UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE MINAS Y ENERGÍA

Assisting mine planning, operation and ore grading by using off-the-shelf machine learning techniques

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Enming Li
Master of Mining Engineering

Madrid, 2025

UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE MINAS Y ENERGÍA

Doctoral Degree in Research, Modeling and Analysis of Environmental Risk

Assisting mine planning, operation and ore grading by using off-the-shelf machine learning techniques

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Presentada para optar al título de Doctor por:

Enming Li
Master of Mining Engineering

Under the supervision of:
Dr. José Ángel Sanchidrián Blanco (Director)
Dr. Pablo Segarra Catasús (Codirector)

Madrid, 2025

Title: Assisting mine planning, operation and ore grading by using off-the-shelf machine learning techniques

Author: Enming Li

Doctoral Programme: Research, Modeling and Analysis of Environmental Risk

Thesis Supervision:

Dr. José Ángel Sanchidrián Blanco, Professor, Universidad Politécnica de Madrid (Director)

Dr. Pablo Segarra Catasús, Associate Professor, Universidad Politécnica de Madrid (Codirector)

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

This thesis has been partially funded by the projects DIGIECOQUARRY (Innovative digital sustainable aggregates systems) and ILLUMINEATION (Bright concepts for a safe & sustainable digital mining future). All funded by the European Union's Horizon 2020 research and innovation program with funding agreements no. 101003750 and 869379, respectively. The PhD candidate was awarded a scholarship by the China Scholarship Council.

Acknowledgement

与大多数人比，我很幸运，也很幸福

回首往昔，如白驹过隙，儿时的一幕幕浮现在我眼前，还记得我出生的那个小房子，是坐落在一个铁轨旁边，经常能听到火车的轰鸣，几年前回去，那个村子已经完全变成了工业区，我的家也变成了废墟。父母为了我的教育，从我出生的地方搬到了另外一个村子，但那个村子离我的小学还是很远，爸爸每天骑着摩托车带我去上学，我记得每次我都是最后一个到学校的，因为路程遥远，但是爸爸妈妈想让我多睡一会儿，所以起床也很晚。后来，为了我离学校近一点，我又搬家了，这次就在我的小学旁边，那个时候，每逢周末，爸爸会背着比较重的电子琴带我去学习弹琴，并且每次都陪我一整个下午，或者是陪我去学英语。到了中学，我的学校是在县城里，家里为了我又把家从开发区搬到了县城。到了高中，我家买了第一台车，爸爸每天送我上下学，妈妈给我准备早餐。高中三年，爸爸妈妈几乎是和我同步休息，我在客厅学习，爸爸则是在旁边把电视声音调到最低，陪我直到我睡觉。终于我考上了一所 985 大学，还多亏爸爸帮我报的志愿。

大学前几年，我几乎是荒废了时光，那时的我很迷茫，以为考上了大学就是终点，还好我在最后一年觉醒，并且幸运的考上了本校研究生，在此要感谢我大学阶段帮助过我的老师和同学。再后来，我幸运的遇见了周健老师，研究生一开始，我很迷茫，也很懈怠，还好他没有放弃我，并且一直帮助我，才有了我在西班牙的故事，我还要感谢我研究生阶段的其他老师和同学。特别致谢送给我可靠的朋友们，喻智，张宁，席斌，邱引贵等人。

其实我做梦都没想到，我会来到西班牙，在此之前我对这个国家的了解仅仅是斗牛，我甚至都不知道他们的火腿也很出名。我同样很幸运的遇见了 José 和 Pablo，一开始来到这个国家，我并不适应，并且生了病，他们带我去了医院，还给我提供了很多生活上的帮助，包括我的同事 Santiago，Alberto，Zahir 和教授 Luis。他们给了我很多包容，帮助我去爱尔兰开会，去奥地利访问，也帮助我发表了很多高质量的文章。

我想说，博士这几年是应该是我人生中最快乐的时光之一，我的收获很大，唯一感到遗憾的就是，我没有好好学习西班牙语，这个国家的人都非常友善，这里的食物很好，风景也非常好，我很喜欢这个国家。

特别感谢我的国家提供奖学金，也要感谢所有的审稿人，要感谢的人太多，我都记在心里，最后，感谢自己吧，多少个日日夜夜的工作，还好你没有放弃，再过五年或者十年，你一定比现在更强大，也能实现更多你内心的目标和理想！

Compared to many, I consider myself truly fortunate and blessed

As I look back, vivid memories of my childhood resurface. I was born in a small house near a railway, now long gone. My parents moved multiple times just so I could be closer to better schools, even riding me long distances on a motorcycle each day, allowing me to sleep a little longer.

Throughout my schooling, they sacrificed silently—carrying heavy instruments for my music lessons, sitting beside me during long study nights, and supporting me unconditionally. I was lucky to enter a top university, and even luckier to stay for my master’s, thanks to the support of my professors and classmates.

Meeting Professor Zhou Jian changed my life—he guided me through confusion, led me into international research, and eventually to Spain. Coming to Spain felt surreal, and I knew little about the country before arriving. I am deeply grateful to José, Pablo, Luis, Santiago, Alberto and Zahir who supported me through illness, conferences, publications, and everyday life. I clearly remember, José, Pablo and Santi stayed with me in the hospital for a long time. Jose and Pablo helped me improve my paper word for word. Alberto accompanied with for my residence permit. You are beyond my professors and colleagues.

These PhD years have been among the happiest of my life. If I have one regret, it's not learning Spanish better. The people here are warm, the food and scenery are wonderful, and I've grown immensely as a person and scholar.

I sincerely thank my country for the scholarship, the reviewers for their feedback, and everyone who helped me along the way. Most of all, I thank myself—for all those tireless days and nights. Don't stop now—five or ten years from now, may you be even stronger and closer to your dreams.

Abstract

By analysing case studies from mining regions worldwide, the thesis provides insights into how machine learning techniques can help in optimizing mining operations. Five main chapters are presented.

The first problem addressed is the evaluation of the median fragment size from blasting. The prediction models involve 19 influential factors by two types of support vector regression (SVR) and five hyper-parameters tuning algorithms. By comparison, it could be found that the v-SVR model optimized by grey wolf optimization showed the best performance. The sensitivity analysis indicated that uniaxial compressive strength is the most influential factor in this case.

In the next chapter, 1024 cases from published literature were used to develop the intelligent prediction models of the gas relative permeability. This chapter proposed five hybrid kernel extreme learning machine (KELM) techniques. The best testing performance was generated by butterfly optimization algorithm when swarm size is equal to 150. By mutual information, gas saturation was identified as the most influential factor, and when used as the sole input, optimized KELM still outperformed the empirical Corey-Brooks model.

While the above two tasks used data collected from published literature, the following chapters include the data collection by field or laboratory experiments to which the author has actively contributed, on which different analysis of images are developed.

For reducing the amount of sampling required and equipment consumption, two new techniques are proposed to evaluate the fluorite grade. 48 drilling chip samples were gathered from six boreholes. Images of borehole walls were captured using a televiewer with white and ultraviolet (UV) light. The cumulative distribution of pixel color intensities (red, green, blue) served as model inputs. A new outlier inspection method was proposed named ‘take one out’ and thus improved model performance. The best predictions were achieved using pixel intensities from the combination of white and UV light scans ($R^2 = 0.83$, RMSE = 3.32%), using optimized SVR.

Based on the success of the televiewer work, a more convenient and lower-cost way is proposed. 494 pellet samples made from drilling chips were photographed using a smartphone. To minimize the impact of environmental factors on image colors, ColorChecker were used for color correction. The average red and blue pixel intensities, along with color uncertainty and texture features, were analyzed using principal component analysis to develop five clustering models. Two clusters (C2 and C3) were qualitatively labeled as waste, while cluster C1 was identified as ore. The spectral clustering (SC) method achieved a waste recognition rate of 94.7% in cluster C3.

Inspired by the fluorite grade work, a new automated photography approach is applied to recognize lithology by an endoscope. Images were automatically extracted from borehole wall videos. Image quality was determined using gray pixel intensity thresholds and no-reference quality metrics. Contrast-Limited Adaptive Histogram Equalization was applied to enhance image quality. A total of 7,583 images were selected. The optimized

Light Gradient Boosting Machine model achieved the best classification performance, with 88.04% accuracy. Feature importance analysis highlighted color counting as the most influential parameter. This method offers a more flexible and cost-effective alternative for lithology identification.

In conclusion, by means of various machine learning techniques, some complicated mining issues can be interpreted or simplified and thus offer insight for the mining design and a better control of the environmental impact.

Resumen

Mediante el análisis de estudios de caso en regiones mineras de todo el mundo, esta tesis ofrece perspectivas sobre cómo las técnicas de aprendizaje automático pueden ayudar a optimizar las operaciones mineras. Se presentan cinco capítulos principales.

El primer problema abordado es la evaluación del tamaño medio de fragmentos tras las voladuras. Los modelos de predicción incluyen 19 factores influyentes mediante dos tipos de regresión por vectores de soporte (SVR) y cinco algoritmos de ajuste de hiperparámetros. Al comparar los resultados, se encontró que el modelo v-SVR optimizado mediante el algoritmo de optimización del lobo gris mostró el mejor rendimiento. El análisis de sensibilidad indicó que la resistencia a la compresión uniaxial es el factor más influyente en este caso.

En el capítulo siguiente, se utilizaron 1024 casos extraídos de la literatura para desarrollar modelos de predicción inteligente de la permeabilidad relativa del gas. Se propusieron cinco técnicas híbridas de KELM (máquinas de aprendizaje extremo con núcleos), y el mejor rendimiento se logró con el algoritmo de optimización de mariposa cuando el tamaño del enjambre fue de 150. Mediante información mutua, se identificó la saturación de gas como el factor más influyente; incluso como única entrada, el modelo KELM optimizado superó al modelo empírico de Corey-Brooks.

Mientras que las dos tareas anteriores usaron datos extraídos de literatura, los capítulos siguientes incluyen recolección de datos en campo o laboratorio, donde el autor participó activamente. Se desarrollaron análisis basados en imágenes.

Para reducir el muestreo necesario y el uso de equipos, se proponen dos nuevas técnicas para evaluar la ley de fluorita. Se recolectaron 48 muestras de detritos de seis sondeos. Se capturaron imágenes de las paredes de los sondeos con un televiewer bajo luz blanca y ultravioleta (UV). La distribución acumulada de intensidades de color (rojo, verde, azul) sirvió como entrada. Se propuso un nuevo método para detectar valores atípicos, denominado “take one out”, que mejoró el rendimiento del modelo. Las mejores predicciones se lograron con intensidades de píxeles combinadas de luz blanca y UV (R^2 = 0.83, RMSE = 3.32%), usando SVR optimizado.

A partir del éxito con el televiewer, se propuso un método más económico y práctico. Se fotografiaron 494 muestras granuladas hechas con detritos de perforación, usando un teléfono móvil. Para minimizar el efecto ambiental en los colores, se usaron ColorChecker para corregir el color. Las intensidades promedio de píxeles rojos y azules, junto con la incertidumbre del color y características de textura, se analizaron mediante análisis de componentes principales para desarrollar cinco modelos de agrupamiento. Dos grupos (C2 y C3) fueron etiquetados como desecho, mientras que el grupo C1 se identificó como mineral. El método de agrupamiento espectral alcanzó una tasa de reconocimiento de desecho del 94.7% en C3.

Inspirado en el trabajo de la ley de fluorita, se aplicó un nuevo enfoque fotográfico automatizado para reconocer litología mediante endoscopio. Se extrajeron automáticamente imágenes de videos de las paredes de los sondeos. La calidad de imagen se determinó mediante umbrales de intensidad en escala de grises y métricas sin referencia. Se usó la ecualización de histograma adaptativa limitada para mejorar la imagen. Se seleccionaron 7,583 imágenes para clasificación litológica. El modelo

optimizado LightGBM logró una precisión del 88.04%. El análisis de importancia de características destacó el conteo de colores como el parámetro más influyente. Este método ofrece una alternativa flexible y rentable para identificar litologías.

En conclusión, mediante diversas técnicas de aprendizaje automático, algunos problemas complejos en minería pueden interpretarse o simplificarse, aportando así ideas para el diseño minero y un mejor control del impacto ambiental.

Table of contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1. Introduction

1.1 Problem statement

In mining engineering, excavation is one of the most important steps. It mainly involves drilling, blasting, ore processing and sorting. The environmental problems caused by these operations deserve attention. Because they can lead to greenhouse gas emissions, water and air pollution soil degradation and other problems. However, these operations are inevitable since the high requirements of resources, the contribution of economy and employment and the dependence of the global supply chain. This thesis provides tools for different mining stages from drilling to rock mining to improve mining operations. It investigates gas relative permeability in reservoir, fragmentation from blasting, fluorite grade discrimination and lithologies assessment through machine learning models.

Blasting fragmentation: Inadequate rock fragmentation in blasting operations often results in excessive ground vibration, dust emissions, and inefficient downstream processes. A number of researchers have proposed and developed various empirical models to predict and estimate the blasting fragment size (Chung and Katsabanis, 2000; Cunningham, 1987; Esen et al., 2003; Gheibie et al., 2009; Ouchterlony and Sanchidrián, 2018; Sanchidrián and Ouchterlony, 2023; Segarra et al., 2018; Spathis, 2004; Thornton et al., 2002). These empirical models were created based on experimental datasets, blasting mechanisms and statistics. They integrated different blast design parameters and rock mass parameters, however, they cannot consider too many influential factors simultaneously. In this dissertation, blasting fragmentation is predicted considering nine factors. Regarding this, the machine learning seems to be a good alternative way to do prediction tasks.

Gas relative permeability in mining operations: Accurate understanding and prediction of gas relative permeability are essential for managing gas flow in underground mines and to predict gas production and recovery factors. High permeability can lead to the release of hazardous gases, increasing the risk of gas explosions or suffocation for workers. Monitoring and predicting permeability helps to ensure proper ventilation and gas extraction, reducing these safety risks. Although it can be effectively measured by laboratory experiments (Esmaeili et al., 2019; Honarpour et al., 2018; Honarpour and

Mahmood, 1988), the preparation of rock samples and strict requirements of laboratory limit its extrapolation and applicability. Various empirical and semi-theoretical functions have been developed (Aigbedion, 2007; Coates and Dumanoir, 1973; Jorgensen, 1991; Timur, 1968). However, these functions can only consider partial influential factors. For complicated conditions, they cannot provide reasonable interpretation. Machine learning can address these issues by providing more accurate and timely predictions, reducing harmful gas emissions and contributing to safer mining practices. However, some current machine learning-based models did not consider the usage of cross validation and thus fail to provide accurate predictive capabilities. In addition, machine learning techniques like hybrid kernel extreme learning machines were rarely applied in this area. Their potentials are worthwhile to be investigated.

Ore grade assessment from in-hole logging images with televiewer: Accurate ore grade estimation is essential for minimizing waste and maximizing resource utilization in mining. Traditional core drilling is costly (Starr and Ingleton, 1992), leading to inefficient extraction strategies and increased environmental degradation. Based on the fact that some minerals present distinct colors and other optical properties. Some studies proposed to utilize the color characteristics from the spectral images to predict ore grade (Berrezueta et al., 2016; Okada et al., 2020; Tanaka et al., 2019), however, these studies need advanced equipment and strict operation environment. To overcome this, some researchers proposed to employ the color parameters from microscopic images (Li et al., 2017) and thin sections (Baykan and Yılmaz, 2010) and mine faces (Desta and Buxton, 2017). These works generally are conducted in the laboratory and thus there is a need to find a way to measure the ore grade in a production environment. For this, the televiewer technique is considered. First, it provides high-resolution, continuous images of the borehole walls, capturing detailed geological structures. This allows for more accurate identification of ore-bearing zones without the need for extensive core sampling. Second, it preserves the borehole while gathering data, unlike traditional coring, which is both time-consuming and costly. Third, compared to conventional photography, the televiewer provides 360-degree coverage and oriented images, making it a more comprehensive tool for understanding the subsurface. This dissertation aims to apply televiewer images to measure the fluorite grade in an underground mine located in Granada province, Spain. By matching the section images and ore grade, the fluorite grade prediction model is established by regression and classification techniques.

Ore grade by smartphone photography: The usage of color characteristics from the televiewer images to predict fluorite grade is promising and is extended to other material layouts, like pellets prepared from drilling chips. This material is commonly used for chemical assessment of fluorite grade using XRF technology to define blending operations. Smartphone is widely available and easy to use under laboratory conditions. In addition, smartphone photography allows for rapid data collection and easy sharing of images with remote teams for collaborative analysis. Meanwhile, to overcome the interference of light sources, a Colorchecker is employed to calibrate the pellet image (Pascale, 2005). The corrected image characteristics are used for the development of new fluorite grade discrimination models by some unsupervised techniques.

Lithology assessment from in-hole images: Precise lithology identification is crucial for mine planning and reducing environmental damage from excessive or misdirected

excavation. Previous studies recognized the rock lithologies according to the physical or chemical properties of rocks (Konaté et al., 2017; Mishra et al., 2022; Sebtosheikh and Salehi, 2015; Wang and Zhang, 2008). However, these methods obtain the prediction from indirect inference and need professional interpretation. Recently, vision-based methods were proposed. These images directly procure the lithology information from thin sections (Faria et al., 2022; Z. Xu et al., 2022), raw rock samples (Xu et al., 2021), hyperspectral-based images (Galdames et al., 2019), drilling core trays (Alzubaidi et al., 2021) and so on. By the employment of machine learning or deep learning techniques, these images can be sufficiently analyzed and provide a good classification for lithologies. However, the preparation of rock samples or costly equipment make them difficult to apply in the practical engineering. The endoscope can fix these drawbacks. In this dissertation, the lithologies (massive limestone, brecciated limestone and high amount of clay) from six blasts in a limestone quarry were measured by endoscope in 79 holes. The purpose is to develop a model that can automatically classify the images at different depths. This tool will serve to assess the quality of the material in the block before blasting and also to plan material hauling and blending depending on the amount of clay. The results could be also used to calibrate Measurement-While-Drilling signals without the intervention of a geologist.

1.2 Objectives

The objectives of this thesis are:

1. Develop new machine learning models to improve the prediction of median fragment size in blasting. Optimizing fragmentation helps reduce excessive energy use, and improve downstream process towards a more sustainable mining operation.

2. Develop new machine learning models to enhance the prediction of gas relative permeability in reservoirs. This involves a better knowledge of the gas deposit in order to perform a better planning of production and improve management of underground storage.

3. Use of televiewer to obtain image of the blasthole walls in a fluorite mine. Then the significant image properties can be extracted for the development of a fluorite grade prediction model.

4. Use of a smartphone as a lower-cost tool to recognize the fluorite grade based on image color analysis. The usage of the smartphone photography provides a non-invasive evaluations of fluorite grade, reducing the number of tests with traditional chemical methods. The ability of Colorchecker to correct the colors can increase the adaptability of the method in adverse environments. The calibrated image properties are used for the discrimination of ore/waste by unsupervised clustering methods.

5. Propose a new system for intelligent prediction and recognition of blasthole lithologies logging with an endoscope. To overcome the inherent shortcomings of the instrument, some user-defined codes are designed to recognize the depth of each lithology displayed in the images. In addition, the image enhance technique

is employed to improve the quality of procured images. By automatically extracting the needed color information from endoscope videos, machine learning models can fast recognize and predict the lithology.

These objectives provide off-the-shelf tools for mining optimization. The first two objectives (Obj. 1 and Obj. 2) are developed using published numerical datasets. The last three objectives (Obj. 3, Obj. 4 and Obj. 5) are achieved from new gathered data. They are based on the colors observed when the rock surfaces are lightened with different sources, mainly white light. Color-based approaches have been validated in two rock masses using different tools (i.e. endoscope, televiewer, and smartphone) that photograph the rock inside the blasthole (Obj. 3 and Obj. 5) or on the pellet surface (Obj. 4).

1.3 Machine learning methods

To facilitate the reading, all machine learning methods used in this thesis have been introduced in this section. The first objective (median blasting fragment) used two regression techniques derived from support vector machine (SVM), i.e., v-SVR and e-SVR, and optimization algorithms (salp swarm algorithm (SSA), grey wolf optimization (GWO), genetic algorithm (GA), particle swarm optimization (PSO) and grid search (GS)). The second objective (gas relative permeability in reservoir) involved kernel extreme learning machine (KELM) and five optimization algorithms including butterfly optimization algorithm (BOA), tunicate swarm algorithm (TSA), multi-verse optimizer (MVO), golden jackal optimization (GJO), Harris hawk's optimization (HHO). The third objective (ore grade prediction by televiewer images) used support vector machine and salp swarm algorithm. The fourth objective (ore grade prediction by smartphone images) used five clustering techniques, i.e., k-means Cluster (K-means), Agglomerative Hierarchical Cluster Tree (AHCF), Gaussian Mixture Model (GMM), Self-Organizing Map (SOM) and Spectral clustering (SC). The fifth objective employed classification techniques, namely Support Vector Classification (SVC), Random Forest (RF), Bagged Tree (BT), Gradient Boost Machine (GBM), Extreme Gradient Boosting (XGB) and Light Gradient Boosting Machine (LGBM). Meanwhile, Bayesian optimization (BY) is applied.

1.3.1 Regression techniques

v-SVR and e-SVR

SVM is a powerful machine learning technique developed by (Vapnik, 2000) to tackle classification problems based on mathematical statistics. SVM can solve regression problems by introducing the ε-insensitive loss function (Cherkassky and Ma, 2004). For the SVR, it can be described as follows.

Suppose that a group of datasets are represented by $\{(x_1, y_1), (x_a, y_a), \dots, (x_{AA}, y_{AA})\}$, where a represents the number of training data, AA denotes the total number of training data, $x \in R^N$ represents that the dimension of input variables is N and $y \in R^1$ indicates an output. If the SVR is used for handling nonlinear regression problems, then it can be written as

$$f(x) = [\boldsymbol{\omega} \cdot \boldsymbol{\theta}(x)] + b \tag{1.1}$$

where $\boldsymbol{\omega}$ is the weight coefficient, $\boldsymbol{\theta}$ is the coefficient used for transforming the nonlinear problem into a linear problem, and b is the model error.

The v-SVR is another branch of SVM (Chang and Lin, 2002; Schölkopf et al., 2000; Thomas et al., 2017). For the v-SVR, there is one important parameter, i.e., v, with the range of [0, 1]. This parameter is used for balancing the number of support vectors and model errors. The primary task of v-SVR is to minimize the Eq. (1.2) to make it subject to the Eq. (1.3):

$$\frac{1}{2}\|w\|^2 + C\left[v\varepsilon + \frac{1}{AA}\sum_{a=1}^{AA}(\xi_a + \xi_a^*)\right] \tag{1.2}$$

$$\begin{cases} [\boldsymbol{\omega}\cdot\boldsymbol{\theta}(x_a)+b]-y_a \le \varepsilon+\xi_a \\ y_a-[\boldsymbol{\omega}\cdot\boldsymbol{\theta}(x_a)+b] \le \varepsilon+\xi_a^* \\ \xi_a,\xi_a^* \ge 0, a=1,2,\cdots,AA, \varepsilon \ge 0 \end{cases} \tag{1.3}$$

where $\|w\|$ is the norm of $\boldsymbol{\omega}$, ξ_a and ξ_a^* represent two positive slack variables and they are used for controlling the distance between the actual values and boundary values, and C is a regularization parameter used for balancing the model errors and model flatness. The ε-insensitive loss function is used for determining if the value of $\boldsymbol{\omega}\cdot\boldsymbol{\theta}(x_a)$ is located in the range of $y \pm \varepsilon$, then the calculation loss can be ignored.

While in the e-SVR (Chang and Lin, 2002), the aforementioned formulation is to minimize the equation (1.4) and make it subject to the equation (1.3).

$$\frac{1}{2}\|w\|^2 + C\frac{1}{AA}\sum_{a=1}^{AA}(\xi_a + \xi_a^*) \tag{1.4}$$

Given the difficulty of selecting ε, (Schölkopf et al., 2000) proposed a new parameter v which can control training errors by controlling the number of support vectors. Therefore, the dual formulations for v-SVR can be described as follows:

$$\begin{cases} minimize: \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T\boldsymbol{HK}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)+\boldsymbol{y}^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) \\ \boldsymbol{EV}^{\mathrm{T}}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)=0, \boldsymbol{EV}^{\mathrm{T}}(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*) \le Cv \\ 0 \le \alpha_a, \alpha_a^* \le \frac{C}{AA}, a=1,2,\cdots,AA \end{cases} \tag{1.5}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$ are the Lagrangian multipliers, $\boldsymbol{HK}=\boldsymbol{\theta}^{\mathrm{T}}(x_a)\boldsymbol{\theta}(x_b)$ is the kernel, $\boldsymbol{EV}$ represents all the vectors, and α_a and α_a^* are the corresponding Lagrangian multipliers of the a_{th} training data.

While in the e-SVR, this formulation is written as:

$$\begin{cases} minimize: \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T\boldsymbol{HK}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)+\boldsymbol{y}^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)+\varepsilon\boldsymbol{EV}^{\mathrm{T}}(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*) \\ \boldsymbol{EV}^{\mathrm{T}}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)=0 \\ 0 \le \alpha_a, \alpha_a^* \le \frac{C}{AA}, a=1,2,\cdots,AA \end{cases} \tag{1.6}$$

Then, the regression function can be approximately described as

$$f(x) = \sum_{a=1}^{AA}(\alpha_a^*-\alpha_a)\,\boldsymbol{\theta}^{\mathrm{T}}(x_a)\boldsymbol{\theta}(x)+b \tag{1.7}$$

In this study, the radial basis function (RBF) kernel is employed and written as

$$k(x_a,x_b) = \exp(-gamma\|x_a-x_b\|^2) \tag{1.8}$$

where *gamma* defines the bandwidth of the RBF.

There are two parameters needed to be selected to control the model performance in the RBF kernel function, i.e. penalty factor *Ct* and RBF kernel deviation *g*. The penalty factor *Ct* is used to balance the model complexity and biasness. A higher *Ct* would probably bring better model fitting while being easier to get stuck into "over-fitting". A smaller *Ct* would decrease the model complexity but also weaken the model performance. The RBF kernel deviation *g* implies the space distribution of mapped data. Generally, larger *g* needs fewer support vectors, while smaller *g* would require more support vectors. The number of support vectors determines the speed of the model development. Two typical data redistributions by hyperplane in SVR are demonstrated in *Figure 1* and *Figure 2*.

Manual selection of *Ct* and *g* costs a long time and thus it is not realistic to obtain the best parameters. Therefore, in this study, optimal algorithms are utilized to opt and optimize these two crucial parameters. Currently, many meta-heuristic algorithms, such as moth flame optimization (Zhou et al., 2021e), equilibrium optimizer algorithm (Yu et al., 2021a) and whale optimization algorithm (Qiu et al., 2022), have been widely used to obtain satisfactory outcomes. In the current study, five types of optimization algorithms, i.e. GS, GA, PSO, SSA and GWO, are chosen to conduct parametric optimization. They are combined with SVR techniques and used to develop different prediction models.

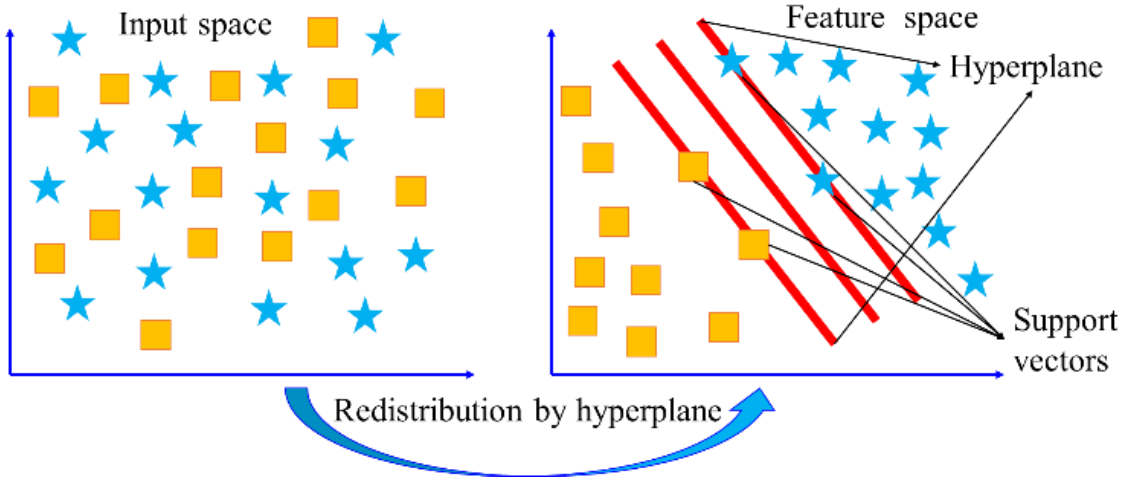


Figure 1. Data redistribution by hyperplane in two-dimensional space.

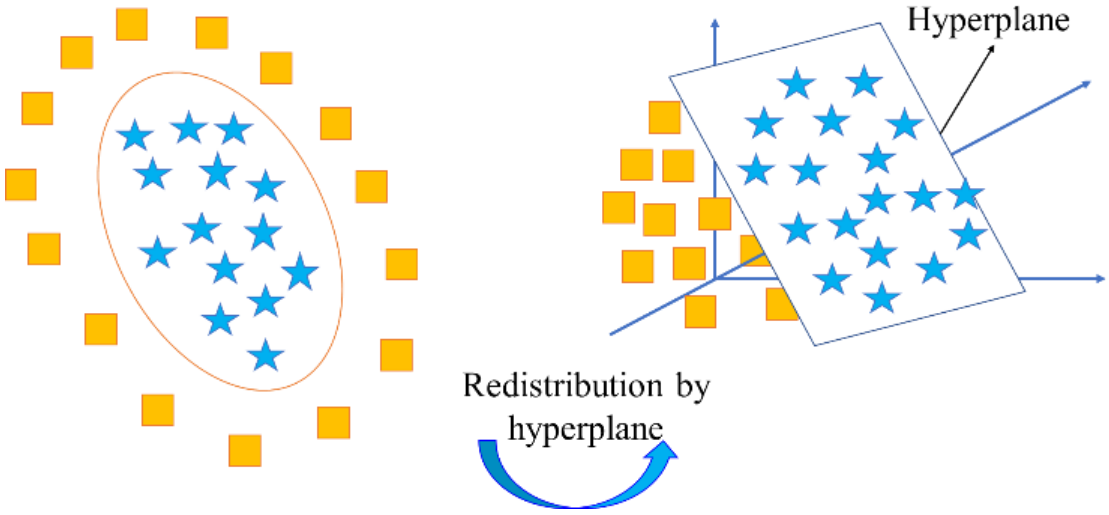


Figure 2. Data redistribution by hyperplane in three-dimensional space.

KELM

KELM is a new learning technique that has emerged in recent years. It evolved from extreme learning machine (ELM), and has been moderately proven to exhibit more honourable generalization in a wide range of practical applications (Wang et al., 2017). ELM is first proposed by Huang et al. (2006) and overcome the limitations of conventional neural network training, including extended training duration and challenges in model tuning. The fundamental concept behind ELM involves assigning random values to the weights and biases of the hidden layer, followed by a single forward propagation step to map the input data to the output space. As a result, ELM offers rapid training and ease of tuning. Nevertheless, when compared to alternative machine learning algorithms, ELM may exhibit reduced interpretability, potentially necessitate a greater number of hidden layer nodes for optimal performance, and occasionally encounter overfitting challenges. To overcome these shortcomings and improve its capability, Huang et al. (2012) proposed kernel-based ELM, i.e., KELM. The fundamental concept of KELM revolves around mapping input data from its original space to a feature space of higher dimensions, aiming to provide a more comprehensive representation of the data's inherent structure and interrelationships. This mapping process is facilitated by kernel functions $K(x, x_n)$ as demonstrated in *Figure 3*, which transform data points from the original space to the feature space while considering their similarity within the feature space. KELM is faster to train because it utilizes a pseudo-inverse matrix or a generalized inverse matrix to compute the output weights (Li and Wang, 2022).

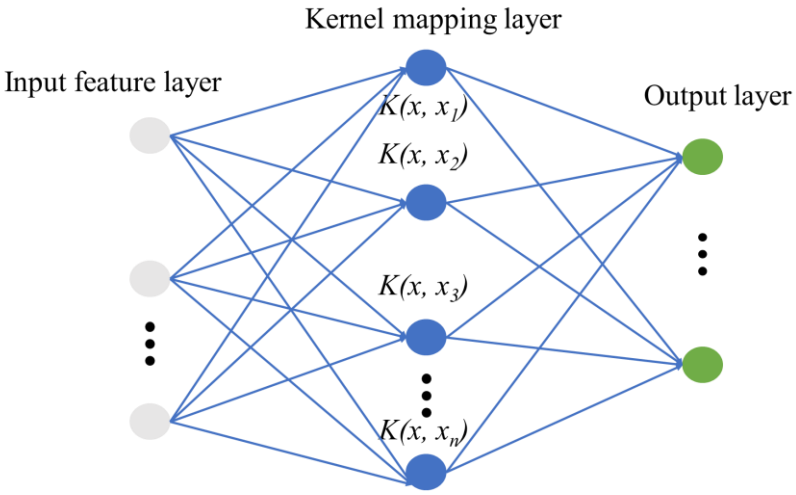


Figure 3. Typical network structure of KELM model.

Different kernel functions have distinct sets of hyper-parameters that influence the KELM behaviour. In this study, the Radial Basis Function kernel was used in this study and there are two significant hyper-parameters in this kernel, i.e., γ and *C* where γ determines the influence of an individual training sample and *C* represents the penalty coefficient which indicates the tolerance for relative errors (Liao et al., 2020). Selecting the appropriate hyper-parameter values in KELM is crucial for capturing the desired characteristics of the data and achieving optimal separation or representation in the transformed feature space. In this study, five meta-heuristic algorithms were employed to carefully choose the hyper-parameters γ and *C* in KELM, ensuring the best regression performance.

1.3.2 Classification techniques

SVC

SVC is a powerful supervised machine learning technique used for classification tasks (Cortes and Vapnik, 1995). It works by finding the optimal hyperplane that best separates data points of different classes in a high-dimensional space. The core idea behind SVC is to maximize the margin between the data points of different categories, ensuring that the model generalizes well to unseen data. SVC can handle both linear and non-linear classification problems using kernel functions. In this study, the radial basis function kernel is employed (Li et al., 2021; Li et al., 2023a).

BT

BT is a bootstrap aggregated ensemble of fine decision trees (Sutton, 2005). Decision trees divide the feature space into smaller regions and make predictions by traversing the tree from the root to the leaf nodes based on feature variants. Bootstrap sampling creates multiple new datasets by randomly selecting samples from the original dataset and allows repeated sampling. A decision tree is built on each of these bootstrapped datasets, resulting in multiple trees. Each tree is trained independently and may have different decision boundaries or predictive patterns due to the variation in the training data. Once the ensemble of trees is constructed, predictions are made by aggregating the individual predictions from each tree. For classification tasks, a majority vote among the trees is used to determine the final prediction; this involves creating multiple variations of a decision tree through bootstrapping and aggregating their predictions in order to make more reliable and accurate predictions.

RF

RF is an ensemble learning method used for classification and regression tasks (Breiman, 2001). It operates by constructing multiple decision trees and outputting the class (for classification) of the individual trees. The core idea is to improve the model's accuracy and robustness by combining the predictions of many trees, reducing overfitting and improving generalization. Each tree in the forest is built using a random subset of the data and features, making the model less sensitive to noise and variations.

GBM

GBM is a powerful ensemble learning technique used for both classification and regression tasks (Natekin and Knoll, 2013). It builds models in a sequential manner, where each new model attempts to correct the errors made by the previous ones. By combining the predictions of multiple weak learners, typically decision trees, GBM gradually improves the model accuracy. The "gradient" aspect refers to the use of gradient descent to minimize the loss function, ensuring each new tree focuses on the areas where the model is underperforming.

XGB

XGB is a high-performance implementation of the Gradient Boosting algorithm (Li et al., 2023b), designed for efficiency, speed, and scalability. It enhances traditional Gradient Boosting by incorporating regularization, which helps control overfitting, and optimizations like parallel processing, handling sparse data, and advanced tree-pruning techniques.

LGBM

LGBM is an advanced, fast and efficient implementation of the Gradient Boosting algorithm designed for large-scale data processing (Xi et al., 2024). It builds decision trees in a leaf-wise manner, focusing on leaves with the highest potential for error reduction, making it much faster and more memory-efficient than traditional boosting methods. LGBM also supports parallel and distributed learning, making it suitable for handling massive datasets with high-dimensional features.

1.3.3 Clustering techniques

K-means

The k-means cluster method, also named Lloyd algorithm (Lloyd, 1982), is an unsupervised machine learning algorithm used for clustering data into different groups based on similarities in the data. It aims to divide a dataset into *k* clusters, the *k* value to be defined a priori. The primary goal of the k-means algorithm is to minimize the within-cluster variance, which measures the sum of squared distances between each data point and the centroid of the cluster it belongs to. The L1 distance (i.e. the sum of absolute differences) is used to evaluate the separation between points; under this distance the clusters centroids are the medians of the PCs of the corresponding pellets. The k-means++ algorithm is used to define the initial cluster centroid positions and the clustering is repeated ten times using new initial cluster seeds to ensure that a global minimum is reached. In this study, it was chosen as a baseline due to its widespread use in various clustering applications.

AHCF

Agglomerative hierarchical clustering is a bottom-up clustering method that builds a hierarchy of clusters by progressively merging smaller clusters into larger ones (Day and Edelsbrunner, 1984). In the initialization stage, each data point starts as its own singleton cluster (i.e., a cluster containing only one data point). In the iteration stage, the two clusters that are the most similar (according to L1 distance) are merged to form a new cluster. And the distances between the new cluster and all other existing clusters are recalculated. This merge and update process is repeated until all data points are grouped into a single cluster, forming a tree-like structure called a dendrogram. The dendrogram represents the nested grouping of data points and the order in which clusters are merged. By cutting the dendrogram at a desired level, a specific number of clusters can be selected. In this study, it is selected to explore the potential existence of hierarchical relationships among the samples.

GMM

The Gaussian Mixture Model (GMM) assumes that the dataset is generated from a mixture of several Gaussian distributions (also known as components) with unknown parameters (Yang et al., 2012). Unlike k-means, which assigns each data point to a single cluster, GMM provides a probability distribution over clusters for each data point, allowing for more nuanced cluster assignments. Each Gaussian component has its own mean vector, covariance matrix, and a mixing coefficient where indicates the proportion of the data points in that component. By employing Expectation-Maximization Algorithm (Moon, 1996), the parameters (means, covariances, and mixing coefficients) would be given initial estimates. Then he probability that each Gaussian component takes for each data point would be calculated. And then, the parameters can be updated using the probabilities calculated. Excepted from the initial stage, the process would be repeated until convergence, typically when the change in the log-likelihood of the data given by the model parameters falls below a threshold. In this study, the number of times to repeat Expectation Maximization Algorithm is defined as one after trial and error. GMM is more flexible than k-means in capturing elliptical and overlapping clusters.

