | 0.91 |
| | test | 88.25 | 0.88 |
| SVM * | validation | 53.50 | 0.53 |
| | test | 49.00 | 0.49 |
| DT | validation | 26.00 | 0.26 |
| | test | 26.00 | 0.26 |
| KNN | validation | 20.25 | 0.17 |
| | test | 20.25 | 0.17 |

* SVM denotes Support Vector Machine; DT denotes Decision Tree; and KNN denotes K-Nearest Neighbor. The division ratio is 0.8:0.1:0.1 for the training (400 images), testing (50 images) and validation (50 images) subsets.

The GoogLeNet model has the highest overall accuracy of 91.25% and the highest overall $F_1$ score of 0.91, both on the validation subset among the four classifiers used, followed by the SVM model with the OA of 53.50% and the OF 0.53 also both on the validation subset. In contrast, the KNN model has the smallest OA and OF, only 20.25% and 0.17, respectively, ether on the validation or test subset. For the test subset, the GoogLeNet model also occupies the first position with the highest OA and OF, 88.25% and 0.88, respectively, and the second is the SVM model with the OA of 49% and OF 0.49, respectively.

According to Figure 6, the GoogLeNet model completely surpasses the other three ML models by at least 70% over the SVM model. The maximum reaches nearly 450% over the KNN model on the validation subset. This result indicates that the GoogLeNet model has a significant advantage over the three ML models, whether from overall accuracy or overall $F_1$ score, even on the test or validation subset. It suggests that the GoogLeNet model is competent in recognizing the GSMs. It is worth noting that the three ML models used could not be used to recognize the GSMs due to their low accuracy for any test or validation. Although the SVM model has a significant advantage over the other two, its OA and OF are too low to meet the essential requirement for classification.

### 3.3. Comparison of Different DL strategies

Generally, there are two DL strategies from scratch and transfer learning with a pretrained network. Transfer learning can help train a network faster for a new classification task, especially on a small dataset, which is popularly used in many AI applications such as historical building detection [49], GDP (gross domestic product) prediction [50], scene classification [51]. Of course, DL from scratch is also widely used in AI with plenty of data. These two DL strategies were carried out in the present study to discriminate which is the best one for GSM recognition. The frozen technique of deep CNN layers was also used in the study to explore its effect on DL results.

It can be evidently found from Table 6 and Figure 7, whatever for validation or test assessment, the accuracy metrics, including OA and OF using transfer learning with the pretrained network (S3 and S4), are both considerably larger than those using the DL from scratch (S1 and S2). The results show that transfer learning with the pretrained DL model has a significant advantage of accuracy over DL from scratch, whether the initial ten layers were frozen.

**Table 6.** Assessment results of different training strategies for deep learning.

| Strategy | Subset | OA * (%) | Overall $F_1$ score |
|---|---|---|---|
| S1 - from scratch, initial ten layers frozen | validation | 55.0 | 0.54 |
| | test | 55.3 | 0.54 |
| S2 - from scratch, no layer frozen | validation | 63.0 | 0.62 |
| | test | 57.5 | 0.67 |
| S3 - pretrained, initial ten layers frozen | validation | 90.8 | 0.90 |
| | test | 89.5 | 0.89 |
| S4 - pretrained, no layer frozen | validation | 93.5 | 0.94 |
| | test | 88.8 | 0.89 |

* OA denotes overall accuracy. The division ratio is 0.8:0.1:0.1 for the training (400 images), testing (50 images) and validation (50 images) subsets. The same parameters for training were set.

The highest overall accuracy is 63.0% for the validation when performing the DL from scratch, while the highest one reaches 93.5% for the validation when performing the transfer learning with the pretrained network (Table 6). The lowest overall accuracy reaches 88.8% for the test when performing the transfer learning with the pretrained network, which is larger than the highest one by increasing the percentage 25.8% when performing the DL from scratch, indicating that transfer learning with a pretrained network could be a preferred choice for the recognition of the GSMs. The highest overall $F_1$ score is 0.67 for the test when performing DL from scratch, while the highest one reaches 0.94 for the validation when performing the transfer learning with the pretrained network. The lowest overall $F_1$ score goes to 0.89 for the test when performing the transfer learning with the pretrained network, which is larger than the highest one by increasing 0.22 when performing DL from scratch, indicating the same advantage from the view of the overall accuracy.

The overall accuracies of the transfer learning validation and test are all largely higher than those of DL by the increased percentage of at least 44%. Meanwhile, all overall $F_1$ scores of the transfer learning validation and test are also tremendously larger than those of DL by the rising percentage of at least 32% (Figure 8). For the overall accuracy, the highest increase percentage reaches 70.00% for the validation between the transfer learning with the pretrained network, no frozen layer, and the DL from scratch with the initial ten layers frozen (S4-S1), while the highest one gets to 61.87% for the test between the transfer learning with the pretrained network and the initial ten layers frozen and the DL from scratch, with the initial ten layers frozen (S3-S1). As for the overall $F_1$ scores, the highest increase percentage goes up to 74.07% for the validation between the transfer learning with the pretrained network and no frozen layer and the DL from scratch with the initial ten layers frozen (S4-S1), while the highest one gets to 64.81% for the test between the transfer learning with the pretrained network and the initial ten layers frozen and the DL from scratch with the initial ten layers frozen (S3-S1). In short, transfer learning with the pretrained GoogLeNet model has a significant advantage of accuracy over DL from scratch, whether the initial ten layers were frozen or not.

### 3.4 ROC and AUC Comparison

Apparently, it can be observed that the mean ROC curve of the GoogLeNet model for the eight types of the GSMs is the steepest among the four models, followed by the SVM model (Figure 9a, 9b). It suggests that the GoogLeNet model is the best among the models used. The mean AUC of 0.9954 of the GoogLeNet model is also the biggest among the four models, indicating its first place. The performance of the GoogLeNet model to classify the eight different types of GSMs was further assessed with the ROC curve on the test subset. The results show that it is difficult to distinguish which type of GSMs has the best performance because the ROC curves of all types are too steep to separate, and the average AUCs of all types almost approach 1 (Figure 9b). It suggests that the GoogLeNet model has high accuracy for recognizing each type of GSMs.

### 3.5. Comparison of Computational Time

#### 3.5.1. Time of Different Models

**Table 7.** Assessment results of computational time of the GoogLeNet and three other ML models.

| Model | $T_{train}$ * (s) | $T_{val}$ (s) | $FPS_{val}$ | $T_{test}$ (s) | $FPS_{test}$ |
|---|---|---|---|---|---|
| GoogLeNet | 299.38 | 1.98 | 25.25 | 1.71 | 29.24 |
| SVM | 121.02 | 9.22 | 5.42 | 7.71 | 6.49 |
| DT | 78.68 | 6.88 | 7.27 | 6.75 | 7.41 |
| KNN | 515.97 | 65.07 | 0.77 | 64.58 | 0.77 |

* $T_{train}$ denotes the total computational time for training with the same parameters; $T_{val}$ is the computational time for validation; and $T_{test}$ is the computational time for testing. $FPS_{val}$ denotes Frames Per Second for validation. $FPS_{test}$ denotes Frames Per Second for the test. SVM denotes Support Vector Machine; DT denotes Decision Tree; and KNN denotes K-Nearest Neighbor. The division ratio is 0.8:0.1:0.1 for the training (400 images), testing (50 images) and validation (50 images) subsets.

It took the longest time for the KNN (515.97 s) model to train on the training images, followed by the GoogLeNet model (299.38 s), and the least did the DT model (78.68 s). It took the longest time for the KNN model to test and validate, followed by the SVM model, and the least did the GoogLeNet model (Table 7). It suggests that the KNN model has the worst performance for it is an unsupervised classifier. Although it took more time for the GoogLeNet model to train than the SVM and DT models, the time for test and validation was both the least, respectively, implying that the GoogLeNet model can predict an image the fastest once it is fully trained. The frames per second of the GoogLeNet model reach 29.24, high enough to perform real-time recognition. Combined with its highest accuracy stated above, the GoogLeNet model has high feasibility to accomplish the task of the recognition of the GSMs.

#### 3.5.2. Time of Different Strategies

Since the GoogLeNet model is the best one among all four models used, to explore the training strategy, four strategies, i.e., S1 - from scratch, initial ten layers frozen, S2 - from scratch, no frozen layer, S3 - pretrained, initial ten layers frozen and S4 - pretrained, no frozen layer, were designed to train the GoogLeNet model on the same data subset with the same parameters. The results show the strategy S4

took the least time (273.20 s) to train, followed by the strategy S3 (279.20 s), the longest (307.73 s) did the strategy S2 (Table 8). As expected, the training time for strategy S1 is shorter than strategy S2, and the training time for strategies S3 and S4 is shorter than that for strategy S1 and S2, respectively. However, it is surprising that the training time for the strategy S3 is longer than that for the strategy S4 because the weights and biases of its initial ten layers are frozen and don't need training. The computational time for the test is in line with the expectation that for validation is not so, the time for the strategy with layers frozen is shorter than that of no frozen layer when using the same strategy from scratch or transfer learning with the same pretrained model. This is probably because the sizes of the input images for training, testing, and validation vary greatly.

**Table 8.** Assessment results of computational time of the GoogLeNet model using different training strategies.

| Strategy | $T_{train}$ * (s) | $T_{test}$ (s) | $T_{val}$ (s) |
|---|---|---|---|
| S1 - from scratch, initial ten layers frozen | 295.06 | 1.50 | 1.89 |
| S2 - from scratch, no layer frozen | 307.73 | 1.52 | 1.97 |
| S3 - pretrained, initial ten layers frozen | 279.20 | 1.64 | 1.98 |
| S4 - pretrained, no layer frozen | 273.20 | 1.68 | 1.84 |

* $T_{train}$ denotes the total computational time for training with the same parameters; $T_{test}$ is the computational time for testing; and $T_{val}$ is the computational time for validation. The division ratio is 0.8:0.1:0.1 for the training (400 images), testing (50 images) and validation (50 images) subsets.