SOM

The Self-Organizing Map (SOM) is an artificial neural network that produces a low-dimensional, representation of the input space of the training samples (Vesanto and Alhoniemi, 2000). The SOM consists of a grid of neurons, each associated with a weight vector of the same dimensionality as the input data. In the initialization stage, the weight vectors of the neurons are generated by random sampling from the input data (known also input vector). The neuron whose weight vector is closest to the input vector can be found using the L1 distance. For the input vector, the distance between it and the weight vectors of all neurons is calculated, and the neuron with the smallest distance is selected as the best matching unit (BMU). The adjustment is done to update the weight vectors and their neighboring neurons to become closer to the input vector. The inputs, BMU identification, and weight updating steps are repeated until convergence. After training, the weight vectors of the neurons form a topological map that preserves the spatial relationships of the input data. Each neuron represents a cluster, input vectors with the

similar distance to a neuron are mapped to the same neuron. SOM is particularly useful in identifying nonlinear data patterns.

SC

Spectral clustering uses the properties of graphs to discover the intrinsic structure of data (Von Luxburg, 2007). Unlike traditional clustering methods such as k-means, which rely on geometric distances in the data space, spectral clustering leverages the relationships between data points to form clusters. First, spectral clustering starts by constructing a graph where each data point is represented as a node. The edges between the nodes indicate the similarity between the data points. To determine such similarity, connecting nodes that are close in distance (e.g., within a certain radius) or connecting each node to its nearest neighbors are measured. The strength of the connection (edge weight) reflects the degree of similarity between the nodes. Once the similarity graph is built, the next step involves creating the Laplacian matrix (Merris, 1994) summarizes how each node is connected to others in the graph. The Laplacian matrix is used to identifying regions of the graph where nodes are more densely connected to each other than to the rest of the graph, indicating potential clusters areas. Spectral clustering then involves computing eigenvectors from the Laplacian matrix. These eigenvectors provide a new representation of the data points, highlighting their connectivity in the graph. By focusing on the most significant eigenvectors, the data is transformed into a lower-dimensional space where the structure of the clusters becomes more apparent. In this reduced-dimensional space defined by the eigenvectors, a conventional clustering algorithm, such as k-means, is applied. This step clusters the data points based on their new representations, which reflect their connectivity in the original similarity graph. Finally, each data point is assigned to one of the clusters identified in the previous step. These clusters correspond to the groups of nodes that were closely connected in the original similarity graph. SC is effective in identifying complex cluster structures, particularly when clusters are not well-separated in the original feature space.

1.3.4 Optimization algorithms

PSO

PSO is a powerful method for solving optimization problems among many evolutionary search methods, prompted by simulating fish schools and bird schools (Kennedy and Eberhart, 1995). In PSO, a herd of birds represents a group of particles, and a food source represents a functional purpose. By sharing and transmitting information about the distance between the bird herds and the food source, the location of the food source can be determined by bird schools. This cooperation allows the entire herd of birds to select the best information about the location of the food source and eventually gather around the food. By implementing these steps, the best food source can be found. In PSO, the particle numbers are initialized by PSO and each particle possesses the same probability of being selected as a candidate solution for the defined problem. In the next step, two crucial properties of each particle need to be determined carefully, i.e. updated speed (V) and iterative position (X) (Poli et al., 2007). The goodness of each particle is appraised by the fitness function and the positions of particle swarms are updated based on the evaluation results of the fitness function. By iterations, particle swarms reach the best position according to the pre-defined target function by users.

Some important parameters in PSO are updated as follows, by which the new position and velocity can be determined:

$$\left.\begin{array}{c} V^{t+1} = \omega V^t + c_1 rand(A)(P_{\text{best}} - X^t) + c_2 rand(B)(G_{\text{best}} - X^t) \\ X^{t+1} = X^t + V^{t+1} \end{array}\right\} \quad (1.9)$$

where t represents the current iteration number; rand(A) and rand(B) signify the random number in the range of (0, 1); P_{best} and G_{best} denote the best position of the single-particle and the whole particle swarm, respectively; c_1 and c_2 are the constants which control the acceleration of particles (Zhou et al., 2013); ω denotes an inertia weight which determines the balance between global optimization and local optimization (Shi and Eberhart, 1998). Generally, the value of ω will decrease with the iteration. It can be determined as follows:

$$\omega^{t+1} = \omega^{max} - \frac{\omega^{max} - \omega^{min}}{Iteration_{max}} t \quad (1.10)$$

where ω^{max} is the maximum inertia weight, $\omega^{min}$is the minimum inertia weight, and $Iteration_{\text{max}}$ is the maximum iteration.

In SVR, the objective of the PSO is to optimize two important hyper-parameters *Ct* and *g*. By updating the velocity and position of the whole particle swarm, the particle swarm will eventually achieve global optimization. The general process of PSO optimization can be interpreted in *Figure 4*.

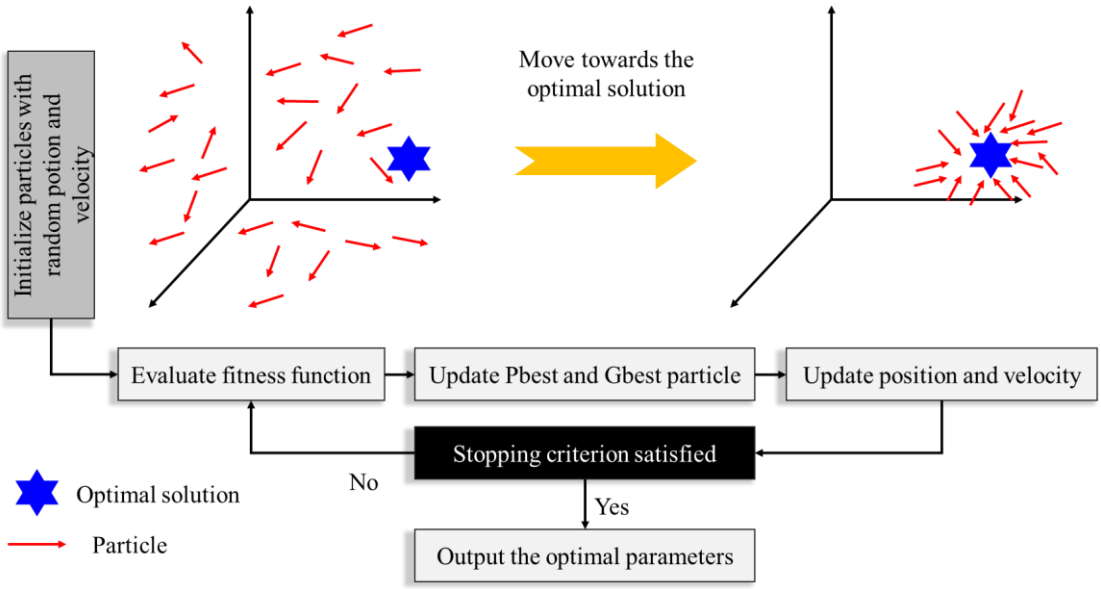


Figure 4. General process of PSO.

GA

GA was first proposed and developed by (Holland, 1992). Inspired by Darwin's theory of evolution, GA reckons that the strongest individual gets more opportunities to reproduce, while undesirable individuals cannot obtain the right to spawn.

In GA, each chromosome is encoded as a candidate solution for a re-defined problem. GA can conduct global optimizations and by implementing the adaptive search process, a better optimal solution replaces the former one, as like other meta-heuristic algorithms.

Generally, there are three steps in GA, i.e. chromosome initialization, ranking and selecting, mating and mutation.

In the beginning, GA randomly produces some chromosomes (individuals) and these individuals constitute a swarm. This step is so-called chromosome initialization. Next, the re-defined function judges the fitness of these individuals and gives corresponding scores. The individual with higher fitness will have the qualification to reproduce, while undesirable individuals will be eliminated. In the mating and mutation stage, those desirable chromosomes will change their genes to generate new chromosomes. Sometimes the mutation will occur to bring some fresh genes to join the reproduction process. The mutation probably is good for reproduction or may not, but it can prevent local convergence. By iteration, the evolutionary process tends to move to the best optimal solution. General working architecture of GA can be found in *Figure 5*.

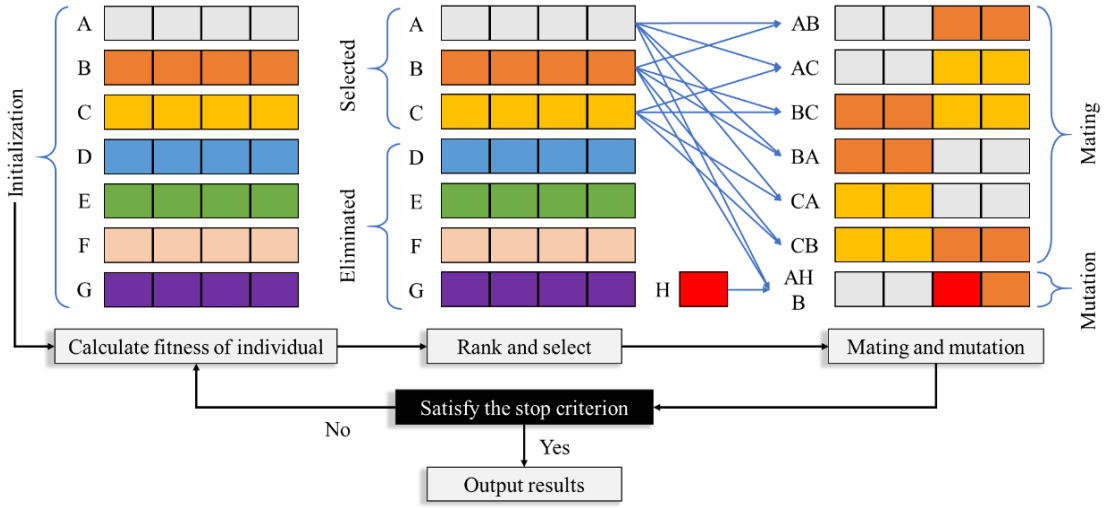


Figure 5. General process of GA.

SSA

Like the above two mentioned algorithms, SSA belongs to a group of swarm intelligence-based optimization algorithms (Mirjalili et al., 2017). It was inspired by the hunting and navigation behaviors of salp swarms. A salp is a kind of marine mollusk with a similar appearance to a jellyfish. They move and hunt together by connecting like social animals. In the salp chains, there is a leader salp and a host of follower salps. The leader moves towards the direction of the food source and leads the movement of followers. This behavior can be regarded as global optimization and the followers explore the food in the local space. This behavior can be regarded as a local optimization. With the cooperation of leader salp and follower salps, falling into local optimum can be largely reduced.

In SSA, there are mainly three steps to optimize the hyper-parameters, i.e. salp swarm initialization, update of leader and update of followers. In the initialization process, the number and position of salp swarm will be defined. In addition, the best food source (the fitness function) will be determined according to the user requirement. In the next step, the optimization process starts in the search domain and this search domain can be represented by a matrix named S_i:

$$S_I = \begin{bmatrix} s_1^1 & s_2^1 & \cdots & s_n^1 \\ s_1^2 & s_2^2 & \cdots & s_n^2 \\ \vdots & \vdots & \cdots & \vdots \\ s_1^m & s_2^m & \cdots & s_n^m \end{bmatrix} \tag{1.11}$$

In this search space, the leader salp will go into action at first so that it can provide the guide for follower salps. The update equation of the position of leader salp is shown as

$$x_m^1 = \begin{cases} F_m + R1\big((ub_m - lb_m)\big)R2 + lb_m, & R3 \geq 0.5 \\ F_m - R1\big((ub_m - lb_m)\big)R2 + lb_m, & R3 < 0.5 \end{cases} \tag{1.12}$$

where x_m^1 represents the position of the leader salp (first salp) in the m_{th} dimension of domain (note that for this case, m=2); F_m denotes the location of the food source of the m_{th} dimension; ub_m and lb_m are the searching upper and lower bounds, respectively; *R2* and *R3* are random variables uniformly distributed in the interval [0,1], and *R1* is calculated from the current iteration number (*lp*) and the total number of iterations (*LP*) as follows:

$$R1 = 2e^{-\left(\frac{4lp}{LP}\right)^2} \tag{1.13}$$

From Eq. (2.13), it can be found that with the increase in iteration, parameter *R1* will decrease.

The position of the i_{th} follower salp in the m_{th} dimension is updated to search the food source in the local range as follows:

For salp followers, they move under the guidance of a salp leader. The new location of salp followers obeys the following equation:

$$x_m^i = \frac{1}{2}\left(x_m^i + x_m^{i-1}\right), i > 1 \tag{1.14}$$

The aforementioned steps will be operated recursively before the optimization reaches the stopping criterion. The optimal process of SSA is shown in *Figure 6*.

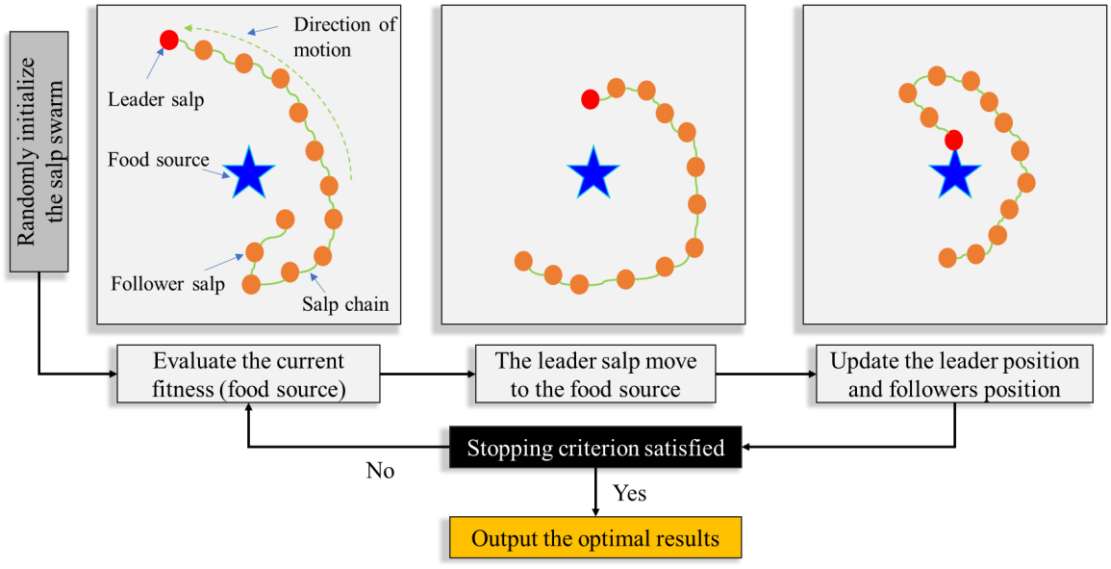


Figure 6. General process of SSA.

GWO

GWO is a kind of novel swarm-based algorithm introduced by (Mirjalili et al., 2014). GWO forms a strict hierarchy structure inspired by the hunting and social behavior of wolves, where different wolves play different roles in a grey wolf group. Every wolf is dominated by the harsh social order. Basically, there are four types of wolves involving:

(1) The alpha wolf (α) may not be the strongest wolf but it plays the manager role in a whole wolf herd. It is responsible for commanding and all other wolves would submit the order from the alpha wolf.

(2) The beta wolves (β) mainly assist the alpha wolves to make decisions and provide advice. They make the action around the alpha wolves and the command from the alpha wolf would be transmitted from the beta wolves to the delta wolves and omega wolves.

(3) The delta wolves (δ) would implement more trivial work like guarding, hunting and caretaking. They obey the order of the alpha wolves and beta wolves.

(4) The omega wolves (ω) are the lowest level of a wolf herd. They have to submit to all other wolves. However, they are still important because without them the wolf herd would have internal problems such as cannibalism.

In the implementation of GWO, the wolf herd updates its distribution and location according to the evaluation results of potential prey coordinates. The wolves make corresponding decisions and approach the prey step by step by iterations. Basically, the social behavior of grey wolves can be divided into three stages: social hierarchy, encircling and hunting.

(1) At the first stage, the hierarchy structure would be established. The grey wolves with the best fitness in the swarm would be labelled as “alpha”, “beta”, “delta” in order. While the rest wolves are regarded as “omega” wolves. During the iteration process, evolution would be directed by the aforementioned three wolves.

(2) The second stage is encircling prey. At this stage, wolves approach and encircle prey and this process can be formulated as

$$\vec{D} = \left|\vec{C} \cdot \vec{X}_{\mathrm{P}}(t) - \vec{X}(t)\right| \tag{1.15}$$

$$\vec{X}(t+1) = \vec{X}_{\mathrm{p}}(t) - \vec{A} \cdot \vec{D} \tag{1.16}$$

$$\vec{A} = 2\vec{a} \cdot \vec{r}_1 - \vec{a} \tag{1.17}$$

$$\vec{C} = 2\vec{r}_2 \tag{1.18}$$

where $\vec{X}_{\mathrm{P}}(t)$ and $\vec{X}(t)$ represent the position of the prey and grey wolf under the *t*-th iteration, respectively; *D* implicates the distance vector between the grey wolf and prey; $\vec{A}$ and $\vec{C}$ are the coefficient vectors; and $\vec{r}_1$ and $\vec{r}_2$ are the random vectors in the range of [0, 1] which control the swarm update. During the iteration process, $\vec{a}$ is a variable that can control the change of $\vec{A}$ and it decreases linearly from 2 to 0 with the increase in iteration.

(3) Hunting stage. We assume that in each iteration, three wolves with the best fitness would be retained, i.e. “alpha”, “beta” and “delta” wolves. These three wolves have the best ability to recognize the position of prey and the other wolves would update their positions. This behavior can be shown as follows:

$$\vec{D}_\alpha = \left|\vec{C}_1 \cdot \vec{X}_\alpha(t) - \vec{X}\right|, \vec{X}_1 = \vec{X}_\alpha(t) - \vec{A}_1 \cdot \vec{D}_\alpha \quad (1.19)$$

$$\vec{D}_\beta = \left|\vec{C}_2 \cdot \vec{X}_\beta(t) - \vec{X}\right|, \vec{X}_2 = \vec{X}_\beta(t) - \vec{A}_2 \cdot \vec{D}_\beta \quad (1.20)$$

$$\vec{D}_\delta = \left|\vec{C}_3 \cdot \vec{X}_\delta(t) - \vec{X}\right|, \vec{X}_3 = \vec{X}_\delta(t) - \vec{A}_3 \cdot \vec{D}_\delta \quad (1.21)$$

The best prey position can be procured by calculating the average of the prey position for "alpha", "beta" and "delta" wolves as follows:

$$\vec{X}(t+1) = \frac{1}{3}\left[\vec{X}_1(t) + \vec{X}_2(t) + \vec{X}_3(t)\right] \quad (1.22)$$

For global optimization, the "alpha", "beta" and "delta" wolves collect the food information respectively and transmit the food information to the whole wolf herd. By employing $|A| > 1$, the wolf individual can conduct a global search. As aforementioned, C is a random vector with the value of [0, 2] which can avoid local optimization. A general diagrammatic sketch of GWO is shown in *Figure 7*.

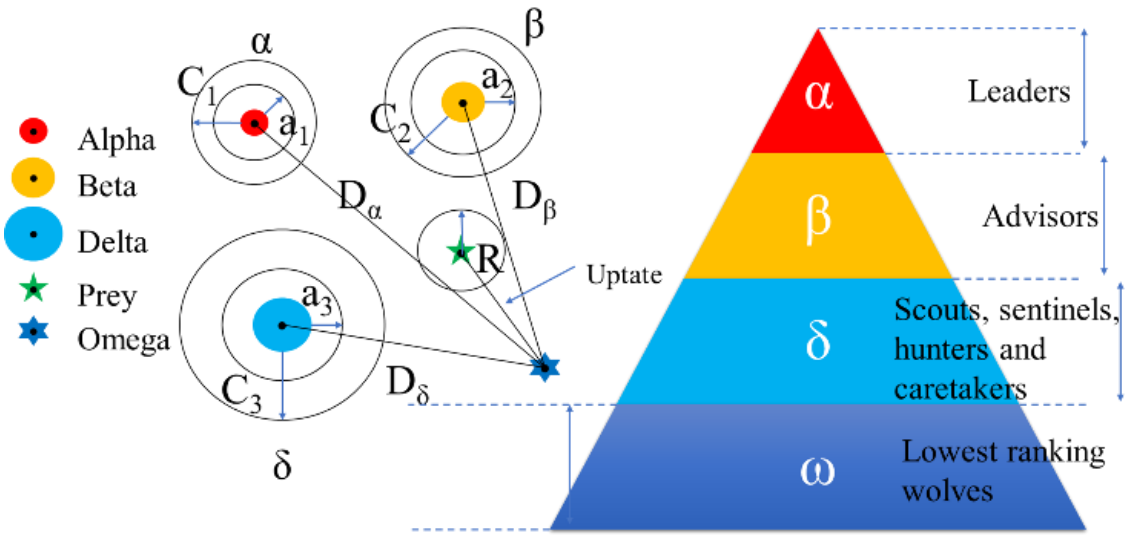


Figure 7. Diagrammatic sketch of GWO.

GS

The GS is a kind of classical parametric optimization method and has been proved to be effective to optimize the model (Zhou et al., 2012). The main idea of this method is to test all combinations of the given parameters and find the most suitable one according to the fitness function. While implementing the GS, the search range and search step need to be determined. A general process of GS optimization is demonstrated in *Figure 8*.

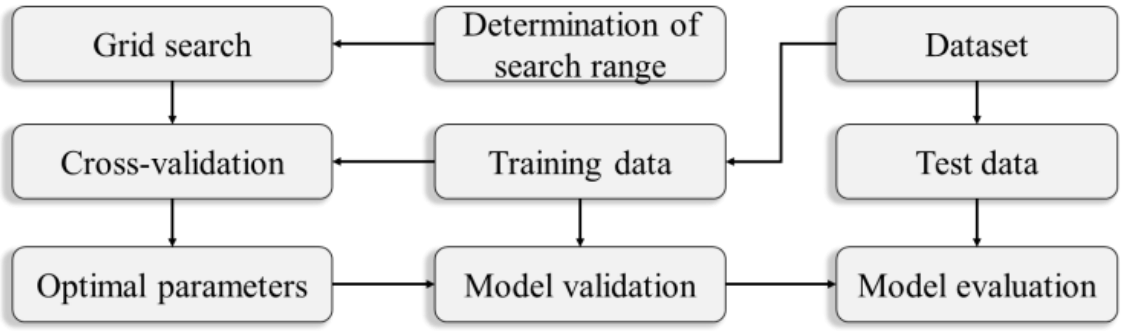


Figure 8. Work flow of GS optimization.

BOA

The BOA draws inspiration from the foraging and mate location behaviours exhibited by butterflies during fly (Arora and Singh, 2019). It has demonstrated efficacy in addressing diverse continuous optimization problems, displaying commendable capabilities in global search and convergence. The fundamental premise underlying the BOA involves emulating the collective behaviour of a population of butterflies, leveraging their communication and cooperative tendencies to identify optimal solutions. The algorithm employs a combination of global and local search strategies, enabling butterflies to engage in stochastic search patterns.

In BOA, the main simulation behaviour from butterflies' scent intensity *f* and it can be formulated as follows:

$$f = CI^a \tag{1.23}$$

where *I* is a user-defined function, *C* denotes the sensory modality and *a* is the power exponent. During the global search, butterflies tend to move towards individuals with stronger scent intensity, as described in Eq. (1.23). However, it is important to recognize that their search paths can be affected by the dynamic natural environment, which is prone to various unexpected events. Therefore, it is crucial to consider the impact of weather fluctuations on butterfly movement and the dispersal of scent signals. In localized search, butterflies ignore scent-based cues and instead adopt random movement patterns in the vicinity of other butterflies. Eqs. (1.24) and (1.25) provide a detailed explanation of this process.

$$y_j^{t+1} = y_j^t + \left(r^2 \cdot k_{opt} - y_j^t\right) \cdot f_j, r > p \tag{1.24}$$

$$y_j^{t+1} = y_j^t + \left(r^2 \cdot y_j^t - y_i^t\right) \cdot f_j, r < p \tag{1.25}$$

where y_j^{t+1} and y_j^t denote the location of jth butterfly during $t+1$ and t iteration. y_i^t is the location of ith butterfly during t iteration, r is a random number within the range of [0, 1], and k_{opt} represents the location of the butterfly with highest fitness. f_j is the scent intensity for j_{th} butterfly. p is used to control the switch probability between global search and local search.

MVO

The Multi-Universe Optimizer (MVO) algorithm draws inspiration from the intricate interactions between black holes, white holes and wormholes in the universe (Mirjalili et al., 2016). This unique algorithm treats each set of variables as a universe, considers each of them as a separate entity, and refines the problem solution space by modelling the gravitational and spacetime curvature of the universe. In this algorithm, by calculating the expansion rate or fitness function value for each universe, the algorithm finds and guides the movement of objects. The expansion rate of the universe is viewed as a defining feature, with high expansion rate white holes transporting their objects to lower expansion rate black holes, thus presenting a dynamic change in the entire cosmic system. This rapid cosmic variability increases the average population expansion rate, effectively enhancing the global search capability of MVOs. Every stage involves the selection of white holes using the next formulas (Eqs. 1.26-1.27):

$$T_n^i = \begin{cases} T_m^i & with\ \alpha < \alpha_{universe_n} \\ T_n^i & with\ \alpha \geq \alpha_{universe_n} \end{cases} \tag{1.26}$$

$$T_n^i = \begin{cases} \begin{cases} T_i + \beta \times (UB - LB) \times d + LB, & c < 0.5 \\ T_i - \beta \times (UB - LB) \times d + LB, & c \geq 0.5 \end{cases} & with\ b < \delta \\ T_n^i & with\ b \geq \delta \end{cases} \tag{1.27}$$

where T_n^i represents the n_{th} object of the i_{th} universe, $\alpha_{universe_n}$ represents the normalized inflation rate of the n_{th} universe, α represents a random value that ranges from 0 to 1, and T_m^i represents the i_{th} object for the m_{th} universe. b, c, and d are random values in the interval [0, 1], T_i represents the i_{th} centre of the best solution, UB and LB indicate the lower and upper boundaries, β indicates the traveling distance rate, and δ denotes the wormhole existence possibility (WEP). The WEP can be calculated according the function:

$$WEP = min + l \times (\frac{max - min}{L}) \tag{1.28}$$

where *min* and *max* denotes the minimum and maximum probability of wormhole existence, respectively, l and L represents the current iteration number and total iteration number, respectively.

The most effective solution is always retained in the MVO process and is applied to improve the other options. In this manner, the wormhole allows the perfect solution to share its knowledge with other solutions.

GJO

Inspired by the hunting behaviour of golden jackals, a kind of biology-inspired optimization algorithm was developed, named GJO (Chopra and Mohsin Ansari, 2022). The golden jackal is medium-sized predator with the similar appearance of wolf and fox. They are distributed in Africa, Europe and Asia. They hunt and rest together and present collaborative behaviours. Generally, there are three steps in their hunting activities. The first one is to search the prey. The second one is to enclose and irritate the prey and the final step is to pounce towards the prey. These steps can be mathematically modelled to

be a new algorithm, i.e., GJO. In the golden jackal group, female and male jackals participate in hunting together and their positions can be updated as follows:

$$P_1(t) = P_M(t) - E \cdot |P_M(t) - rl \cdot Prey(t)| \quad (1.29)$$

$$P_2(t) = P_{FM}(t) - E \cdot |P_{FM}(t) - rl \cdot Prey(t)| \quad (1.30)$$

where $Prey(t)$ represents the prey's coordinate at iteration *t*. $P_M(t)$ and$P_{FM}(t)$ denote the coordinate of male and female jackals at iteration *t*, respectively. The updated male and female jackal positions are $P_1(t)$ and$P_2(t)$. *rl* is a random number related to Levy movement (Auger-Méthé et al., 2011). E indicates the evading energy of prey and can be defined as:

$$E = E_1 * E_0 \quad (1.31)$$

where E_1 denotes the decreasing energy of the prey and E_0 represents the initial energy of the prey. They can be calculated as follows:

$$E_0 = 2 * r - 1 \quad (1.32)$$

$$E_1 = c1 * (1 - (t/T)) \quad (1.33)$$

where *r* is a random number in the range of [0, 1], *t* and *T* denotes the current and total iteration number, *c1* is a coefficient used for describing the energy decrease of the prey. And then the jackal positions can be obtained by averaging $P_1(t)$ and $P_2(t)$.

As the prey is chased by jackals, the jackal pair encloses the prey gradually. The following equations describe the actions of male and female jackals when pouncing and devouring the prey:

$$P_1(t) = P_M(t) - E \cdot |rl \cdot P_M(t) - Prey(t)| \quad (1.34)$$

$$P_1(t) = P_{FM}(t) - E \cdot |rl \cdot P_{FM}(t) - Prey(t)| \quad (1.35)$$

where $P_1(t)$ and $P_2(t)$ reflect the updated locations of the jackal pair while attacking the prey.

TSA

The TSA is inspired by the behaviour and characteristics of tunicates, a type of marine organism known for their self-organization and cooperative nature in achieving efficient feeding and survival (Kaur et al., 2020). TSA emulates the collective behaviour of tunicates by utilizing a population of virtual tunicates to solve optimization problems. Each virtual tunicate represents a potential solution within the search space, and the algorithm operates iteratively, allowing the virtual tunicates to interact with one another and adjust their positions based on their fitness values and information gathered from neighbouring tunicates. At the core of TSA lies the concept of "pharyngeal pumping," which refers to the filtration mechanism employed by tunicates to extract nutrients from the surrounding water. This concept is adapted into the optimization context, where virtual tunicates exchange information and assimilate knowledge from their neighbours to enhance their own solutions.

In the process of searching, the position of each tunicate can be indicated by a vector $\vec{A}$ and $\vec{A}$ is influenced by the gravity force, social force and water flow advection, and then $\vec{A}$ can be represented as:

$$\vec{A} = \frac{\vec{G}}{\vec{M}} = \frac{c_2 + c_3 - 2 \times c_1}{\lfloor P_{min} + c_1 \times P_{max} - P_{min} \rfloor} \tag{1.36}$$

where $\vec{G}$ denotes the gravity force, $\vec{M}$ denotes the social forces between different search individuals, c_1, c_2 and c_3 are random numbers in the range of [0, 1], P_{min} and P_{max} signify the initial and subordinate speeds of making social interaction. In the current work, the values of P_{min} and P_{max} are 1 and 4, however, more sensitivity analysis of these values can be seen in Kaur et al. (2020). The movement of searching individuals can be determined as:

$$\overrightarrow{PD} = \left| \overrightarrow{FS} - R \times \overrightarrow{P_P(x)} \right| \tag{1.37}$$

where $\overrightarrow{PD}$ is the distance between the optimal solution and searching individual, *x* represents the current iteration, $\overrightarrow{FS}$ and $\overrightarrow{P_P(x)}$ represents the position of optimal solution and searching individual, respectively. *R* is a random number in the range of [0, 1]. After the determination of the movement of searching tunicate, it will move towards the direction of a better solution. To maintain this kind of movement trend, the updated position of tunicate $\overrightarrow{P_P(x')}$ can be shown as:

$$\overrightarrow{P_P(x')} = \begin{cases} \overrightarrow{FS} + \vec{A} \times \overrightarrow{PD}, \ if \ R \geq 0.5 \\ \overrightarrow{FS} - \vec{A} \times \overrightarrow{PD}, \ if \ R \leq 0.5 \end{cases} \tag{1.38}$$

HHO

The HHA is an optimization algorithm that draws inspiration from the hunting behavior of Harris's hawks, a type of raptor found in the Americas (Heidari et al., 2019). HHA imitates the cooperative hunting strategies employed by these birds to efficiently search for and capture prey. In HHA, a population of virtual hawks is employed to solve optimization problems. Each virtual hawk represents a potential solution within the search space. Similar to the hunting behavior of Harris's hawks, the algorithm enables the virtual hawks to collaborate and communicate with each other to enhance their search efficiency. The central concept behind HHA is the cooperative hunting behavior observed in Harris's hawks. These birds work together, with some hawks serving as leaders and others as followers. The leaders guide the hunting process, while the followers assist in capturing prey. In the optimization context, this concept is adapted, and the virtual hawks exchange information and coordinate their actions to improve their solutions. Generally, three stages are involved in HHA, i.e., exploration phase, transition phase, exploitation phase. In the first phase, the position of a Harris hawk can be updated as:

$$\mathrm{X(t+1)} = \begin{cases} X_{hawk}(t) - r_1 |X_{hawk}(t) - 2 r_2 X(t)|, & q \geq 0.5 \\ X_{rabbit}(t) - X_m(t) - r_3 \big(LB + r_4 (ub - lb)\big), & q < 0.5 \end{cases} \tag{1.39}$$

where t represents the iteration number, $X_{hawk}(t)$ and $X_{rabbit}(t)$ denote the position of the prey and a random Harris hawk based on the current population, r_1, r_2, r_3, r_4 and q are random values in the range of [0, 1]. *ub* and *lb* defines the upper searching bound and

lower search bound, respectively. In addition, $X_m(t)$ is defined as the average coordinate position in a population of Harris hawks. In the transition stage, the escaping energy of prey changes according to the following function:

$$E(t) = FE_0(1 - \frac{t}{T}) \tag{1.40}$$

where E_0 and $E(t)$ denotes the initial escaping energy and the current escaping energy and T is the maximum iteration number, F represents the factor to show the decreasing energy of rabbit. In the exploitation stage, the Harris hawk attempts to approach and attach the prey and the prey will attempt to escape. According to the theory of HHA, there are four escaping strategies where a random parameter r and escaping energy E is used to determine the specific situation. Different situations would cause different movement and convergence behaviors. The detailed explanations can be seen in Yu et al. (2020b). To initiatively interpret the flow of meta-heuristic algorithms, aforementioned five optimization algorithms are presented in a chart as shown in *Figure 9*. In addition, the control parameters in these five optimization algorithms have been presented in *Table 1*.

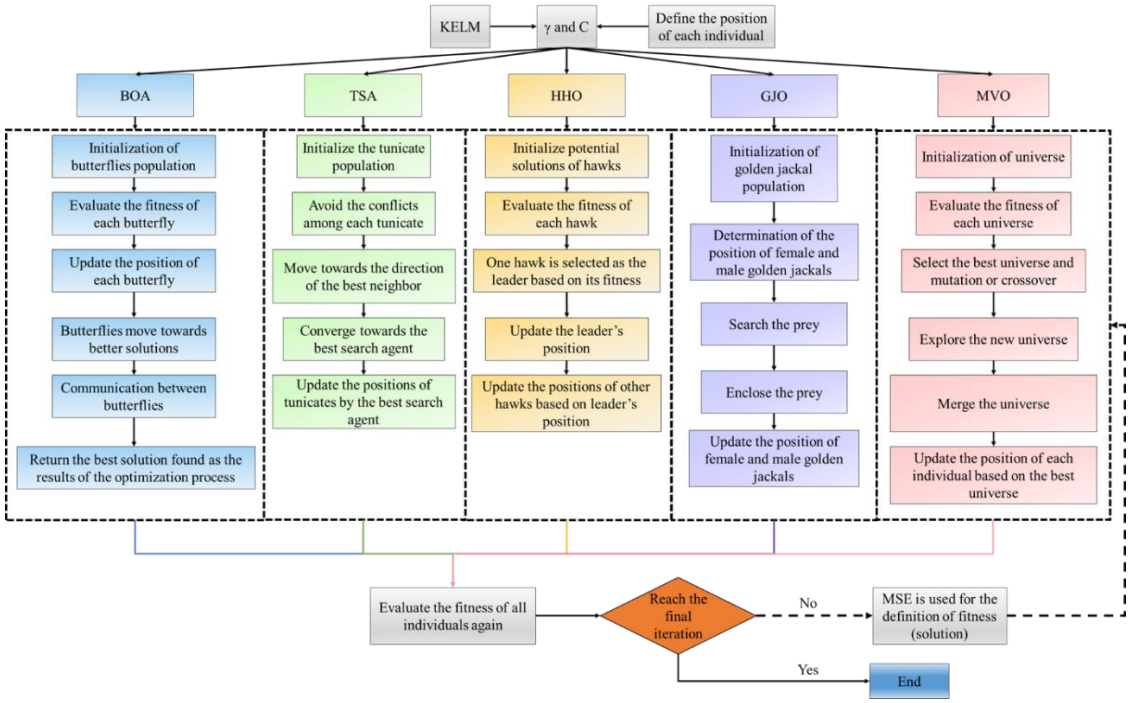


Figure 9. General flow of five meta-heuristic algorithms.

BY

Bayesian optimization algorithm is an optimization technique used to find the minimum or maximum of an objective function that is expensive to evaluate (Zhou et al., 2021). In this study, it is used for tuning the hyper-parameters in the aforementioned classification models. It builds a probabilistic model, typically using a Gaussian Process, to predict the function's behavior and guide the search towards promising areas (Zhou et al., 2021).

Table 1. Control parameters in different optimization algorithms.

Optimization algorithm	Parameters	Value
BOA	Power exponent	0.1
	Probability switch	0.8
	Sensory modality	0.1
MVO	Minimum of wormhole existence probability	0.2
	Maximum of wormhole existence probability	1
TSA	Initial speed of making social interaction	1
	Subordinate speeds of making social interaction	4
GJO	Coefficient used for the definition of the energy decrease of the prey	1.5
HHO	Factor to show the decreasing energy of rabbit	2

1.4 Thesis structure

The present dissertation is organized following a flowchart (Figure 10) that encompasses the key structure, process and achievements for the development of the predictive models for the mining-topics covered in this work. Figure 10 serves as the foundation upon which the subsequent chapters are structured, following a chronological order that reflects the finished work and the progress of the thesis.

At first, the blasting median fragment was predicted using a small-size (79) dataset. The successful application of support vector machine and several optimization algorithms indicate the strong fitting capabilities of machine learning techniques. Therefore, in the next, a larger gas permeability dataset (1024) was employed to examine the predictive abilities of machine learning algorithms. However, on the one hand, these two tasks were implemented based on the published data. On the other hand, the physical mechanism of blasting fragment and gas relative permeability is known more or less. To further validate the robustness and feasibility of machine learning, in the next chapters, two more complicated tasks were considered, i.e., the recognition of ore grade and lithology based on color properties. Three new methods were proposed. Initially, the borehole wall images were collected using a televiewer. The extracted color characteristics were used as inputs to classify the fluorite grade. Then a lower-cost way was proposed, i.e., the usage of smartphone to procure optical properties from fluorite pellets. The success of these two works has encouraged us to consider applying machine learning techniques to a broader range of image-related domains, i.e., lithology recognition. Finally, the borehole wall images obtained from an endoscope were extracted to reflect the lithology distribution. The significant color information was extracted as inputs to discriminate three different lithologies.