### 3.6 Comparison of Accuracies of Different Types of GSMs

As demonstrated above, the GoogLeNet model obtained the best accuracy and performance among the four selected models. Hence, we analyzed the confusion matrix and assessment metrics of the eight types of GSMs from the GoogLeNet model on the testing subset. The results are shown in Table 9 and Table 10, respectively.

It can be found that the precision (95.92%) of the geyser type outperforms those of the other seven types and the precision (82.67%) of the NG type is the least among all types (Table 9). The geyser type's recall (94.00%) is the best among all types, and the least falls to the fumarole type. The warm spring type was predicted with 88.00% for both the precision and recall, resulting in an $F_1$ score of 0.88 (Table 10).

For the 50 WS samples, they were predicted to be one HS, one MP, two HA, and two CL types, confused by four types of GSMs, while there were four HS and two NG predicted to be WS. The HS samples were predicted to be four WS, one MP, and two NG, confused by three types of GSMs, while four types (WS, FU, MP, and HA) of samples were predicted as HS, resulting in an $F_1$ score of 0.84 which is the least among those of all types. The GE samples were confused solely by two FU samples, while the FU samples were confused by three GE samples. The GE type has the best precision and recall and the best $F_1$ score (0.95). However, the FU type has a relatively high precision but the lowest recall, leading to an $F_1$ score of 0.87.

The 46 MP samples were predicted positively, and two to HS, one to HA and one to NG negatively, bringing about a high recall of 92.00%, while one WS, one HS, two FU, and two NG samples were predicted falsely to be the MP type, causing a precision of 88.46%. The $F_1$ score for the MP type is thus calculated as 0.90. Accordingly, the $F_1$ score for the HA, CL, and NG types is calculated as 0.91, 0.92, and 0.84, respectively. The overall precision, recall, accuracy, and $F_1$ score reach 89.14%, 89.00%, 89.00%, and 0.89, respectively, for the GoogLeNet model, 50 image samples each type on the testing subset.

These confusions of different types are probably caused by the coexisting of various GSMs (e.g., Figure2a) in a photograph and some human mistakes resulting from the difficulty of visual interpretation and classification. The trained model was confused by some ambiguous photographs with coexisting of two or more types of GSMs. Therefore, the discriminative features extracted by the model were also confused to some extent. It suggests that it should be very important to create a high-quality dataset for the feature extraction of an AI task.

**Table 9.** Confusion matrix of the eight types of geothermal surface manifestations for the GoogLeNet model on the testing subset.

| Type | | Predicted label | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1.WS | 2.HS | 3.GE | 4.FU | 5.MP | 6.HA | 7.CL | 8.NG | Recall (%) |
| Actual label | WS * | 44 | 1 | | | 1 | 2 | 2 | | 88.00 |
| | HS | 4 | 43 | | | 1 | | | 2 | 86.00 |
| | GE | | | 47 | 3 | | | | | 94.00 |
| | FU | | 2 | 2 | 41 | 2 | 1 | | 2 | 82.00 |
| | MP | | 2 | | | 46 | 1 | | 1 | 92.00 |
| | HA | | 4 | | | | 46 | | | 92.00 |
| | CL | | | | | | | 46 | 4 | 92.00 |
| | NG | 2 | | | | 2 | 1 | 2 | 43 | 86.00 |
| Precision (%) | | 88.00 | 82.69 | 95.92 | 93.20 | 88.46 | 90.20 | 92.00 | 82.67 | |

* WS is short for Warm Spring; HS is short for Hot Spring; GE is short for GEyser; FU is short for FUmarole; MP is short for Mud Pot; HA is short for Hydrothermal Alteration; CL is short for Crater Lake; and NG is short for None Geothermal surface manifestation.

**Table 10.** Assessment matrics of the eight types of geothermal surface manifestations for the GoogLeNet model on the testing subset.

| No. | Type | Precision (%) | Recall (%) | Accuracy (%) | $F_1$ score |
|---|---|---|---|---|---|
| 1 | WS * | 88.00 | 88.00 | | 0.8800 |
| 2 | HS | 82.69 | 86.00 | | 0.8431 |
| 3 | GE | 95.92 | 94.00 | | 0.9495 |
| 4 | FU | 93.20 | 82.00 | | 0.8723 |
| 5 | MP | 88.46 | 92.00 | | 0.9020 |
| 6 | HA | 90.20 | 92.00 | | 0.9109 |
| 7 | CL | 92.00 | 92.00 | | 0.9200 |
| 8 | NG | 82.67 | 86.00 | | 0.8431 |
| 9 | Overall | 89.14 | 89.00 | 89.00 | 0.8901 |

* WS is short for Warm Spring; HS is short for Hot Spring; GE is short for GEyser; FU is short for FUmarole; MP is short for Mud Pot; HA is short for Hydrothermal Alteration; CL is short for Crater Lake; and NG is short for None Geothermal surface manifestation.

# 4. Discussion

The present study confirmed that the deep transfer learning model outperforms the three traditional ML models. So the effects of different data division methods, data augmentation, and hyperparameter optimization on the DL accuracy metrics are emphatically discussed below without mentioning the three ML models. The limitations and future work are explained and analyzed as well.

### 4.1. Effect of Data Division

Generally, the larger the number of examples, the better the model's performance. As shown in Table 11, it is clearly seen that scenario A (a ratio of 0.8:0.1:0.1 for the divisions of training, testing, and validation) outperforms the other two, B (a ratio of 0.6:0.2:0.2) and C (a ratio of 0.4:0.3:0.3) for both overall accuracy and overall $F_1$ score, both validation and test, indicating the more the training images, the better the accuracy of the model. Data augmentation and more images with high quality should be used to improve the accuracy of the GSM-Net further.

**Table 11.** Accuracy comparison of the training results of three scenarios of data division by deep learning with the pretrained GoogLeNet model.

| Scenario | Subset | Overall accuracy (%) | Overall $F_1$ score |
|---|---|---|---|
| A-ratio 0.8:0.1:0.1 (400: 50: 50) | validation | 90.50 | 0.9047 |
| | test | 89.75 | 0.8975 |
| B-ratio 0.6:0.2:0.2 (300:100:100) | validation | 88.50 | 0.8849 |
| | test | 88.75 | 0.8871 |
| C-ratio 0.4:0.3:0.3 (200:150:150) | validation | 86.25 | 0.8624 |
| | test | 88.25 | 0.8827 |

### 4.2. Effect of Data Augmentation

Data augmentation is commonly used to improve the accuracy of a model in the DL area. There are many image processing methods such as filp, rotation, and translation for data augmentation. We investigated the effect of data augmentation of image flip to verify the finding stated above in the 4.1 section by training three times to avoid errors of randomness. The division ratio of the JYU-GSM dataset is set to 0.8:0.1:0.1 for the training (400 images), testing (50 images), and validation (50 images) subsets for this investigation. All models were tested and validated on the same test and validation subsets, respectively.

As shown in Table 12 and Figure 10, it can be evidently found that the vertical flip (2X) has a positive effect to increase the accuracy of both test and validation while the horizontal flip (2X) and the horizontal flip after the vertical flip (4X) methods have probably negative effects on the cross-validation accuracies. The overall accuracy rises up from (88.75 ± 0.90)% and (88.83 ± 1.66)% of no augmentation to (89.33% ± 0.63)% and (89.67 ± 0.52)% of the augmentation 2X (the vertical flip) method for validation and test (Table 12), increasing 0.94% and 0.89%, respectively; A similar trend appears in the overall $F_1$ score (Figure 10b). The standard deviations of both overall accuracy and overall $F_1$ score metrics become

smaller, indicating that the vertical flip (2X) augmentation could increase the model's prediction robustness and improve its performance. By contrast, a negative effect occurs on the accuracy metrics of the test and validation results when using horizontal and vertical flips (4X), which is beyond our expectations. The overall accuracy of the augmentation 4X method for validation and test goes down to (86.50 ± 1.39)% and (88.17 ± 0.88)%, decreasing 1.49% and 1.43%, respectively. And a similar trend also appears in the overall $F_1$ score for both validation and test (Figure 10b). With the increase of the training images with data augmentation, the overall accuracy and overall $F_1$ score will not always go up. This is probably because some types of images of the training dataset, such as hot springs, geysers, and fumaroles, should be input into the model vertically for training. Otherwise, they could influence the accuracy performance of the model negatively.

To verify this speculation, the additional data subsets of the vertical flip (1X), horizontal flip (1X), and horizontal flip (2X, including the original training subset) were trained to obtain three DL models, respectively. And these three models were validated and tested on the same validation and test data subsets as the original ones. As shown in Table 12, the accuracy metrics for validation and test vary a little when using the vertical flip (1X) augmentation while those for validation both drop a lot when using the other two augmentation methods with the horizontal flip, those for test drop a lot when using the horizontal flip (1X), and those for test rise a little when using the horizontal flip (2X). This suggests that the horizontal flip augmentation has a negative effect on the accuracy of the DL model used.

**Table 12.** Assessment metric comparison of the cross-validation results with image transformation and data augmentation.

| No. | Augmentation | Subset | Overall accuracy (%) | Overall $F_1$ score |
|---|---|---|---|---|
| 0 | OR (1X) * | validation | 88.75 ± 0.90 | 0.8881 ± 0.0089 |
| | | test | 88.83 ± 1.66 | 0.8881 ± 0.0167 |
| 1 | VF (1X) | validation | 88.83 ± 0.95 | 0.8886 ± 0.0099 |
| | | test | 87.67 ± 0.80 | 0.8761 ± 0.0081 |
| 2 | HF (1X) | validation | 76.50 ± 3.12 | 0.7627 ± 0.0318 |
| | | test | 79.75 ± 0.66 | 0.7961 ± 0.0061 |
| 3 | VF (2X) | validation | 89.33 ± 0.63 | 0.8937 ± 0.0059 |
| | | test | 89.67 ± 0.52 | 0.8960 ± 0.0054 |
| 4 | HF (2X) | validation | 86.58 ± 0.38 | 0.8659 ± 0.0040 |
| | | test | 89.50 ± 1.50 | 0.8944 ± 0.0158 |
| 5 | HF_VF (4X) | validation | 86.50 ± 1.39 | 0.8650 ± 0.0134 |
| | | test | 88.17 ± 0.88 | 0.88815 ± 0.0084 |

* The division ratio of the JYU-GSM dataset is 0.8:0.1:0.1 for the training (400 images), testing (50 images), and validation (50 images) subsets. OR (1x) denotes the original training subset; VF (1X) denotes the vertical flip (1X) of the OR (1X); HF (1X) denotes the horizontal flip (1X) of the OR (1X); VF (2X) denotes the OR (1X) and its vertical flip (2X); HF (2X) denotes the OR (1X) and its horizontal flip (2X); HF_VF (4X) denotes the OR (1X), its vertical flip, and their horizontal flips. All models were tested and validated on the same test and validation subsets, respectively. The overall accuracy and overall $F_1$ score is the average of training three times, respectively, including its standard derivation followed the average value.

The $F_1$ scores of the eight types of GSMs were further analyzed to reveal the effects of different data augmentation methods for a single type of GSMs. It can be found from Figure 11 that the vertical flip (2X) augmentation has a positive effect on the inference accuracy of the DL model used for validation for almost all types of GSMs except mud pods, and it also has a positive one for test for the eight types of GSMs except warm springs and geysers. However, the horizontal flip (2X) and horizontal flip after vertical flip (4X) methods decrease the inference accuracies for validation for all types of GSMs (Figure 11a), but the effects of these two augmentation methods are uncertain for the inference accuracies for test for all types of GSMs. For the test (Figure 11b), the prediction accuracies of WS and NG descend while those of HS, GE, FU, HA, and CL rise when using the horizontal flip (2X), and those of WS, GE, FU, CL, and NG go down while those of HS, MP, and HA rise up when using the horizontal flip after vertical flip (4X) method. Therefore, it is necessary to optimize the JYU-SGM dataset further to reduce the uncertainty caused by imbalances of data distribution.

### 4.3. Effect of Hyperparameter Optimization

First, we tested the effect of the number of epochs. We set 20 epochs to train the GoogLeNet model with the strategy S2 while the other hyperparameters keep the same. Surprisingly, the model stopped early at the $13^{th}$ epoch and got an overall accuracy of 60.50% for validation and 59.75% for test. The overall $F_1$ scores were less than 60% for both validation and test. It suggests that DL from scratch and with no weights of GoogLeNet cannot satisfy the accuracy requirements for the task of recognizing the GSMs. Therefore, we used transfer learning to retrain the weighted GoogLeNet model for better results in the study. For this, six epochs were found to be enough to get a good model after retraining the GoogLeNet model with more epochs many times.

Second, we tested the effect of the initial learning rate with different values. The results show that the initial learning rate set to be 5e-4 is an optimal value for good enough accuracy and performance. When the initial learning rate was set to be larger or smaller, the accuracy metrics would go down and be less than 5e-4.

Third, the mini-batch size was tested with 8, 16, and 32, respectively. It is found that the mini-batch size of 16 is the best suitable value for the GoogLeNet model to train to get better accuracy and performance.

The other parameters were also finetuned in our study. Their effects were less than those of the initial learning rate and epoch. So they were not discussed in detail here.

### 4.4. Limitations and Future Work

In the AI domain, big data, algorithms, and computing powers are considered as the three core driving forces of AI. Because of the great advances of these three forces, plenty of significant progress has been made to resurge the breakthrough of AI in the recent decade. In the present study, we first created a new image dataset, equipped a suitable experimental environment, and then selected empirically four models to perform comparative analysis to push forward the advance of geothermal AI, so-called Geothermal 4.0

[6]. The present study is a challenge for DL in the geothermics domain because there is a lack of a suitable dataset with high quality to train and develop a DL model for the recognition of the GSMs. Despite hard work, there are inevitable limitations of the dataset creation, DL model selection and design, and hyperparameter optimization. These issues are eternal topics in the field of AI.

First of all, the data preparation and preprocessing method would require a tremendous amount of manual work, with a corresponding increase in the subjectivity of the assessment and a corresponding increase in the likelihood of inaccuracies. It took us plenty of effort and time to construct the JYU-GSM, which contains 500 images in each class of the eight categories. The key challenge for this dataset task is manually downloading and labeling for visual classification of sample images of the eight GSM classes with great effort. The biggest difficulty lies in the symbiosis of different types of GSMs stated above in Section 3.1. This makes it very difficult to accurately discriminate the unique features of each type and classify the GSMs manually with visual interpretation. Although we have double-checked and confirmed the classification of the JYU-GSM, there might be some misclassifications that could result in the decline of accuracy. Hence, it deserves further study on improving the JYU-GSM and specific design of DL architecture that fits with the problem of GSM for better accuracy.

It can be clearly found from Table 6, Figure 8, and Figure 10 that all assessment results for the validation and test subsets have minor gaps between them, implying that the distribution characteristics of the JYU-GSM should be a little unbalanced. This phenomenon would influence the overall accuracy when shuffling the input training images in each epoch. As a matter of fact, the validation and test results will fluctuate slightly up and down at 90% for the OA or 0.9 for the OF. It is suggested that the JYU-GSM should be updated with higher quality so that it could be trained to get a more optimal DL model with better performance, which deserves further study in detail.

It is worthy to point out that the frozen-layer strategy may have few effects on the model's accuracy, especially for transfer learning whose accuracies vary no more than 5% (Figure 8, S4-S3), whether using the strategy or not. It is observed that the retrained models using the strategy of DL from scratch have not converged or gotten a good enough result. In our next work, more epochs and more images need to be adopted to train and test for better performance.

Further optimization selection of the ML algorithms could also be a way to fulfill the task of the recognition of the GSMs. In addition, more DL pretrained models such as VGG16, ResNet-50, and SqueezeNet, different CNN architectures used widely in image recognition, deserve further research on exploring their feasibilities for recognition of the GSMs in the future. Furthermore, how to use these CNN models for other tasks such as geothermal anomaly and target area detection in the geothermics domain will be our next major work in the coming years.

A recent study indicates that $^{3}He/^{4}He$ analysis of thermal springs locates the mantle suture in a continental collision [52]. State-of-the-art methods like GIS and remote sensing, ML and DL (e.g., ANN, CNN, and transformer), and other photographic technology could improve geothermal data analysis and

analytical model building methods. They could make geothermal exploration and evaluation more efficient and deserve comprehensive integration investigations in the future to quickly and accurately find more GSMs such as thermal springs to help promote broader applications in similar areas of earth sciences.

# 5. Conclusions

In the study, we investigated the application of DL in the geothermics domain compared with traditional ML, specifically aiming to explore the feasibility of DL in recognition of GSMs. We created a new image dataset of the JiaYing University Geothermal Surface Manifestation photographs, namely the JYU-GSM dataset, and compared the accuracy and performance of one DL model, GoogLeNet, with three traditional ML models, i.e., SVM, DT, and KNN. The results show that the GoogLeNet model outperforms the SVM, DT, and KNN models significantly. The model retrained by using the pretrained GoogLeNet can be suitable for the recognition of the GSMs for its high accuracies, while the three traditional ML models are not suitable enough to fulfill this task due to their relatively low accuracies. In conclusion, it is very feasible to use deep transfer learning to recognize GSMs in the geothermics domain once a high-quality GSM dataset is available.

# Declarations

**Author Contributions:** Conceptualization, Y.X., and M.Z.; methodology, Y.X., and M.Z.; software, Y.X. and K.H.; validation, M.Z., Y.L., and J.L.; formal analysis, Y.X., and K.H.; investigation, Y.X., M.Z., Y.C., and J.L.; resources, Y.C. and J.L.; data curation, Y.X., Y.L., J.L. and M.Z.; writing—original draft preparation, Y.X., and M.Z.; writing—review and editing, Y.X., K.H., and M.Z.; visualization, Y.L.; supervision, Y.X.; project administration, Y.X. and M.Z.; funding acquisition, Y.X., M.Z., and K.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Guangdong Natural Science Foundation, grant number 2017A030307040, the Guangdong Province Special Project in Key Fields for Universities (New Generation Information Technology), grant number 2020ZDZX3044, the Ordinary University Characteristic Innovation Project of Guangdong Province, grant number 2020KTSCX140, the Research Ability Improvement Project of Key Construction Disciplines in Guangdong Province, grant number 2021ZDJS073, and the Intangible Cultural Heritage Research Base Project of Guangdong province, grant number 17KYKT13. This study was partly supported by the NSFC, grant number 61976104.

**Data Availability Statement:** The JYU-GSM dataset supporting reported results can be found and downloaded for free at https://pan.baidu.com/s/1vYbtSALH-bUaSs6P-TTUHw (code: w6au) or https://doi.org/10.5281/zenodo.6220526. The JYU-GSM dataset and the MATLAB codes for the four machine learning algorithms (SVM, DT, KNN, and GoogLeNet) used during the current study are also available from the corresponding author (Y.X.) upon reasonable request.

**Acknowledgments:** Y.X. would like to acknowledge the support from the China Scholarship Council (Grant number 201808440171). Y.X. would like to thank the University of Hawaii at Manoa for sharing the usage of MATLAB licenses. The authors would like to thank Baidu Inc. (https://image.baidu.com), Microsoft Inc. (https://en.bing.com), and a great deal of tourism service websites and tourists' personal blogs for providing the photographs to download for free. The authors would like to thank the editors and reviewers for their valuable comments and suggestions that greatly improve our manuscript and also to thank Research Square for posting our manuscript preprint online at https://doi.org/10.21203/rs.3.rs-1377072/v2.