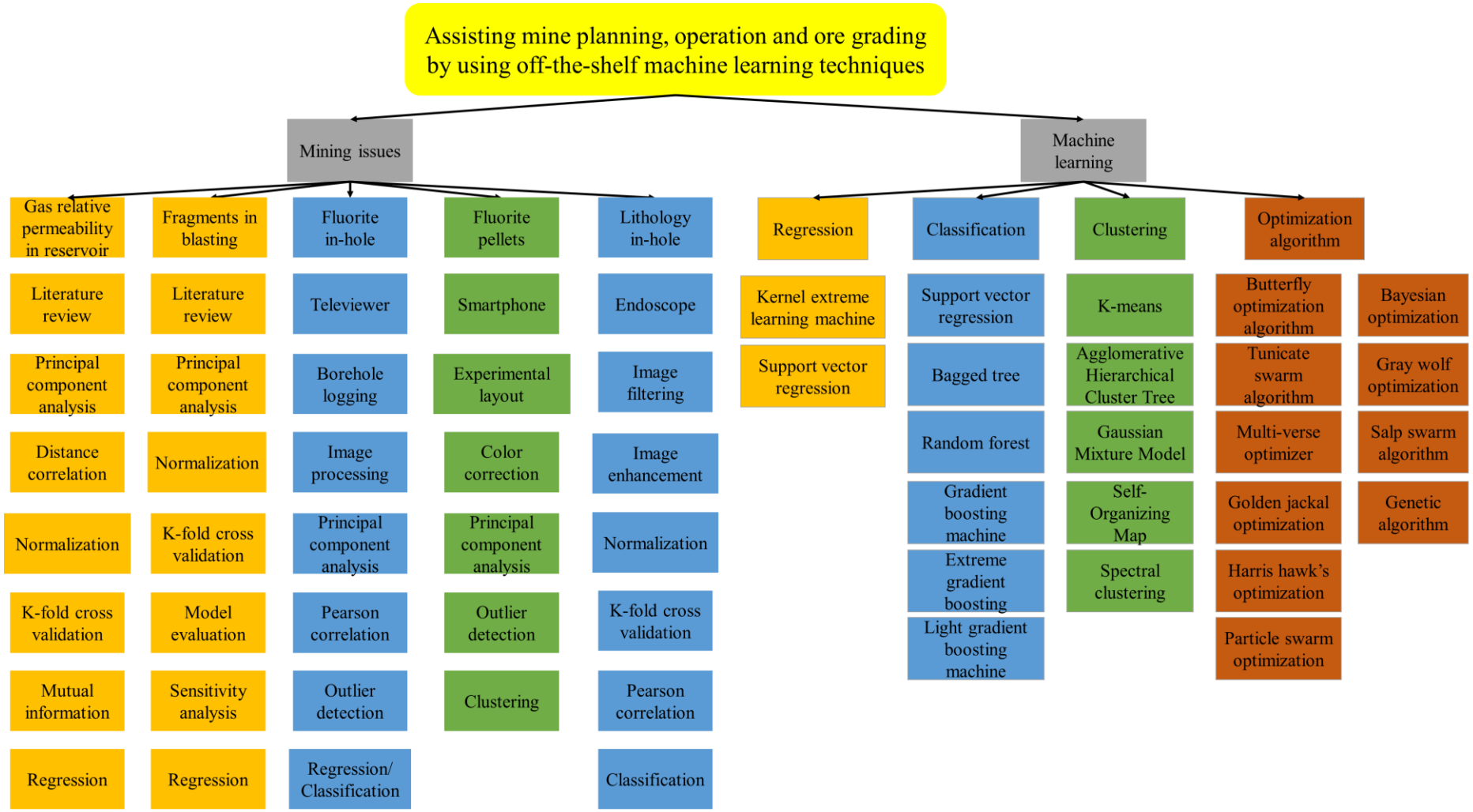


Figure 10. The general thesis structure.

The dissertation is composed of seven chapters and eight appendixes. The first chapter offers a comprehensive description of this dissertation with special emphasis in the dissemination and the papers that resulted from the work done.

Chapters 2 to 6 delve into the application of machine learning techniques to predict different stages of mining in three deposits, namely coal, gas, fluorite and limestone. They follow the same structure, that consists of an introduction, literature review, data analysis, model development, discussion, limitation and conclusion. These literature survey reviews state-of-the-art models, from classical models to artificial intelligence-based models, from traditional measurement methods and more sophisticated equipment. Their characteristics, advantages, and disadvantages are discussed to emphasize the significance of the current work. One the one hand, Chapter 2 and Chapter 3 can be considered as the groundwork of this dissertation where the datasets are obtained directly from the published work and they are also the complimentary for the existing work. From Chapters 4 to 6, new ideas are proposed and the data was collected from the field measurements and laboratory experiments.

Chapter 2 investigates median fragmentation in blasting through optimized support vector regression techniques. Two types of support vector regression and five hyper-parameters tuning techniques are compared. An overall evaluation system proposed by Zorlu et al. (2008) is used to assess the model performance. The sensitivity analysis is adopted to measure the importance of blasting parameters.

Chapter 3 describes a tool to predict the gas relative permeability in reservoir. The predictive capabilities of five optimization techniques are compared by an overall evaluation index named GI and Taylor diagram. The mutual information is employed to measure the impact of individual input variable on the model development. The most significant input is selected to model the gas relative permeability by kernel extreme learning machine-based model and traditional Corey-Brooks Model.

Chapter 4 introduces the process of sampling, borehole logging with televiewer and image processing in a fluorite underground mine. In the measurements, two types of light resources are used, i.e., white light and ultraviolet light. Pearson correlation and principal component analysis are used for extracting the significant image characteristics. As a result, the significant colour intensities under the illumination of different light sources are used as inputs to predict the amount of fluorite in the images of the blasthole walls. The fluorite grade is forecasted by the salp swarm algorithm and support vector machine. As a novelty, the occurrence of outliers during the sampling is investigated using a ‘take one out’ method.

Chapter 5 is a follow up of Chapter 4 in which the fluorite grade is assessed from pellet images of the drilling chips used for chemical analysis by mine. In this chapter, the production of fluorite pellets, the experimental layout and necessary equipment are introduced. The pellet images are taken by a smartphone and calibrated by a Colorchecker. Color intensities and texture parameters are used to develop the ore/waste discrimination model by five clustering techniques. The resulting model is helpful to reduce the amount of chemical analysis made by the mine.

Chapter 6 validates the use of in-hole images to investigate the lithology. For this, a low-cost tool like an endoscope is employed to map the lithologies in borehole images collected in a limestone quarry. This involves extracting the images from the videos, digitalizing the depths shown in the videos and processing the images using the Contrast-limited adaptive histogram equalization. Similarly, as in Chapter 5, some color intensities and texture properties are extracted and used as inputs. Six optimized machine learning techniques by Bayesian optimization is devoted to the development of and intelligent lithology recognition models.

Chapter 7 summarizes the main findings, the practical implications and future development directions. Additionally, eight appendixes are included after the 'References' chapter, they include supplementary information and the resulting papers in which the work has been disseminated.

The work described in this thesis is addressed in five journal papers (see Table 2). Two of them have been already published in Q1 journals and one has been published in a Q2 journal indexed in Journal Citation Report and the other two are under review. The link between each paper and the thesis objectives are shown in Table 3. It facilitates a better understanding of the research presented. Based on these activities, the contributions by author to each paper are shown in Table 4. The candidate has played a significant role in all of them, leading the research, analysis and the paper preparation.

Table 2. Literature generated in this thesis.

	Appendix	Type	Reference
Paper	**A**	J	Li, E., Yang, F., Ren, M., Zhang, X., Zhou, J., & Khandelwal, M. (2021). Prediction of blasting median[a] fragment size using support vector regression combined with five optimization algorithms. Journal of Rock Mechanics and Geotechnical Engineering, 13(6), 1380-1397. https://doi.org/10.1016/j.jrmge.2021.07.013
	B	J	Li, E., Zhang, N., Xi, B., Yu, Z., Fissha, Y., Taiwo, B. O., ... & Zhou, J. (2024). Analysis and modelling of gas relative permeability in reservoir by hybrid KELM methods. Earth Science Informatics, 1-28. https://doi.org/10.1007/s12145-024-01326-2
	C	J	Li, E., Segarra, P., Sanchidrián, J. A., Gómez, S., Fernández, A., Navarro, R., & Bernardini, M. (2023). Application of percentile color intensities of borehole images for automatic fluorite grade assessment. Ore Geology Reviews, 105790. https://doi.org/10.1016/j.oregeorev.2023.105790
	D	J	Li, E., Segarra, P., Sanchidrián, J. A., Gómez, S., Iglesias, L., Fernández, A., & Navarro, R. Fluorite ore recognition using spectral clustering and smartphone digital images calibrated with a ColorChecker: A case study at the Lujar underground mine, Spain. Minerals Engineering. (minor revision)
	E	J	Li, E., Catalán, I., Segarra, P., Ahmed, Z., Sanchidrián, J. A., Gómez, S., & Fernández, A. Lithology identification using borehole images by contrast-limited adaptive histogram equalization (CLAHE) and machine learning models. Journal of Rock Mechanics and Geotechnical Engineering. (under review)

[a]There was an error in the title that has been corrected

Table 3. Relevance of appended papers to thesis objectives.

	Paper				
	JCR Journal				
Objective	**A**	**B**	**C**	**D**	**E**
1-Fragments in blasting	X				
2-Gas relative permeability		X			
3-Ore grade			X	X	
4-Lithology					X

Table 4. Authors' contribution.

	Paper				
Authors	**A**	**B**	**C**	**D**	**E**
Ahmed, Zahir	-	-	-	-	2,3
Bernardini, Maurizio	-	-	2	-	-
Catalán, Ignacio	-	-	-	-	3,4
Fernández, Alberto	-	-	2	5	5
Gómez, Santiago	-	-	5	5	5
Iglesias, Luis	-	-	-	2	-
Li, Enming	1-5	1-5	1,3-5	1-5	1-5
Navarro, Rafael	-	-	2	2	-
Sanchidrián, José. A.	-	-	3,4,5	5	4,5
Segarra, Pablo	-	5	3,4,5	3,4,5	4,5

[1]Responsible for the work described in the paper.
[2]Collection of data.
[3]Analysis of data and results.
[4]Preparation of the manuscript.
[5]Revision and final approval of manuscripts.

Chapter 2. Prediction of blasting median fragment size using support vector regression

Nomenclature

ACO (Ant colony optimization)	MVRA (Multivariate regression analysis)
AI (Artificial intelligence)	PCA (Principal component analysis)
ANFIS (Adaptive network-based fuzzy inference system)	NH (Number of holes)
ANN (Artificial neural network)	NR (Number of rows)
B (Burden)	P (Joint persistency)
B.D (Burden/Hole diameter)	PI (Point load index)
BGAM (Boosted generalized additive model)	PSO (Particle swarm optimization)
BI (Blastability index)	QB (Quantity of blasted rock pile)
BPNN (Back propagation neural network)	HL (Hole length)
BRT (Boosted regression tree)	q (Specific charge or powder factor)
BS (Bench slope)	Q_e (Total explosives charge)
B.S (Burden/Spacing)	R^2 (Coefficient of determination)
CSO (Cat swarm optimization)	RBF (Radial Basis Function)
D (Hole diameter)	RMSE (Root mean square error)
De (Linear explosive density)	RMR (Rock mass rating)
DJ (Density of joint)	RQD (Rock quality designation)
DR (Ratio of total delays per number of rows)	RSE (Relative strength of effects)
E (Elastic modulus)	S (Spacing)
FFA (Firefly algorithm)	S.B (Spacing/Burden)
FIS (Fuzzy inference system)	Sch (Schmidt hammer rebound number)
GA (Genetic algorithm)	S/D (Spacing/Hole diameter)
GPR (Gaussian process regression)	SD (Specific drilling)
GS (Grid search)	SSA (Salp swarm algorithm)
GSI (Geological strength index)	ST (Stemming)
GWO (Grey wolf optimization)	ST.B (Stemming/Burden)
H (Bench height)	SVM (Support vector machine)
H.B (Stiffness ratio)	SVR (Support vector regression)
ICA (Imperialist competitive algorithm)	T (Tensile strength)
J.B (Average sub grade drilling/Burden)	TC (Total charge per delay)
JPO (Joint plane orientation)	TR (Delay between the rows)
JS (Joint spacing)	UCS (Uniaxial compressive strength)
L (Average length)	VAF (Variance accounted for)
L.Wd (Length/width)	Wd (Average width)
MC (Charge per delay)	WD (Water depth)
MH (Maximum holes per delay)	x_{50} (50% passing or median size)
MI (Mutual information)	X_B (In situ block size)
MR (Multiple regression)	ρ (Rock density)
MSE (Mean square error)	

2.1 Introduction

Blasting is widely used in many engineering operations including mining engineering, civil engineering and tunneling (Hu et al., 2020; Sanchidrián et al., 2007; Wang et al., 2018a, 2018b; Zhou et al., 2022a, 2022b). The primary purpose of the blasting operation is fragmentation and displacement of rock mass in mining engineering. Although significant developments have been achieved in explosive technology, only 20%-30% explosive energy is used for the actual fragmentation and displacement of rock mass (Ebrahimi et al., 2016; Khandelwal and Monjezi, 2013), and the rest of the energy is dissipated in the ground or air and produces various hazardous effects, such as backbreak (Esmaeili et al., 2014; Monjezi et al., 2012), ground vibration (Zhou et al., 2020) and flyrock (Armaghani et al., 2014). In mining operations, a blast pattern must be designed with due care to obtain optimum fragment size. Fragment size plays a crucial role in subsequent crushing and grinding operations. It is well-known that there is a tradeoff between fragment size and economic benefit. Large fragments need secondary blasting to reduce boulder size; however, finer fragments will increase the cost of mining due to higher explosive charge. Therefore, controlling and predicting the blast fragment size is important in mining operations. Median fragment size (x_{50}) is a crucial index that measures the goodness of blasting designs. Over the past decades, various models have been proposed to evaluate and predict blasting fragmentation. Among these models, artificial intelligence (AI)-based models are becoming more popular due to their outstanding prediction results for multi-influential factors. In this study, support vector regression (SVR) techniques are adopted as the basic prediction tools, and five types of optimization algorithms, i.e. grid search (GS), grey wolf optimization (GWO), particle swarm optimization (PSO), genetic algorithm (GA) and salp swarm algorithm (SSA), are implemented to improve the prediction performance and optimize the hyper-parameters. The prediction model involves 19 influential factors that constitute a comprehensive x_{50} evaluation system based on AI techniques. Three types of mathematical indices, i.e. mean square error (MSE), coefficient of determination (R^2) and variance accounted for (VAF), are utilized for evaluating the performance of different prediction models. Finally, sensitivity analysis is performed to understand the influence of input parameters on the median size.

2.2 Literature review

Blast fragmentation is influenced by a number of blast design parameters. It is difficult to propose a fitting equation considering all the influential factors simultaneously. In such complex circumstances, AI-based methods which can develop a complicated relationship among various input and output variables, attract the attention of scholars worldwide. Over the past few decades, AI-based methods have brought satisfactory prediction results in mining engineering and various geo-engineering fields (Gordan et al., 2016; Hajihassani et al., 2015; Moayedi and Jahed Armaghani, 2018; Zhou et al., 2016). For instance, Shi et al. (2012) applied a support vector machine (SVM) to assessing rock fragmentation distribution and compared their prediction abilities with artificial neural network (ANN), multivariate regression analysis (MVRA) and Kuznetsov methods. Gao et al. (2018) used Gaussian process regression (GPR) with five kernel functions to

propose various rock fragmentation prediction models. The GPR model performance was evaluated by the coefficient of determination (R^2) and also compared with other AI-based models. It was found that the GPR-squared exponential model showed the best prediction performance. Monjezi et al. (2009) applied a fuzzy inference system (FIS) to predicting rock fragmentation and compared their results with regression analysis. They found that FIS-based models performed much better than the regression models. Hasanipanah et al. (2018) explored the feasibility of particle swarm optimization-adaptive network-based FIS (PSO-ANFIS) technique to estimate rock fragmentation and compared the prediction capability of PSO-ANFIS model with SVM, ANFIS and nonlinear multiple regression (MR) models. They found that the PSO-ANFIS model outperformed other models and its prediction capability was assessed with R^2 and root mean square error (RMSE). Asl et al. (2018) adopted ANN and firefly algorithm (FFA) as prediction tools to optimize the flyrock distance and rock fragmentation distribution and found that the obtained models can offer favorable evaluation results for flyrock and fragment size evaluation. More pertinent work about blast fragmentation prediction using AI methods is tabulated in Table 5 and Table 6. These models predict a percentile fragment size (e.g. median size or 80-percentile size); note, however, that in many of these works the median size or x_{50} or is wrongly called mean size (Ouchterlony, 2016).

Table 5. Previous work about blast fragmentation prediction using AI techniques: Part 1.

Reference	Technique	Input	Data No.	R^2
Monjezi et al. (2009)	FIS	*B, S, ST, SD, PF, HL, ρ, MC*	415	0.96
Monjezi et al. (2010a)	BPNN	*B.S, D, ST, TC, PF, NR, MH, PI, TR*	132	0.985
Monjezi et al. (2010b)	ANN	*D, HL, B.S, ST, N, PF, ρ, MC*	250	0.98
Kulatilake et al. (2010)	BPNN	*S.B, HL/B, B/D, ST/B, PF, X_B, E*	91	0.941
(Bahrami et al. (2011)	BPNN	*D, HL, B, S, PF, RMR, BI, SD, ST, MC*	220	0.97
Kulatilake et al. (2012)	ANN	*S.B, HL/B, B/D, ST/B, PF, X_B, E*	109	0.94
Shi et al., (2012)	SVM	*S.B, H/B, B/D, ST/B, PF, E, X_B*	102	0.962
Sayadi et al. (2013)	ANN	*B, S, HL, SD, q*	103	0.85
Enayatollahi et al. (2014)	ANN	*HL, PF, SD, BS, S.B, WD, ST, MC, NR, RQD, T, B*	70	0.98
Monjezi et al. (2014)	ANN	*B, S, PF, NR, D, MC, ST, H*	135	0.92-0.95
Esmaeili et al. (2015)	SVM	*DR, q, ST, D, BI, S.B*	80	0.83
	ANFIS			0.89
Shams et al. (2015)	FIS	*B, S, D, Sch, DJ, PF, ST*	185	0.922
Ebrahimi et al. (2016)	BPNN	*B, S, ST, HL, PF*	34	0.78
Ghaeini et al. (2017)	MI	*UCS, P, RQD, JS, ρ, q, B, ST, S/D, JPO*	36	0.81
Asl et al. (2018)	ANN	*B, S, HL, SD, ST, MC, PF, GSI*	200	0.94
Dimitraki et al. (2019)	ANN	*BI, PF, QB*	100	0.8

Table 6. Previous work about blast fragmentation prediction using AI techniques: Part 2.

Reference	Technique	Input	Data No.	R^2
Hasanipanah et al. (2018)	PSO-ANFIS	*B, S, ST, q, MC*	72	0.89
	SVM			0.83
	ANFIS			0.81
Gao et al. (2018)	GPR	*B, S, ST, PF, MC*	72	0.948
	SVM			0.83
	ANFIS			0.81
	PSO-ANFIS			0.89
Sayevand et al. (2018)	ICA	*MC, B, S, ST, PF, RMR*	80	0.947
(Mojtahedi et al., 2019)	FFA-ANFIS	*B, S, ST, PF, MC*	72	0.98
Huang et al. (2022)	CSO	*q, B, RMR, MC, ST, S*	75	0.985
Fang et al. (2021)	FFA-BGAM	*PF, MC, S, ST, B, H*	136	0.98
Zhang et al. (2020)	ACO-BRT	*PF, MC, S, ST, B, H*	136	0.962
Zhou et al. (2021a)	FFA-ANFIS	*B, S, ST, PF, MC, RMR*	88	0.981
	GA-ANFIS			0.989

Note: ACO - Ant colony optimization; ANFIS - Adaptive network-based fuzzy inference system; ANN – Artificial neutral network; B - Burden; BGAM - Boosted generalized additive model; BI - Blastability index; BPNN - Back propagation neural network; BRT - Boosted regression tree; BS - Bench slope; B.S - Ratio of burden to spacing; CSO - Cat swarm optimization; D - Hole diameter; DJ - Density of joint; DR - Ratio of total delays per number of rows; E - Elastic modulus; FFA - Firefly algorithm; FIS - Fuzzy inference system; GA - Genetic algorithm; GSI - Geological strength index; H - Bench height; HL - Hole depth; JPO - Joint plane orientation; JS - Joint spacing; MC - Charge per delay; MH - Maximum holes per delay; MI - Mutual information; NR - Number of rows; P - Joint persistency; PF - Powder factor; PI - Point load index; PSO – Particle swarm optimization; q - Specific charge; QB - Quantity of blasted rock pile; RMR - Rock mass rating; RQD - Rock quality designation; S - Spacing; S.B – Ratio of spacing to burden; Sch - Schmidt hammer rebound number; S/D - Ratio of borehole spacing to diameter; SD - Specific drilling; ST - Stemming; SVM – Support vector machine; T - tensile strength; TC - Total charge per delay; TR - Delay between rows; UCS - Uniaxial compressive strength; WD - Water depth; X_B - In situ block size; ρ - Rock density;

Research gap: A remarkable outcome could be achieved by implementing these models; however, there are still some shortcomings that need to be addressed. For instance, the selection of hyper-parameters is not determined by standard methods and probably ignores more effective hyper-parameters. In addition, cross-validation was not implemented and thus the prediction accuracy may not be scientific. Apart from that, a limited number of influential parameters were taken into consideration in the past studies to develop the blast fragmentation prediction models. From the literature review, it can be found that various significant factors were ignored and thus the prediction models failed to provide convincing results.

In this study, support vector regression (SVR) models including e-SVR and v-SVR are utilized as the main prediction tools for fragments in blasting combined with five optimization algorithms, i.e. GA, PSO, grey wolf optimization (GWO), salp swarm algorithm (SSA) and grid search (GS) method. Meanwhile, cross-validations are employed to examine the prediction capability of different models. Three mathematical indices, i.e. R^2, mean square error (MSE) and variance accounted for (VAF), are used to assess the prediction performance. Finally, the sensitivity analysis is implemented to understand the sensitivity of each input parameter on blasting median fragment size. A general working framework of this study is demonstrated in Figure 11.

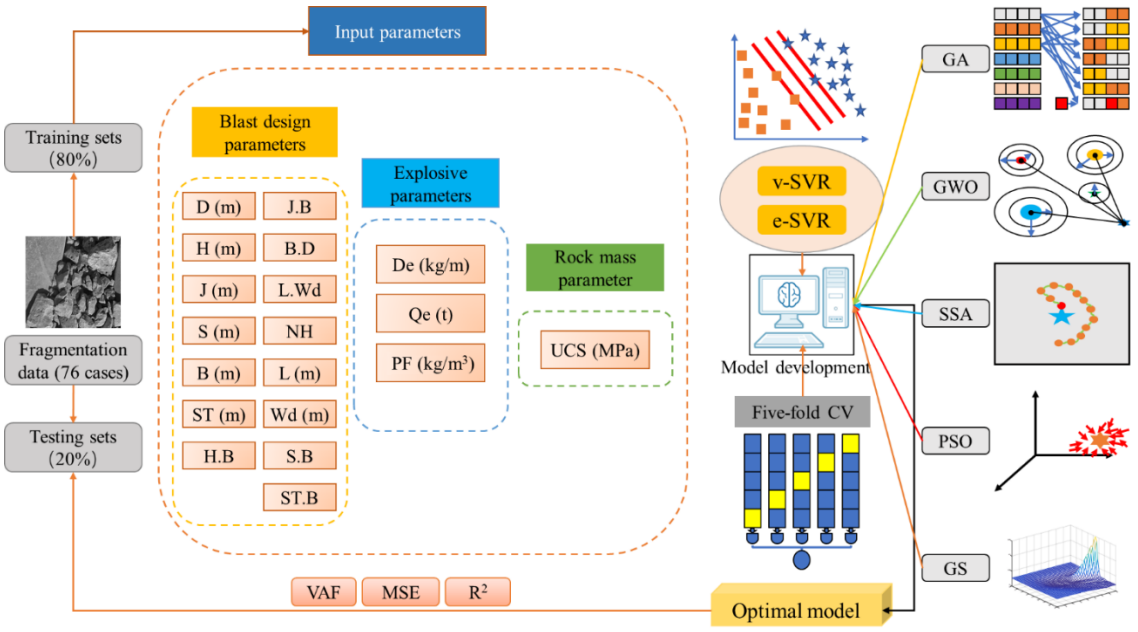


Figure 11. A framework of SVR-based model for blasting fragment size evaluation.

2.3 Data description

In this study, a total of 76 groups of blasting datasets are utilized (Kumar Sharma and Rai, 2017). The datasets involve 19 influential factors and one output parameter (median fragment size (x_{50}) (m)). These influential factors are categorized into three groups, i.e. blast design parameters, explosive parameters and rock mass parameters. Compared with the past research, more influential parameters are used for developing prediction models. The blast design parameters are blast hole diameter (D) (m), average bench height (H) (m), average sub-grade drilling (J) (m), average spacing (S) (m), average burden (B) (m),

average stemming (*ST*) (m), average length (*L*) (m), average width (*Wd*) (m), *S*/*B* ratio (*S.B*), *ST*/*B* ratio (*ST.B*), stiffness ratio (*H.B*), *J*/*B* ratio (*J.B*), *B*/*D* ratio (*B.D*), length/width ratio (*L.Wd*), and number of holes (*NH*), whereas explosive parameters are total explosive amount (*Qe*) (t), linear explosive density (*De*) (kg/m) and powder factor (*PF*) (kg/m^3). The rock mass parameter is reflected by the uniaxial compressive strength (UCS) (MPa).

General data distribution of each of the parameters is shown in Figure 12 based on the violin plot. A violin plot consists of a boxplot and a density plot, where a black spot is a median value. The black box represents the range from the lower quartile to the upper quartile. The black line indicates the 95% confidence interval. The outer shape of the box represents the density estimation of the data. The original datasets are divided into two parts at a ratio of 4:1. The 80% datasets are used to establish the training networks (Esmaeili et al., 2015). The remaining 20% of the datasets are not used in the model development but are employed to test the prediction performance of the network.

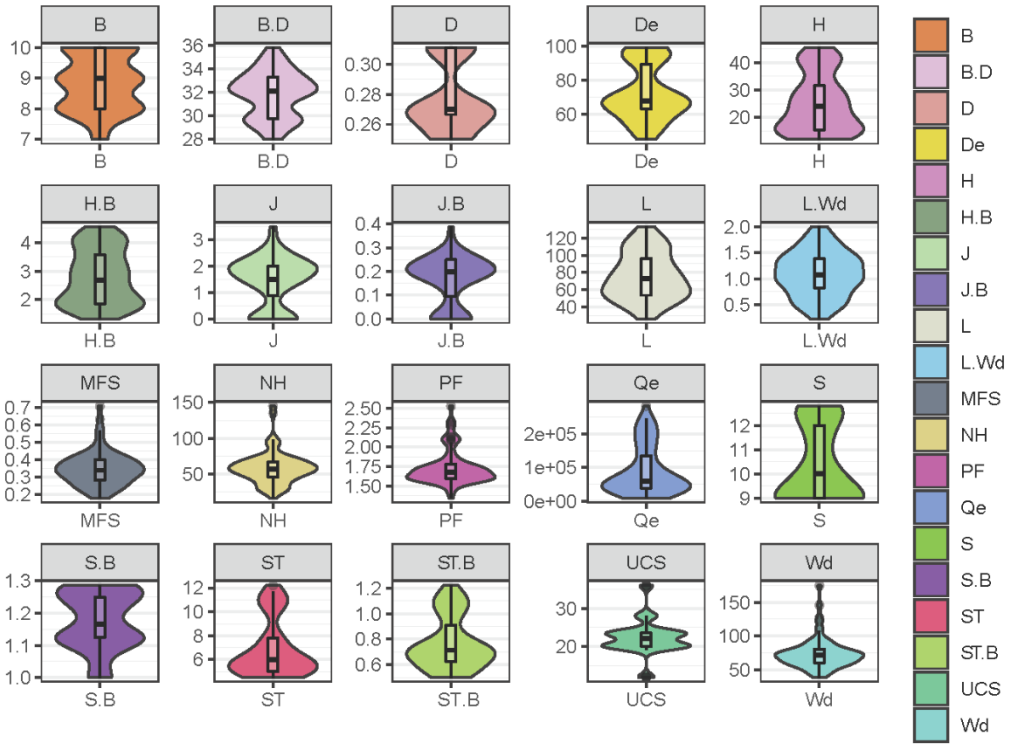


Figure 12. Data distribution of all parameters used for developing prediction models.

2.4 Model development

2.4.1 Principal component analysis and cross-validation

A principal component analysis (PCA) is used for extracting and recognizing the significant information from a group of multidimensional variables (Wold et al., 1987). By matrix transformation, original data are mapped into a new dimension. The data with large dimensions is presented with fewer dimensions. However, data information is discarded by such operations. The original data are compressed and good for calculation.

By implementing the PCA, 19 variables are transmitted into 7 new variables and the overfitting problems can be improved.

Cross-validation is a useful method for assessing the model robustness and generalization. It can avoid over-fitting and under-fitting to some extent. In this study, the *k*-fold cross-validation method is utilized (Fushiki, 2011). Original training set is equally separated into *k* sets. *k*–1 sets are treated as new training sets and the remaining sets are used for validation. This process is operated for *k* times. The average accuracy of these *k* models is used as the performance index of the regression model. In this study, 5-fold cross-validation is utilized according to the scale of the datasets. A general display of 5-fold cross-validation is shown in Figure 13. In this figure, P1, P2, P3, P4 and P5 represent the prediction results of the corresponding fold, respectively.

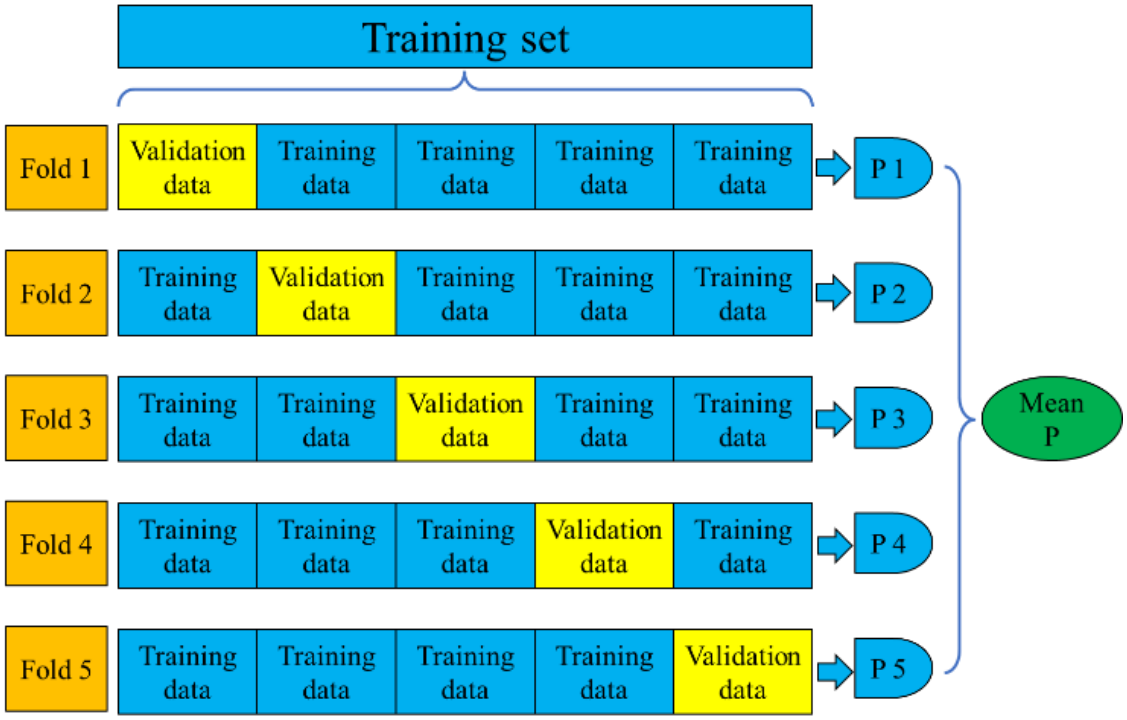


Figure 13. Schematic diagram of 5-fold cross-validation.

2.4.2 Evaluation metrics

To control and compare the performance of proposed algorithms, three mathematical evaluation metrics are adopted, i.e., R^2 (Eq. (2.1)), MSE (Eq. (2.2)) and VAF (Eq. (2.3)). For different optimization algorithms, they would output different prediction results and thus these values can be used for determining the best optimization method. Meanwhile, for the synthetical assessment performance of these five algorithms, a comprehensive evaluation system is employed (Zorlu et al., 2008). In this system, the best prediction performance is endowed with a higher score and inferior performance is given a lower score. By adding all scores, each model obtains cumulative scores and the one with the highest scores is considered as the best.

$$R^2 = 1 - \frac{\sum_k^N (y_k - y_k')^2}{\sum_k^N (y_k - \overline{y_k})^2} \tag{2.1}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{k}^{N}(y_k - \mathrm{y}'_k)^2 \tag{2.2}$$

$$\mathrm{VAF} = \left(1 - \frac{vaf(y_k - \mathrm{y}'_k)}{vaf(y_k)}\right) \times 100\% \tag{2.3}$$

In the given equation, k represents the current data sample number under evaluation, y_k signifies the measured median fragments, $\overline{y_k}$ denotes the average value of the measured median fragments, y'_k represents the corresponding predictive value, N stands for the total number of data sample, and *vaf* indicates the variance function. For these indicators, achieving better performance entails lower MSE values, as well as higher R^2 and VAF values.

2.4.3 Parameter configurations

To develop e-SVM and v-SVM-based x_{50} prediction models, four types of meta-heuristic algorithms and GS methods are tested and ten models are established and compared. The aforementioned 19 influential factors are utilized as the input parameters and the x_{50} is used as the output parameter. In meta-heuristic algorithms, many parameters influence the optimal effect. Among these parameters, the swarm size and the number of iterations have significant impacts on optimization performance (Koopialipoor et al., 2019; Li et al., 2021b; Yu et al., 2021c). Therefore, in this study, these two parameters are discussed and compared. Detailed parameter configurations are listed in Table 7.

Table 7. Parameter configurations of four meta-heuristic algorithms.

Meta-heuristic algorithm	Parameter	Value
GWO	*a*	Decreasing linearly from 2 to 0
	Ub	100
	Lb	0.01
PSO	c_1	2
	c_2	2
	k_p	0.6
	wV	1.2
	wP	1
GA	*ggap*	0.9
SSA	*Ub*	100
	Lb	0.01

Note: *Ub* - Upper boundary; *Lb* - Lower boundary; c_1, c_2 - Acceleration constant; *ggap* - Probabilities of crossover and mutation; k_p - Parameter determining the relationship between particle velocity and movement; *wV* - Elastic coefficient in the velocity update formula; *wP* - Elastic coefficient in population update formula.

2.4.4 PSO-SVR optimization

To achieve the best performance of blasting x_{50} prediction, several parameters including velocity equation, number of iterations and swarm size, which have a significant influence on PSO are carefully selected. Among these parameters, the number of iterations and swarm size bring a significant impact on optimization performance. Generally, with the increase of iteration, the optimization performance tends towards stability but the calculation time also increases. For judging the optimization performance, the MSE value is used in this section.

During testing, when the number of iterations is 500, the prediction performance is stable. Therefore, the number of iterations is determined to be 500 and the swarm sizes equal to 60, 70, 80, 90 and 100 are tested to choose the best optimal parameters. As Figure 14 depicts, the optimal process of each swarm size is different and with the increase of iteration, MSE value decreases. Each model obtains the lowest MSE value, when it reaches the end of optimization.

To compare the prediction performance of different swarm sizes, three evaluation performance metrics are referenced in this study and each performance is given a corresponding score based on the comprehensive evaluation system (Zorlu et al., 2008). In this grading system, better performance is assigned a higher score and by adding all scores, each parameter configuration obtains comprehensive scores. The one with the highest score is regarded as the best parameter combination.

Table 8. Performance and scores of different swarm sizes of PSO-e-SVR.

Swarm size	Performance and score						Final score
	Training set			Testing set			
	R^2	MSE	VAF	R^2	MSE	VAF	
60	0.8442 (3)	0.00512 (3)	80.41 (3)	0.8116 (2)	0.01388 (5)	79.48 (3)	19
70	0.8267 (2)	0.00583 (2)	77.63 (2)	0.8175 (4)	0.01625 (2)	75.79 (2)	14
80	0.8931 (5)	0.00299 (5)	88.38 (5)	0.8331 (5)	0.01413 (4)	81.13 (5)	29
90	0.8713 (4)	0.00367 (4)	85.81 (4)	0.7999 (1)	0.01413 (4)	79.99 (4)	21
100	0.8202 (1)	0.00609 (1)	76.64 (1)	0.8121 (3)	0.0169 (1)	74.79 (1)	8

Note: The value in the round brackets represents the corresponding score.

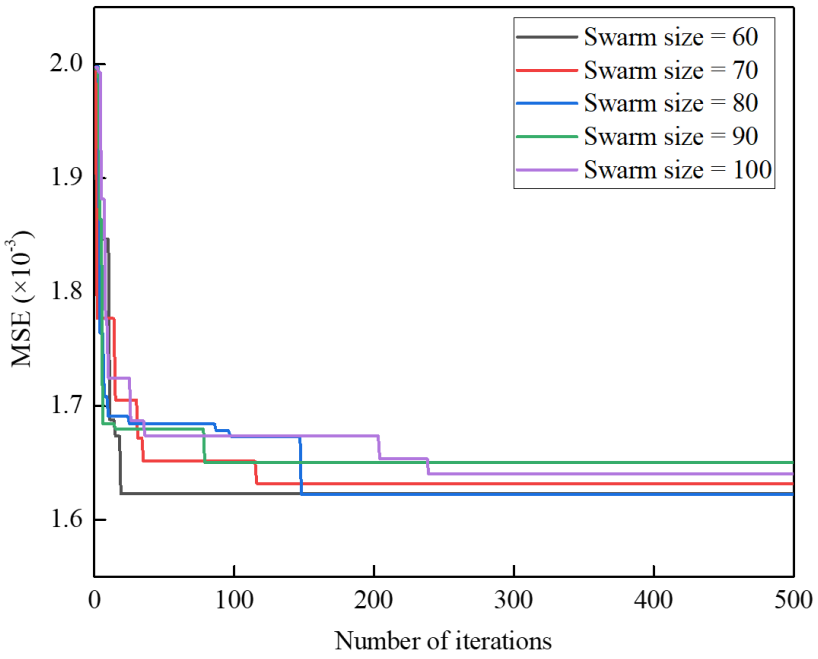
Table 9. Performance and scores of different swarm sizes of PSO-v-SVR.