**Conflicts of Interest:** The authors declare no conflict of interest.

# Abbreviations

The following abbreviations are used in the text of this manuscript.

| | |
|---|---|
| AI | Artificial Intelligence |
| ANN | Artificial Neural Network |
| AUC | Area Under the Curve |
| CL | Crater Lake |
| CNN | Convolutional Neural Network |
| CT | Computational Time |
| DL | Deep Learning |
| DT | Decision Tree |
| FPS | Frames Per Second (fps) |
| FU | Fumarole |
| GE | Geyser |
| GIS | Geographic Information System |
| GPU | Graphic Processing Unit |
| GSM | Geothermal Surface Manifestation |
| GSM-Net | Geothermal Surface Manifestation deep neural Network model |
| HA | Hydrothermal Alteration |
| HOG | Histogram of Oriented Gradient |
| HS | Hot Spring |
| JYU-GSM | JiaYing University Geothermal Surface Manifestation |
| KNN | K-Nearest Neighbor |
| ML | machine learning |
| MP | Mud Pod |
| NG | None GSM (Geothermal Surface Manifestation) |
| OA | Overall Accuracy |
| OF | Overall $F_1$ score |
| ROC | Receiver Operating Characteristic |
| SVM | Support Vector Machine |
| WS | Warm Spring |

# References


1. Gupta, H.; Roy, S., *Geothermal Energy: An Alternative Resource for the 21st Century*. Elsevier Science: 2007; p 279.

2. Huang, S., Geothermal energy in China. *Nat Clim Change* **2012**, 2, (8), 557-560.

3. Pang, Z.; Huang, S.; Hu, S.; Zhao, P.; He, L., Geothermal studies in China Progress and prospects 1995-2014. *Chin J Geol* **2014**, 49, (3), 719-727.

4. Wang, J., *Geothermics ant Its Applications*. Science Press: Beijing, 2015; p 548.

5. Duoji; Wang, G.; Zheng, K., *Study on the development and utilization strategy of geothermal resources in China*. Science Press: Beijing, 2017; p 148.

6. Muther, T.; Syed, F.I.; Lancaster, A.T.; Salsabila, F.D.; Dahaghi, A.K.; Negahban, S., Geothermal 4.0: AI-enabled geothermal reservoir development- current status, potentials, limitations, and ways forward. *Geothermics* **2022**, 100, 102348.

7. Sedara, S.O.; Alabi, O.O., Heat flow estimation and quantification of geothermal reservoir of a basement terrain using geophysical and numerical techniques. *Environ Earth Sci* **2022**, 81, (3), 70.

8. Zhang, X.; Hu, Q., Development of Geothermal Resources in China: A Review. *J Earth Sci-China* **2018**, 29, 452-467.

9. Zhou, W.; Hu, X.; Yan, S.; Guo, H.; Chen, W.; Liu, S.; Miao, C., Genetic Analysis of Geothermal Resources and Geothermal Geological Characteristics in Datong Basin, Northern China. *Energies* **2020**, 13, (7), 1792.

10. Peng, C.; Pan, B.; Xue, L.; Liu, H., Geophysical survey of geothermal energy potential in the Liaoji Belt, northeastern China. *Geotherm Energy* **2019**, 7, (1), 14.

11. He, L.; Chen, L.; Dorji; Xi, X.; Zhao, X.; Chen, R.; Yao, H., Mapping the Geothermal System Using AMT and MT in the Mapamyum (QP) Field, Lake Manasarovar, Southwestern Tibet. *Energies* **2016**, 9, (10), 855.

12. Zhang, G.; Liu, C.; Liu, H.; Jin, Z.; Han, G.; Li, L., Geochemistry of the Rehai and Ruidian geothermal waters, Yunnan Province, China. *Geothermics* **2008**, 37, (1), 73-83.

13. Du, J.; Liu, C.; Fu, B.; Ninomiya, Y.; Zhang, Y.; Wang, C.; Wang, H.; Sun, Z., Variations of geothermometry and chemical-isotopic compositions of hot spring fluids in the Rehai geothermal field, southwestern China. *J Volcanol Geoth Res* **2005**, 142, (3–4), 243-261.

14. Minissale, A.A., A simple geochemical prospecting method for geothermal resources in flat areas. *Geothermics* **2018**, 72, 258-267.

15. Chan, H.; Chang, C.; Dao, P.D., Geothermal Anomaly Mapping Using Landsat ETM+ Data in Ilan Plain, Northeastern Taiwan. *Pure Appl Geophys* **2018**, 175, (1), 303-323.

16. Calvin, W.M.; Littlefield, E.F.; Kratt, C., Remote sensing of geothermal-related minerals for resource exploration in Nevada. *Geothermics* **2015**, 53, 517-526.

17. Coolbaugh, M.F.; Kratt, C.; Fallacaro, A.; Calvin, W.M.; Taranik, J.V., Detection of geothermal anomalies using Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) thermal infrared images at Bradys Hot Springs, Nevada, USA. *Remote Sens Environ* **2007**, 106, (3), 350-359.

18. Hellman, M.J.; Ramsey, M.S., Analysis of hot springs and associated deposits in Yellowstone National Park using ASTER and AVIRIS remote sensing. *J Volcanol Geoth Res* **2004**, 135, (1–2), 195-219.

19. Xiong, Y.; Chen, F.; Huang, S., Application of remote sensing technique to the identification of geothermal anomaly in Tengchong area, southwest China. *J Chengdu Univ Technol (Sci Technol)* **2016**, 43, (1), 109-118.

20. Zhang, Y.; Zhang, Y.; Yu, H.; Li, J.; Xie, Y.; Lei, Z., Geothermal resource potential assessment of Fujian Province, China, based on geographic information system (GIS) -supported models. *Renew Energ* **2020**, 153, 564-579.

21. Tende, A.; Aminu, M.; Gajere, J., A spatial analysis for geothermal energy exploration using bivariate predictive modelling. *Sci Rep-Uk* **2021**, 11, 19755.

22. Wardoyo, G.; Pratama, H.; Sutopo, T.; Ashat, A.; Yudhistira, Y., Application of Artificial Intelligence in Forecasting Geothermal Production. *IOP Conf Series: Earth Environm Sci* **2021**, 732, 012022.

23. Assouline, D.; Mohajeri, N.; Gudmundsson, A.; Scartezzini, J., A machine learning approach for mapping the very shallow theoretical geothermal potential. *Geotherm Energy* **2019**, 7, (1), 19.

24. Gentana, D.; Sulaksana, N.; Sukiyah, E.; Yuningsih, E., Morphotectonics of Mount Rendingan Area Related To the Appearances of Geothermal Surface Manifestations. *Indones J Geosci* **2019**, 6, (3), 291-309.

25. Freski, Y.R.; Hecker, C.; van der Meijde, M.; Setianto, A., The effects of alteration degree, moisture and temperature on laser return intensity for mapping geothermal manifestations. *Geothermics* **2021**, 97, 102250.

26. Dramsch, J.S., 70 years of machine learning in geoscience in review. *Adv Geophys* **2020**, 1–55.

27. Gangwani, P.; Soni, J.; Upadhyay, H.; Joshi, S., A Deep Learning Approach for Modeling of Geothermal Energy Prediction. *Int J Comput Sci Inf Secur* **2021**, 18, (1), 62-65.

28. Shahdi, A.; Lee, S.; Karpatne, A.; Nojabaei, B., Exploratory analysis of machine learning methods in predicting subsurface temperature and geothermal gradient of Northeastern United States. *Geotherm Energy* **2021**, 9.

29. Yang, W.; Xiao, C.; Zhang, Z.; Liang, X., Identification of the formation temperature field of the southern Songliao Basin, China based on a deep belief network. *Renew Energ* **2022**, 182, 32-42.

30. Xu, S.; Guo, Y., *The Basis of Geothermics*. Science Press: Beijing, 2009; p 207.

31. Wohletz, K.; Heiken, G., *Volcanology and Geothermal Energy*. University of California Press: Berkeley, 1992.

32. White, D.E., Characteristics of geothermal resources. In United States, 1973; Vol. 54:4 Annual Meeting of the American Geophysical Union, Washington, DC, 16-20 Apr 1973.

33. Donti, P.L.; Kolter, J.Z., Machine Learning for Sustainable Energy Systems. *Annu Rev Env Resour* **2021**, 46, (1), 719-747.

34. Ribeiro, A.M.N.C.; Do Carmo, P.R.X.; Endo, P.T.; Rosati, P.; Lynn, T., Short- and Very Short-Term Firm-Level Load Forecasting for Warehouses: A Comparison of Machine Learning and Deep Learning Models. *Energies* **2022**, 15, (3), 750.

35. F., M.; L., B., Classification of hyperspectral remote sensing images with support vector machines. *Ieee T Geosci Remote* **2004**, 42, (8), 1778-1790.

36. Lizarazo, I., SVM-based segmentation and classification of remotely sensed data. *Int J Remote Sens* **2008**, 29, (24), 7277-7283.

37. Xiong, Y.; Zhang, Z.; Chen, F., Comparison of Artificial Neural Network and Support Vector Machine Methods for Urban Land Use/Cover Classifications from Remote Sensing Images: A Case Study of Guangzhou, South China. In *Proc. 2010 Int Conf Comp Appl System Model (ICCASM 2010)*, IEEE Xplore: Taiyuan, China, Oct. 22-24, 2010; pp V13-52-56.

38. Turker, M.; Koc-San, D., Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping. *Int J Appl Earth Obs* **2015**, 34, 58-69.

39. Li, J.; Chen, Y.; Schuster, G., Separation of Multi-mode Surface Waves by Supervised Machine Learning Methods. *Geophys Prospect* **2019**, 68.

40. Li, W.; Liu, L.; Gong, W., Multi-objective uniform design as an SVM model selection tool for face recognition. *Expert Syst Appl* **2011**, 38, (6), 6689-6695.

41. Garg, A.; Vijayaraghavan, V.; Mahapatra, S.S.; Tai, K.; Wong, C.H., Performance evaluation of microbial fuel cell by artificial intelligence methods. *Expert Syst Appl* **2014**, 41, (4, Part 1), 1389-1399.

42. Chi, W.H.; Chi, J.L., A comparison of methods for multiclass support vector machines. *IEEE T Neural Net* **2009**, 13, (2), 415-425.

43. Xiong, Y.; Wang, R.; Li, Z., Extracting land use/cover of mountainous area from remote sensing images using artificial neural network and decision tree classifications: A case study of Meizhou, China. In *Proc 2010 Int Symp Intel Infor Proc Trus Comp (IPTC 2010)*, IEEE Computer Society: Huanggang, China, Oct. 28 - 29, 2010; pp 133-136.

44. Otukei, J.R.; Blaschke, T., Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. *Int J Appl Earth Obs* **2010**, 12, (S1), S27-S31.

45. Tooke, T.R.; Coops, N.C.; Goodwin, N.R.; Voogt, J.A., Extracting urban vegetation characteristics using spectral mixture analysis and decision tree classifications. *Remote Sens Environ* **2009**, 113, (2), 398-407.

46. LeCun, Y.; Bengio, Y.; Hinton, G., Deep Learning. *Nature* **2015**, 521, 436-444.

47. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A., Going Deeper with Convolutions. *arXiv:1409.4842 [cs.CV]* **2014**.

48. Stanford Vision and Learning Lab, ImageNet. https://image-net.org (accessed on 20 January 2022).

49. Xiong, Y.; Chen, Q.; Zhu, M.; Zhang, Y.; Huang, K., Accurate detection of historical buildings using aerial photographs and deep transfer learning. In *2020 IEEE Int Geosci Remote Sens Symp (IGARSS 2020)*, Sep 26 - Oct 2, 2020 • Waikoloa, Hawaii, USA, 2020; pp 1592-1595.

50. Kumar, S.; Muhuri, P.K., A novel GDP prediction technique based on transfer learning using $CO_2$ emission dataset. *Appl Energ* **2019**, 253, 113476.

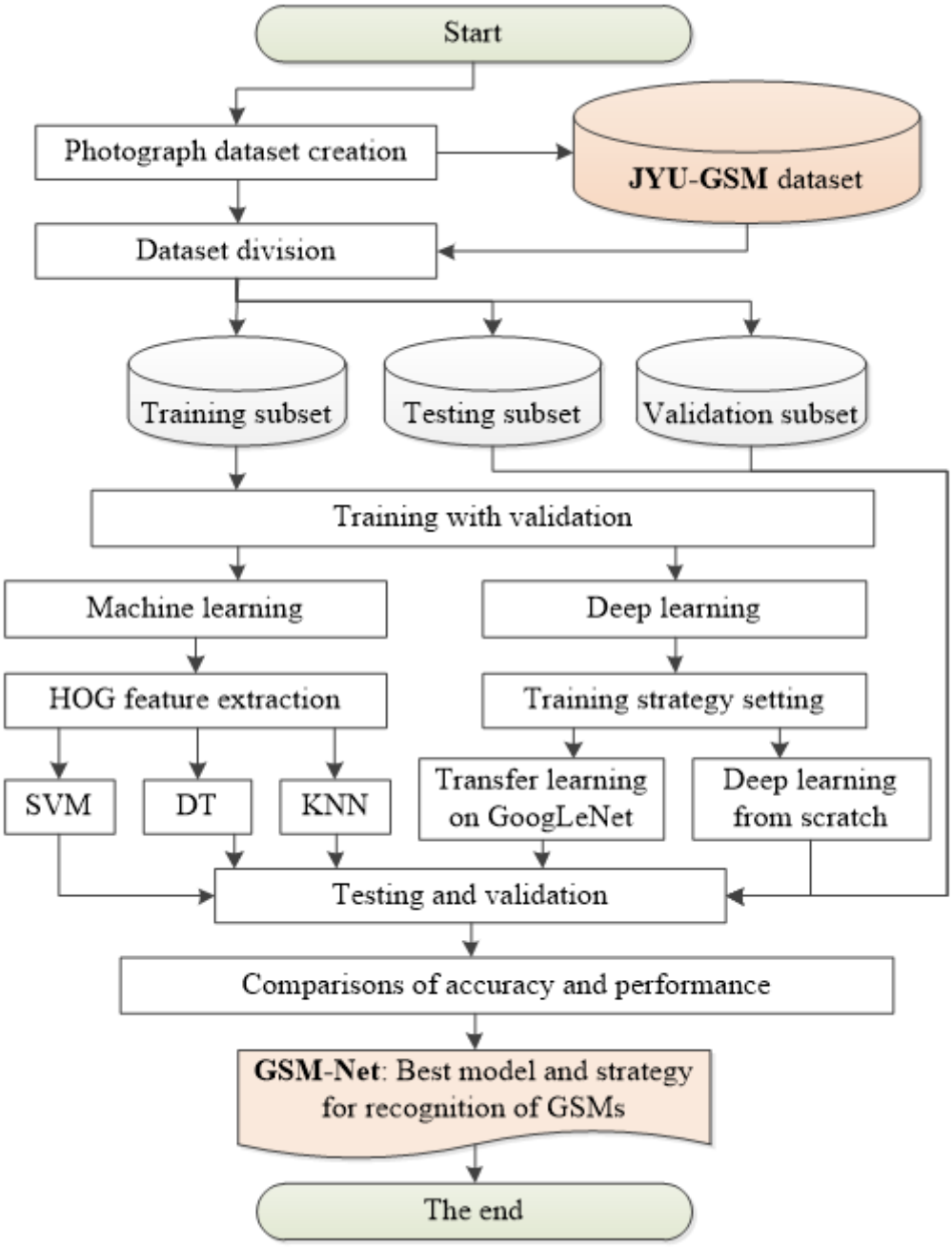
51. Pires De Lima, R.; Marfurt, Convolutional Neural Network for Remote-Sensing Scene Classification: Transfer Learning Analysis. *Remote Sens-Basel* **2019**, 12, 86.

52. Klemperer, S.L.; Zhao, P.; Whyte, C.J.; Darrah, T.H.; Crossey, L.J.; Karlstrom, K.E.; Liu, T.; Winn, C.; Hilton, D.R.; Ding, L., Limited underthrusting of India below Tibet: $^{3}He/^{4}He$ analysis of thermal springs locates the mantle suture in the continental collision. *PNAS* **2022**, 119, (12), e2113877119.

# Tables

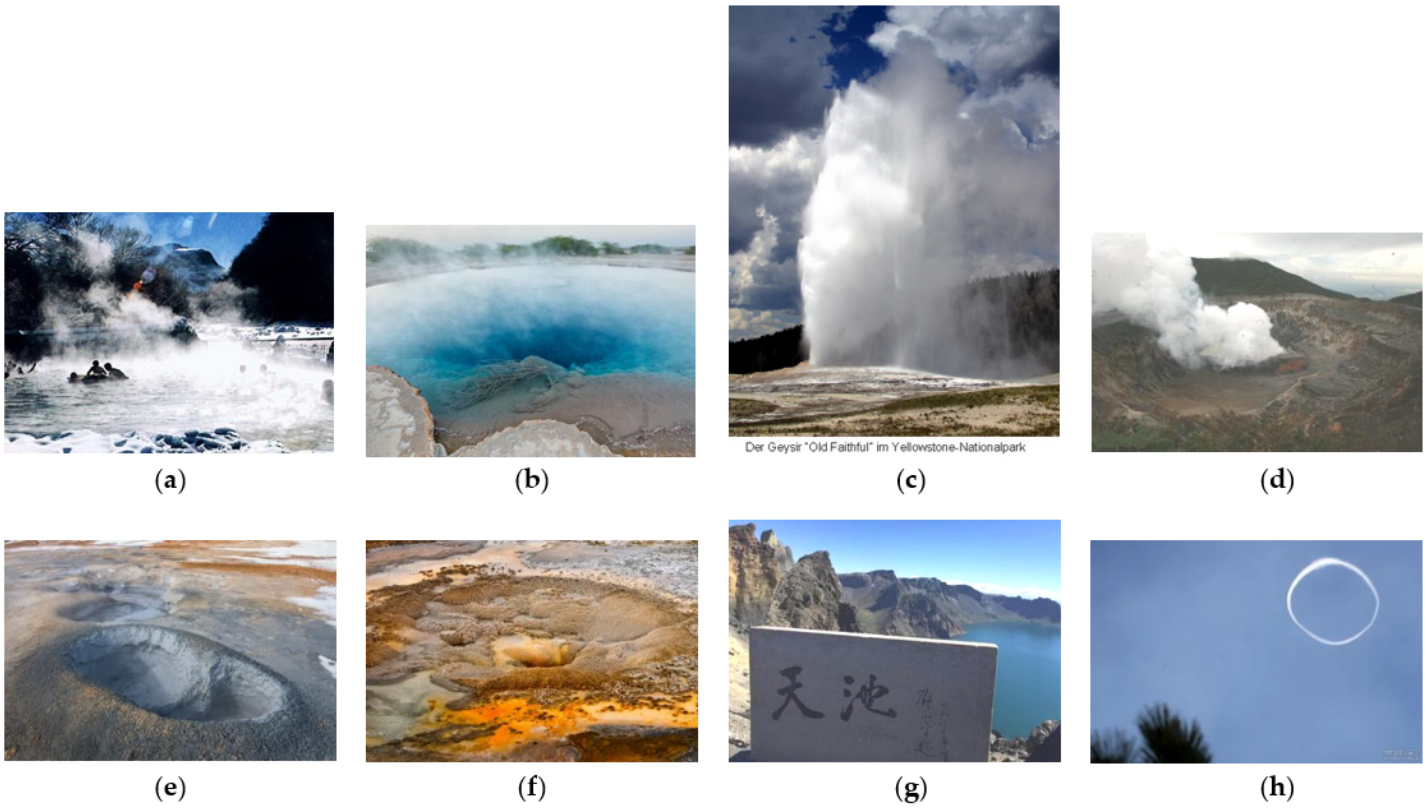
Table 1 is in the supplementary files section.