Swarm size	Performance and score						Final score
	Training set			Testing set			
	R^2	MSE	VAF	R^2	MSE	VAF	
60	0.8743 (5)	0.00362 (5)	85.99 (5)	0.8313 (2)	0.01313 (3)	82.35 (4)	24
70	0.8353 (2)	0.00517 (1)	80.22 (1)	0.8404 (5)	0.0137 (2)	80.39 (1)	12
80	0.8726 (4)	0.00365 (4)	85.86 (4)	0.8073 (1)	0.01415 (1)	80.64 (2)	16
90	0.8417 (3)	0.00482 (3)	81.56 (3)	0.8335 (3)	0.01283 (4)	82.07 (3)	19
100	0.831 (1)	0.00509 (2)	80.48 (2)	0.8386 (4)	0.01233 (5)	82.64 (5)	19

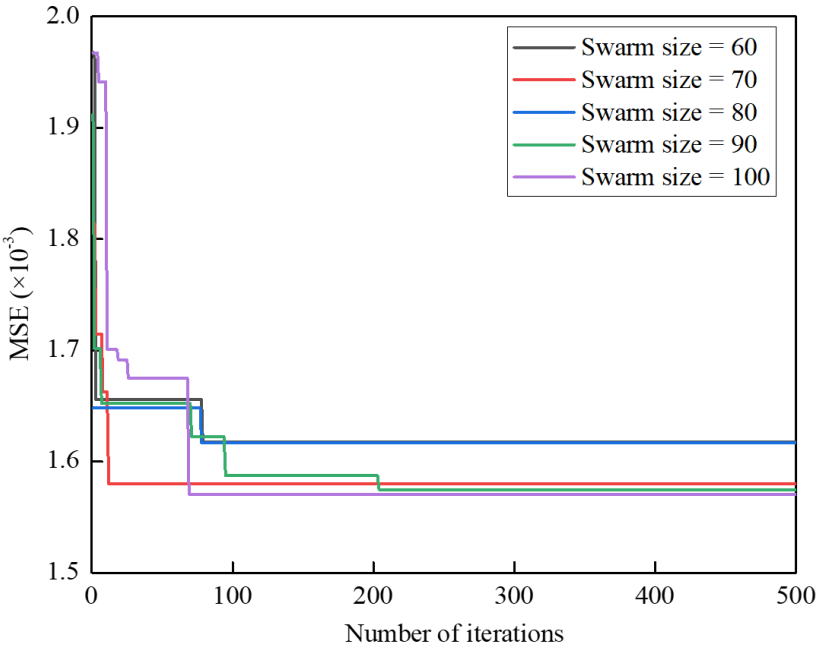
From Table 8, when the swarm size is equal to 80 for the PSO-e-SVR, the established model obtains the best prediction performance, where R^2, MSE and VAF values are 0.8931, 0.00299 and 88.38, respectively for the training sets, whereas those are 0.8331, 0.01413 and 81.13, respectively for the testing sets. In addition, when the swarm size is 80, the evaluation performance is superior to the other models except for the MSE of the testing sets. For the PSO-v-SVR, when the swarm size is 60, the developed model procures the best prediction performance, where R^2, MSE and VAF values are 0.8743, 0.00362 and 85.99, respectively for the training sets, whereas those are 0.8313, 0.01313 and 82.35, respectively for the testing sets. When the swarm size is 60, the prediction model shows strong fitness abilities in training sets, while for the testing sets, the model with the swarm size equal to 100 seems to have stronger fitness abilities. Detailed results of the training and testing sets can be found in Table 9. Then, the accumulation bar charts are shown in Figure 86 and Figure 87 in Appendix 1 to help to interpret the ranking results.

2.4.5 GA-SVR optimization

As mentioned earlier, GA can bring a positive effect on parameter selection and optimization. Before implementation, the most effective GA parameters that are utilized to govern GA-SVR prediction models should be opted. Similar to the development of PSO-SVR models, the impacts of the number of iterations and swarm size should be discussed. For the sake of comparison, the same number of iterations and swarm size are employed in GA-SVR. The optimization process of GA-e-SVR and GA-v-SVR can be found in Figure 15. By calculating the performance scores, it can be observed that when the swarm size is equal to 100 for GA-e-SVR, the proposed model can obtain the best prediction performance. At this time, R^2, MSE and VAF values are 0.8474, 0.00496 and 81.05, respectively for training sets, whereas those are 0.8174, 0.01488 and 77.97, respectively for testing sets. From Table 10, when the swarm size is equal to 100, all evaluation metrics obtain the best performance. Therefore, for the GA-e-SVR, it can be concluded that the model with a swarm size equal to 100 has the best robustness in this study.



(a) PSO-e-SVR



(b) PSO-v-SVR

Figure 14. Optimization performance with different swarm sizes: (a) PSO-e-SVR and (b) PSO-v-SVR.

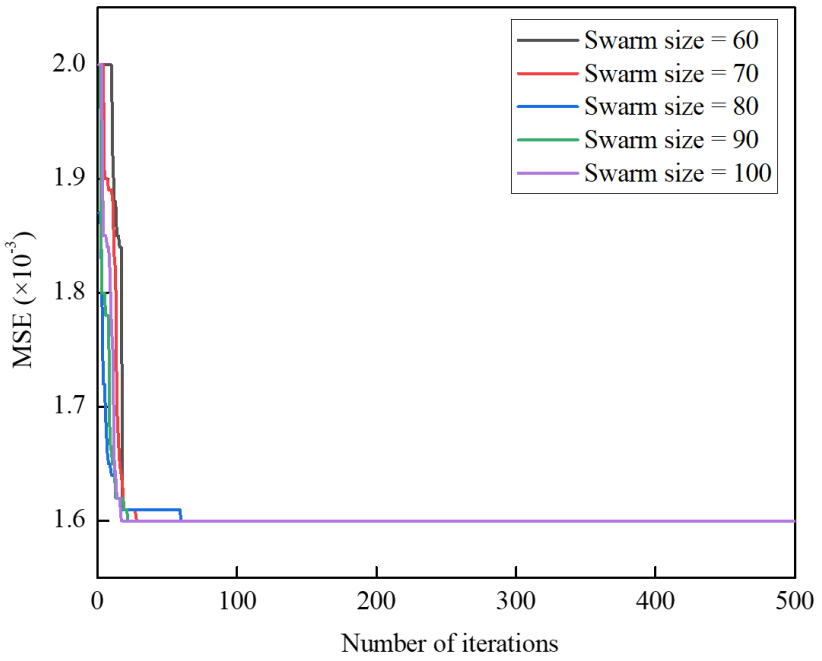
Table 10. Performance and scores of different swarm sizes of GA-e-SVR.

Swarm size	Performance and score						Final score
	Training set			Testing set			
	R^2	MSE	VAF	R^2	MSE	VAF	
60	0.8461 (4)	0.00501 (4)	80.86 (4)	0.8159 (4)	0.01495 (4)	77.87 (4)	24
70	0.8449 (1)	0.00506 (1)	80.67 (1)	0.8144 (1)	0.01496 (3)	77.82 (2)	9
80	0.8457 (2)	0.00502 (3)	80.81 (3)	0.8151 (2)	0.01498 (1)	77.81 (1)	12
90	0.8458 (3)	0.00502 (3)	80.81 (3)	0.8154 (3)	0.01496 (3)	77.83 (3)	18
100	0.8474 (5)	0.00496 (5)	81.05 (5)	0.8174 (5)	0.01488 (5)	77.97 (5)	30

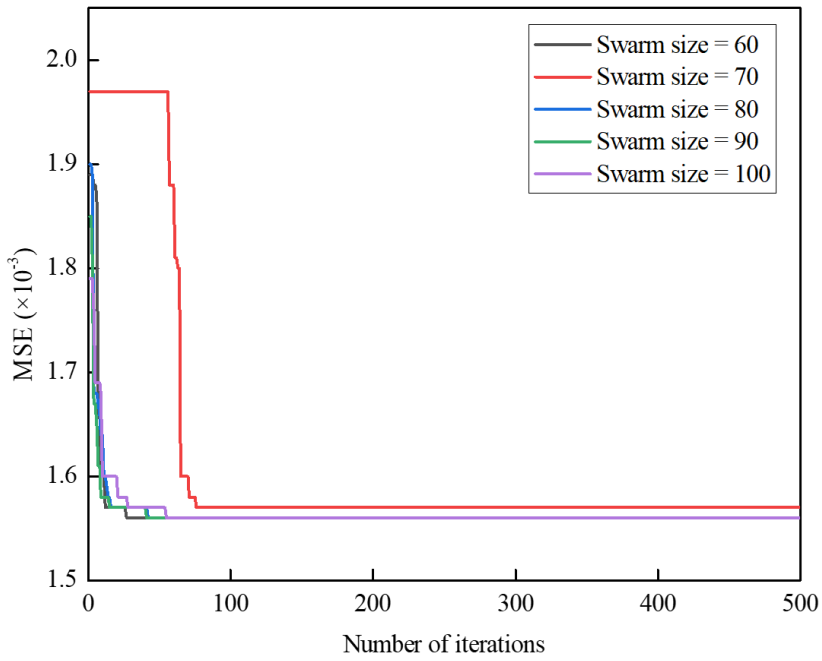
Table 11. Performance and scores of different swarm sizes of GA-v-SVR.

Swarm size	Performance and score						Final score
	Training set			Testing set			
	R^2	MSE	VAF	R^2	MSE	VAF	
60	0.8347 (4)	0.005 (4)	80.84 (4)	0.8363 (2)	0.01252 (3)	82.41 (4)	21
70	0.8377 (5)	0.00486 (5)	81.35 (5)	0.8323 (1)	0.01261 (1)	82.4 (3)	20
80	0.8347 (3)	0.005 (4)	80.84 (4)	0.8368 (3)	0.0125 (4)	82.44 (5)	23
90	0.8325 (1)	0.0051 (1)	80.46 (1)	0.8391 (4)	0.01249 (5)	82.36 (2)	14
100	0.8329 (2)	0.00509 (2)	80.47 (2)	0.8398 (5)	0.01253 (2)	82.3 (1)	14

For the GA-v-SVR, when the swarm size is 80, the procured model brings the best comprehensive prediction results, with R^2, MSE and VAF values of 0.8347, 0.005 and 80.84, respectively for the training sets, and 0.8368, 0.0125 and 82.44, respectively for the testing sets. From Table 11, although the model with swarm size equal to 80 does not obtain prominent performance in each single evaluation metric, its comprehensive performance is better than the other models. For the model with the swarm size equal to 70, it produces a good prediction effect in training sets but fails to bring desirable feedback in the testing sets. As for models with swarm sizes equal to 90 and 100, they obtain superior scores in R^2 and MSE in testing sets but show mediocre performance in the other metrics. The intuitive scoring results of GA-e-SVR and GA-v-SVR are shown in Figure 88 and Figure 89 in Appendix 1, respectively.



(a) GA-e-SVR



(b) GA-v-SVR

Figure 15. Optimization performance with different swarm sizes: (a) GA-e-SVR and (b) GA-v-SVR.

2.4.6 SSA-SVR optimization

As a kind of novel swarm intelligence algorithm, SSA shows its superiority and feasibility in various optimal problems. SSA is easy to implement and adjust. For the sake of comparison, the same hyper-parameter values are also employed in the SSA-based models. The optimization process of SSA-based models with different swarm sizes is shown in Figure 16.

By comparing the model performance, it can be observed that when the swarm size is equal to 80 for SSA-e-SVR, the proposed model can obtain the best prediction performance according to Table 12. For the SSA-e-SVR, R^2, MSE and VAF values are 0.8463, 0.00139 and 80.9, respectively for the training sets, whereas those are 0.8157, 0.00415 and 77.88, respectively for the testing sets. From Table 12, when the swarm size is equal to 80, the prediction model produces excellent prediction performance and each evaluation metric generates the best feedback.

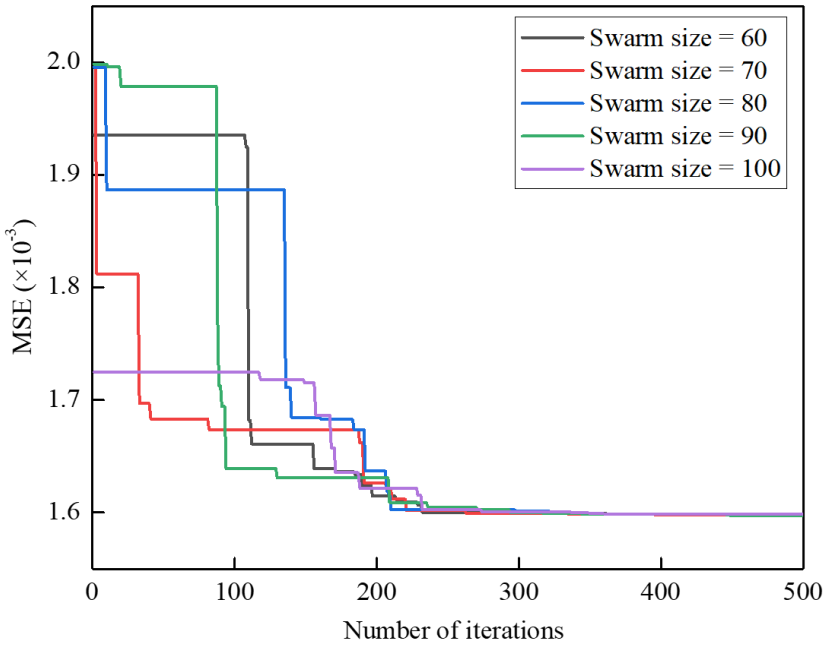
Table 12. Performance and scores of different swarm sizes of SSA-e-SVR.

Swarm size	Performance and score						Final score
	Training set			Testing set			
	R^2	MSE	VAF	R^2	MSE	VAF	
60	0.846 (3)	0.00139 (5)	80.84 (2)	0.8155 (2)	0.00416 (4)	77.8 (1)	17
70	0.8461 (4)	0.00139 (5)	80.85 (3)	0.8157 (5)	0.00415 (5)	77.88 (5)	27
80	0.8463 (5)	0.00139 (5)	80.90 (5)	0.8157 (5)	0.00415 (5)	77.88 (5)	30
90	0.846 (3)	0.00139 (5)	80.86 (4)	0.8157 (5)	0.00415 (5)	77.85 (3)	25
100	0.8457 (1)	0.0014 (4)	80.8 (1)	0.8152 (1)	0.00416 (4)	77.84 (2)	13

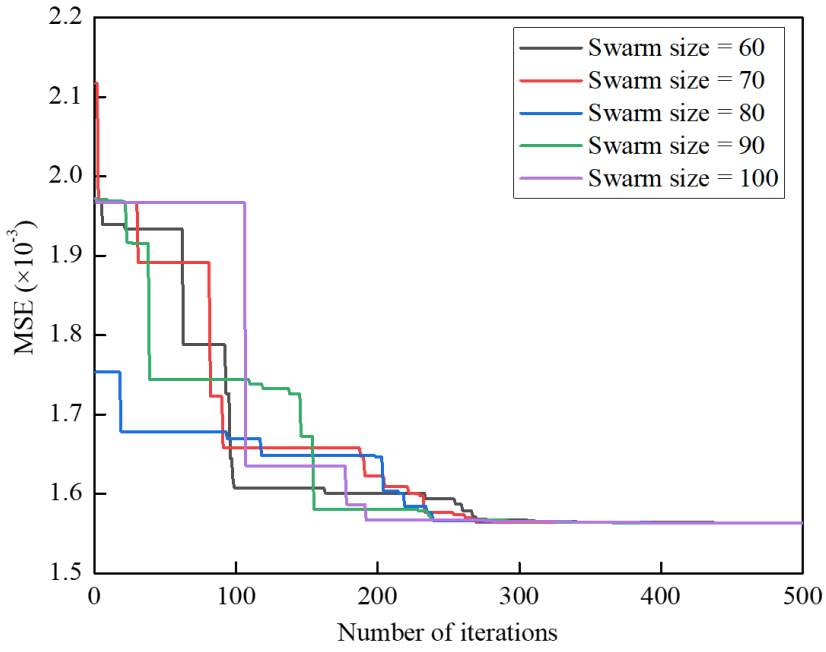
Table 13. Performance and scores of different swarm sizes of SSA-v-SVR.

Swarm size	Performance and score						Final score
	Training set			Testing set			
	R^2	MSE	VAF	R^2	MSE	VAF	
60	0.8332 (1)	0.00141 (1)	80.59 (1)	0.839 (5)	0.00346 (5)	82.43 (4)	17
70	0.8369 (5)	0.00137 (5)	81.09 (5)	0.8348 (1)	0.0035 (1)	82.33 (1)	18
80	0.8352 (4)	0.00138 (4)	80.9 (4)	0.8364 (2)	0.00348 (2)	82.43 (4)	20
90	0.8343 (2)	0.00139 (3)	80.76 (2)	0.8374 (4)	0.00347 (4)	82.43 (4)	19
100	0.8345 (3)	0.00139 (3)	80.81 (3)	0.8369 (3)	0.00347 (4)	82.44 (5)	21

From Table 13, when the swarm size is 100, the proposed model produces the best synthetical prediction results for the SSA-v-SVR with R^2, MSE and VAF values of 0.8345, 0.00139 and 80.81, respectively for the training sets, and 0.8369, 0.00347 and 82.44, respectively for the testing sets. Although the model with a swarm size of 70 generates outstanding performance in training sets, it does not work well for the testing sets. Therefore, its generalization and robustness are poor under this situation. Finally, the intuitive ranking results of SSA-e-SVR and SSA-v-SVR are depicted in Figure 90 and Figure 91 in Appendix 1.



(a) SSA-e-SVR



(b) SSA-v-SVR

Figure 16. Optimization performance with different swarm sizes: (a) SSA-e-SVR, (b) SSA-v-SVR.

2.4.7 GWO-SVR optimization

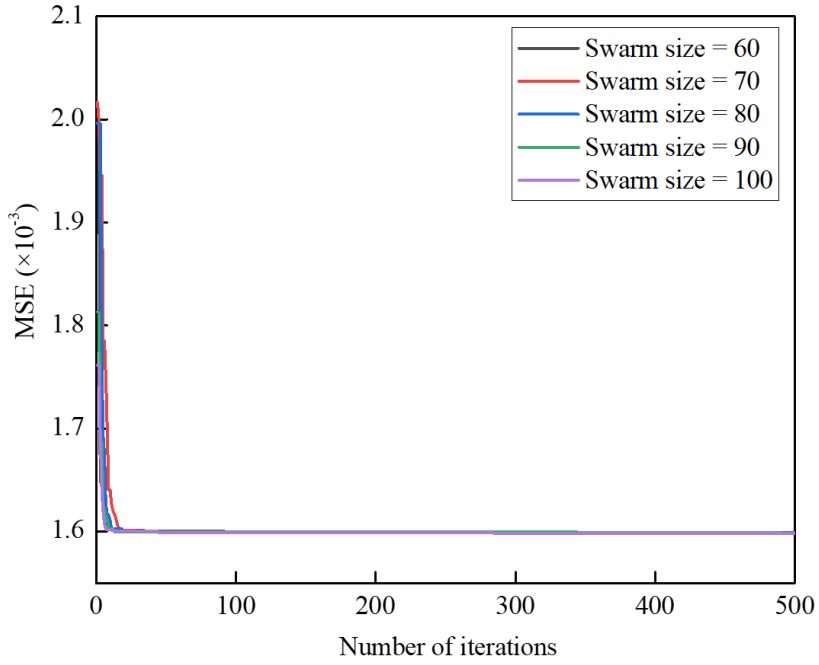
The GWO is easy to interpret as a kind of effective optimization strategy. In the GWO, three kinds of wolves control the optimization process. Similarly, the number of iterations is set to be 500 and the swarm (wolf) sizes are set to be 60, 70, 80, 90 and 100, respectively. From Figure 17, it can be seen that GWO optimization has a faster

convergence velocity, which reflects its eminent optimization abilities. By comparing the model performance in Table 14, it can be observed that when the swarm size is equal to 90 for GWO-e-SVR, the proposed model can obtain the best prediction performance. For the GWO-e-SVR, R^2, MSE and VAF values are 0.8463, 0.00139 and 80.89, respectively for the training sets, whereas those are 0.8156, 0.00414 and 77.9, respectively for the testing sets. It should be noticed that although the R^2 value in testing sets is a little bit inferior to the one with the swarm size of 100, the model with the swarm size of 90 has the higher comprehensive score in any other evaluation metrics.

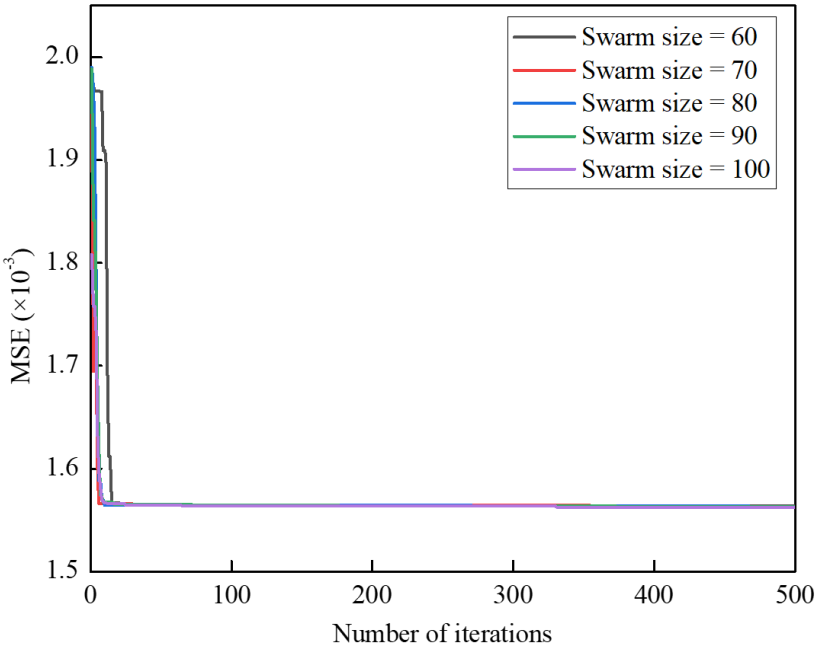
For the GWO-v-SVR, the proposed model produces the best comprehensive prediction feedback when the swarm size is 100, as shown in Table 15, with R^2, MSE and VAF values of 0.8355, 0.00138 and 80.98, respectively for the training sets, and 0.8353, 0.00348 and 82.41, respectively for the testing sets. It is worth noting that although the model with swarm size equal to 100 has a superior performance in training sets, it fails to procure similar performance in testing sets. That is to say, the generalization of this model is worthwhile to be improved and discussed in the future. Perhaps, by changing and testing different swarm sizes and iteration numbers, it produces more persuasive scenarios. Similarly, the ranking results of GWO-e-SVR and GWO-v-SVR are displayed in Figure 92 and Figure 93 in Appendix 1, respectively.

Table 14. Performance and scores of different swarm sizes of GWO-e-SVR.

Swarm size	Performance and score						Final score
	Training set			Testing set			
	R^2	MSE	VAF	R^2	MSE	VAF	
60	0.8457 (2)	0.0014 (2)	80.78 (2)	0.8151 (2)	0.00416 (3)	77.81 (1)	12
70	0.8451 (1)	0.0014 (2)	80.71 (1)	0.8149 (1)	0.00415 (4)	77.84 (4)	13
80	0.8462 (4)	0.00139 (5)	80.86 (4)	0.8153 (3)	0.00416 (3)	77.84 (4)	23
90	0.8463 (5)	0.00139 (5)	80.89 (5)	0.8156 (4)	0.00414 (5)	77.9 (5)	29
100	0.8458 (3)	0.00139 (5)	80.82 (3)	0.8157 (5)	0.00416 (3)	77.83 (2)	21



(a) GWO-e-SVR



(b) GWO-v-SVR

Figure 17. Optimization performance with different swarm sizes: (a) GWO-e-SVR, (b) GWO-v-SVR.

Table 15. Performance and scores of different swarm sizes of GWO-v-SVR.

Swarm size	Performance and score						Final score
	Training set			Testing set			
	R^2	MSE	VAF	R^2	MSE	VAF	
60	0.8326 (1)	0.00141 (1)	80.52 (1)	0.8392 (5)	0.00346 (5)	82.43 (5)	18
70	0.8348 (3)	0.00139 (4)	80.86 (3)	0.8365 (3)	0.00348 (3)	82.42 (4)	20
80	0.8337 (2)	0.0014 (2)	80.66 (2)	0.838 (4)	0.00347 (4)	82.39 (1)	15
90	0.8351 (4)	0.00139 (4)	80.87 (4)	0.8362 (2)	0.00348 (3)	82.41 (3)	20
100	0.8355 (5)	0.00138 (5)	80.98 (5)	0.8353 (1)	0.00348 (3)	82.41 (3)	22

2.4.8 GS-SVR optimization

In the GS optimization, there are mainly two parameters that need to be tuned, i.e. grid step and grid bound. In this study, different grid bounds are changed and tested and it is found that grid bound almost does not have any influence on the optimization effect. Therefore, the search bound of *Ct* and *g* is set to (2^{-8}, 2^{8}). In the next step, five grid steps are tested, i.e. 0.2, 0.4, 0.6, 0.8 and 1 in the GS-e-SVR and GS-v-SVR models. From Table 16 and Table 17, it can be said that finer grids always do not bring better prediction performance. It is due to the limitation of the GS method. It is almost impossible to examine the performance of all nodes in specified grid bound, unless the grid step is set to be ultra-small. Therefore, some more effective parameter combinations perhaps are missed. As per Figure 18 and Figure 19, one grid node represents one kind of parameter combination and the longitudinal axis represents the corresponding MSE value. With the grid color changes from yellow to dark blue, a lower MSE value is obtained. When the grid step equals 0.2, 0.4 and 0.6, the GS-e-SVR method would obtain the best comprehensive performance, and for the GS-v-SVR method, the most suitable grid step is 0.8.

Table 16. Performance and scores of different grid steps of GS-e-SVR.

Grid step	Performance and score						Final score
	Training set			Testing set			
	R^2	MSE	VAF	R^2	MSE	VAF	
0.2	0.8464 (4)	0.00499 (5)	80.93 (5)	0.8154 (5)	0.01487 (5)	77.98 (5)	29
0.4	0.8464 (4)	0.00499 (5)	80.93 (5)	0.8154 (5)	0.01487 (5)	77.98 (5)	29
0.6	0.8464 (4)	0.00499 (5)	80.93 (5)	0.8154 (5)	0.01487 (5)	77.98 (5)	29
0.8	0.7858 (1)	0.00754 (1)	71.21 (1)	0.7151 (2)	0.02666 (2)	59.61 (1)	8
1	0.8563 (5)	0.0051 (2)	80.54 (2)	0.7139 (1)	0.02685 (1)	60.16 (2)	13

Table 17. Performance and scores of different grid steps of GS-v-SVR.

Grid step	Performance and score						Final score
	Training set			Testing set			
	R^2	MSE	VAF	R^2	MSE	VAF	
0.2	0.8391 (5)	0.00502 (4)	80.78 (4)	0.8394 (2)	0.01335 (2)	81.03 (2)	19
0.4	0.8308 (3)	0.00523 (3)	79.95 (3)	0.8434 (3)	0.01286 (3)	81.67 (3)	18
0.6	0.8214 (2)	0.00546 (2)	79.02 (2)	0.8461 (4)	0.01252 (5)	82.11 (4)	19
0.8	0.8347 (4)	0.00488 (5)	81.25 (5)	0.8298 (1)	0.01256 (4)	82.49 (5)	24
1	0.7989 (1)	0.00647 (1)	74.98 (1)	0.8462 (5)	0.01744 (1)	75.1 (1)	10

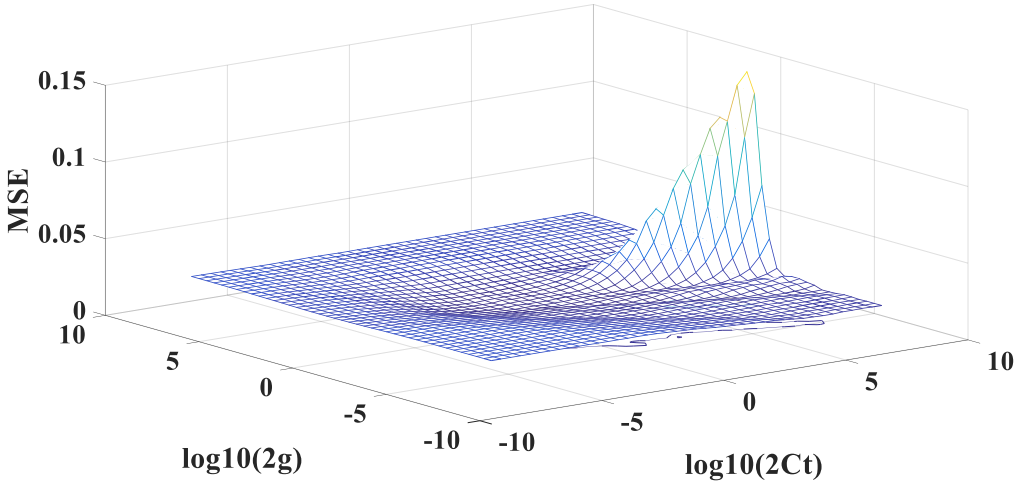


Figure 18. GS-e-SVR optimization curve with grid step equal to 0.4.

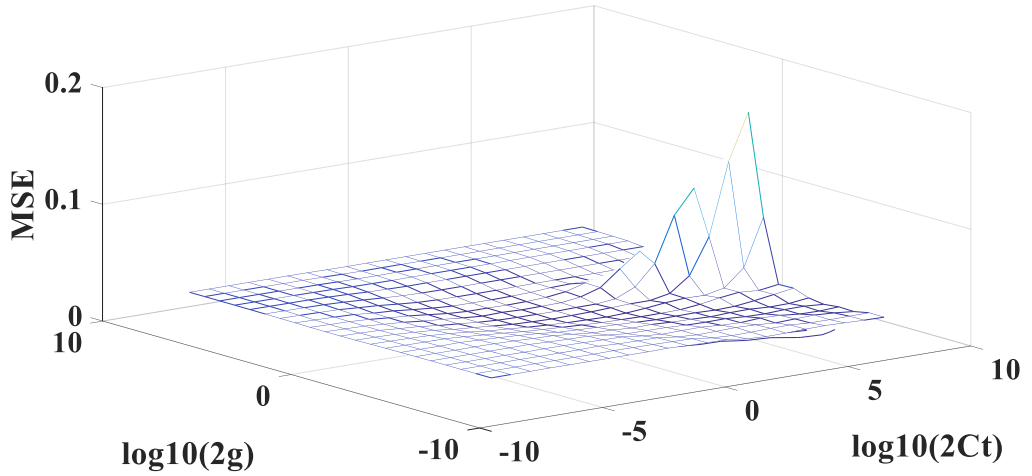


Figure 19. GS-v-SVR optimization curve with grid step equal to 0.8.

2.5 Comparison of optimal algorithms

Based on the previous discussions, the model with the highest comprehensive scores of each optimal algorithm is selected and compared. The detailed data are listed in Table 18. Among the e-SVR-based x_{50} prediction models, PSO-e-SVR obtains the best comprehensive scores. In addition, PSO-e-SVR also obtains the best performance in R^2 and VAF. Only considering the performance metric MSE, GWO-e-SVR outperforms the other algorithms for both training and testing sets. SSA-e-SVR has prominent performance with the metric MSE when being tested in training sets.

Table 18. Comparison of optimal algorithms of e-SVR.

Algorithm	Swarm size (grid step)	Performance and score						Final score
		Training set			Testing set			
		R^2	MSE	VAF	R^2	MSE	VAF	
PSO-e-SVR	80	0.8931 (5)	0.00299 (3)	88.38 (5)	0.8331 (5)	0.01413 (3)	81.13 (5)	26
GA-e-SVR	100	0.8474 (4)	0.00496 (2)	81.05 (4)	0.8174 (4)	0.01488 (1)	77.97 (3)	18
SSA-e-SVR	80	0.8463 (2)	0.00139 (5)	80.9 (2)	0.8157 (3)	0.00415 (4)	77.88 (1)	17
GWO-e-SVR	90	0.8463 (2)	0.00139 (5)	80.89 (1)	0.8156 (2)	0.00414 (5)	77.9 (2)	17
GS-e-SVR	0.4	0.8464 (3)	0.00499 (1)	80.93 (3)	0.8154 (1)	0.01487 (2)	77.98 (4)	14

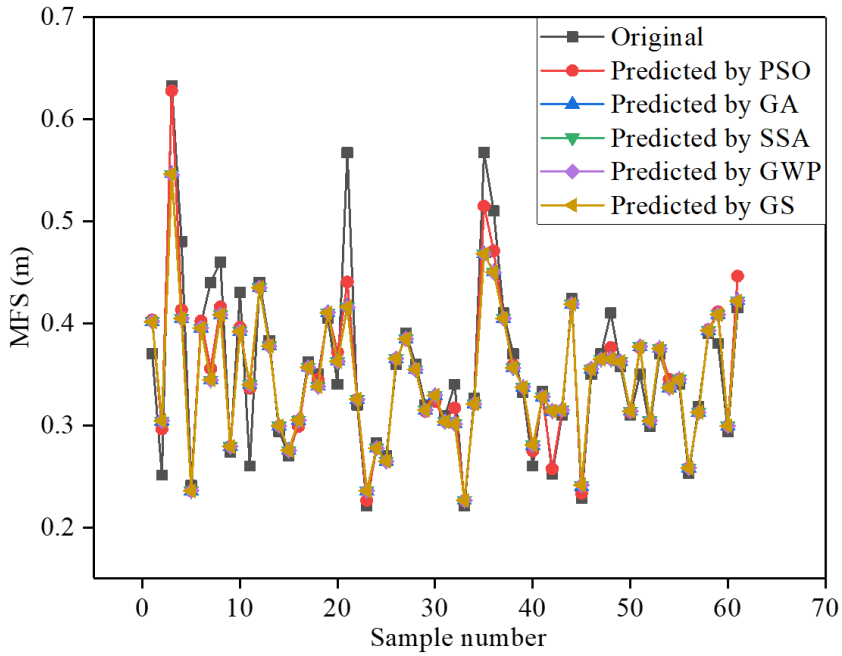
For the v-SVR-based x_{50} prediction, GWO-v-SVR procures the best comprehensive scores in Table 19. However, SSA-v-SVR is also competent because it has an excellent performance in the testing sets which means this algorithm has strong generalization ability based on the current data. Apart from that, the ability of PSO-v-SVR is worthwhile to be declared because it obtains the best scores for R^2 and VAF in the training set.

Finally, e-SVR and v-SVR-based models with the highest synthetical scores are screened and compared. Given the aforementioned discussions, it can be observed that PSO-e-SVR and GWO-v-SVR obtain the best scores, respectively and GWO-v-SVR shows better performance in four metrics, i.e. MSE of the training set, R^2, MSE and VAF of testing set. While PSO-e-SVR only outperforms in two metrics, i.e. R^2 and VAF of the training set. The GWO-v-SVR model shows superior performance in the testing set which proves that it has better generalization and robustness capabilities. For AI-based models, subtle advantages can be magnified for large-scale data. Therefore, GWO-v-SVR can be regarded as the best method for the median size prediction in this study. Corresponding swarm size and number of iterations are 100 and 500, respectively. Finally, predicted x_{50}

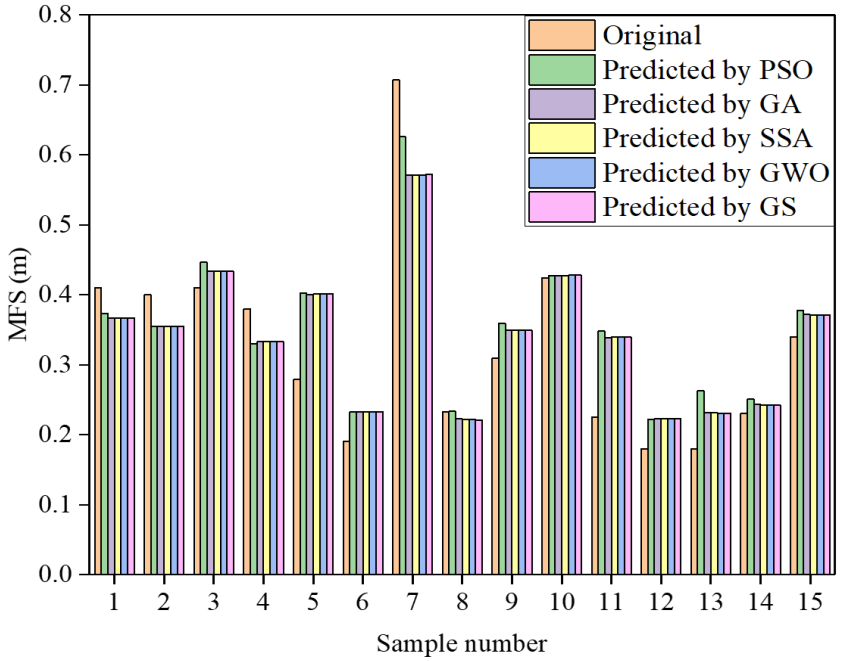
values along with their actual values are demonstrated in Figure 20 and Figure 21 for e-SVR and v-SVR-based models, respectively.

Table 19. Comparison of optimal algorithms of v-SVR.