# Figures



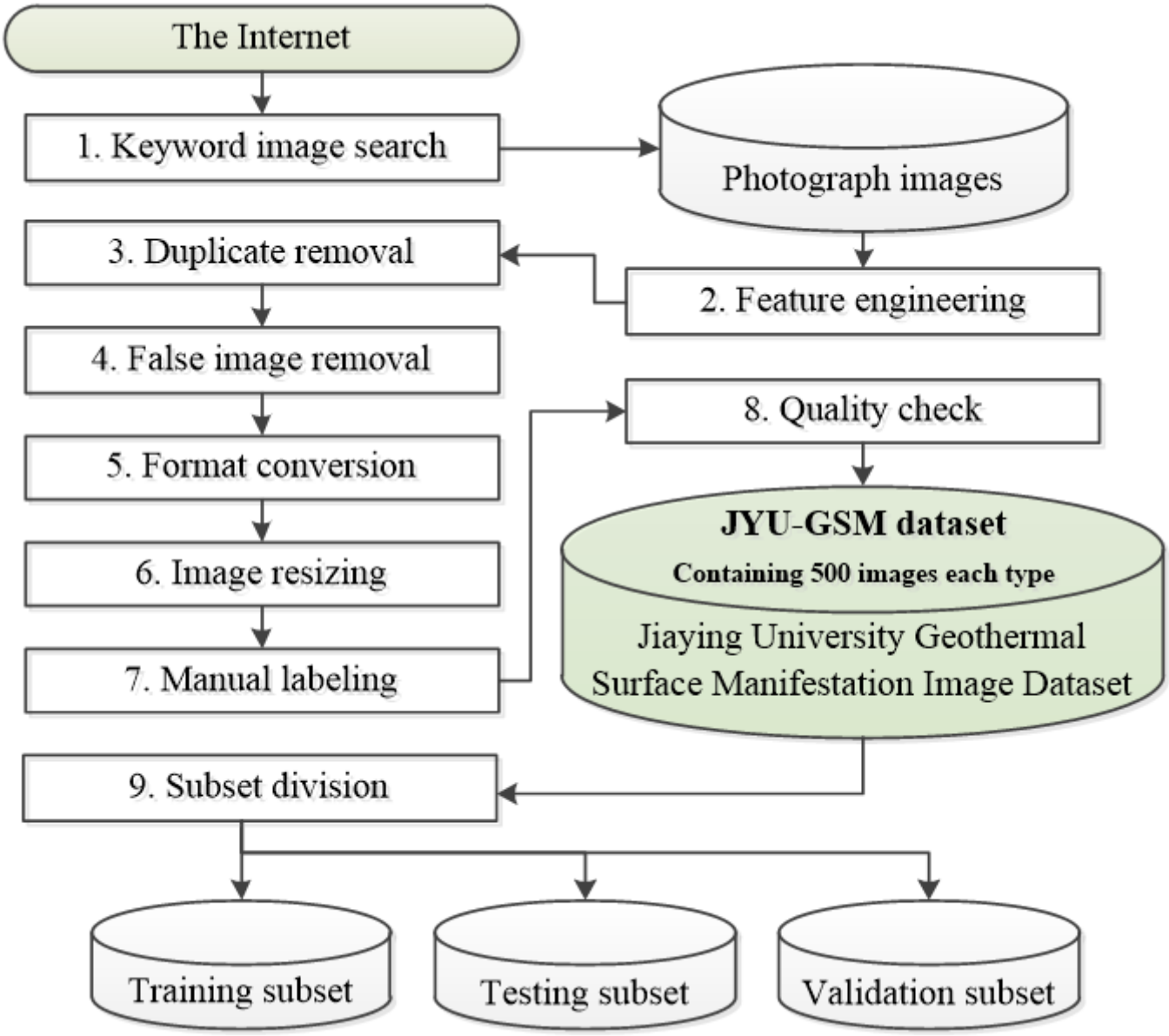

## Figure 1

Overall workflow chart of the present study. JYU-GSM is the abbreviation of the JiaYing University Geothermal Surface Manifestation; HOG is short for Histogram of Oriented Gradients, SVM is short for Support Vector Machine, DT is short for Decision Tree, and KNN is short for K-Nearest Neighbor; GSM-Net is the abbreviation of the Geothermal Surface Manifestation deep neural Network model retrained from GoogLeNet; GSMs denotes Geothermal Surface Manifestations.




## Figure 2

Example images of the eight types of geothermal surface manifestations (GSMs). (a) warm spring; (b) hot spring; (c) geyser; (d) fumarole; (e) mud pot; (f) hydrothermal alteration; (g) crater lake; and (h) none of GSM.

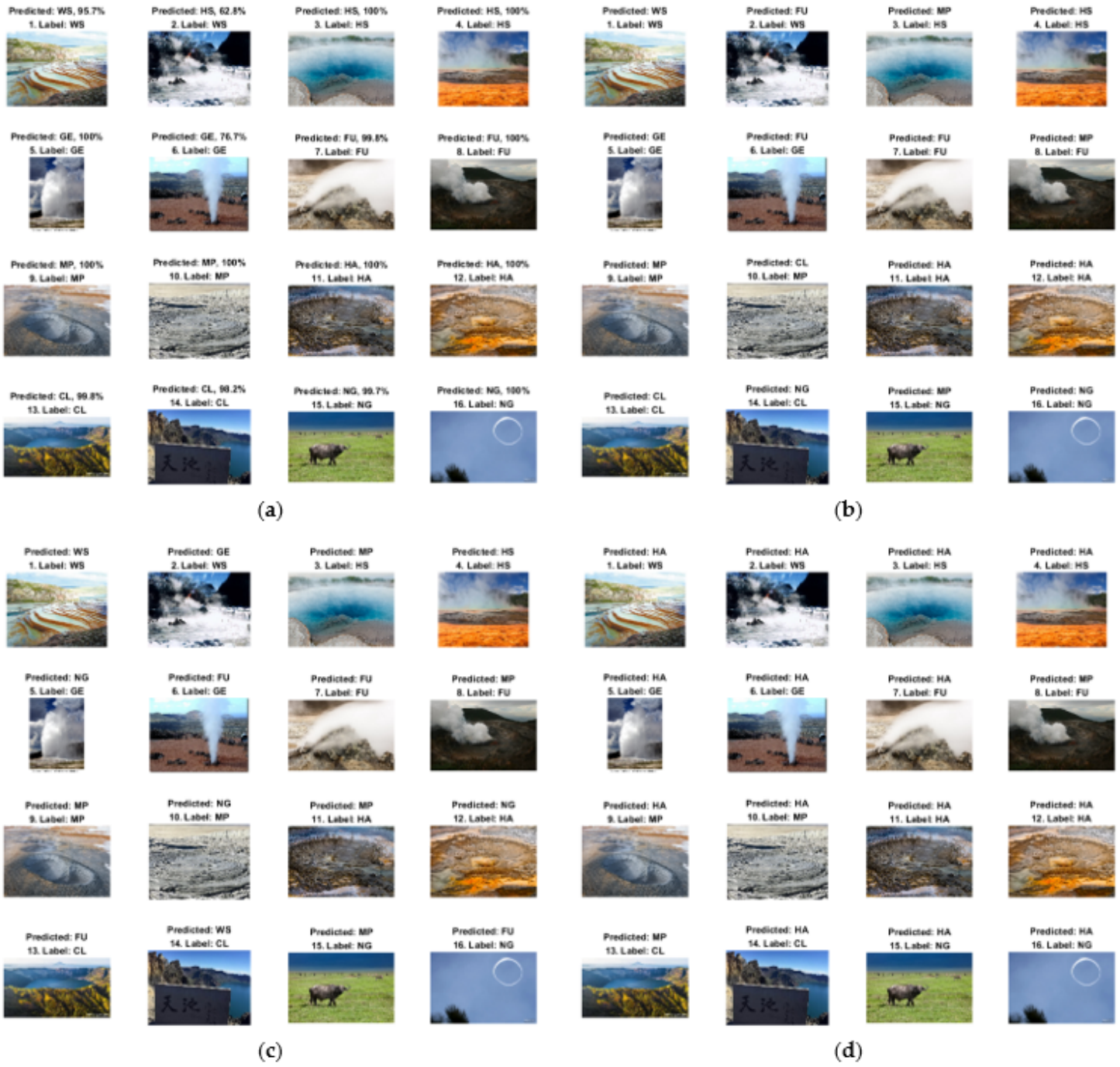



## Figure 3

Workflow chart of the data preparation and preprocessing. JYU-GSM is the abbreviation for the JiaYing University Geothermal Surface Manifestation photographs dataset.