Algorithm	Swarm size (grid step)	Performance and score						Final score
		Training set			Testing set			
		R^2	MSE	VAF	R^2	MSE	VAF	
PSO-v-SVR	60	0.8743 (5)	0.00362 (3)	85.99 (5)	0.8313 (2)	0.01313 (1)	82.35 (1)	17
GA-v-SVR	80	0.8347 (3)	0.005 (1)	80.84 (2)	0.8368 (4)	0.0125 (3)	82.44 (4)	17
SSA-v-SVR	100	0.8345 (1)	0.00139 (4)	80.81 (1)	0.8369 (5)	0.00347 (5)	82.44 (4)	20
GWO-v-SVR	100	0.8355 (4)	0.00138 (5)	80.98 (3)	0.8353 (3)	0.00348 (4)	82.41 (2)	21
GS-v-SVR	0.8	0.8347 (3)	0.00488 (2)	81.25 (4)	0.8298 (1)	0.01256 (2)	82.49 (5)	17

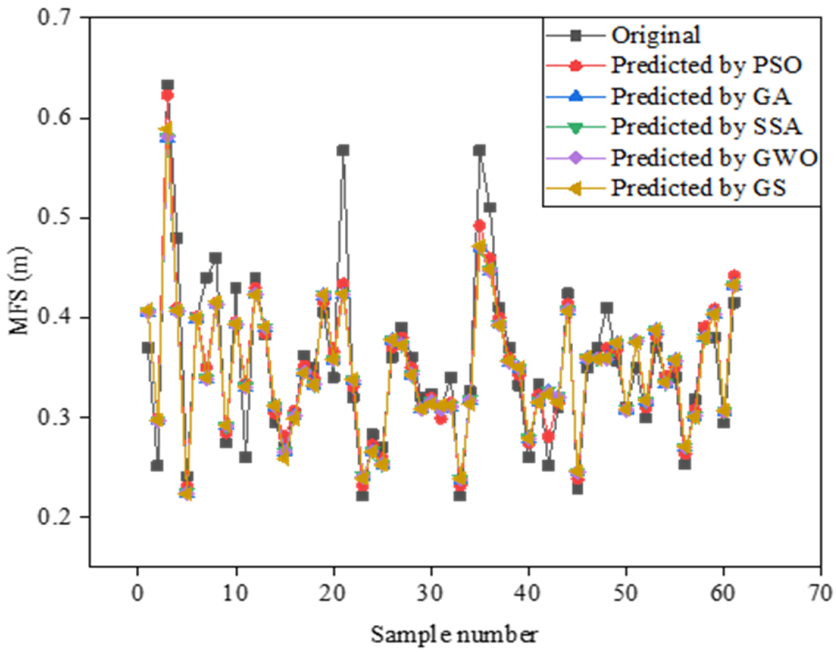


(a) training set

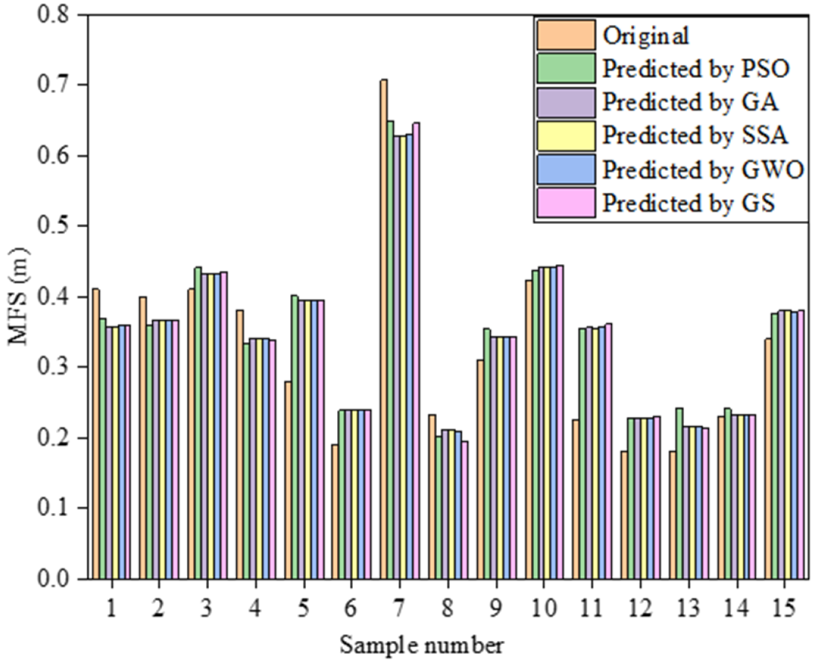


(b) testing set

Figure 20. Predicted results using e-SVR-based models: (a) training set, (b) testing set.



(a) training set



(b) testing set

Figure 21. Predicted results using v-SVR-based models: (a) training set, (b) testing set.

2.6 Sensitivity analysis

For identifying and comparing the sensitivity of different influential factors on x_{50}, the cosine amplitude method proposed by (Yang and Zhang, 1997) is utilized. Each input and output are considered as a column matrix. Then total of 20 column matrices are obtained as follows:

$$x_a = \{x_{a1}, x_{a2}, \cdots, x_{an}\} \tag{2.4}$$

where the length of each matrix is equal to the number of datasets and then the sensitivity of different influential factors on x_{50} can be analyzed by:

$$S_{ab} = \frac{\sum_{n=1}^{76} x_{an} x_{bn}}{\sqrt{\sum_{n=1}^{76} x_{an}^2} \sqrt{\sum_{n=1}^{76} x_{bn}^2}} \tag{2.5}$$

where x_{an} represents one kind of input variable; x_{bn} represents the output variable; and n represents the sequence number of the data, and for each variable, total of 76 groups of data are evaluated.

By implementing the cosine amplitude method, the relative strength of effects (RSE) of each variable can be evaluated. It is found that the most sensitive factor is *UCS* for x_{50}, as shown in Figure 22. This result is reasonable because when the violent blasting shock wave occurs, the dynamic stress on the rock will also increase sharply and the strength of the rock itself counteracts this effect. Therefore, this indicates that *UCS* has a positive influence on the fragment size. The significance of *UCS* on the fragment size has also been reported in some previous researches (Jug et al., 2017; Nainggolan et al., 2018).

The sensitivity of each factor not less than 0.8 indicates that selected factors indeed contribute to x_{50}. Finally, the sensitivity of different parameters on blasting the median size based on available data can be sorted in ascending order as: *Qe*, *J*, *J.B*, *L.Wd*, *H*, *L*, *ST*, *H.B*, *NH*, *De*, *ST.B*, *B*, *S*, *PF*, *B.D*, *Wd*, *D*, *S.B* and *UCS*. It is noted that different data would bring different results. However, current analysis results could provide some references for blasting designs with similar conditions. Interestingly, the powder factor that plays a relevant role in fragmentation does not have the largest influence on the median size. The reason for this may be due to: at first, other factors have more directly control how energy is distributed and how cracks propagate, such as rock mass structure, drilling parameters and rock mechanical properties. Secondly, RSE of power factor is obtained based on the current dataset and mathematical function. Therefore, this value doesn't consider the internal blasting mechanism and deviation induced by small dataset.

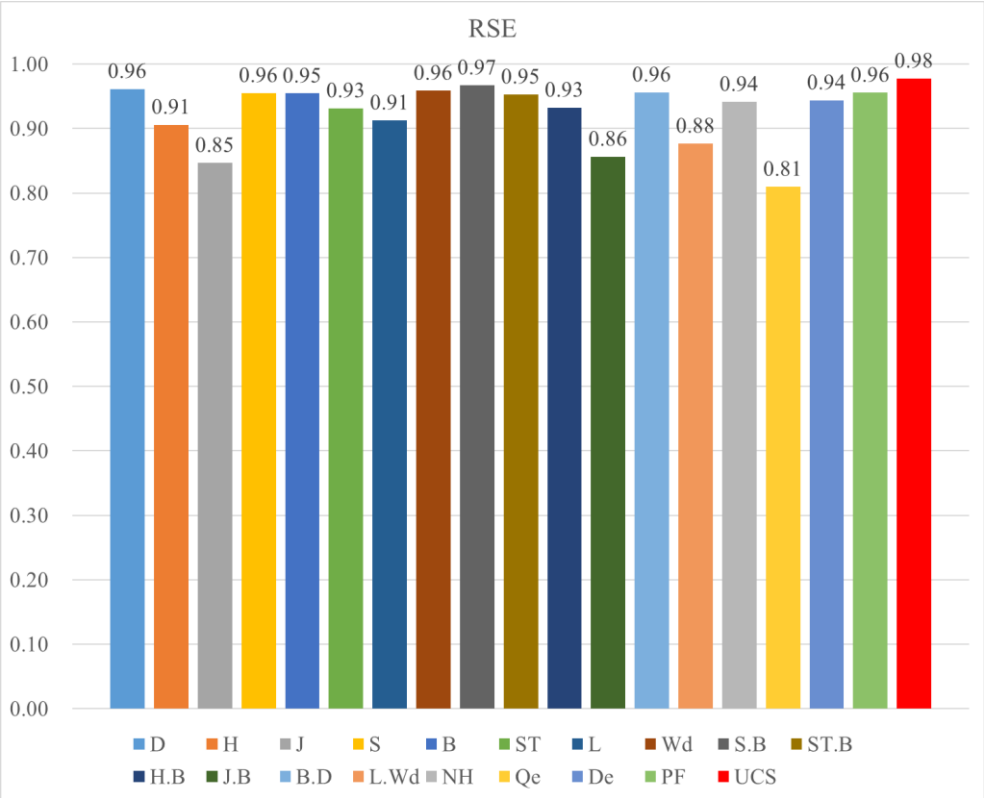


Figure 22. Sensitivity analysis of different factors on the median size.

2.7 Limitation

By utilizing SVR as the predominant strategy to predict the median size, satisfactory prediction accuracies are procured. However, there are still some drawbacks and limitations that need to be improved in future work. Firstly, the scale of data used to establish the evaluation models is still small and only a several dozens of samples are collected. It tends to be easy to mine more useful information from larger datasets and thus the performance of prediction models remains to be strengthened. Secondly, deeply analyzing blasting mechanisms is significant for mining more related factors of the median size. Thirdly, more advanced meta-heuristic algorithms are worthwhile to be combined with SVR prediction models to improve the prediction accuracy. Apart from that, the potential of some other powerful regression tools such as random forest (Zhou et al., 2020), extreme gradient boosting (Li et al., 2023b), and gradient boost machine (Natekin and Knoll, 2013) are not investigated and compared in this study. Despite that a sensitivity analysis provides the relative contribution of each parameter (predictor) in the model, the physical meaning of each parameter in the final fragmentation is unknown.

2.8 Conclusions

The prediction and evaluation of the median size are always influenced by many factors and the relationship between these factors and fragment size is elusive. Therefore, it is hard to be evaluated by means of a certain function. AI-based techniques can simulate sophisticated relationships between influential factors and output targets compared to the

conventional functions. In this study, a group of blasting fragment datasets which contains 19 types of influential factors are adopted. To control and estimate x_{50}, two types of SVR-based techniques, i.e. v-SVR and e-SVR, are employed. Then, five different types of optimal algorithms are combined with SVR to optimize the hyper-parameters. Three types of mathematical indices combined with a comprehensive ranking system are utilized to evaluate the model performance. As a result, it is found that GWO-v-SVR obtains the most comprehensive prediction performance with R^2, MSE and VAF values of 0.8355, 0.00138 and 80.98, respectively for the training set, and 0.8353, 0.00348 and 82.41, respectively for the testing set.

According to the sensitivity analysis results, it is found that the UCS plays the most important role in influencing the median size. Compared with other AI-based studies, this research takes a higher number of influential factors into consideration. On the one hand, the prediction results can provide significant guidance for complicated blasting design. On the other hand, the feasibility of AI-based techniques is validated again in open-pit mining fields. Finally, this study offers a more accurate prediction approach for multi-parameter coupling blasting evaluation of the median size compared with the original study (Kumar Sharma and Rai, 2017).

Chapter 3. Analysis and modelling of gas relative permeability in reservoir by hybrid KELM methods

Nomenclature

AGS (Average grain size)	LIGHTGBM (Light gradient boosting machine)
ANFIS (Adaptive neuro-fuzzy inference system)	LM (Levenberg–Marquardt)
ANN (Artificial neural network)	LSSVM (Least square support vector machine)
APBD (Absolute permeability before desalination)	MAE (Mean absolute error)
BK (Bulk density)	MSE (Mean squared error)
BOA (Butterfly optimization algorithm)	MFR (Microspherical focused resistivity)
CALI (Caliper logs)	MLPNN (Multilayer perception neural network)
CALO (Calorimetry)	MVO (Multi-verse optimizer)
COA (Cuckoo optimization algorithm)	NNAW (Neural network adaptive wavelet)
COLOR (Color)	NT (Neutron)
CP (Core porosity)	NTP (Neutron porosity)
CWT (Time of compression wave travel)	P (Porosity)
CWV (Velocity of compression wave)	PBD (Porosity before desalination)
DEN (Density)	PC (Principal component)
DENC (Density correction)	PCA (Principal component analysis)
DENL (Density log)	PEF (Photoelectric log)
DENTR (Density tool reading)	R^2 (Coefficient of determination)
DEP (Depth)	RF (Random forest)
DEPH (Depth horizon)	RL (Resistivity log)
DEPI (Depth interval)	RMSE (Root mean squared error)
DR (Deep resistivity)	RVR (Relevance vector regression)
ELM (Extreme learning machine)	SaDE (Self-adaptive differential evolution)
ER (Electrical resistivity)	SC (Salt concentration)
FBD (Formation bottom depth)	Sg (Gas saturation)
FIS (Fuzzy interface system)	SGB (Stochastic Gradient Boosting)
FR (Focused resistivity)	Sgc (Critical gas saturation)
FTD (Formation top depth)	SOP (Secondary porosity)
GA (Genetic algorithm)	Sorg (Residual oil saturation)
GI (Overall evaluation index)	SSD (Social ski-driver)
GJO (Golden jackal optimization)	STT (Sonic transit time)
GR (Gamma ray)	SVM (Support vector machine)
GWO (Grey wolf optimization)	Swc (Connate water saturation)
HGAPSO (Hybrid genetic algorithm and particle swarm optimization)	TOP (Total porosity)
HHO (Harris hawk's optimization)	TRR (True resistance)
K (Derived core permeability)	TSA (Tunicate swarm algorithm)
KELM (Kernel extreme learning machine)	VAF (Variance accounted for)
Krg (Relative gas permeability)	WS (Water saturation)

3.1 Introduction

Accurate determination of the petrophysics of reservoirs rocks is crucial for mineral exploration, engineering construction, geohazard assessment, and environmental protection. Among these properties, flow-related phenomena play significant roles in porous media since the properties offer valuable insights into rock composition, strength, stability, and permeability, enabling decision makers and professionals to devise suitable strategies (Ghassemi, 2012; Hu and Huang, 2017; Regnet et al., 2019; Sander et al., 2017; Zhang et al., 2022). However, the intricate porous structure of rocks introduces complexity and diversity in the flow behaviour of fluids such as water, oil, and gas within them. Consequently, further research and analysis is imperative for comprehending and predicting fluid behaviour in rocks. Permeability is one of the key physical properties among various fluid characteristics (Li et al., 2020; Marathe et al., 2012; Qiao et al., 2022; Tian et al., 2021; Xu et al., 2022). It represents the ability of a fluid to flow through the interconnected system of pores and fractures within a rock or formation, providing insights into the rock's resistance to fluid flow (Ahmed et al., 1991). Increased permeability in a reservoir leads to a greater rate of hydrocarbon flow production from the reservoir. One the one hand, the permeability is closely linked to the volume of fluid that can be contained within the pore space and thus possesses the ability to flow. On the other hand, understanding the fluctuations of this parameter throughout the reservoir area is crucial for identifying the best placement for production wells and selecting the most effective trajectory for them. In the practical engineering, the interaction between rocks and fluid such as connate water saturation, residual oil saturation, critical gas saturation, porosity and other factors linked with well conditions or formation damage increase the difficulty of the determination.

3.2 Literature review

Traditionally, scholars have conducted laboratory experiments to measure the relative permeability of porous rocks (Honarpour and Mahmood, 1988). For example, the steady-state method consists of introducing two immiscible phases into a porous medium sample at a constant rate, allowing for one-dimensional flow until the pressure drop and fractional flow reach a stable state (Esmaeili et al., 2019). The effective permeability of each phase is determined using Darcy's law, based on the measured pressure drop and flow rates. In contrast, the unsteady-state method employs a displacement technique whereby a continuous injection of a displacing fluid at a constant rate takes place within an initially saturated core sample of the displaced fluid. The fluctuating dynamics of pressure drop and the resulting volumes of the phases are periodically monitored and recorded. Furthermore, the Capillary pressure approach and Centrifuge approach are also utilized to assess permeability (Honarpour et al., 2018). While laboratory measurements are widely regarded as the most direct and accurate method for determining permeability, they still possess certain limitations that can result in significant errors. One such limitation arises from the preparation and cutting process of the test sample, which has the potential to disrupt the original pore structure of the rock and adversely impact the accuracy of permeability measurements.

Additionally, ensuring a completely sealed boundary for the rock during the experiment poses a significant challenge, as any leakage or bypassing of the flow channel can introduce errors. Moreover, the experimental approach is often time-consuming and costly. The intricate procedures involved in sample preparation, data collection, and analysis contribute to lengthy experimental processes. Furthermore, variations in experimental techniques among different researchers can yield inconsistent results, further exacerbating measurement errors.

Moreover, some empirical, semi-theoretical functions have been developed to the determination of the permeability of a reservoir rock. For instance, Jorgensen (1991) proposed an empirical function to predict the permeability of sandstone in which the porosity is considered as the main parameter to indicate the permeability. The porosity needs to be obtained from borehole-geophysical logs. By the analysis of 155 laboratory experiments of sandstone samples, Timur (1968) developed an empirical function considering porosity and residual fluid saturation. With the aid of support of core and log studies, Coates and Dumanoir (1973) related irreducible water saturation to intrinsic permeability, and also explored the application of the proposed technique on non-irreducible conditions. Aigbedion (2007) presented a correlation function between the logarithmic permeability and porosity. These functions can provide fast evaluation for permeability. However, it can be found that they can only integrate a few potential factors related to the permeability and fail to achieve desirable performance for complicated conditions.

In the past few years, artificial intelligence is undergoing rapid development, and machine learning methods have demonstrated successful applications in various geotechnical fields, yielding promising outcomes (Anifowose et al., 2017; Zhou et al., 2023). These techniques possess the capability to handle intricate data structure, uncover concealed patterns and correlations, and detect non-linear relationships and complex patterns within the data. Furthermore, machine learning methods offer automation and optimization of tasks, thereby enhancing efficiency and reducing human resource costs. Therefore, machine learning seems to be a good alternative for modelling reservoir rock characteristics. For instance, Gholami et al. (2012) applied support vector machine (SVM) to predict the horizontal permeability in three wells based on 175 digital well log and core log data. The comprehensive predictive capability of SVM was compared with general regression neural network (GRNN). It can be found that SVM presented faster modelling speed and higher prediction accuracy. Ahmadi (2015) proposed to utilize LSSVM (Least Squares Support Vector Machine) and genetic algorithm to predict relative gas/oil permeability. The high-performing and low-uncertainty of LSSVM model demonstrated remarkable capability in predicting gas/oil relative permeability in petroleum reservoirs. Erofeev et al. (2019) employed conventional core analysis results along with data from production layer coring depths and top and bottom depths. Seven machine learning techniques were utilized to predict salinity in the characterized cores, as well as the porosity and permeability of the reservoir after desalination caused by drilling mud or water injection.

Recently, Kumar et al. (2022) developed an XG-Boost-based modelling method to forecast two-phase oil/water relative permeability. In this study, more than one thousand data points were collected from published literature and considered the influence of

temperatures. Two rock formations were considered, i.e., unconsolidated sand and sandstone. The success of the XG-Boost approach proposed more alternatives for modelling reservoir properties. Mahdaviara et al. (2020) employed different machine learning techniques to test predictive ability for the relative permeability of the gas. It could be found that multilayer perceptron with Levenberg-Marquardt Algorithm (LMA) optimized approach surpassed other algorithms and five traditional literature methods. Song et al. (2021) recognized the significant influential factors for the two-phase relative permeability. This is achieved by conducting microfluidic experiments and image recognition for saturation extraction. And then the deep neural network was used to predict the two-phase permeability. It could be indicated that the prediction performance of the deep neural network is lower than 0.05 for mean squared error (MSE). Moreover, Nazari and Hajizadeh (2023) proposed a relevance vector regression optimized by grey wolf optimization algorithm to forecast permeability based on different well logs and eight log parameters, such as caliper logs, computed gamma ray, density correction and so on. It could be found that optimized relevance vector regression produced better prediction performance than relevance vector regression. Matinkia et al. (2023) presented a new multilayer perception neural network to predict permeability in which social ski-driver algorithm was used to tune the neural network and achieved the highest coefficient of determination (0.9928) compared with genetic algorithm and particle swarm optimization. More work about the prediction of permeability can be seen in Table 20 where the authors, the inputs, the methods, the dataset number, and main prediction performance have been presented.

Research gap: Machine learning has strong predictive capability and the mechanism of the permeability of reservoir is complicated. Therefore, employing machine learning techniques to predict permeability seems to be promising. However, the prediction potential of kernel extreme learning machine (KELM) on the permeability is rarely investigated.

Regarding this, this section employed kernel extreme learning machine to be the benchmark to develop the prediction models for predicting the gas relative permeability in reservoir. Besides, five novel meta-heuristic algorithms were employed including butterfly optimization algorithm (BOA), tunicate swarm algorithm (TSA), Multi-verse optimizer (MVO), Golden jackal optimization (GJO), Harris hawk's optimization (HHO) algorithm to tune the hyper-parameters in KELM. Five-fold cross validation was employed to enhance the generalization of hybrid KELM model. An extensive dataset from the experiments which contain 1024 data from the literature were taken to develop models. Four classical statistical indicators were used to measure the model performance, i.e., root mean squared error (RMSE), coefficient of determination (R^2), variance accounted for (VAF) and mean absolute error (MAE). In addition, two comprehensive manners, overall evaluation index (GI) and Taylor Diagram, were evaluated to provide overall model assessments. Therefore, five new methods were proposed to predict the gas relative permeability and then the mutual information was calculated to measure the influence of potential factors on the prediction model uncertainty. The most influential factor was selected and used as the only input to re-evaluate the reliability of KELM-based models. Finally, a few classical machine learning algorithms were also used and compared with the hybrid KELM algorithms. The main novelty of this chapter is that five

new machine learning methods were proposed to predict the gas relative permeability and performed better than other empirical or machine learning techniques.

3.3 Data analysis and pre-processing

To model and analyse the relative gas permeability (Krg), a broad experimental dataset contains rock properties and flow characteristics were utilized from the published literature presented by Seyyedattar et al. (2022). In this dataset, connate water saturation (Swc), residual oil saturation (Sorg), critical gas saturation (Sgc), derived core permeability (K), porosity (P), formation type and gas saturation (Sg) are considered as potential influential factors for gas permeability. Among these factors, formation type is considered as discrete variable and labelled to 1, 2 and 3 for sandstone, limestone, and dolomite, respectively. The other factors are continuous variables. To describe the variable distribution, the rock formation type is demonstrated in Figure 23 by a pie chart. Other influential factors are demonstrated by boxplots as seen in Figure 24. Some key statistical indicators for inputs and an output have been presented in Table 21 (Abad et al., 2022). To develop the hybrid KELM models, the whole dataset is divided into a training set and a testing set where 80% of the data (820 data) is used as training model and the remaining 20% (204 data) is used as testing model performance (Yu et al., 2021b).

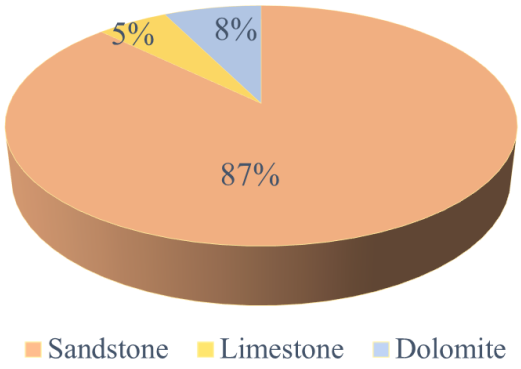


Figure 23. Rock formation distribution.

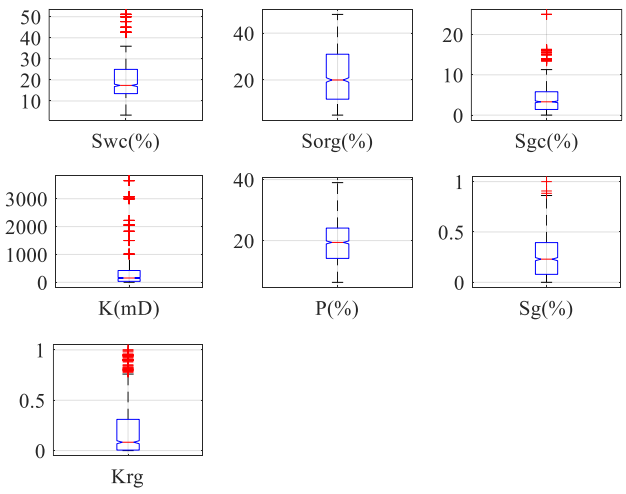


Figure 24. General data distribution by boxplots.

Table 20. The main work about the permeability prediction during the last ten years.

Authors	Method	Inputs	Data No.	R^2
Ahmadi et al. (2014)	GA-FIS	DENTR, STT, TOP, BK	1000	0.9627
	LSSVM	DENTR, STT, TOP, BK	1000	0.994
Ahmadi (2015)	GA-LSSVM	Swc, Sorg, Sgc, K, P, formation, Sg	1024	0.9941
Shokooh Saljooghi and Hezarkhani (2015)	NNAW	DEP, ER, GR, WS, TOP, formation, SOP	270	0.90
Elkatatny et al. (2018)	ANN	BK, NTP, mobility index	1223	0.95
Moussa et al. (2018)	SaDE-ANN	MFR, DR, NTP, BK, GR	743	0.979
Ahmadi and Chen (2019)	HGAPSO-LSSVM	STT, DENTR, BK, TOP,	1000	0.9966
Erofeev et al. (2019)	SVR	SC, FTD, FBD, PBD, APBD, DEP, DEN, AGS, COLOR, DEPH	102	0.856
Urang et al. (2020)	ANN	CP, K, NT, DENL, WS	1199	0.9753
Subasi et al. (2022)	SGB	GR, ER, DEN, NT, DEP.	1140	0.7625
Subasi et al. (2022)	SGB	GR, ER, DEN, NT, DEP.	2600	0.6332
Okon et al. (2021)	LM-ANN	DEPI, GR, DENL, RL	955	0.9624
Matinkia et al. (2023)	SSD-MLPNN	DEP, CALO, corrected GR, PEF, NTP, DEN, TRR, CWT	1614	0.9928
Nazari and Hajizadeh (2023)	RVR-GWO	CALI, computed GR, DENC, CWV, NT, PEF, DEN, spectral GR	2506	0.9215
Sheykhinasab et al. (2023)	COA-ELM	CALI, DEP, DENL, NTP, PEF, spectral GR, CWT, FR	1269	0.9931

Note: AGS - Average grain size; ANN – Artificial neural network; APBD - Absolute permeability before desalination; BK - Bulk density; CALI - Caliper logs; CALO – Calorimetry; COA - Cuckoo optimization algorithm; COLOR – Color; CP - Core porosity; CWV - Velocity of compression wave; CWT - Time of compression wave travel; DEN – Density; DENC – Density correction; DENL – Density log; DENTR - Density tool reading; DEP – Depth; DEPH - Density correction; DEPI - Depth interval; DR - Deep resistivity; ELM - Extreme learning machine; ER - Electrical resistivity; FBD - Formation bottom depth; FIS – Fuzzy inference system; FR – Focused resistivity; FTD - Formation top depth; GA – Genetic algorithm; GR - Gamma ray; GWO – Grey wolf optimization; HGAPSO - Hybrid genetic algorithm and particle swarm optimization; K - Derived core permeability; LM - Levenberg–Marquardt; LSSVM - Least square support vector machine; MFR - Microspherical focused resistivity; MLPNN - Multilayer perception neural network; NNAW - Neural network adaptive wavelet; NT – Neutron; NTP - Neutron porosity; P – Porosity; PBD - Porosity before desalination; PEF - Photoelectric log; RL – Resistivity log; RVR - Relevance vector regression; SaDE - Self-adaptive differential evolution; Sg - Gas saturation; Sgc - Critical gas saturation; SC - Salt concentration; SGB - Stochastic Gradient Boosting; Sorg - Residual oil saturation; SOP - Secondary porosity; SSD - Social ski-driver; STT - Sonic transit time; SVR – Support vector regression; Swc - Connate water saturation; TOP - Total porosity; TRR - True resistance; WS – Water saturation;

Table 21. Key statistical indicators for inputs and the output.

Statistical indicators	Swc (%)	Sorg (%)	Sgc (%)	K (mD)	P (%)	Sg (%)	Krg
Minimum	3.3	5	0	1.48	6.3	0	0
Maximum	51.1	48	25	3650	39	1	1
Standard deviation	10.03	10.63	4.03	655.75	7.17	0.21	0.26
Average	20.03	21.55	4.36	386.70	19.52	0.26	0.20
Skewness	1.15	0.19	1.69	3.06	0.38	0.73	1.41
Kurtosis	4.18	1.88	6.51	12.65	2.53	2.85	3.99
Median	17.40	20	3.30	154.36	19.40	0.2300	0.08
Mode	29.90	30	1	1.48	16	0	0

To measure the correlation between potential influential factors and gas permeability, the non-linear "Distance Correlation" was employed (Székely and Rizzo, 2014, 2009). According to the analysis from "Distance Correlation" shown in Figure 25, it can be found that Sg is the most correlated. In addition, Swc and Sorg present higher correlation compared with other influential factors. As for Sgc, K and P, they only have very slight correlation with gas permeability. To warrant a justification for the trade-off between dimensionality reduction and the loss of interpretability, the principal component analysis (PCA) is conducted to reduce the complexity of input data (Bro and Smilde, 2014) for the training data. PCA is a technique used to reduce the number of dimensions in a large dataset, while preserving as much of the original variation in the data as possible. This is done by identifying and extracting the underlying structure or patterns in the data using linear algebra. The main idea of PCA is to find a new coordinate system that represents the most important information in the data. The first coordinate axis corresponds to the direction of greatest variance, and the second axis corresponds to the direction of the second-greatest variance, and so on. These new axes are called principal components (PC), and their number is equal to the number of dimensions in the original data. By projecting the data onto these principal components, PCA transforms the original high-dimensional data into a lower-dimensional space that captures most of the important information. In this section, 95% is set to be threshold to capture the most important information.

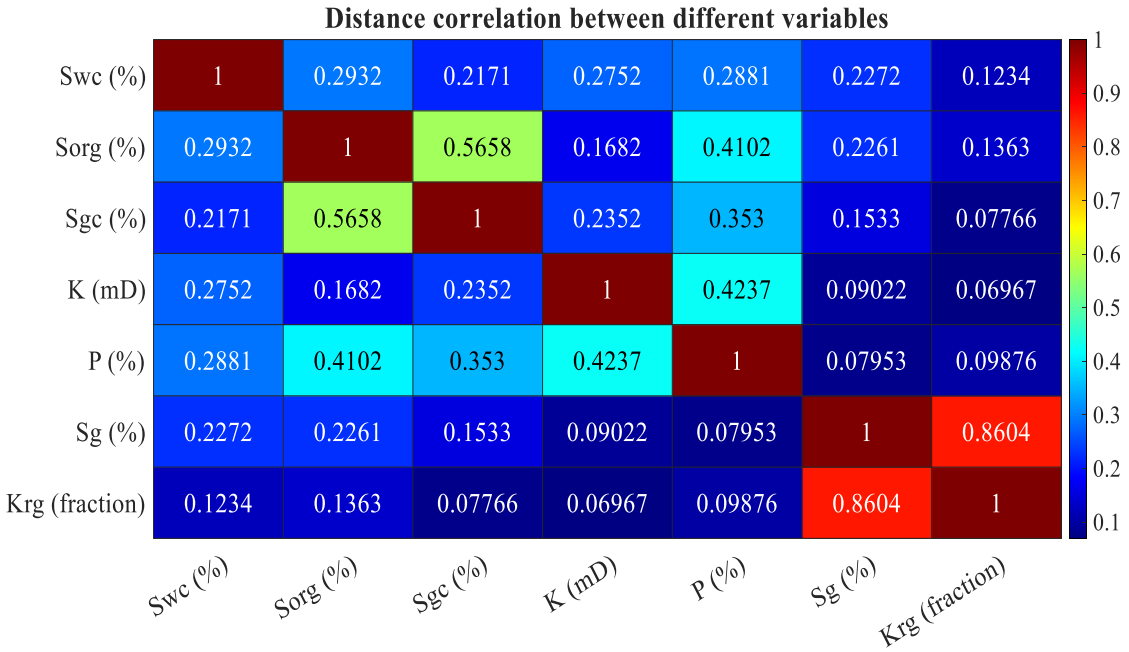


Figure 25. Distance Correlation between different variables.

The PCA results can be found in Table 22, where the first six PCs are retained and used as inputs. The explained percent of six PCs is 28.07, 23.10, 16.94, 11.22, 10.06, and 5.89, respectively. The variance of six PCs is 0.09, 0.08, 0.06, 0.04, 0.03 and 0.02, respectively. In addition, the eigenvalues of covariance matrix (latent) of each PC have been listed. It can be found that the most significance PC corresponds to the highest eigenvalue. In addition, to ensure that features are on a similar scale, preventing features with larger values from dominating those with smaller scales, the normalization was processed for the input data according the function (3.1), and then all inputs were normalized into to the scale [0, 1]. Where $input_{scale}$ represents the input after normalization, $input_{min}$ and $input_{max}$ represent the minimum and maximum value of this input, respectively and *input* denotes the original input.

$$input_{scale} = \frac{input - input_{min}}{input_{max} - input_{min}} \tag{3.1}$$

Table 22. Main information from PCA results.

Principal component (PC)	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7
Explained percent	28.10	22.92	17.28	11.37	9.76	5.93	4.64
Cumulative percent	28.10	51.02	68.31	79.67	89.43	95.36	100.00
Latent	0.09	0.08	0.06	0.04	0.03	0.02	0.02

3.4 Proposed Methodology

To develop the prediction models for evaluating the gas relative permeability in reservoir, this study employed KELM to be the benchmark. And five meta-heuristic algorithms were employed including BOA, TSA, MVO, GJO, HHO algorithm to tune the hyper-parameters in KELM. Meanwhile, a five-fold cross validation was used to enhance the model reliability and robustness. The significance of application of these techniques in this study includes: at first, the prediction ability of KELM on the has relative permeability is rarely studied. Secondly, the employment of five optimization algorithms can achieve semi-automatic adjustment of hyper-parameters in KELM. Thirdly, the cross validation reduces the possibility of over-fitting and increase the robustness of prediction models. In the next section, some brief introductions about these techniques would be given. Meanwhile, a general flow of this study is shown in Figure 26.

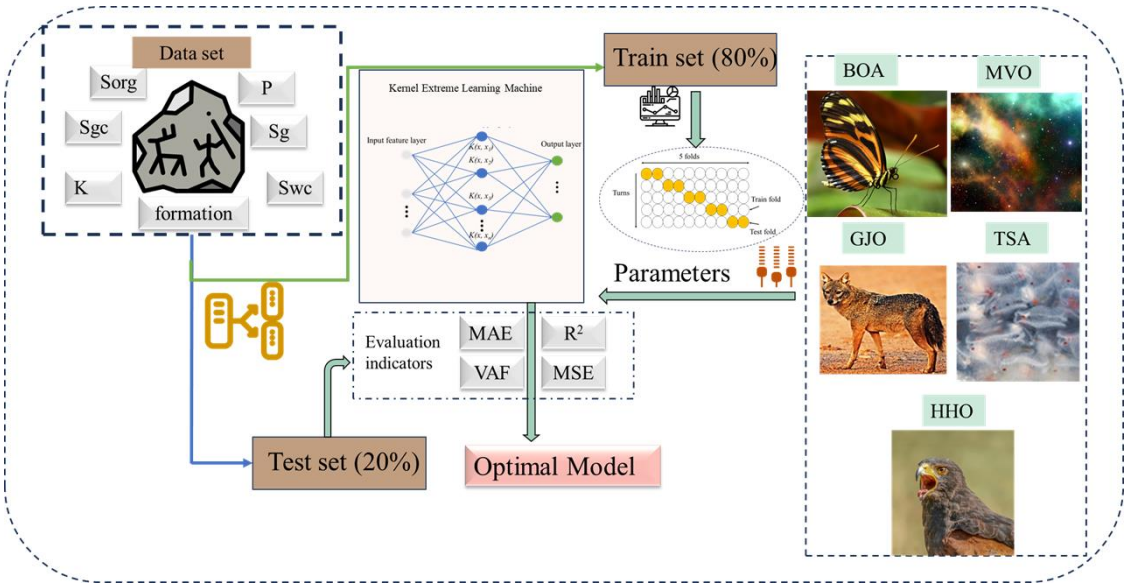


Figure 26. General working framework of intelligent prediction of the gas relative permeability.

3.5 Model development and evaluation metrics

To assess the gas permeability, hybrid KELM techniques were employed to model intelligent prediction models, i.e., HHO-KELM, GJO-KELM, MVO-KELM, TSA-KELM and BOA-KELM. Five meta-heuristic algorithms were used to tune the hyper-parameters in KELM. Generally, swarm-based meta-heuristic algorithms relied on two key parameters, namely swarm size and iteration number, to optimize performance. After conducting some experiments and comparisons, it was observed that an iteration number of 200 offered stable optimization and less optimization time. When the iteration number continues to increase, the optimization time would also increase. Consequently, this value was applied to each gas permeability prediction model. Similarly, extensive testing was conducted on different swarm sizes, leading to the realization that slight variations in evaluation indicators occurred. Limited by the workload, it is almost impossible to consider all combinations of swarm size and iteration number. In this study, swarm sizes

of 25, 50, 75, 100, 125, 150, 175, and 200 were employed in each method to meticulously select the most suitable size. Four classical mathematical indicators (Xie et al., 2021; Zhou et al., 2022c), namely R^2 (see Eq. (2.1)), VAF (see Eq. (2.2)), RMSE (Eq. (3.2)), and MAE (Eq. (3.3)), were utilized to assess the model performance.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{k}^{N}(y_k - \mathrm{y}'_k)^2} \tag{3.2}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{k}^{N}|y_k - \mathrm{y}'_k| \tag{3.3}$$

where achieving better performance entails lower RMSE and MAE values, as well as higher R^2 and VAF values. While for a specific model, sometimes it cannot obtain the best performance for each indicator. In order to comprehensively evaluate and compare the predictive capabilities of hybrid models, an overall evaluation index named *GI* is suggested (Zhang et al., 2023) where a larger GI indicates a better overall prediction capability. This index combines the four performance evaluation metrics as follows:

$$GI = \frac{\frac{VAF}{100}+R^2}{RMSE+MAE} \tag{3.4}$$

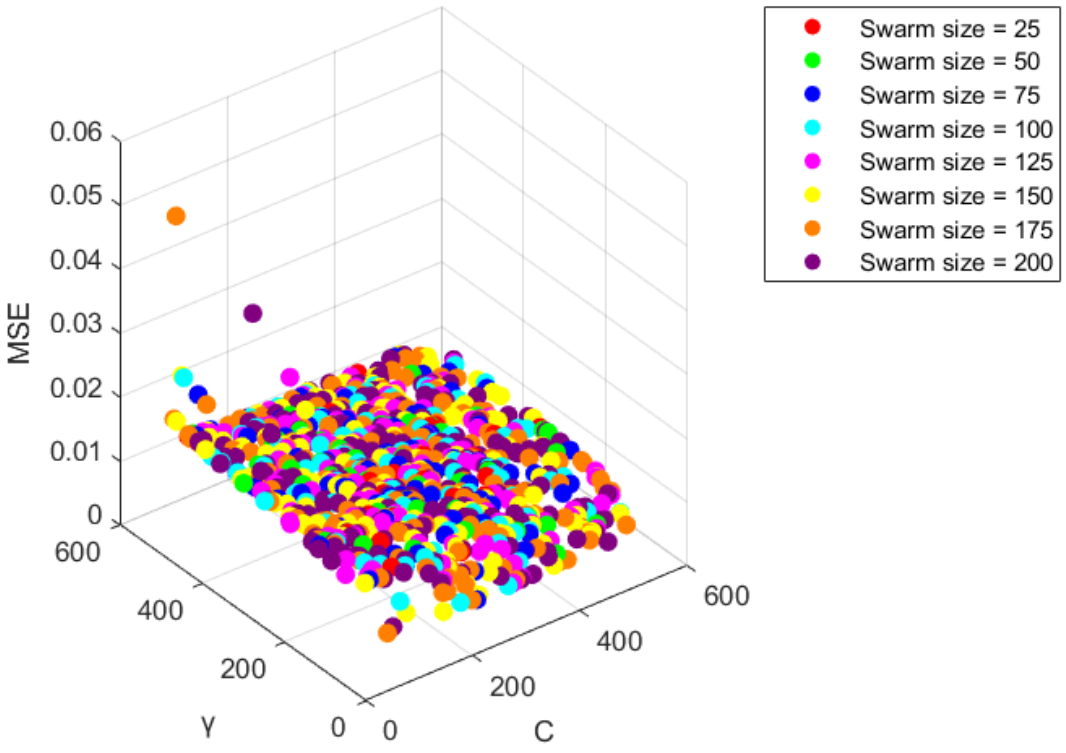


Figure 27. Initial hyper-parameters distribution and corresponding fitness value (MSE) for different swarm size.

Table 23. Best hyper-parameter pair at the initial stage and corresponding initial and final fitness from different swarm size.

Swarm size	Initial fitness ($\times 10^{-3}$)	Final fitness ($\times 10^{-3}$)					Initial C	Initial γ
		HHO	MVO	GJO	TSA	BOA		
25	13.7	4.88	4.37	4.35	4.36	4.88	211.61	46.14
50	9.94	4.91	4.47	4.35	4.36	4.88	361.25	8.16
75	8.46	4.88	4.44	4.35	4.43	4.88	207.47	1.44
100	8.58	4.88	4.36	4.35	4.38	4.46	265.82	1.44
125	8.61	4.89	4.35	4.35	4.35	4.39	276.23	2.87
150	8.67	4.88	4.36	4.35	4.38	4.37	146.32	3.07
175	8.63	4.88	4.35	4.35	4.37	4.78	199.5	3.07
200	8.64	4.88	4.36	4.35	4.37	4.49	312.01	2.87

For the same swarm size, the same hyper-parameter pair (C and γ) were generated for the fairness of comparison. The MSE value (explained in Eq. (2.2)) from five-fold cross validation was used to be the fitness value (Jack Feng et al., 2005). And then the hyper-parameters achieving the lowest fitness value were selected as the initial fitness value. All hyper-parameter pairs and their fitness can be seen in Figure 27. The initial fitness value for each swarm size and corresponding hyper-parameters can be seen in Table 23. The optimization process of five hybrid KELM-based models can be seen in Figure 94 to Figure 98 in Appendix 2. It can be found that for each method, it can achieve lower fitness value which means that the optimization process is efficient. In addition, it can be seen that the GJO-KELM, HHO-KELM and TSA-KELM produced faster convergence speed. MVO-KELM presented differential optimization process after 125^{th} iteration. BOA-KELM presented differential optimization process among the whole iteration. The detailed final fitness has been demonstrated in Table 23.

3.6 Results and Discussion

3.6.1 Results

Since the physical boundary of the output (Krg) is between 0 and 1, the predicted values over than 1 are constraint to be 1 and the ones less than 0 are to be 0. The detailed prediction performance of proposed five hybrid gas permeability as well as optimized hyper-parameters can be seen in Table 41-Table 50 in Appendix 2. All prediction scenarios can achieve R^2 larger than 0.9810 for the training set. In addition, all prediction scenarios can achieve R^2 larger than 0.9745 for the testing set which further indicated the powerful generalization of using hybrid KELM models to predict gas permeability. It can be also indicated that the prediction performance between the testing set and training set is close and thus the over-fitting phenomenon did not occur in these models.

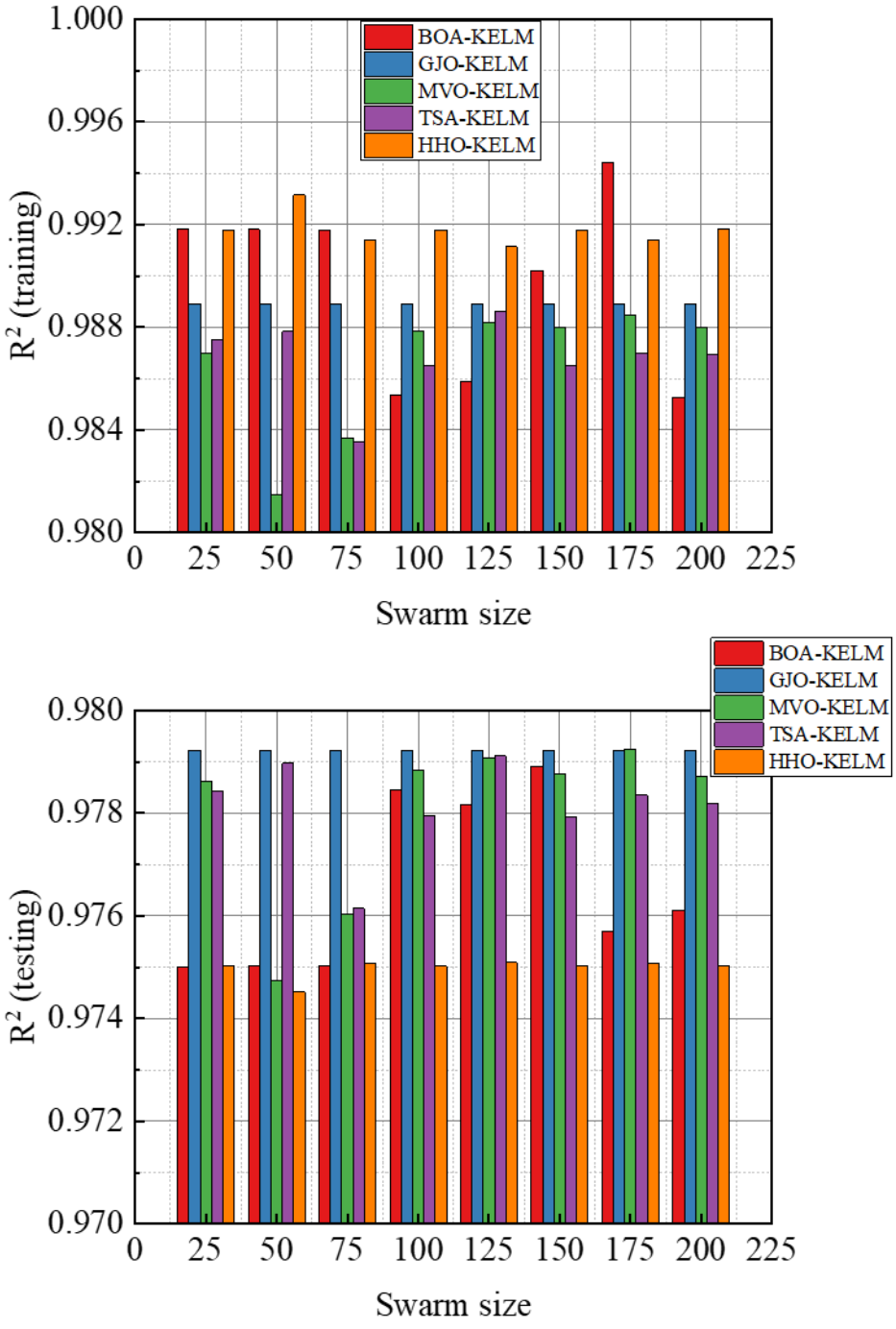


Figure 28. Comparison of R^2.

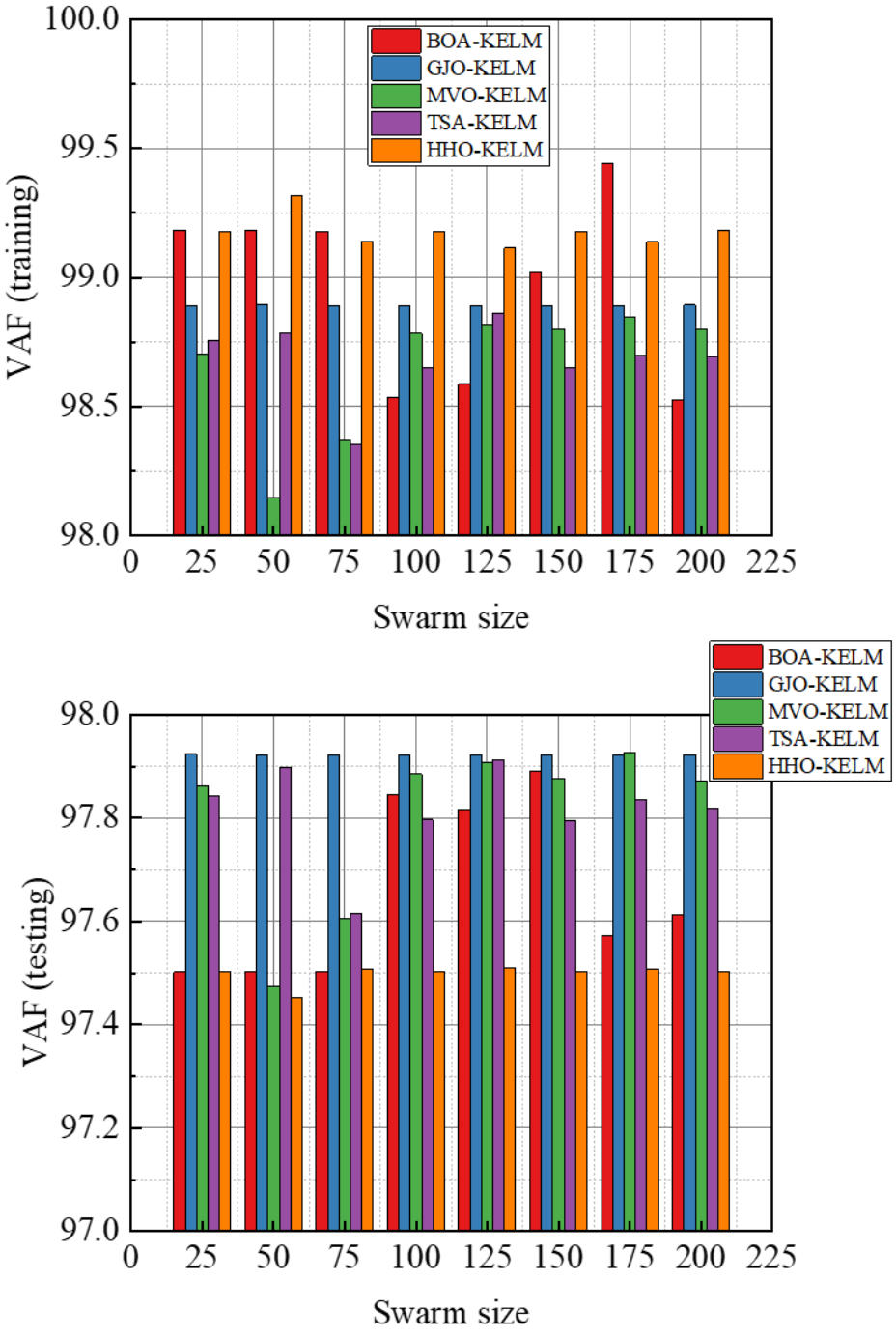


Figure 29. Comparison of VAF.

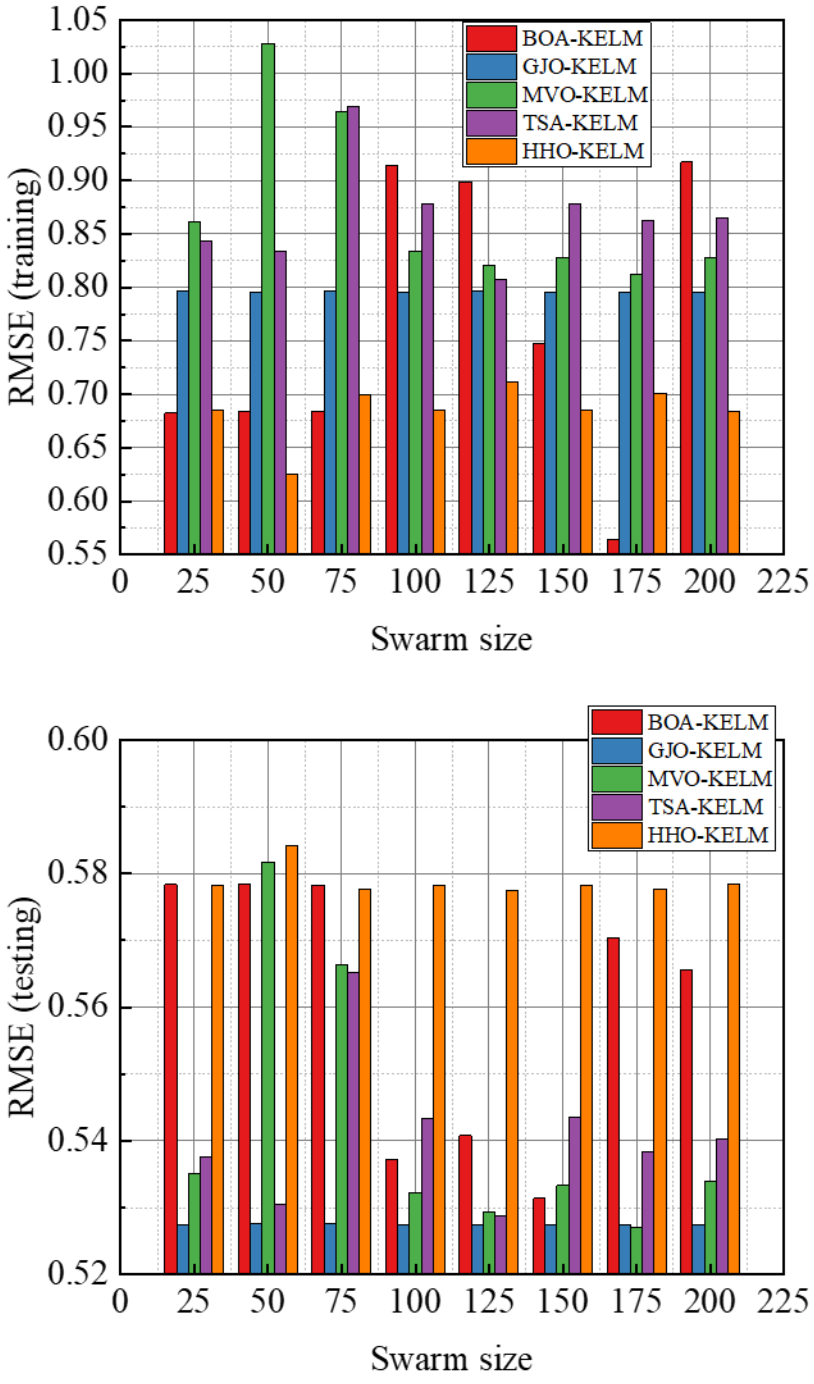


Figure 30. Comparison of RMSE.

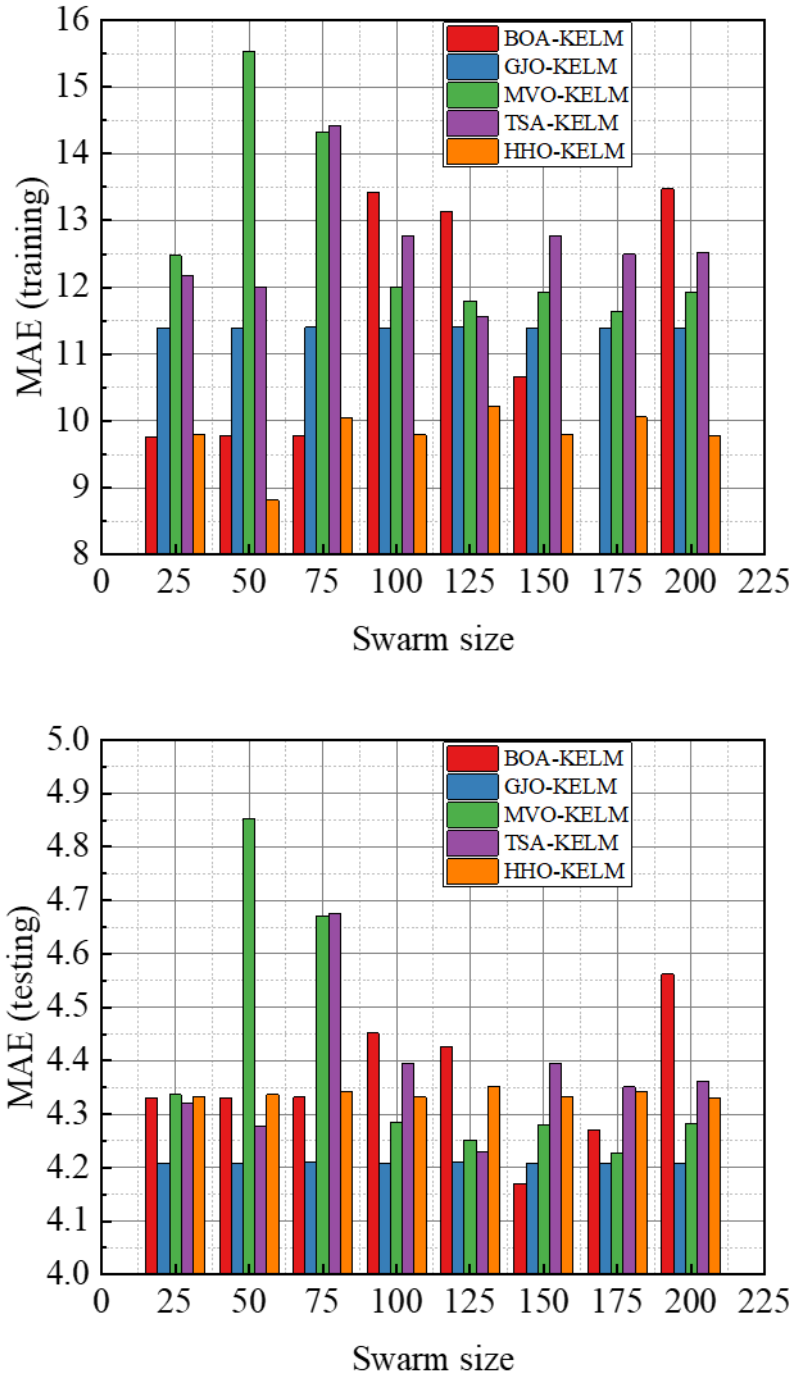


Figure 31. Comparison of MAE.

For intuitively displaying the prediction performance difference between different models, the comparison of R^2, RMSE, VAF and MAE by some bar charts has been shown in Figure 28 to Figure 31. For the training R^2, it can be seen that the best R^2 is achieved by BOA-KELM with swarm size 175. For the testing R^2, HHO-KELM generally produces lower than R^2 than other four hybrid KELM models. GJO-KELM generates competitive testing R^2 for different swarm size. The best testing R^2 is achieved by MVO-KELM equal to 0.9793 with swarm size 175. For the training VAF, the best performance is from BOA-

KELM model with swarm size 175. In addition, it can be seen that HHO-KELM model also present competitive performance for swarm size. For the testing VAF, the best value is equal to 97.9233 and produced by GJO-KELM with swarm size 25. For the training RMSE, the best performance is from BOA-KELM model with swarm size 175, i.e., 0.5646. HHO-KELM also presents desirable results for different swarm size. For the testing RMSE, it can be seen that GJO-KELM models present competitive performance for different swam size. However, the lowest RMSE is from GJO-KELM with swarm size 175 and it is equal to 0.5271. For the training MAE, the best result is from HHO-KELM method with swarm size 50 and it is equal to 8.8178. For the testing MAE, the BOA-KELM model with swarm size 150 performs the best, i.e., 4.1706. For the individual evaluation metric, it can be indicated that it is produced by different KELM models. Therefore, it is hard to conclude which scenario is the best.

To measure the overall performance, the *GI* value was adopted here. Evaluation results by the *GI* value have been shown in Figure 32. It can be found that the best swarm size for each method is 175, 200, 175, 125 and 50 for BOA-KELM, GJO-KELM, MVO-KELM, TSA-KELM and HHO-KELM, respectively for the training set. It can be found that the best swarm size for each method is 150, 200, 175, 125 and 200 for BOA-KELM, GJO-KELM, MVO-KELM, TSA-KELM and HHO-KELM, respectively for the testing set. The detailed results about GI can be seen in Table 41 to Table 50 in Appendix 2. Since the testing set is more suitable to reflect the model generalization, therefore, only the performance of GI on the testing set is considered to compare the superiority of prediction models. According to the *GI* assessment on the testing set, the generalization performance ranking of the five hybrid KELM-based models from highest to lowest is as follows: BOA-KELM, GJO-KELM, MVO-KELM, TSA-KELM and HHO-KELM. The predicted results from these superior models are shown in Figure 33 to Figure 37 where the modelling results from the training set and testing set are presented, respectively. It can be seen that most data concentrate near the perfect fitting line ($y=x$) while several data is out of the fitting lines $y=1.2x$ and $y=0.8x$. Finally, it can be concluded that the BOA-KELM model with swarm size 150 would be recommended to predict the Krg. In the practical application, a dataset with the same inputs as in this study is needed. These inputs can be input into the BOA-KELM model and then the predicted Krg would be obtained and thus save a lot of experiment and measurement process.

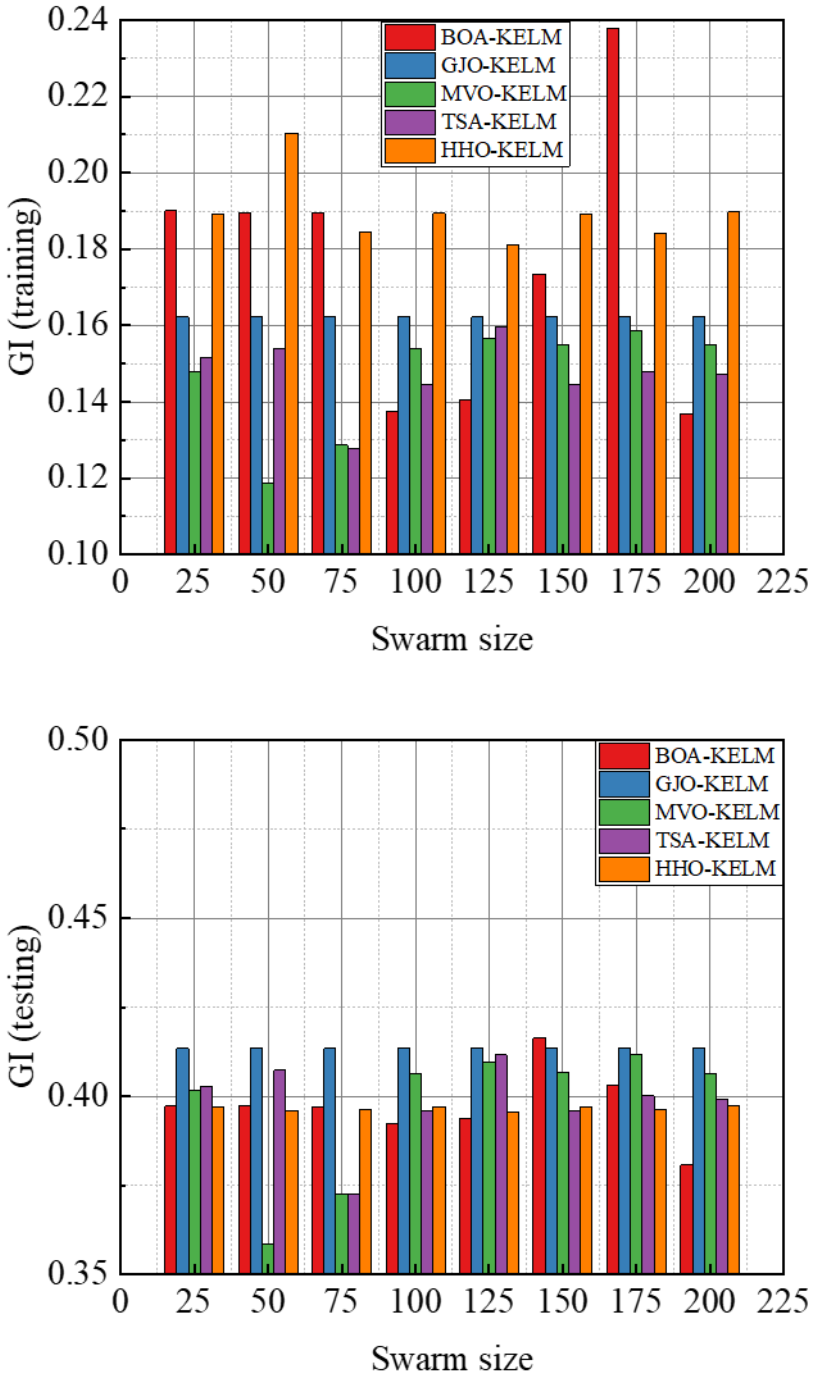


Figure 32. GI evaluation for different swarm sizes and methods.

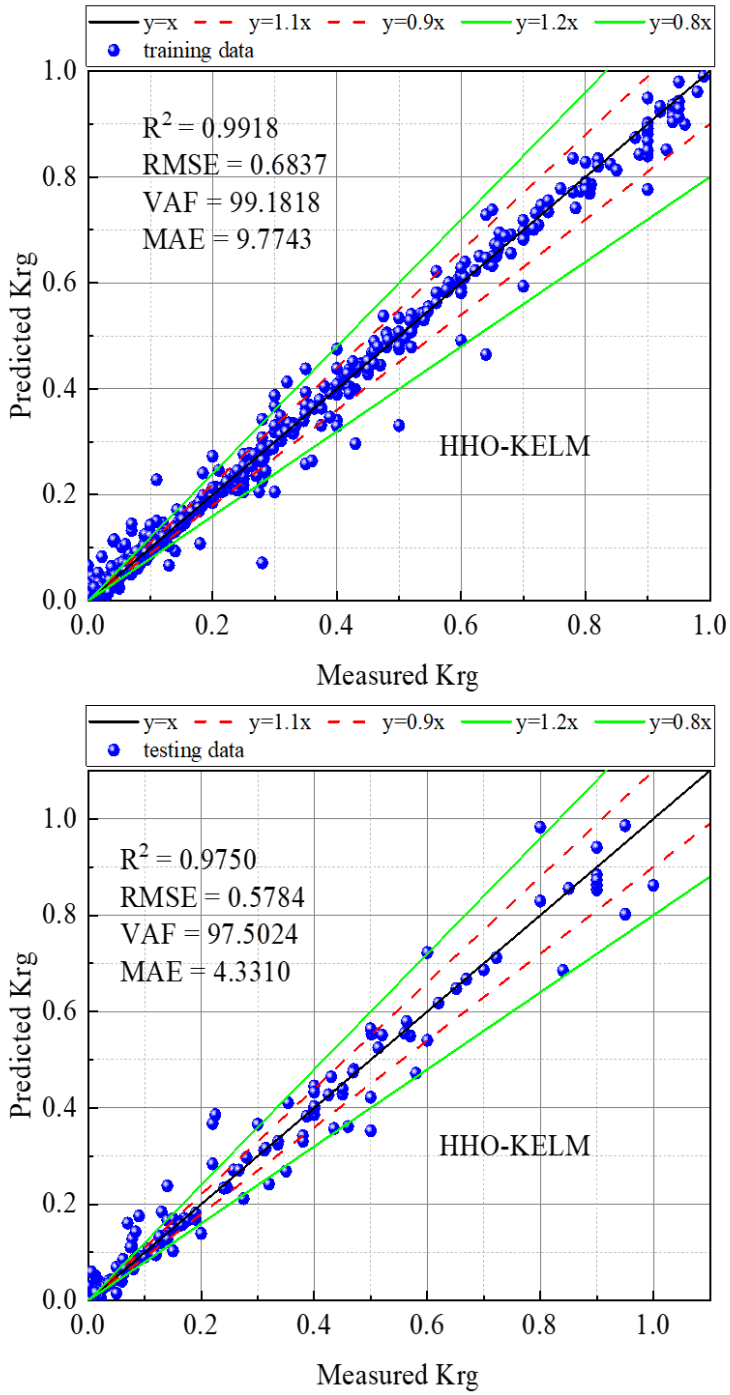


Figure 33. Comparisons between measured and predicted gas permeability by HHO-KELM.

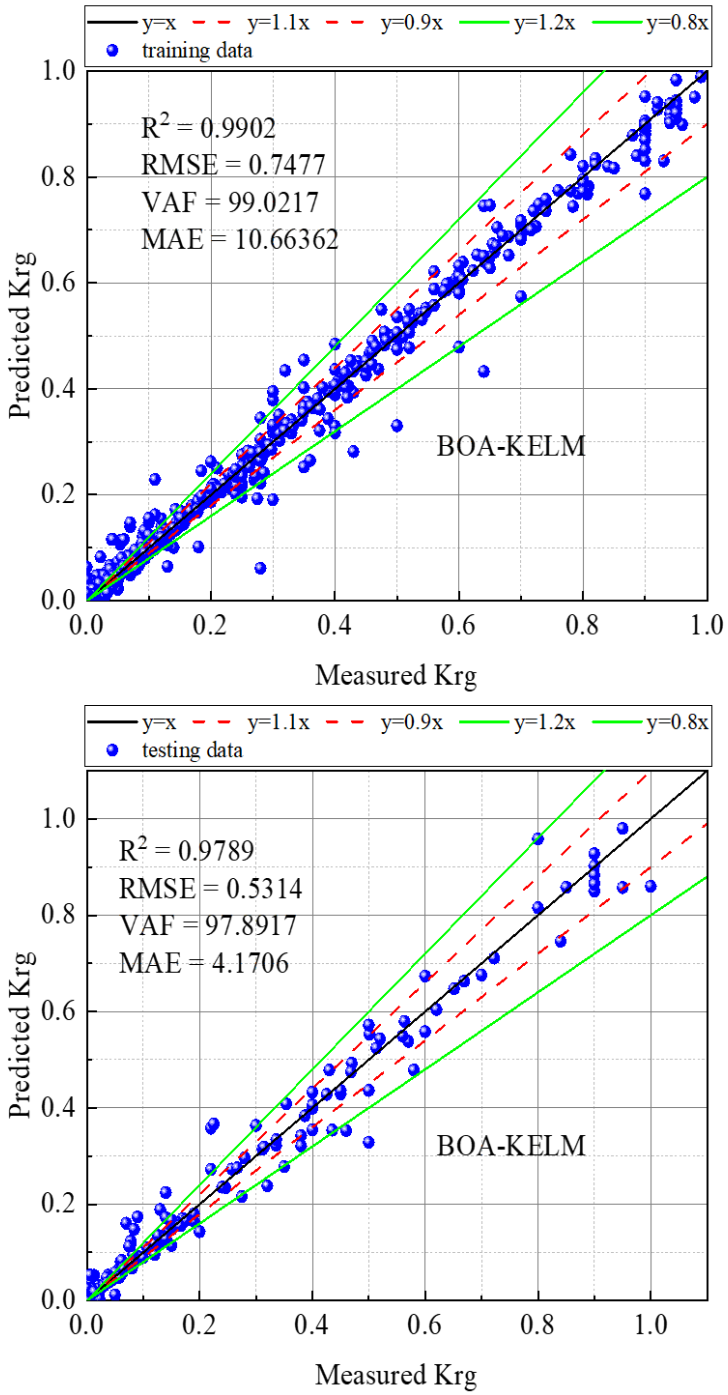


Figure 34. Comparisons between measured and predicted gas permeability by BOA-KELM.

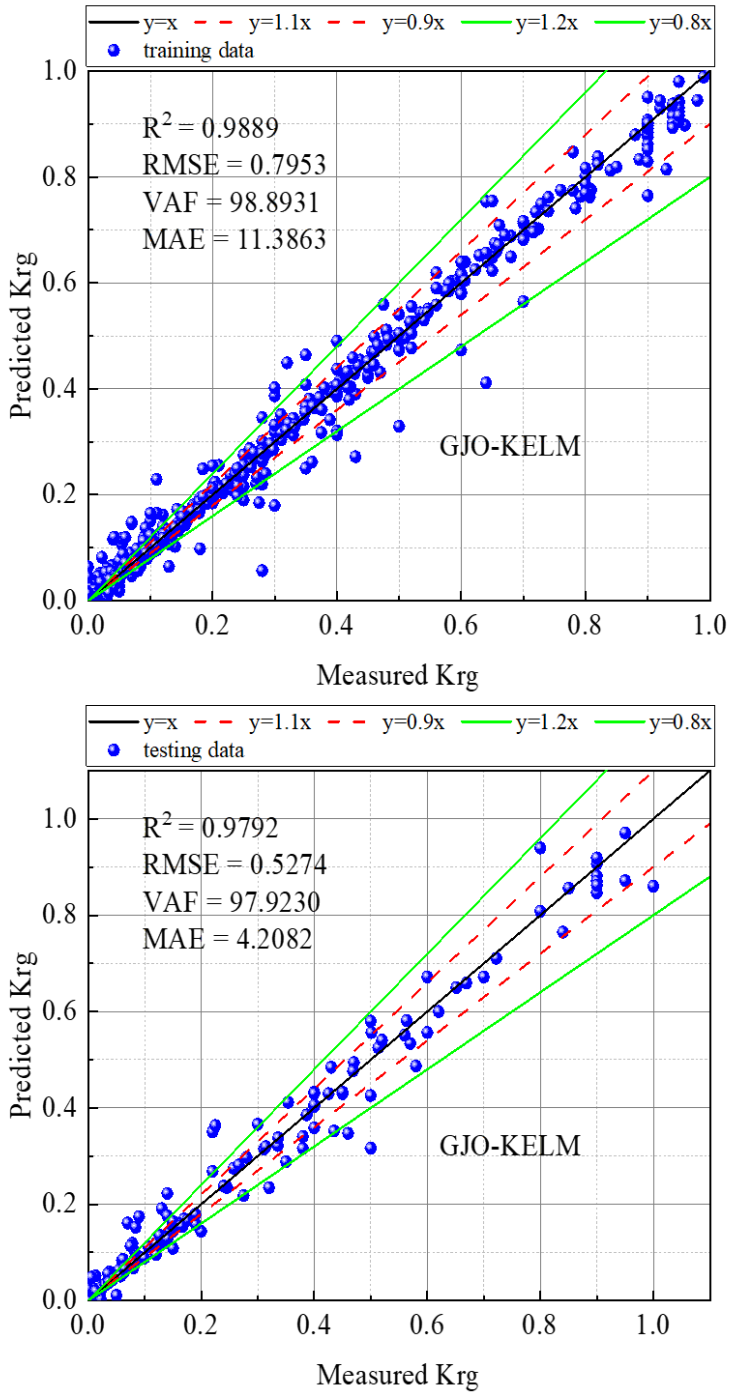


Figure 35. Comparisons between measured and predicted gas permeability by GJO-KELM.

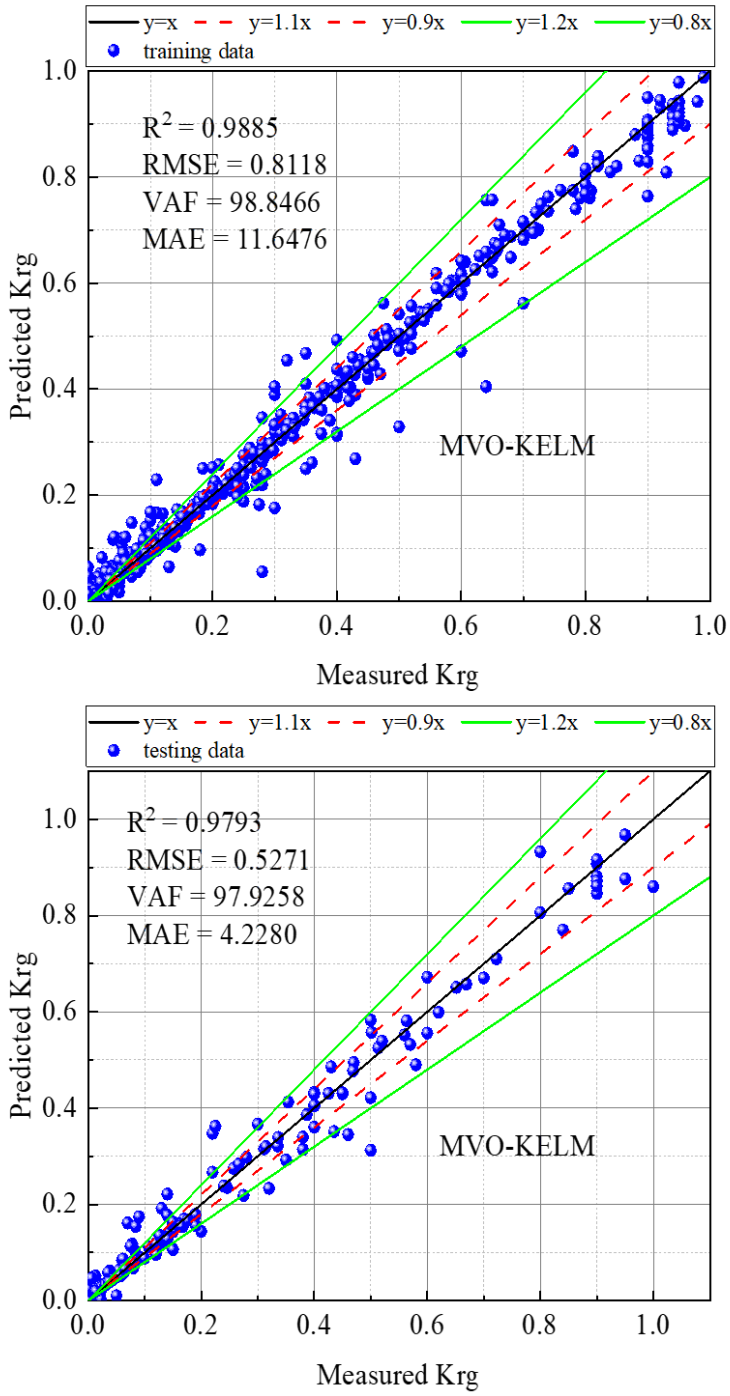


Figure 36. Comparisons between measured and predicted gas permeability by MVO-KELM.