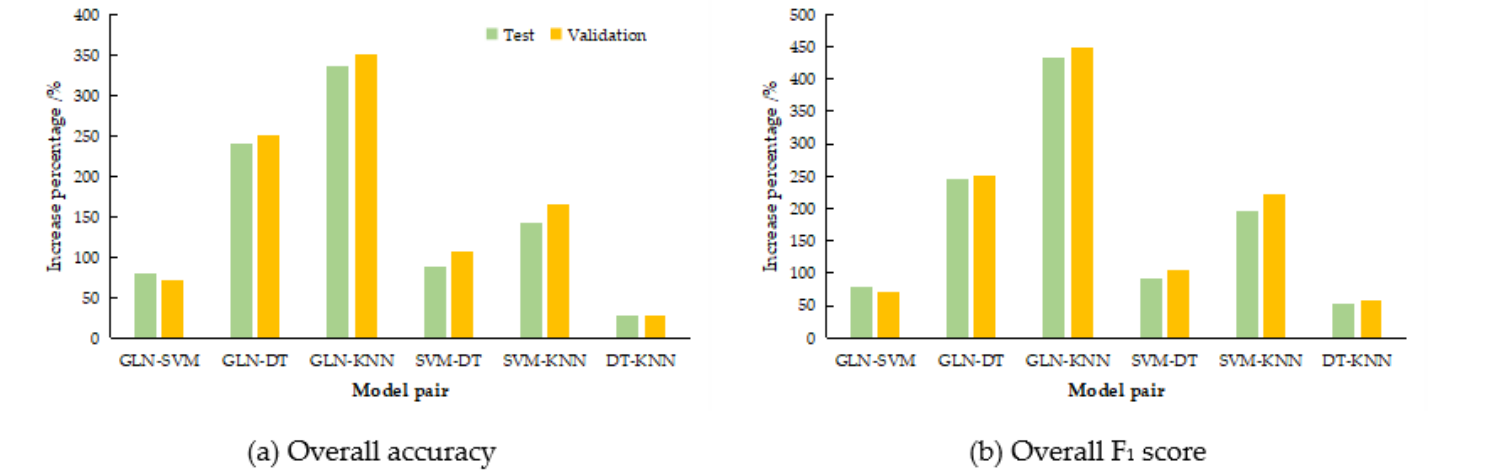

## Figure 4

Training progress screenshot graph of the GoogLeNet model in MATLAB.



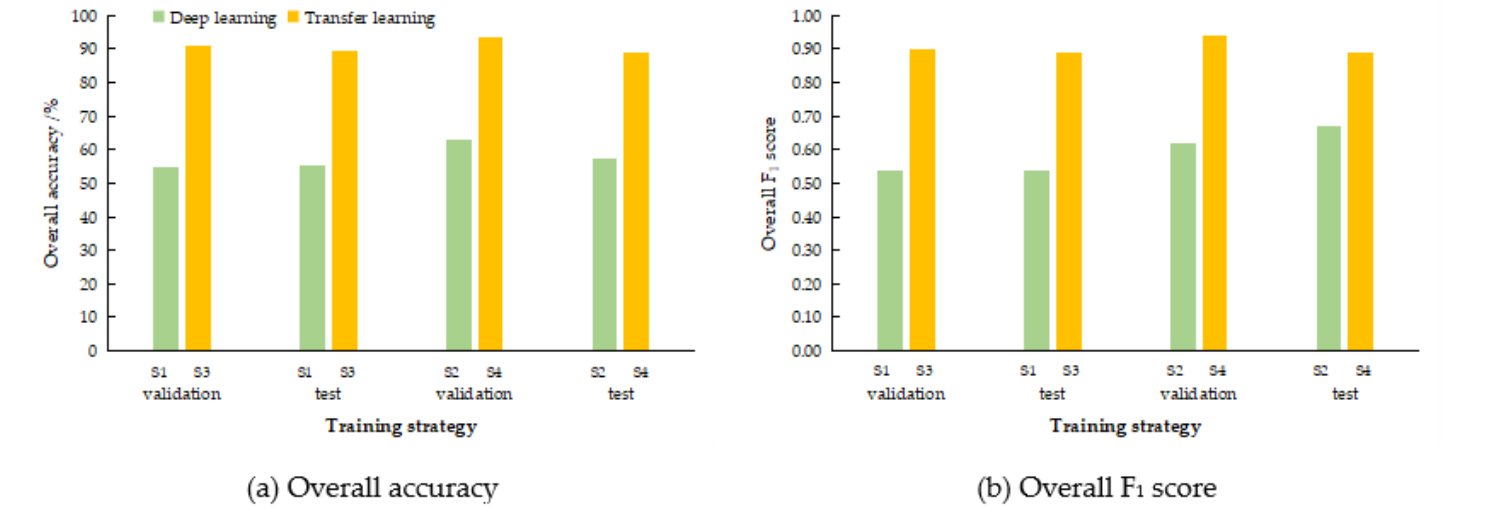

## Figure 5

Test image presentations of different classifier models. (a) The GoogLeNet model; (b) The Support Vector Machine (SVM) model; (c) The Decision Tree (DT) model; (d) The K-Nearest Neighbor (KNN) model. WS is short for Warm Spring; HS is short for Hot Spring; GE is short for GEyser; FU is short for FUmarole; MP is short for Mud Pot; HA is short for Hydrothermal Alteration; CL is short for Crater Lake, and NG is short for None of the Geothermal surface manifestation. The maximum probability of an image predicted to be a type is labeled after the abbreviation by the GoogLeNet model. The digital number before “Label” is the numerical order of an image.



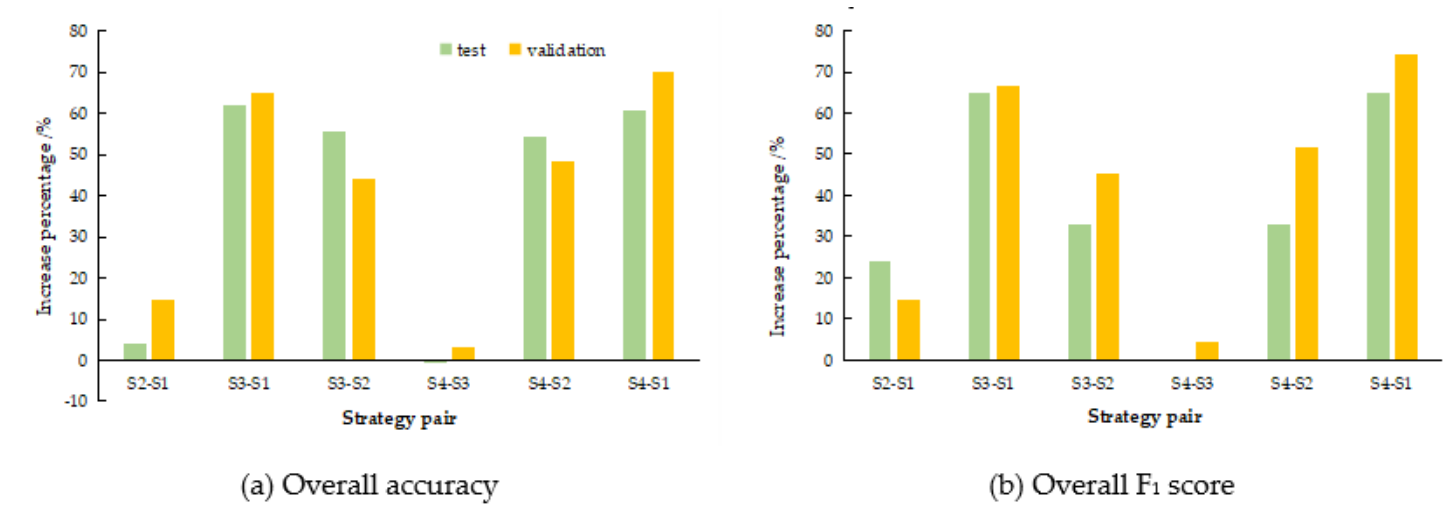

## Figure 6

Assessment metric change comparison of different pairs of models. (a) Overall accuracy; (b) Overall $F_1$ score. GLN denotes the GoogLeNet model; SVM denotes the Support Vector Machine model; DT denotes the Decision Tree model; and KNN denotes the K-Nearest Neighbor model. GLN-SVM means the difference of the accuracy metrics for GLN and SVM, the others as the same as this meaning.



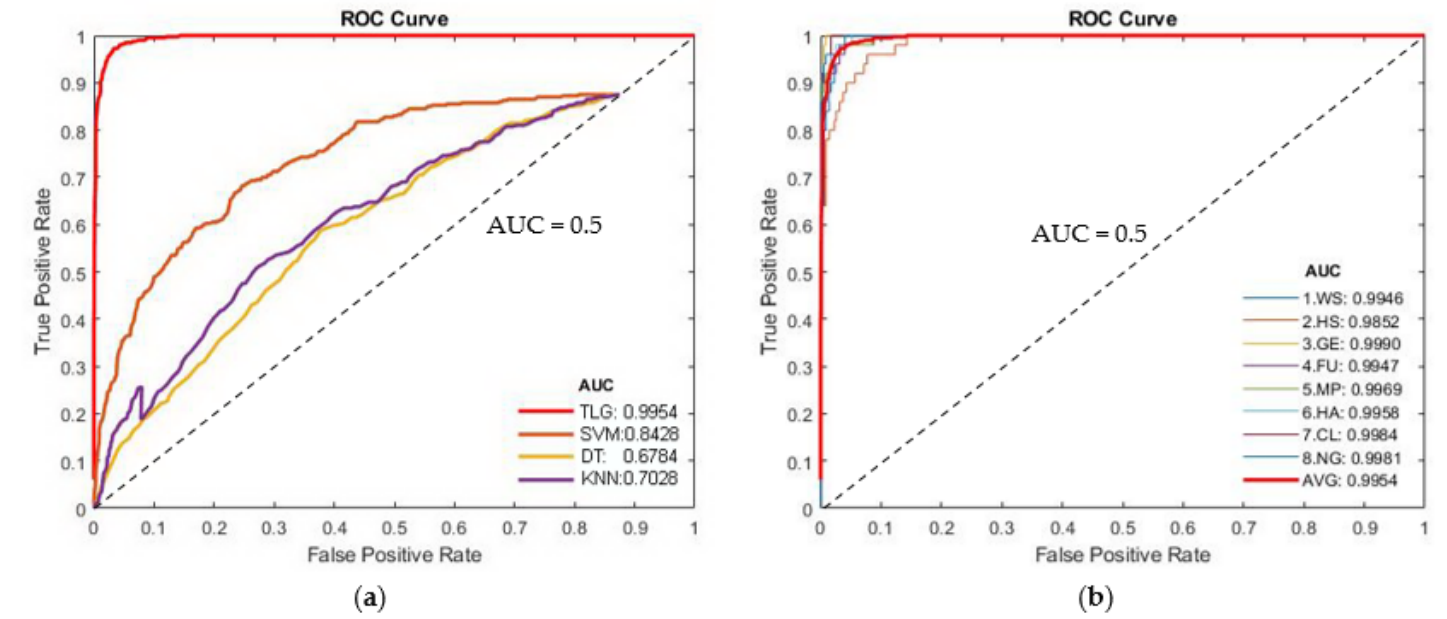

## Figure 7

Assessment metric comparison of different training strategies. (a) Overall accuracy; (b) Overall $F_1$ score. S1 denotes the training strategy from scratch and with the initial ten layers frozen; S2 denotes the training strategy from scratch and with no frozen layer; S3 denotes the training strategy of the pretrained network and with the initial ten layers frozen; and S4 denotes the training strategy of the pretrained network and no frozen layer.