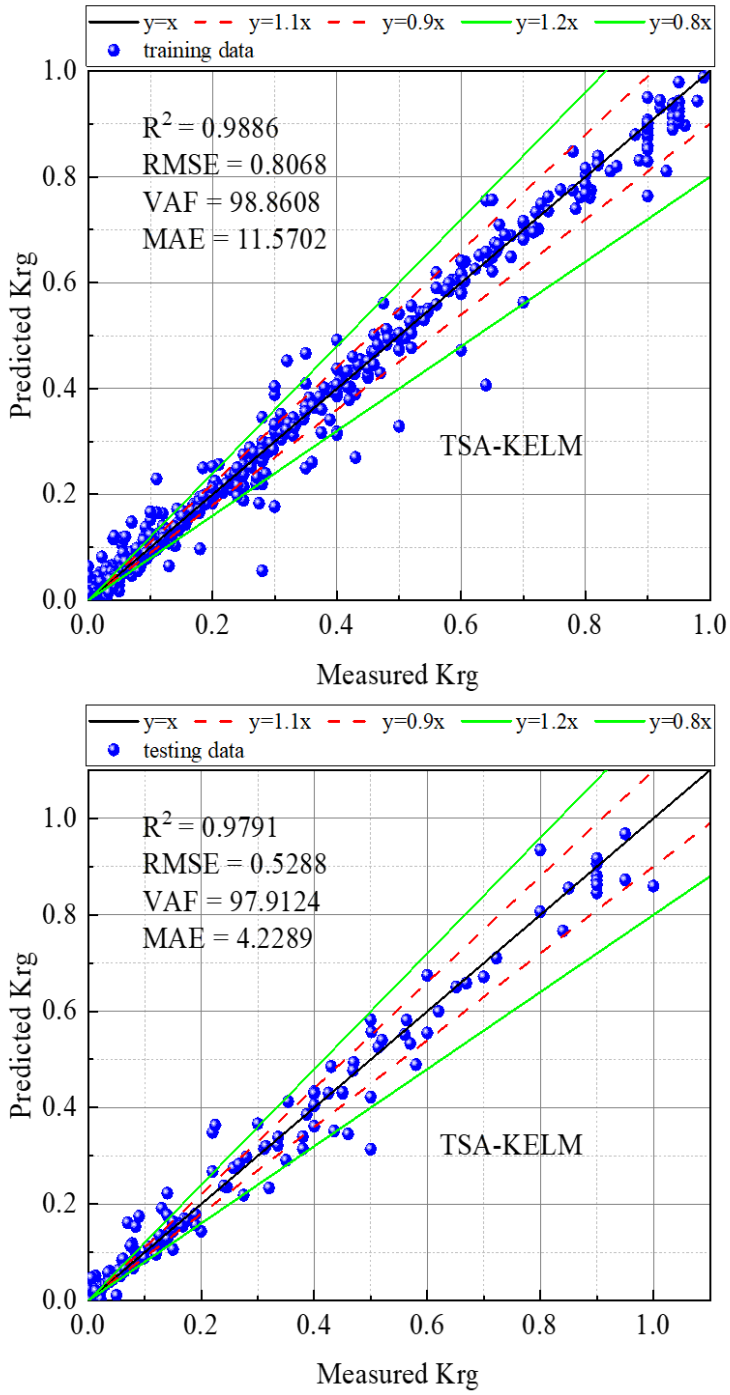


Figure 37. Comparisons between measured and predicted gas permeability by TSA-KELM.

3.6.2 Comparisons of proposed models and other models

To validate the superiority of developed five models in this study, several classical machine learning techniques, i.e., support vector machine (SVM), random forest (RF), ELM, (ANN) and a kind of novel algorithm, i.e., Light gradient boosting machine (LIGHTGBM) (Fan et al., 2019) were employed with the same data pre-processing and dataset. The Taylor Diagram was used to check their performance. The Taylor Diagram

utilizes three statistical metrics, i.e., RMSE, Pearson correlation coefficients, and standard deviation, to provide a comprehensive assessment (Zhou et al., 2021c). A shorter distance between the predicted and reference points (represented by a black dot) signifies superior prediction performance as Figure 38 shows. The findings reveal that hybrid KELM models performed much better than classical models and slightly superior than novel machine learning model for the training set and testing set. In conclusion, the results indicate that developed five hybrid KELM models are promising for gas permeability prediction and worthwhile to be considered to be new approach to integrate complicated gas permeability cases.

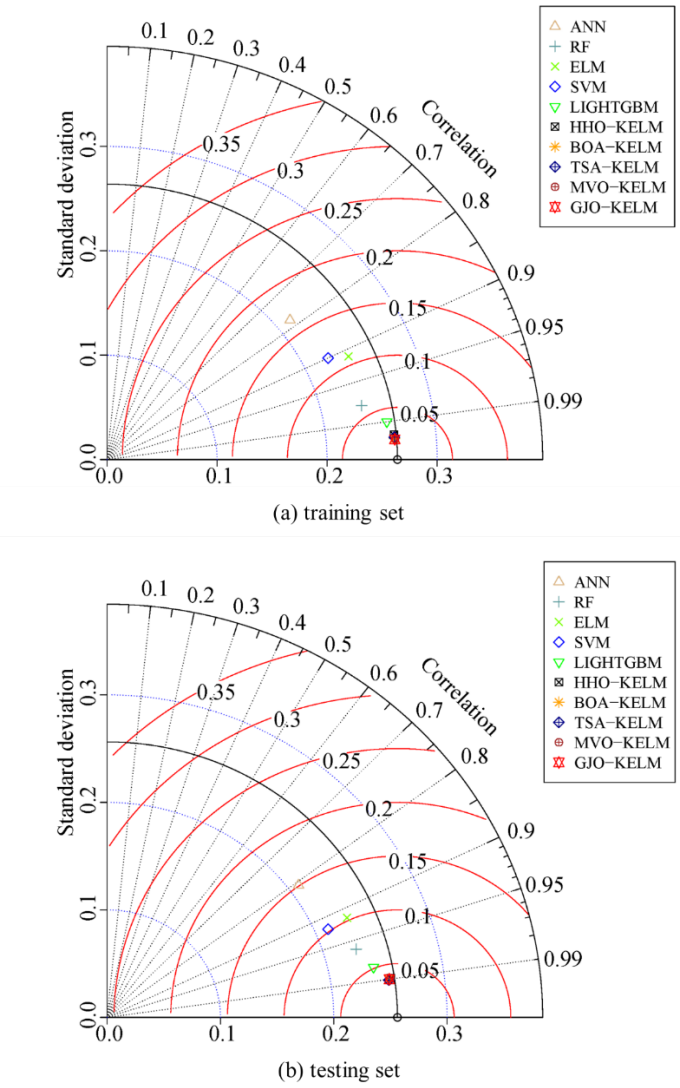


Figure 38. Taylor diagram for the assessment of model performance: (a) training set; (b) testing set.

3.6.3 Mutual information and single input modelling

As discussed before, proposed five hybrid KELM models have presented stronger prediction abilities than some classical and novel models for the gas permeability in reservoir. However, they cannot reach a perfect prediction which means that the prediction models cannot interpret the complicated relationship with influential factors.

Regarding this, the mutual information techniques was employed (Kraskov et al., 2004). Mutual information measures the interdependence between two random variables by quantifying the information that one variable reveals about the other. It encompasses both linear and nonlinear relationships and can be applied to diverse data types. A higher mutual information value suggests a stronger connection, while lower values indicate weaker or independent associations. The mutual information between influential factors and gas permeability for the whole original dataset and testing results from five hybrid models are shown in Table 24. The calculation function of mutual information is as follows (Cover and Thomas, 2005):

$$\mathrm{I(X;Y)} = \int y \int x \, P_{(X,Y)}(x,y) \log\left(\frac{P_{(X,Y)}(x,y)}{P_X(x)P_Y(y)}\right) dxdy \tag{3.5}$$

where $P_{(X,\,Y)}$ is the joint probability density function of X and Y, and P_X and P_Y are the marginal probability density functions of X and Y, respectively. It can be found that the ranking of mutual information from high to low is Sg, Swc, K, Sgc, Sorg and P for the original dataset. For the prediction models, the mutual information between gas permeability and Sg decreased a lot compared with the results from other inputs. These situations indicate that although the models exhibit good predictive performance overall, on the one hand, its explanatory power on the factor Sg is relatively weak; on the other hand, the factor Sg is more sensitive to the gas relative permeability. To explore the influence of Sg on the model performance, the Sg was used separately as an input factor and five hybrid models were employed again. Considering that there is only one input in the new scenario, the iteration number and swarm size is set to be 100 and 30, respectively, to save the optimization time for each method. For the purpose of comparison, several other machine learning techniques were also tested for the single-input scenario including SVM, RF, ANN, ELM and LIGHTGBM. Their prediction performance can be seen in Table 25. The corresponding optimization process can be seen in Figure 39. It can be found that the overall prediction performance decreased than the previous multi-inputs scenario, however, it is still acceptable which indicates the robustness of proposed five hybrid KELM models for modelling gas relative permeability. Moreover, a classical empirical function which also only considers the Sg was used to validate the superiority of proposed five methods, named Corey-Brooks Model (Brooks and Corey, 1964). The Corey-Brooks model is a widely used empirical function that correlates Sg and Krg in reservoir rocks. It assumes a power-law relationship:

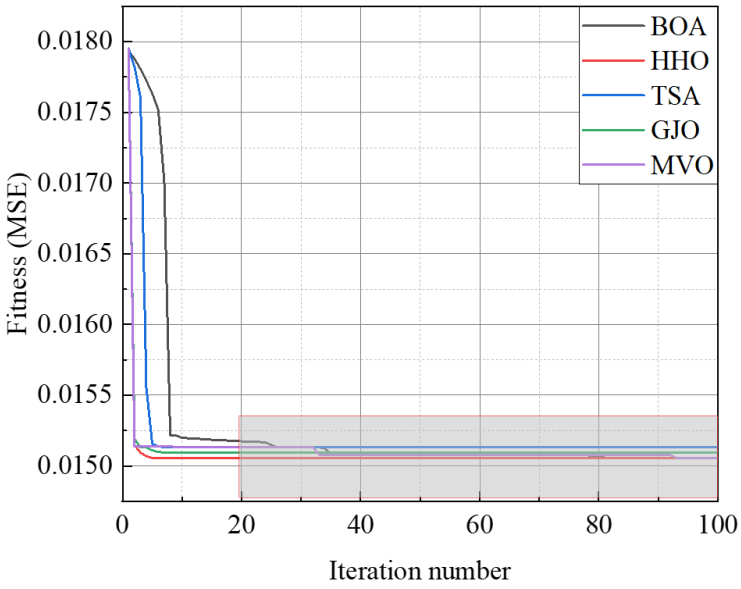
$$Krg = k_0 \times Sg^m \tag{3.6}$$

Where k_0 and m are fitting parameters determined through regression analysis using experimental data or well log data. To obtain these two fitting parameters, the "Curve Fitting Tool" was used based on the environment Matlab 2021a. And then the function of Corey-Brooks model can be obtained as follows:

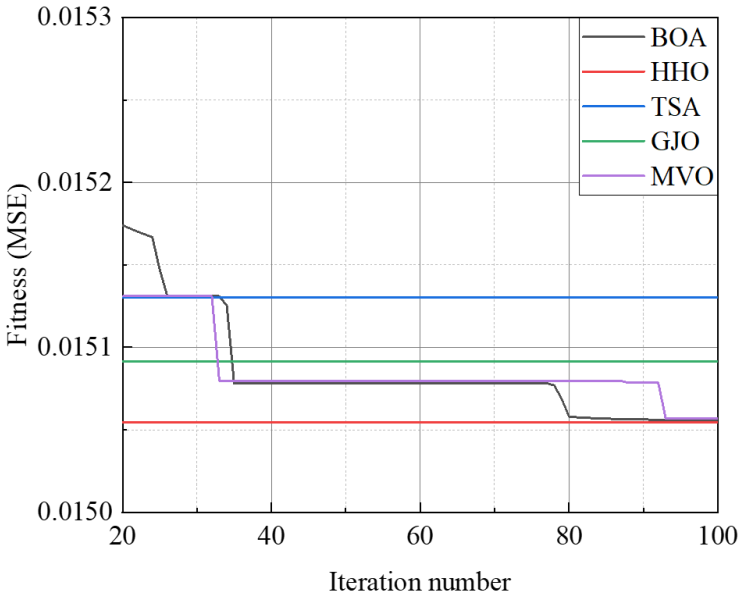
$$Krg = 1.186 \times Sg^{1.432} \tag{3.7}$$

The fitting performance by Corey-Brooks model has also been shown in Table 25. It can be seen that proposed five hybrid KELM methods produced better R^2 and VAF than Corey-Brooks model and other machine learning models. Among all methods, MVO-KELM brought the best R^2 and VAF for the training set and testing set. However, Corey-Brooks model and other machine learning methods presented better performance on

RMSE and MAE than hybrid KELM methods. In addition, it can also be found that RF generated competitive performance and it is worthwhile to be investigated in the future study. In the practical application, assuming that there is a Krg dataset related to Sg, then we can consider using MVO-KELM method on the condition that we need to pursue a higher R^2 and VAF.



(a) Overall optimization process



(b) Axes magnification

Figure 39. Optimization process of five hybrid KELM models based on the gas saturation: (a) Overall optimization process; (b) Axes magnification.

Table 24. Mutual information between influential factors and gas permeability from original dataset and developed models.

	Swc (%)	Sorg (%)	Sgc (%)	K (mD)	P (%)	Sg (%)
Original dataset	0.2424	0.1459	0.1813	0.2057	0.1109	1.7028
HHO-KELM	0.2393	0.1553	0.1742	0.2057	0.1266	1.564
BOA-KELM	0.2393	0.1506	0.1736	0.2056	0.1283	1.5953
TSA-KELM	0.2389	0.1499	0.1718	0.2054	0.1248	1.5972
MVO-KELM	0.2384	0.1497	0.1716	0.2054	0.1249	1.574
GJO-KELM	0.2378	0.15	0.1721	0.2054	0.1258	1.594

Table 25. Prediction performance based on gas saturation.

Method	Testing set				Training set			
	R^2	RMSE	VAF	MAE	R^2	RMSE	VAF	MAE
BOA	0.84	1.47	83.85	13.85	0.79	3.47	78.96	61.12
HHO	0.84	1.47	83.80	13.87	0.79	3.47	78.93	61.18
TSA	0.84	1.46	84.09	13.66	0.79	3.42	79.48	59.83
GJO	0.84	1.46	84.09	13.67	0.79	3.42	79.50	59.77
MVO	0.94	0.91	93.97	7.37	0.93	2.01	93.11	32.19
Corey-Brooks Model	0.75	0.13	75.45	0.08	0.83	0.11	83.23	0.07
ANN	0.81	0.11	81.13	0.08	0.76	0.13	76.18	0.09
ELM	0.84	0.10	83.85	0.08	0.82	0.11	82.37	0.08
RF	0.92	0.07	91.87	0.05	0.95	0.06	94.63	0.04
SVM	0.84	0.10	84.04	0.08	0.81	0.12	80.74	0.08
LIGHTGBM	0.83	0.11	82.98	0.07	0.83	0.11	83.37	0.07

3.7 Limitation

Some of the shortcomings in this study is that the data used for developing the hybrid KELM Krg models is from published literature and some outliers perhaps exist. In a future study, new measurement should be conducted to obtain more data for validating the hybrid KELM models. Other machine learning techniques, such as genetic algorithm, particle swarm optimization and multiple extreme learning machine techniques have potentials for predicting Krg and worthwhile to be investigated (Abad et al., 2022; Farsi et al., 2021). Finally, the prediction capability of hybrid KELM models have not been tried in other reservoir properties.

3.8 Conclusion

Monitoring and measuring the gas relative permeability in reservoir has always been a challenging task due to the anisotropy and heterogeneity of reservoir. Meanwhile, it has been influenced by various factors and the relationship between these factors and gas relative permeability is always elusive. Regarding these, modelling an effective and convenient tool to predict gas relative permeability seems to be promising. For the gas relative permeability cases affected by multiple factors, using a specific empirical function to integrate complicated influential factors is almost impossible.

The machine learning techniques have strong abilities to fit the complicated factors with targets. Over the past few years, various machine learning techniques have been tried to predict the gas relative permeability, however, the research about the potential of KELM is few. Regarding this, this study proposed five new hybrid KELM scenarios to consider connate water saturation, residual oil saturation, critical gas saturation, permeability, porosity, formation type and gas saturation to predict the gas relative permeability. Five novel meta-heuristic algorithms named butterfly optimization algorithm, tunicate swarm algorithm, Multi-verse optimizer, Golden jackal optimization and Harris hawk's optimization were adopted to tune the KELM hyper-parameters. In addition, the five-fold cross validation was also used to improve the model generalization. To ensure the fairness, all models were set to be the iteration number equal to 200 and the model with the same swarm size was given the same initial hyper-parameters. And four classical mathematical indicators were used, i.e., R^2, RMSE, MAE and VAF, to evaluate the model performance and GI was used to provide an overall evaluation.

According to the performance of GI for the testing set, it could be indicated that BOA-KELM model with swarm size 150 performed better than other hybrid KELM models with training set and testing set. The Taylor Diagram was used to compare the prediction ability of hybrid KELM models with some other machine learning techniques. It could be found that hybrid KELM performed better than other machine learning algorithms. To check the influence of inputs on the model interpretation, the mutual information technique was used and it can be found that the gas saturation has a larger influence on the hybrid KELM models. To validate the model robustness, the gas saturation was individually used as an input and then it can be found that proposed hybrid KELM models still can obtain acceptable prediction performance and MVO-KELM produced the best R^2 (0.94) for the testing set among other machine learning techniques and a traditional

Corey-Brooks model. The main novelty of this study is to propose five new machine learning techniques to predict the gas relative permeability and thus provide new insights for the prediction of reservoir properties. In addition, the proposed KELM-based methods presented better prediction performance than some classical machine learning methods and worthwhile to be applied to some other reservoir fields.

Chapter 4. Application of percentile color intensities of borehole images for automatic fluorite grade assessment

Nomenclature

AcT (Classification accuracy for total samples)	R^2 (Coefficient of determination)
AcW (Classification accuracy for waste)	R (Red)
AcLG (Classification accuracy for low grade)	RMSE (Root mean squared error)
AcMG (Classification accuracy for medium grade)	S (Saturation)
B (Blue)	SSA (Salp swarm algorithm)
CA (Correlation analysis)	SVC (Support vector classification)
CDF (Cumulative distribution function)	SVM (Support vector machine)
G (Green)	SVR (Support vector regression)
H (Hue)	TR (Training sets)
L (Lightness)	TS (Testing sets)
LG (low grade)	UV (Ultraviolet)
MAPE (Mean absolute percentage error)	VAF (Variance accounted for)
MG (Medium grade)	W (White)
MWD (Measurement while drilling)	WS (Waste)
PCA (Principal component analysis)	XRF (X-Ray Fluorescence)
PCI (Percentiles of color intensities)	

4.1 Introduction

The accuracy of reserves evaluation and the distribution of ore grades are key aspects in mining economics, planning and design. The prediction and evaluation of mineral grades plays a crucial role in the mining industry in its struggle to stay competitive under volatile prices, variable chemical and mineralogical composition, and declining ore grades. Swift grade determination of the ore that is being mined is instrumental to mining efficiency, hence methods for providing information on ore grade in an inexpensive and efficient way are of great interest for the mining industry. The harsh environment of underground works often limits the applicability of sophisticated and expensive analysis equipment, and the on-site implementation of complex analytical methods.

Two types of direct methods for determining the ore grade are core drilling and drill cuttings analysis. Core drilling is costly resulting in a limited dataset from which an ore grade model is inferred (Starr and Ingleton, 1992). On the contrary, the drill cuttings analysis from ordinary drilling for blasting or roof support allows for a dense sampling net; however, this method often has a limited accuracy as only average values per borehole are determined (Boesch and Rabalais, 1987). When the number of sampled holes is large, chemical analysis of drill cuttings can also be very demanding in terms of labor and assaying costs.

Minerals with different grades and components may present different colors and other optical properties. Some researchers have analyzed the optical properties of minerals at microscopic scale (Donskoi et al., 2015, 2013; Lane et al., 2008). Tanaka et al. (2019) reported a method for the recognition of acidic alteration zones in a deposit by distinguishing the intrinsic absorption peaks in the short-wavelength infrared region from various alteration minerals. Donskoi et al. (2007) combined a series of image analysis and mineral measurement techniques to distinguish the minerals with similar composition and texture. Berrezueta et al. (2016) proposed a new method by combining multispectral and color image analysis from microscopic observations to identify and quantify parameters related with geometallurgical performance such as ore grade, grain size and mineral liberation. Okada et al. (2020) proposed a quick and non-destructive technique to identify mineral types before mineral processing by utilizing RGB (red, blue and green) pixels information, hyperspectral imaging and deep learning techniques. Liu et al. (2019) show that deep learning, transfer learning, clustering algorithms and supervised learning techniques provide a more effective mineral recognition than traditional ones.

Despite rock recognition from spectral information being promising for ore grade assessment, the procurement of these data needs advanced equipment and also benign, friendly working environment. Inexpensive spectrometers are sometimes prone to errors in wavelength shift requiring frequent calibration. In some cases, optimum acquisition of images or segmentation techniques must be applied to guarantee an accurate recognition or classification. In view of this, color parameters of ore images seem to be a good approach to characterize the mineral characteristics due to its accessibility, low cost and convenience (Marschallinger, 1997; Thompson et al., 2001). For instance, Ramil et al. (2018) proposed an automatic identification system of in situ granite minerals based on artificial neural networks and RGB values of pixels of images of small-scale slabs. Li et al. (2017) developed a novel classification method of sandstone microscopic images

named Festra based on gray levels. Baykan and Yılmaz (2010) identify minerals with the aid of artificial neural networks using color information such as RGB, hue, saturation and lightness (HSL) of thin sections from a rotating polarizing microscope equipped with a digital camera. Desta and Buxton (2017) acquired in-situ georeferenced RGB images from the mine faces to interpret the distribution of minerals. Unsupervised learning techniques allowed to distinguish five mineral types with an accuracy of nearly 80 %. In summary, it is apparent that the color properties of mineral images have a great potential to provide reliable information for evaluating mineral grades or for mineral recognition.

4.2 Literature review

Recently, machine learning techniques have been successfully applied to address mineral grade prediction (Dumakor-Dupey and Arya, 2021; Jafrasteh and Fathianpour, 2017; Jooshaki et al., 2021; Kaplan and Topal, 2020; Mery and Marcotte, 2022; Sun et al., 2019). As an example, Patel et al. (2019) apply SVR of color intensities of images taken over a lab scale conveyor belt to monitor the quality of iron ore. Zhang et al. (2018) use back propagation artificial neural network to offline assess phosphate grade of flotation concentrate samples. RGB, or other color features, seem to be good alternatives to identify mineral grades. Perez et al. (2011) employed principal component analysis to RGB representation to extract color features and combined with texture features from five different rock samples including massive sulfide, disseminated sulfide, “net textured”, gabbro and peridotite to recognize the composition. Chatterjee et al. (2010b) utilized 189 features extracted from segmented images of a limestone mine and a neural network model to identify the grade attributes of limestone (CaO, Al_2O_3, Fe_2O_3 and SiO_2). Among these 189 features, 112 color features were involved and 42 features were gray level moments.

Research gap: However, in order to apply these approaches for ore grade control in mine planning and production quality assessment, it must be shown that the complex relationship between color parameters and mineral grades generally found from microscopic images at lab scale still applies to macro photography of rock outcrops taken in a production environment. Regarding this, this chapter describes a practical application of support vector machine (SVM) for ore grading in an underground fluorite mine. A televiewer is applied to obtain the optical information from borehole walls, which is rarely used for the determination of ore grade.

It covers all aspects from the inception of the experiments, data collection, input preparation, model description and results. Forty-eight drilling chips samples are collected while drilling six pseudo-horizontal boreholes at depth intervals of half a meter and their chemical composition determined through X-Ray fluorescence; the response of the drill rig is used to accurately define the depth of each sample along the blasthole. Images of the blasthole walls are collected with an optical televiewer with white and ultraviolet (UV) illumination. The color information of the images is characterized by the cumulative distribution of pixel color intensities of red, green and blue, used as inputs. A well-known metaheuristic algorithm is used to calibrate the SVM hyperparameters. Repeated k-fold cross validation is applied to increase the prediction performance due to

the small-size of the dataset. The utility of the proposed methodology can reduce the amount of lab analysis in ore grade control.

4.3 Data collection and description

The tests were carried out in the Lújar underground mine located in Órgiva (Granada province, Spain). A fluor-lead deposit composed by fluorite, galena and dolomite as gangue is mined. The host rock is mainly dark massive dolomicrite and limestone in which fluorite occurs as dark and white-purple crystals that may develop zebra patterns in some cases. The fact that the ore appears as fault-related veins or as irregular strata bound bodies with typical grades in the order of 15 % in fluorite complicates the in-situ ore recognition (Amor and Navarro, 2016; Ilin et al., 2019).

Six pseudo-horizontal boreholes were drilled in the same mine area by an Atlas Copco 282 jumbo equipped with a measurement-while-drilling (MWD) system. The holes had an approximate length of 3.5 m, a diameter of 102 mm and an upward inclination of 5°. This allowed their cleaning by the injection of water in order to improve the quality of the in-hole images. An endoscope inspection was carried out to verify the wall cleanliness and make sure that no faults were crossing the holes. A PVC pipe was used to push the logging tool, with the wireline in its axis, inside the hole. The tool was then pulled back to surface by means of constant-velocity winch.

4.3.1 Drill chips assaying

Drilling chips were collected by means of a tray placed below the borehole collar. Drilling was stopped at approximate intervals of 0.5 m to collect the detritus and place a new tray. Eight samples were collected for every hole making up 48 samples in total. Drops in percussive pressure and rotation pressure recorded in the drilling logs were used to identify the drilling stops and obtain the corresponding initial and final depths of each sample, see the dashed black lines in the drilling records shown as an example in Figure 40. The rotation pressure is sensitive to other effects related to the characteristics of the rock mass, such as the presence of structural discontinuities, as is the case of the red dashed rectangle in Figure 40, but these drops are easily distinguished from the drill stop ones.

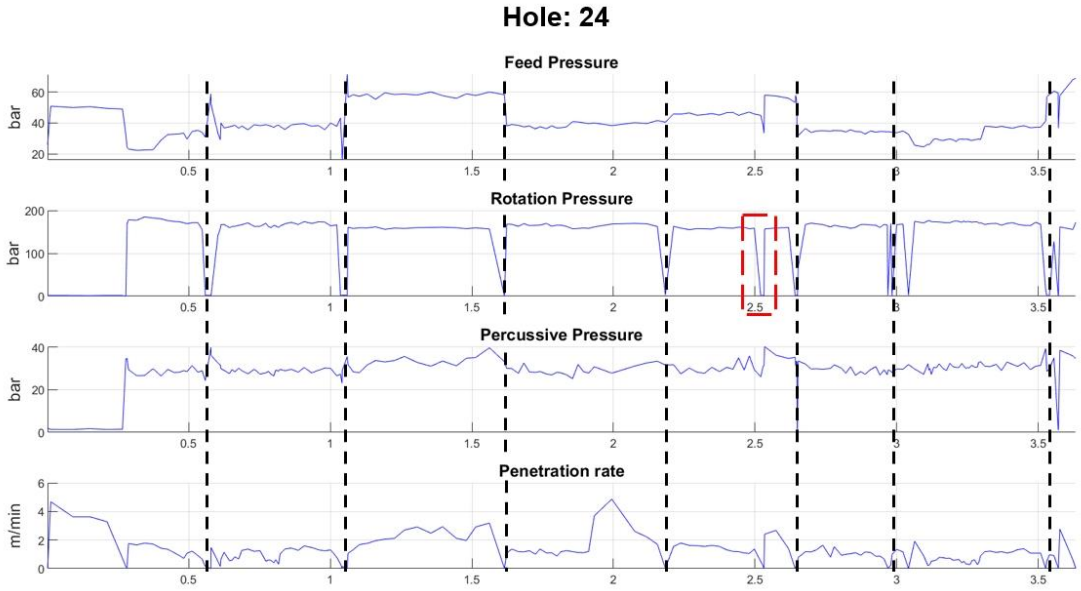


Figure 40. Drill log of borehole H24. Drilling stops are marked by black dashed lines; a potential rock mass discontinuity is highlighted by a red rectangle.

The drilling chips from each sample, approximately 1.3 kg, were quartered to 1/8 of the initial mass with a sample splitter with eight slots of 40 mm of aperture. This material was dried at 90°C during 24 h, and ground in a vibratory disc mill (Restch RM100) to a size below 80 μm. After this, the sample was quartered to 1/16 to obtain a 10 g sample from which powder pellets were prepared. The pellets were analyzed in an X-Ray Fluorescence (XRF) Thermo Scientific ARL OPTIM'X WDXRF 50 kV analyzer, composed by Rhodium anode, crystals LiF200, InSb and AX06, and standard patterns of Thermo Fisher Sci; the software Oxsas 2.2 of Thermo Fisher Sci was used. The amount of compounds, mainly CaF_2, $CaCO_3$, $CaMg(CO_3)_2$, SiO_2, Fe_2O_3 and Al_2O_3, is obtained through stoichiometric balance of the composition provided by the XRF analyzer. Figure 41 shows for each borehole, the initial and final depth of each drilling chips sample and the corresponding fluorite percentage; the composition of the rock in each of these sections is assumed to be uniform and it is classified as function of the fluorite content as waste (WS, CaF_2<10%, blue in Figure 41), low grade ore (LG, 10≤CaF_2<20%, green) and medium grade ore (MG, 20≤CaF_2<45%, red).

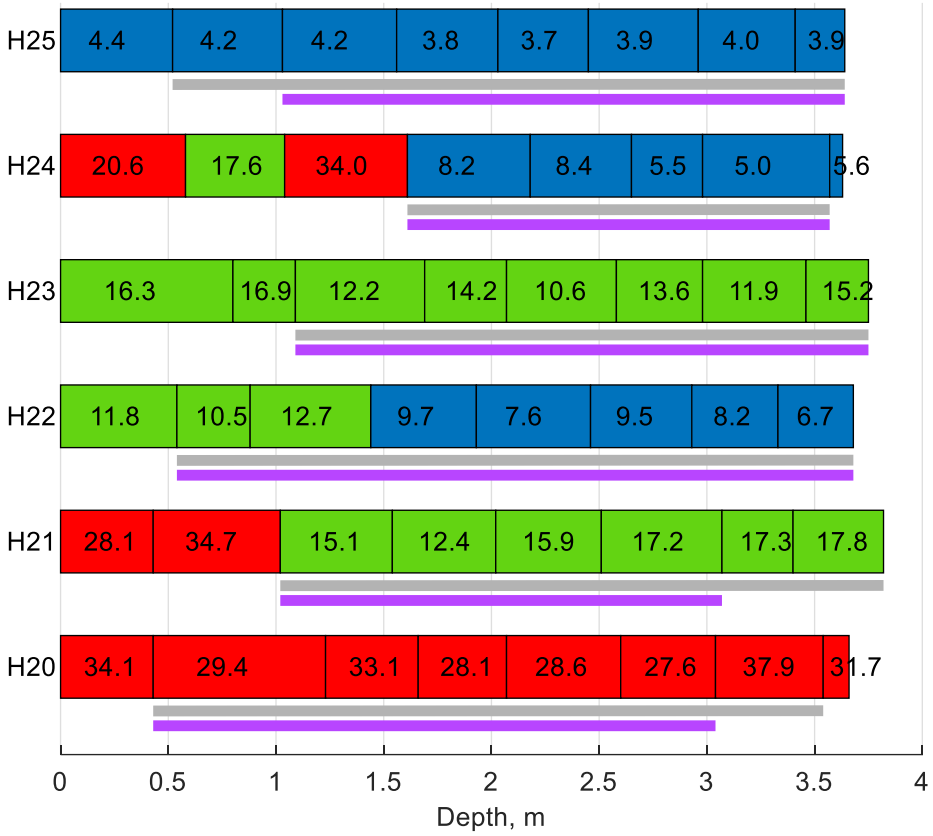


Figure 41. Ore grades of drilling chip samples. The quantity in each section is the percentage of fluorite content and the color indicates the rock classification: blue for waste, green for low grade ore and red for medium grade ore. The lengths scanned are indicated by grey (white light) and violet (UV light) lines.

4.3.2 Borehole logging

Boreholes were logged with an optical televiewer manufactured by ALT composed by a QL40 OBI-2G logging tool of 1.5 m length, a data acquisition system, a mini-winch that pulls the logging tool at constant velocity and a computer to set-up the tool, display, and record the images of the borehole walls. The logging tool has 3-axis accelerometers and magnetometers in its central part to survey the borehole path and a digital image sensor at the bottom, with an active pixel array of 1.2 Mpx and fisheye matching optics (see left image in Figure 42). It incorporates two LED series for lighting the internal walls: one emits white light and the other UV light. The latter has a wavelength range 340 to 400 nm with the emission peak at 365 nm. Each borehole was logged first with white light to record the natural colors of the rocks and second with ultraviolet light to outline the fluorescence of the main rock types.

The logging tool was centered with respect to the borehole axis with two centralizers mounted at the top (rear) and bottom (forward) parts of the probe; the rear one can be seen outside the borehole in Figure 42. It was pushed with two plastic rigid pipes until the bottom of the hole though this was not always reached when high resistance was encountered to avoid damage to the optical system. From the end position, the mini-winch

pulled the probe outwards while the borehole wall was scanned with an axial and circumferential resolution of 0.36 and 0.33 mm/px, respectively. Since the optical sensor is positioned at the end of the probe it was necessary to manually sustain the probe until the sensor was near the collar (see Figure 42). Despite this, some 0.5 to 1.5 m of the borehole length in the collar section, depending on the scan, could not be scanned, see Figure 41. The sections of the borehole scanned with white and UV lights considered for the analysis are coincident with the initial and final depths of the drilling chips samples. The actual lengths covered are shown with grey and violet bars below the fluorite compositions for each borehole in Figure 41. Borehole sections associated to each drilling sample that are not fully scanned with the televiewer from the initial and final depths of each sample are discarded, as it is unknown whether the scanned part of the section is representative of the chemical composition of the actual drilling chips sample. This reduces the number of sections that can be correlated with rock images at the corresponding depths to 36 for scans with white light and 32 for those with UV light.

Despite that the size of the resulting database is relatively small to develop a model that could be generalized to other geological conditions or operations, we hope it is enough to validate the methodology proposed, in which the percentile color intensities of borehole images are used to assess automatically the ore grade.

Figure 42. Bottom part of the logging tool (left) and final stages of borehole surveying with the forward centralizer and the glass tube inside the hole (right).

4.3.3 Image processing

Figure 43 shows a typical example of televiewer logs from scans with white and UV lights at a length interval in which fluorite was defined visually by an experienced geologist; the assay of drilling chips classifies the rock in that section as medium grade ore (see Figure 41), which is compatible with the presence of dolomitic breccia observed. For each scan, the RGB value of each pixel is calculated by WellCAD (ALT, 2020) and it is represented by a colour palette in the three columns on the right of the corresponding

image of the borehole walls in Figure 43. This gives, for each light type, three 2D matrices for red, green and blue colors with intensities in the range 0-255. The corresponding mean color intensity at each depth is the white curve (three 1D arrays). Note that no apparent difference is observed visually between fluorite and dolomitic breccia.

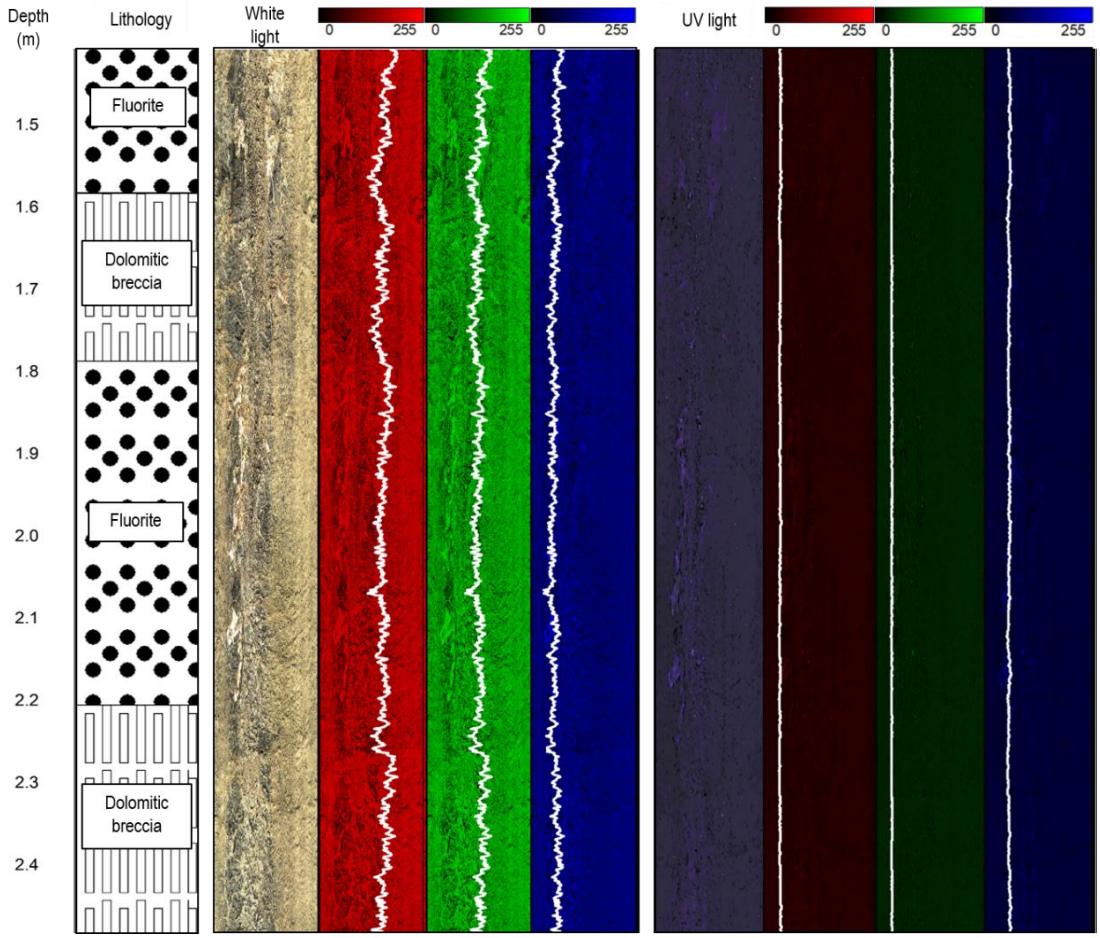


Figure 43. Section of borehole H20. From left to right: lithology, televiewer processed image and RGB logs with white and UV illumination.