## Figure 8

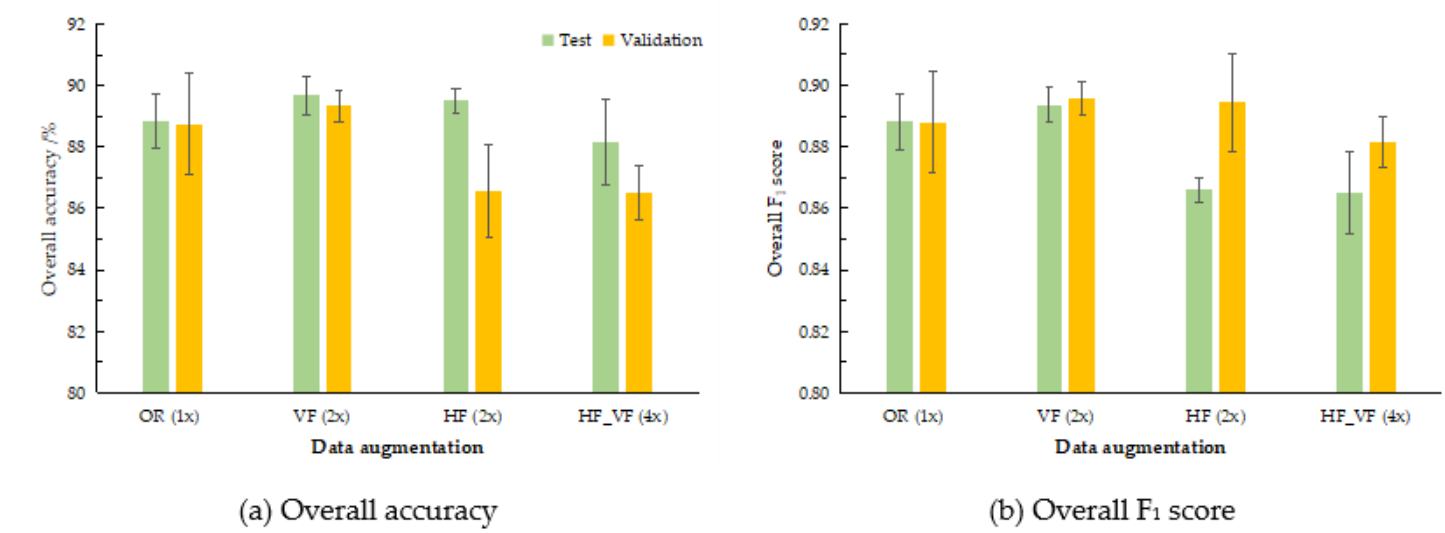
Assessment metric change comparison of different pairs of training strategies. (a) Overall accuracy; (b) Overall $F_1$ score. S1 denotes the training strategy from scratch and with the initial ten layers frozen; S2 denotes the training strategy from scratch and with no frozen layer; S3 denotes the training strategy of the pretrained network and with the initial ten layers frozen; and S4 denotes the training strategy of the pretrained network and no frozen layer. S2-S1 means the difference of the accuracy metrics for S2 and S1, the others as the same as this meaning.




## Figure 9

ROC (Receiver Operating Characteristic) curve. (a) Test assessment for the four classifier models. (b) Test assessment for different types of geothermal surface manifestations (GSMs) by the GoogLeNet model. AUC means the Area Under the Curve. TLG denotes Transfer Learning with the pretrained GoogLeNet;

SVM denotes the Support Vector Machine algorithm; DT denotes the Decision Tree algorithm; and KNN denotes the K-Nearest Neighbor algorithm. WS denotes Warm Spring; HS denotes Hot Spring; GE denotes Geyser; FU denotes Fumarole; MP denotes Mud Pot; HA denotes Hydrothermal Alteration; CL denotes Crater Lake; NG denotes None of GSMs; and AVG denotes the Average of all AUC values of the eight types of GSMs.




(a) Overall accuracy

(b) Overall $F_1$ score

## Figure 10

Assessment metric comparison of the validation and test results with data augmentation. (**a**) Overall accuracy; (**b**) Overall $F_1$ score. The division ratio of the JYU-GSM dataset is 0.8:0.1:0.1 for the training (400 images), testing (50 images), and validation (50 images) subsets. OR (1x) denotes the original training subset; VF (2x) denotes the OR (1x) and its vertical flip (2x); HF (2x) denotes the OR (1x) and its horizontal flip (2x); HF_VF (4x) denotes the OR (1x), its vertical flip, and their horizontal flip. All models are tested and validated on the same test and validation subsets, respectively.

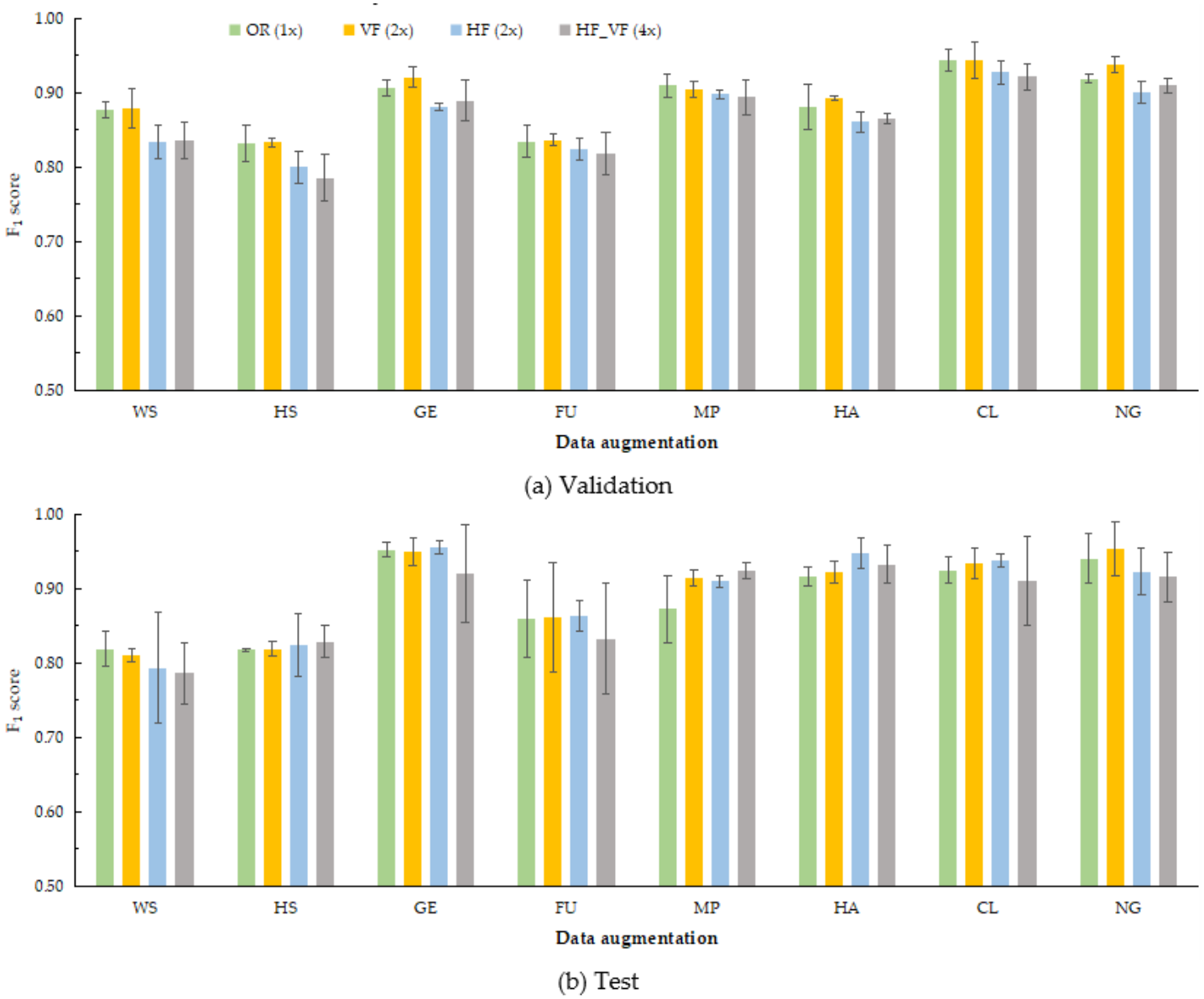



**Figure 11**

$F_1$ score comparison of different GSMs with data augmentation. (**a**) Validation; (**b**) Test. The division ratio of the JYU-GSM dataset is 0.8:0.1:0.1 for the training (400 images), testing (50 images), and validation (50 images) subsets. OR (1x) denotes the original training subset; VF (2x) denotes the OR (1x) and its vertical flip (2x); HF (2x) denotes the OR (1x) and its horizontal flip (2x); HF_VF (4x) denotes the OR (1x), its vertical flip, and their horizontal flip. All models are tested and validated on the same test and validation subsets, respectively. WS is short for Warm Spring; HS is short for Hot Spring; GE is short for GEyser; FU is short for FUmarole; MP is short for Mud Pot; HA is short for Hydrothermal Alteration; CL is short for Crater Lake; and NG is short for None Geothermal surface manifestation.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.docx