The image of each section of the televiewer, defined by the initial and final depths of the drill chips samples collected for assaying, is formed by approximately 1500 sets of red, green and blue color intensities. To characterize this color information, percentiles 10, 20, …, 100 of the distribution of pixel intensities of red, green and blue colors in that section are calculated; this leads to two triplets of percentiles of color intensities (PCI) of pixels, $(W_{R,p}, W_{G,p}, W_{B,p})$ and $(UV_{R,p}, UV_{G,p}, UV_{B,p})$ for each hole section and p percentile for the white and UV light scans (p=10, 20, …, 100), making up 30 PCI for each hole section and type of light scan. The cumulative distribution functions (CDFs) of pixel intensities of red, green and blue from each scan type are shown in Figure 44. They are colored as function of the fluorite content (hot colors represent medium ore grade; cold ones correspond to waste).

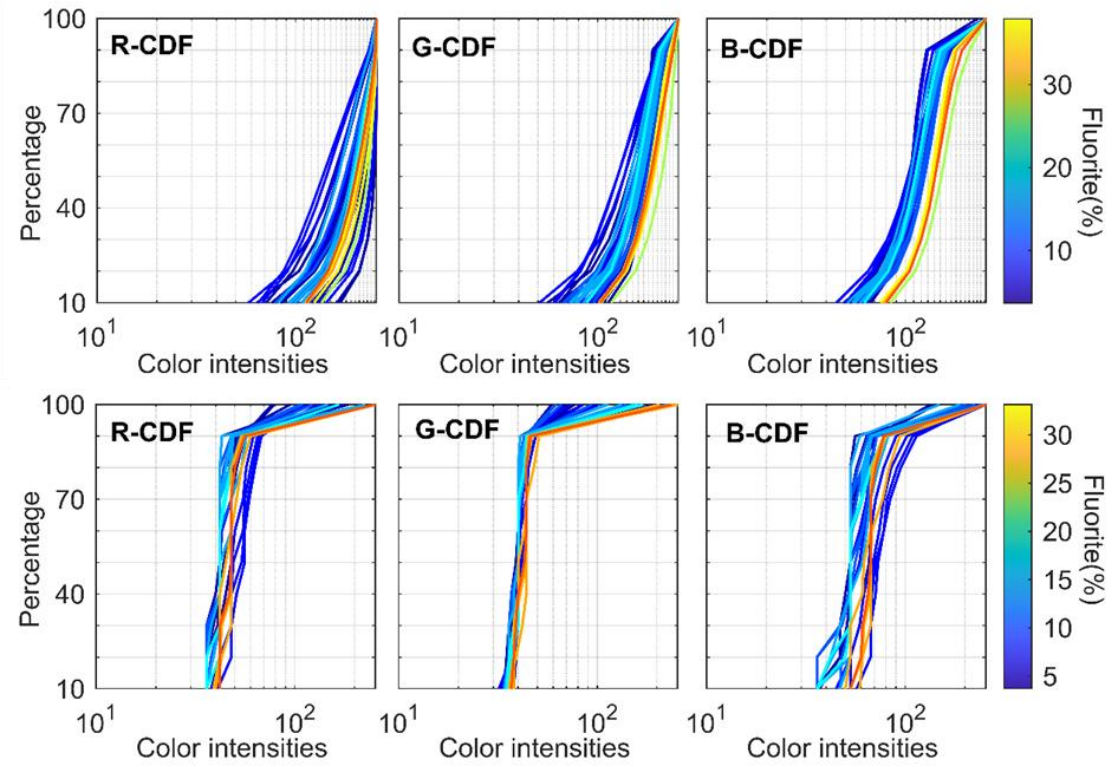


Figure 44. Cumulative distribution functions of color intensities for white (top graphs) and UV light (bottom graphs) scans; data correspond to all six boreholes.

Although some trend between PCI and CaF_2 content can be observed, mainly for green and blue colors from white light scans (see central and right graphs in Figure 44), these relations are not easily described with simple analytical functions of the PCI that could be applied in a classical multivariant analysis. The resort to advanced machine learning techniques to explore such complex relationship appears to be natural.

4.4 Data pre-processing

The sets of triplets, $(W_{R,p}, W_{G,p}, W_{B,p})$ and $(UV_{R,p}, UV_{G,p}, UV_{B,p})$, p=10,20,…,100, for the sections considered are selected as input parameters; the respective fluorite content and grade classification are taken as the output for the machine learning algorithms.

The combination of the three possible input sets (i.e. PCI from white light, from UV light and both) and two types of feature extraction techniques (i.e. principal component analysis and correlation analysis) to reduce the size of the input dimension are proposed. This makes up six different modeling scenarios that are summarized in Table 26; they are identified with two letters, the first one describes the light source (W for white light scans, UV for UV light scans and WUV for the combination of color characteristics for white and UV light scans) and the second one describes the feature extraction technique (PCA for principal component analysis and CA for correlation analysis).

Table 26. Summary of scenarios according to input parameters and feature extraction.

Scenario	Input parameters[a]	Feature extraction	No. of training / testing cases
W_{PCA}	$(W_{R,p}, W_{G,p}, W_{B,p})$	PCA	30/6
W_{CA}		CA	30/6
UV_{PCA}	$(UV_{R,p}, UV_{G,p}, UV_{B,p})$	PCA	26/6
UV_{CA}		CA	26/6
WUV_{PCA}	$(W_{R,p}, W_{G,p}, W_{B,p}, UV_{R,p}, UV_{G,p}, UV_{B,p})$	PCA	26/6
WUV_{CA}		CA	26/6

[a] p=10,20,…,100.

4.4.1 Dataset partition

The original dataset needs to be divided into training and testing sets. The first is used for developing the model and the second is used for verifying its generalization and robustness. Generally, the ratio of training to testing set cases is 8:2 or 7:3, which can be tuned according to the scale of the data. The cases of valid scans are 36 for W light and 32 for UV light, as explained in Section 2.2. They are divided in training/testing 30/6 and 26/6 respectively, as shown in Table 26. For the combined use of the W and UV light, data from both scans must be available, so the number of valid scans is in this case equal to the number of UV valid scans i.e. 32. The cases of waste are 16 and 15 for W and UV light scans, respectively; sections of low grade are 14 and 12 for W and UV light scans, respectively, and sections of medium grade are scarce, 6 and 5 for white and UV light scans. This unbalanced number of cases of the three grades may lead to uncorrelated training and testing data sets, on which a weak generalization of the supervised learning would be obtained. For selecting the training and testing sets, one case was randomly selected from each borehole (hence six cases are selected, see Table 26) and used for developing the testing set and the other cases constituted the training set. Although they don't fully meet the 8:2 rule, they are still reasonable in view of the number of scan data. For eliminating the adverse effects caused by unbalanced data division, a k-fold cross-validation (Fushiki, 2011; Rodriguez et al., 2010) is applied. It randomly separates the original training set into *k* equal-size subsets where *k*-1 subsets are used as a new training set and the remaining subset is used for validation. The algorithm searches for a model that leads to the best fitness value for the *k* sets of training samples. According to some studies (Marcot and Hanea, 2021; Yadav and Shukla, 2016), 5 or 10-fold cross-validation works well. Considering the scale of the training set, 5-fold cross-validation has been employed.

The prediction ability of machine learning techniques is often assessed from only one random division of the dataset into training and testing sets. This prediction performance may not properly reflect the overall goodness of the dataset and the prediction ability of the model. The relatively small size of our datasets and their unbalanced nature may cause unstable prediction results as different partitions of the original dataset would lead to

different prediction accuracy. In order to account for this, a repeated *k*-fold cross validation technique was employed where thirty random divisions of the dataset are implemented to produce thirty different combinations of training and testing sets. For each training set, the 5-fold cross validation is implemented. The average results from the thirty training/testing combinations provide a more robust evaluation of the prediction ability.

4.4.2 Feature extraction

For each section, the original inputs from W and UV scans are 30 color intensities respectively which encompass a large input dimension compared with the size of the dataset. To reduce the complexity of calculation and preserve as much statistical information as possible, principal component analysis and correlation analysis are considered.

Principal component analysis

PCA (Wold et al., 1987) finds new uncorrelated variables, or principal components, that are linear combinations of the original variables that maximize the variance between them. Substituting the original variables by a few principal components reduces the input dimension and simplifies the model fit. The cumulative variance is shown in Table 27; a percent of the total variance higher than 95% is considered to define the number of components retained, this being 3 principal components for W and 8 principal components for UV datasets. The eigenvalues (i.e. variance size) of each principal component are also presented.

Table 27. Percentage of the cumulative total variability in the data explained by each principal component.

Light source	1	2	3	4	5	6	7	8
W[a]	65.1	93.1	97.3	-	-	-	-	-
W-latent	1.02	0.44	0.07	-	-	-	-	-
UV	62.6	73.9	82.3	87.6	90.9	92.9	94.7	95.7
UV-latent	1.67	0.30	0.23	0.14	0.09	0.05	0.05	0.03

[a]Percentiles with constant intensity are discarded for the analysis

Correlation analysis

The results of a Spearman correlation analysis between PCI of pixels and the fluorite content are presented for each light source in Table 27; no results are shown when the intensity colors for a given percentile are constant, as occurs for the 100 percentiles from red and green with W light illumination. PCI with significant correlation (coefficient $|r| \geq 0.3$ and p-value≤ 0.05) have been selected as inputs (highlighted in bold in Table 28). For W light, this applies to green percentiles 10-40 and 70-90 and most of the blue percentiles, while no red percentile meets the significance condition; for UV light, these are one percentile for red (100) and blue (30), and five percentiles for green (10, 40, 50,

90 and 100). The different significances of correlations of PCI and fluorite content for different illuminating sources may indicate some differential optical response from the materials.

Table 28. Correlation coefficients between PCI and fluorite content.

	W light			UV light		
Percentile	Red	Green	Blue	Red	Green	Blue
10	0.16	**0.35**	**0.4**	0.17	**0.51**	0.12
20	0.12	**0.33**	**0.39**	0.28	0.25	0.29
30	0.13	**0.37**	**0.41**	0.35	0.25	**0.39**
40	0.12	**0.36**	**0.39**	0.22	**0.42**	0.3
50	0.06	0.33	**0.41**	0.15	**0.45**	0.2
60	-0.01	0.31	**0.45**	0.11	0.3	0.21
70	-0.04	**0.37**	**0.47**	0.18	0.2	0.26
80	-0.1	**0.42**	**0.48**	0.1	0.27	0.31
90	-0.1	**0.4**	**0.45**	0.12	**0.37**	0.27
100	-	-	0.15	**0.62**	**0.53**	0.17

Note: Bold numbers are $|r| \geq 0.3$ and p-value≤ 0.05. It is noted that some inputs have $|r| \geq 0.3$ but p-value>0.05 and thus are not bold.

4.5 The model

The support vector machine will be used in this study as the benchmark tool to predict fluorite grade (Support vector regression, SVR) or classify the rock into waste, low grade ore, and medium grade ore (Support vector classification, SVC), as function of an n-dimensional set of input variables (linear combinations of PCI defined from the PCA or the most relevant PCI obtained from CA). SVM was developed initially for tackling classification issues, and it can also be extended to solve regression problems (Quan et al., 2022; Vapnik, 2000). SVM is very appropriate for analyzing small databases with large-dimension input data, as is the case of this work, compared with other classical approaches like artificial neural network and K-nearest neighbors (Qi and Tang, 2018). The main idea of SVC is to find the optimal separating hyperplane that correctly partitions the training dataset with the largest geometric separation, while SVR aims to find a function that deviates from every output by no more than a certain error for each training data point. Details of the SVM optimization can be found in (Smola and Schölkopf, 2004).

Similar to the introduction of support vector machine in Chapter 2, two key hyper-parameters named penalty factor C and radial based function kernel deviation γ would be optimized (Quan et al., 2022). In order to optimize these parameters, a myriad of metaphor-based optimization algorithms of artificial intelligence-based models is available (Li et al., 2021a; Li et al., 2021b; Xi et al., 2024a; Xi et al., 2024b; Zhou et al., 2021e). A common practice in these ML-based papers is to rate some of these algorithms according to the resulting performance to solve a specific problem; however, the prediction performance by different meta-heuristic algorithms is very similar. Regarding

this, to address the performance of a bio-inspired meta-heuristic algorithm is not the motivation of this study, but to discuss the potential of percentile color intensity of images to assess the rock composition and the ore grade in particular by ML-based models. The salp swarm algorithm (SSA) (Mirjalili et al., 2017) that has proved to be effective in solving different optimization problems in various domains (Li et al., 2021a; Li et al., 2021b), and also been proved to be more powerful in Chapter 3, therefore, it is used here in combination with SVM to select suitable support vector parameter combinations so as to prevent local optima. To facilitate reading, the detailed introduction of SSA would not be given here. Alternatively, a general sketch of the optimization process of SSA for C and γ is shown in Figure 45. However, the swarm size (i.e. the number of salps) and the maximum number of iterations must be chosen, both being in fact significant optimization parameters.

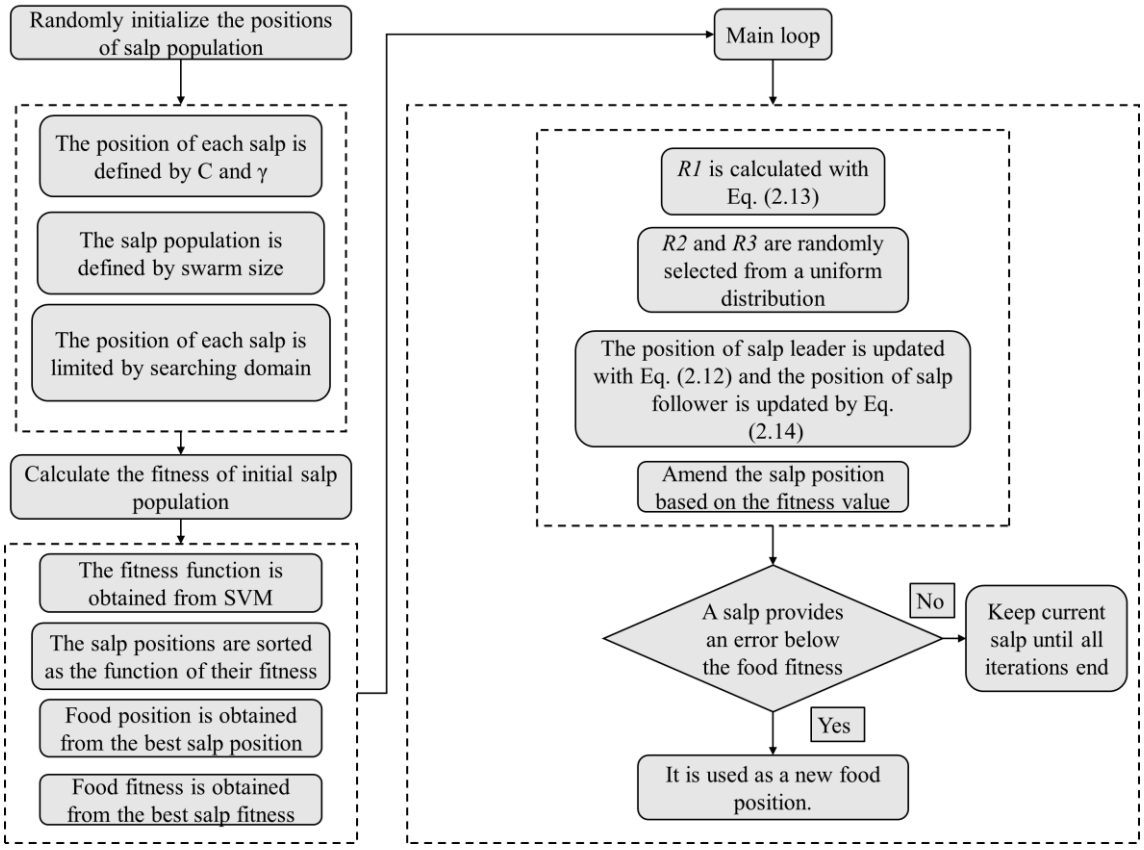


Figure 45. General optimization process of the selection of SVM hyper-parameters by SSA.

4.6 Results and discussion

Regression and classification patterns are developed and evaluated for six scenarios of input parameters (see Table 26). The model was programmed in a Matlab (MATLAB, 2022) environment, with support vector machine code from LIBSVM (Chang and Lin, 2011). Input parameters are the principal components selected in Section 4.4.1 for PCA and in Section 4.4.2 for CA.

For regression, the output is the percentage of fluorite and for classification, it is the fluorite grade class. Both regression and classification models utilize the same SSA parameters (swarm size and iteration number) and inputs. As explained in Section 4, similar to other swarm-based heuristic algorithms, swarm size and maximum iterations have a key impact on speed and prediction performance (Yu et al., 2020a; Zhou et al., 2021b, 2021f). Different iteration numbers and swarm sizes were tested. When swarm size and maximum iterations were small, the prediction performance was generally not stable and over-fitting or under-fitting sometimes occurred. A swarm size of 50 and a maximum iteration number of 200 were found to yield efficient optimization results and, when increased, the prediction performance did not significantly improve while the computational time increased. Those values were thus adopted in all models.

For regression, four classical indicators have been employed: coefficient of determination (R^2), root mean squared error (RMSE), variance accounted for (VAF); they are shown in Eqs. (2.1), (3.2), (2.3), respectively, and mean absolute percentage error (MAPE). For classification, the classification accuracy for total samples (AcT), waste (AcW), low grade (AcLG) and medium grade (AcMG) have been used (see Table 29**)**. As mentioned before, 30 random divisions of training and testing sets are carried out and for each of them, a prediction model is built; the same divisions are used for regression and classification for all scenarios. A summary of the main statistics of the metrics for regression scenarios W_{PCA}, W_{CA}, UV_{PCA} and UV_{CA} can be seen in Table 30 for training (TR) and testing (TS) sets. Taking the R^2 as a specific research objective, the mean value for the training sets is excellent while it is lower for the testing sets.

Table 29. Summary of the performance metrics.

Regression	Formula[a]	Classification	Formula[b]
R^2	Eq. (2.1)	Total	$AcT = \frac{T_W + T_{LG} + T_{MG}}{N_W + N_{LG} + N_{MG}}$
VAF	Eq. (2.3)	WS	$AcW = \frac{T_W}{N_W}$
RMSE	Eq. (3.2)	LG	$AcLG = \frac{T_{LG}}{N_{LG}}$
MAPE	$\text{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left\|\frac{y_i - y'_i}{y_i}\right\|$	MG	$AcMG = \frac{T_{MG}}{N_{MG}}$

Note: [a]y_i and y_i' denotes the measured and predicted fluorite content; [b]*Ac* means classification accuracy. T_c is the number of true positives of the *c* category (*c* is *W* for waste, *LG* for low grade and *MG* for medium grade). N_c is the number of sections of the *c* category.

Table 30. Main prediction performance statistics from full-data for white and UV light scans.

Light Source		TR				TS			
		R^2	VAF	MAPE	RMSE	R^2	VAF	MAPE	RMSE
W_{PCA}	Mean	0.92	92.60	0.21	2.37	0.63	68.82	0.47	5.08
	Min.	0.87	87.94	0.04	0.34	-0.27	-3.91	0.23	2.94
	Max.	1.00	99.86	0.32	3.09	0.87	97.00	0.63	9.06
	Std.	0.03	3.00	0.07	0.74	0.25	23.87	0.11	1.53
W_{CA}	Mean	0.94	94.54	0.16	1.76	0.62	67.61	0.46	4.99
	Min.	0.82	82.29	0.04	0.33	-0.42	-34.89	0.27	2.73
	Max.	1.00	99.87	0.40	3.58	0.92	93.72	0.68	8.97
	Std.	0.05	5.26	0.12	1.17	0.37	34.37	0.09	1.69
UV_{PCA}	Mean	0.94	94.09	0.13	1.67	0.54	57.89	0.45	5.40
	Min	0.84	84.30	0.03	0.29	0.14	14.61	0.23	2.47
	Max	1.00	99.88	0.33	3.32	0.91	92.93	0.79	7.98
	Std.	0.06	5.79	0.09	1.14	0.23	22.82	0.13	1.48
UV_{CA}	Mean	0.80	80.30	0.25	3.62	0.60	66.55	0.38	4.96
	Min	0.71	71.50	0.04	0.80	0.00	0.23	0.15	1.89
	Max	0.99	99.10	0.33	4.41	0.95	95.74	0.81	8.32
	Std.	0.07	6.78	0.07	0.83	0.27	24.84	0.15	1.67

Note: Min: minimum; Max: maximum; Std.: standard deviation

Some low, or even negative R^2 occur (see the minimum values in Table 30) which could be due to outliers in the data. For detecting such outliers, one section is removed from the dataset and the remaining sections are used for developing the regression models. This is repeated until all sections have been individually removed. We can assume that if one section is an outlier, then the determination coefficient will increase significantly when it is not included in the calculation. In order to assess whether the improvement is significant from a statistical point of view, the 95-percentile of the determination coefficient is employed as criterion. All four scenarios are tested, i.e., W_{PCA}, W_{CA}, UV_{PCA} and UV_{CA}. If the case removed improves the prediction performance in most scenarios, then it can be considered a candidate for outlier. The 'take one out' method involves: i) feature extraction (PCA and CA) from the new dataset; ii) random divisions of the dataset into 30 sets of training and testing data (the testing set is always formed by 6 samples and the training by 29 for white light and 25 for UV light), and iii) train and test the prediction model for each of the random divisions. This operation is repeated 36 times and 32 times

for W- and UV-based scenarios, respectively, until all models leaving out one section each are built.

When an outlier is removed from the dataset, significantly better prediction performance will be procured and this will result in a significantly higher R^2. Figure 46 shows the mean R^2 of the 30 random divisions for all regression models with one section removed; the numbering of sections is (see Figure 41) from collar to bottom, hole H20 to H25, so that section 1 is 29.4 % CaF_2 and the last section (#36) is 3.9 % CaF_2. Sections #6, #11, #12 and #30 do not exist in the UV dataset. The horizontal lines in Figure 46 show the 95 percentiles of R^2. The following sections removal score above this percentile: W_{PCA}: #5, #15; W_{CA}: #5; #32; UV_{PCA}: #5, #31; UV_{CA}: #1, #32.

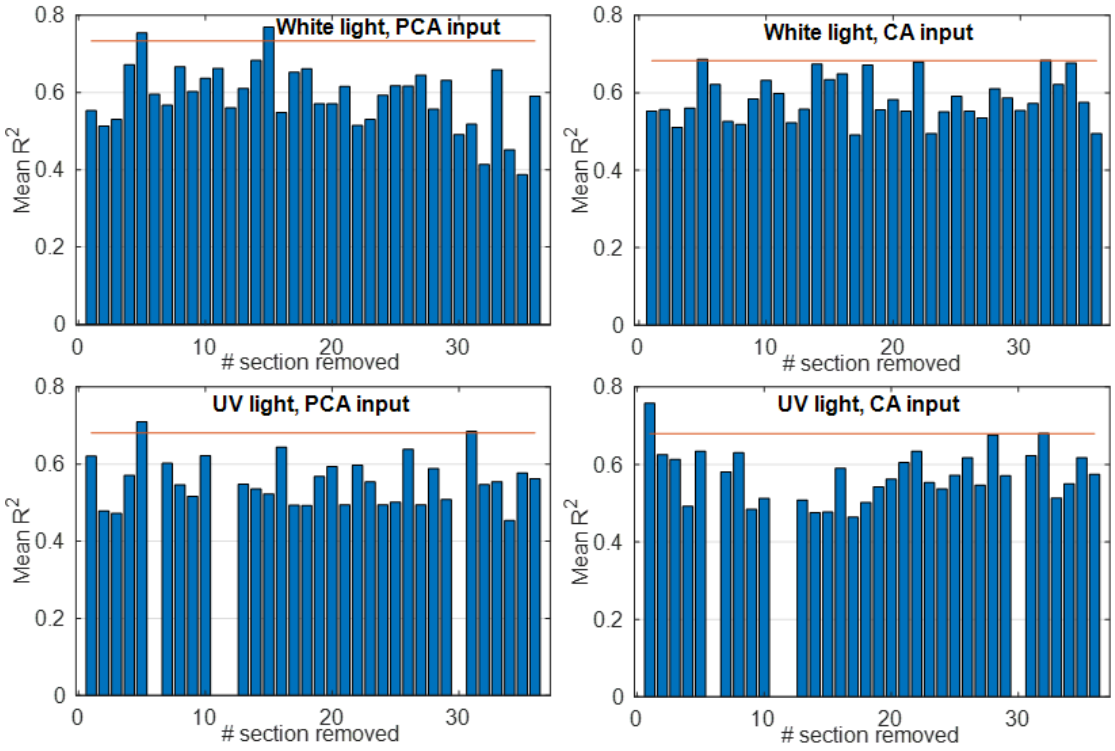


Figure 46. Average R^2 from 30 divisions for "take one out" models.

The removal of section #5 improves outstandingly the results of three out of four scenarios (it scores at percentiles 95.8, 98.6, 98.5 and 88.9) so this section could be an outlier. Section #32 could also be a candidate although its R^2 only scores above 95 % in two cases, while in the other two is nowhere near that threshold. For this reason, only section #5 (i.e. fifth section of borehole H20, with 27.6 % CaF_2) was removed. We suspect that the reason for its offending behavior is uneven sampling with an uncertain grade rating, probably with lower fluorite percentage than the actual one; higher fluorite percentage has a slight right shift tendency for G- and B-CDF plots as Figure 44 shows, where the 27.6 % CaF_2 section lies in the rightmost position, though not being the section with the highest-grade. Interestingly, the same procedure was implemented for classification scenarios but the classification results were not noticeably influenced by any single section removal, which can be explained because of the relatively broad classification ranges. For the relevant case of Section 5, this sample has a measured grade of 27.6 %, and is usually wrongly predicted in the regression with a grade much in excess

of that value, although still falling in the medium grade ore class, that covers the range 20≤CaF2<45%, so being correctly classified.

After removal of section #5, the PCA and CA selection is redone with the remaining 35 white and 31 UV light scans and also with combined WUV data. The cumulative total variability explained by each principal component is different but the number of components considered as inputs are the same as in Table 27 for W and UV sets, while one less component is required for WUV. For CA results, significantly correlated PCI are fewer for both white light and UV light scans compared with the results before removing section #5; however, significant correlations are still obtained for most of the blue percentiles from white scans. The detailed results can be seen in Table 51 and Table 52 in the Appendix 3. A summary of the main metrics statistics (where 30 new random divisions were employed for each scenario) can be seen in Table 53 and Table 54 in the Appendix 3. It appears as if principal component inputs are more sensitive to outliers in the UV scenario, both W and UV light (left graphs in Figure 46) than the straight variables selected as having a better correlation with grade (right graphs in Figure 46). In the W_{CA} analysis, section 5 barely exceeds the 95 % threshold, see Figure 46 upper right plot, while in the UV_{CA} analysis, the removal of section 5 does not relevantly affect R^2. Figure 47 and Figure 48 shows boxplots of the distributions of the metrics for SVR, and Figure 49 and Figure 50 for SVC where the training and testing set performance is presented by green and blue box, respectively.

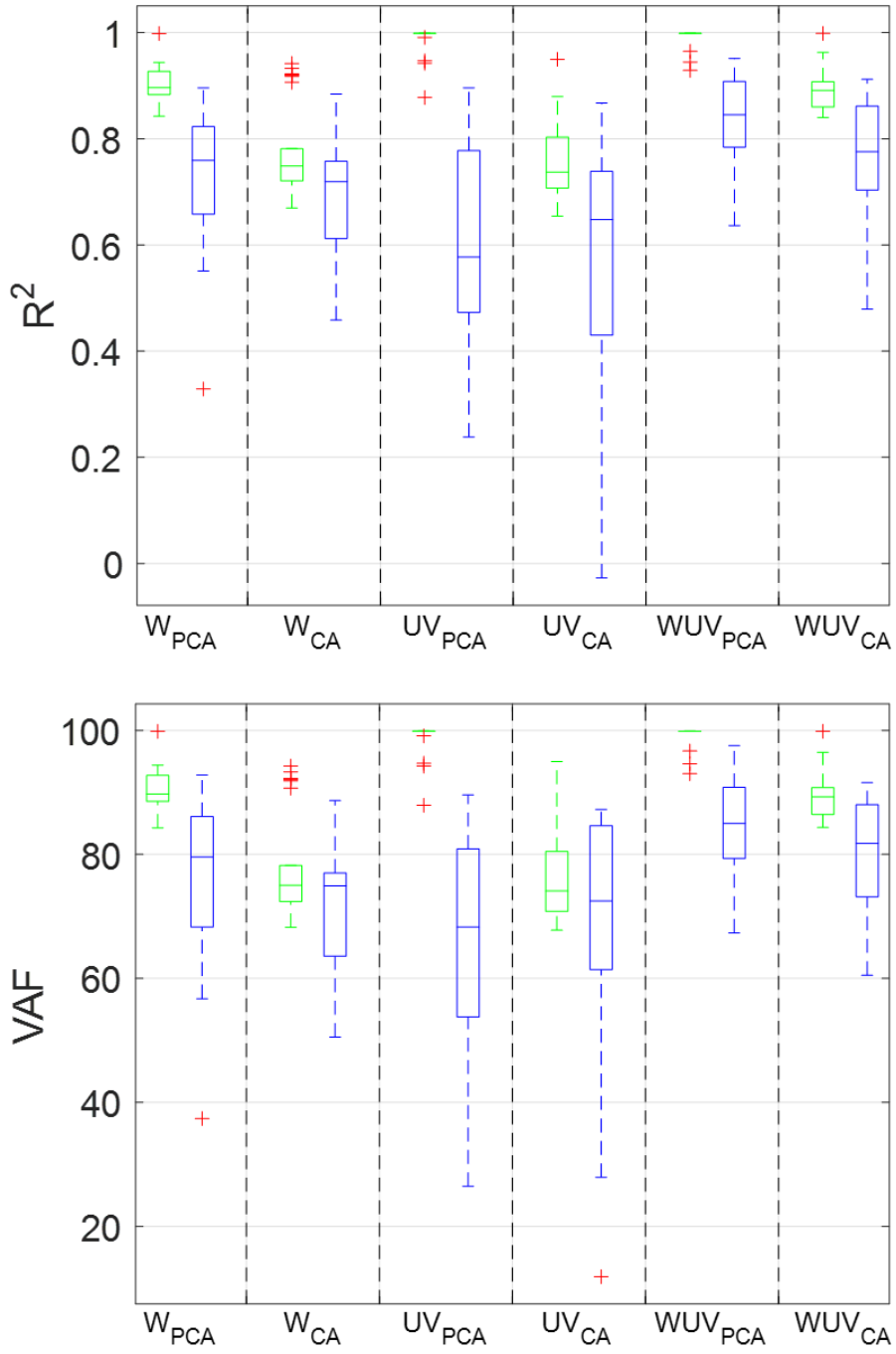


Figure 47. Summary of the prediction performance of R^2 and VAF for the six scenarios considered after removing section #5; green boxes correspond to training and blue boxes to testing; refer to Table 29 for descriptions of the metrics.

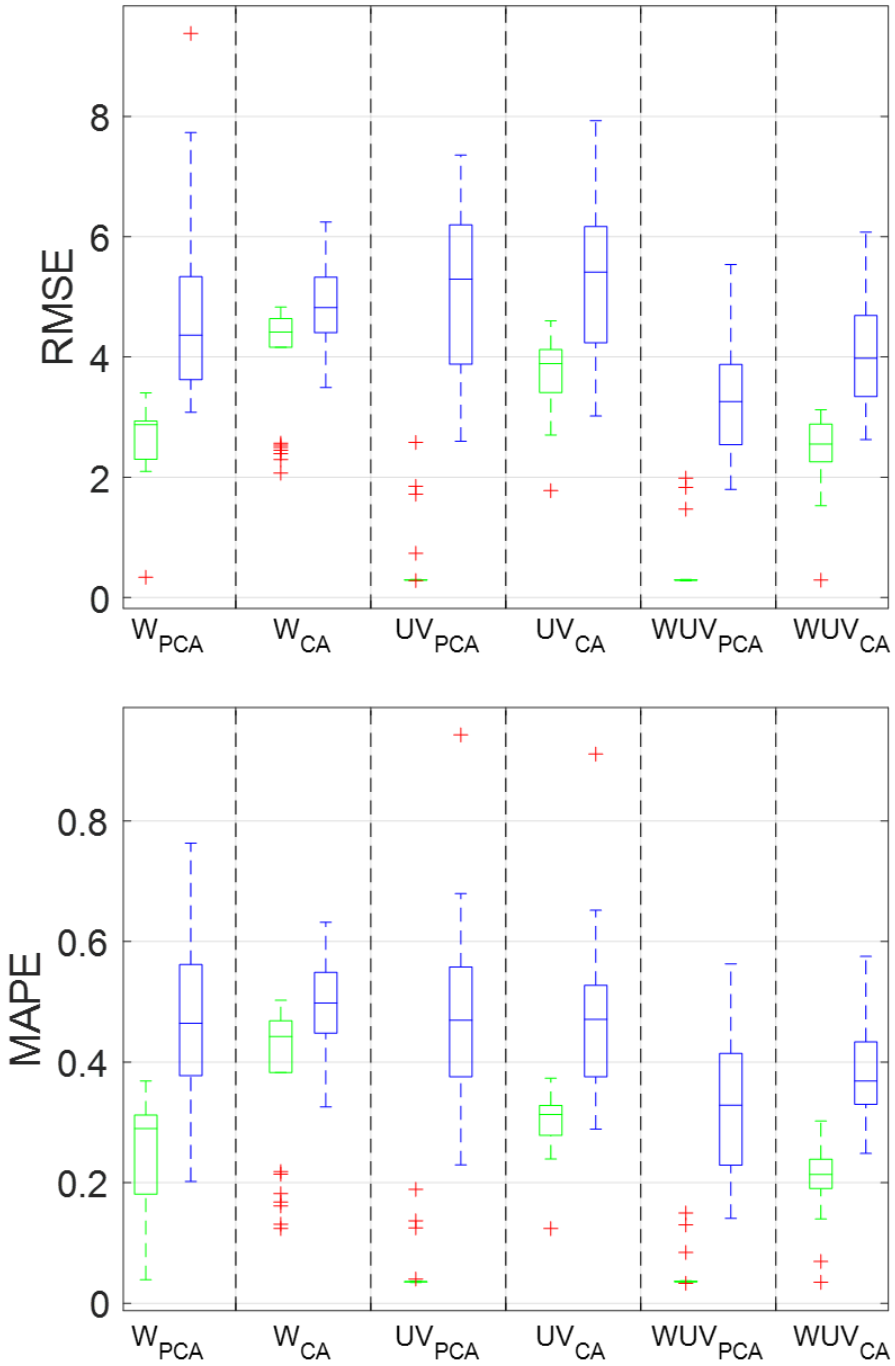


Figure 48. Summary of the prediction performance of RMSE and MAPE for the six scenarios considered after removing section #5; green boxes correspond to training and blue boxes to testing; refer to Table 29 for descriptions of the metrics.

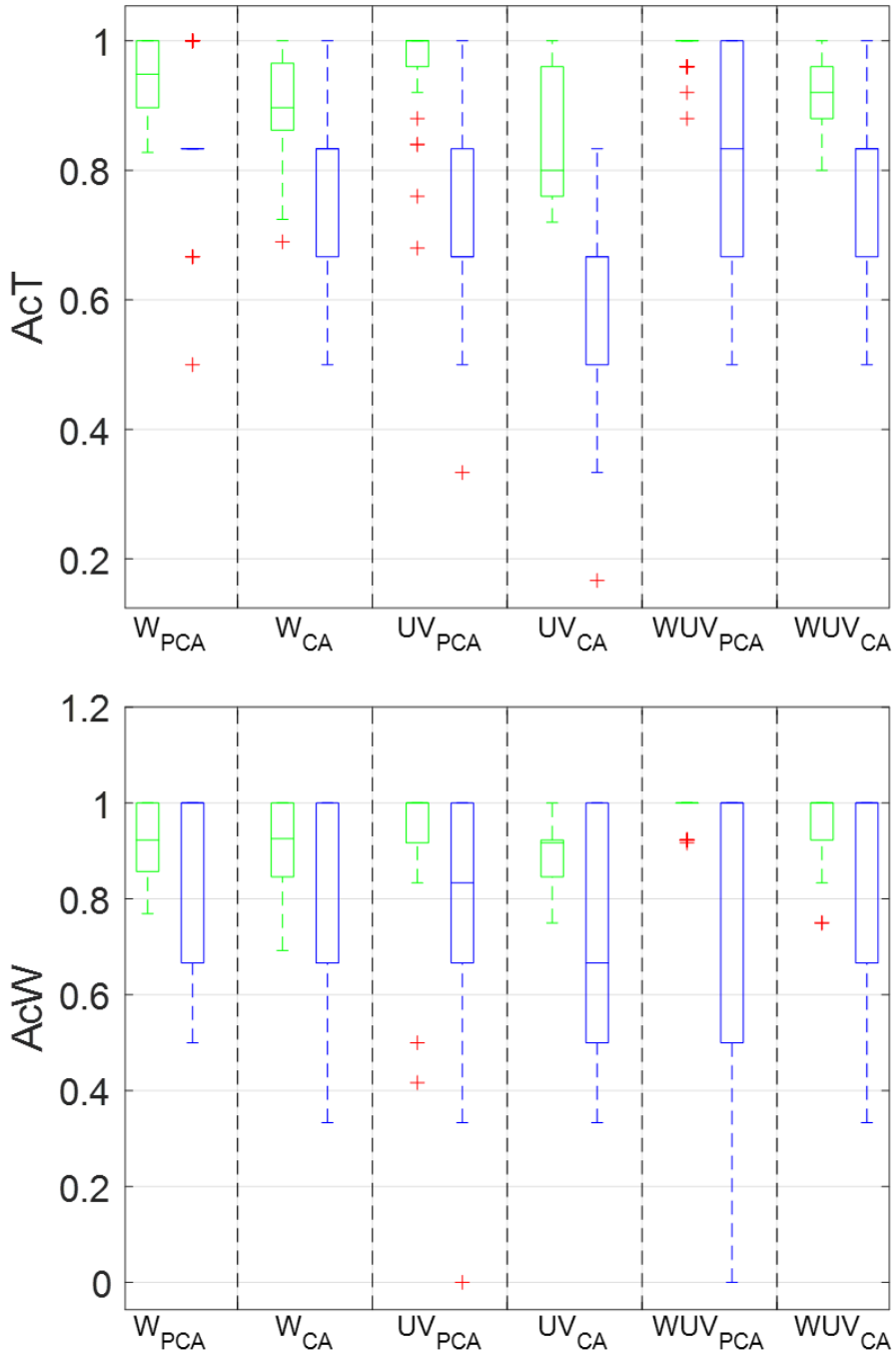


Figure 49. Summary of the prediction performance of AcT and AcW for the six scenarios considered after removing section #5; green boxes correspond to training and blue boxes to testing; refer to Table 29 for descriptions of the metrics.

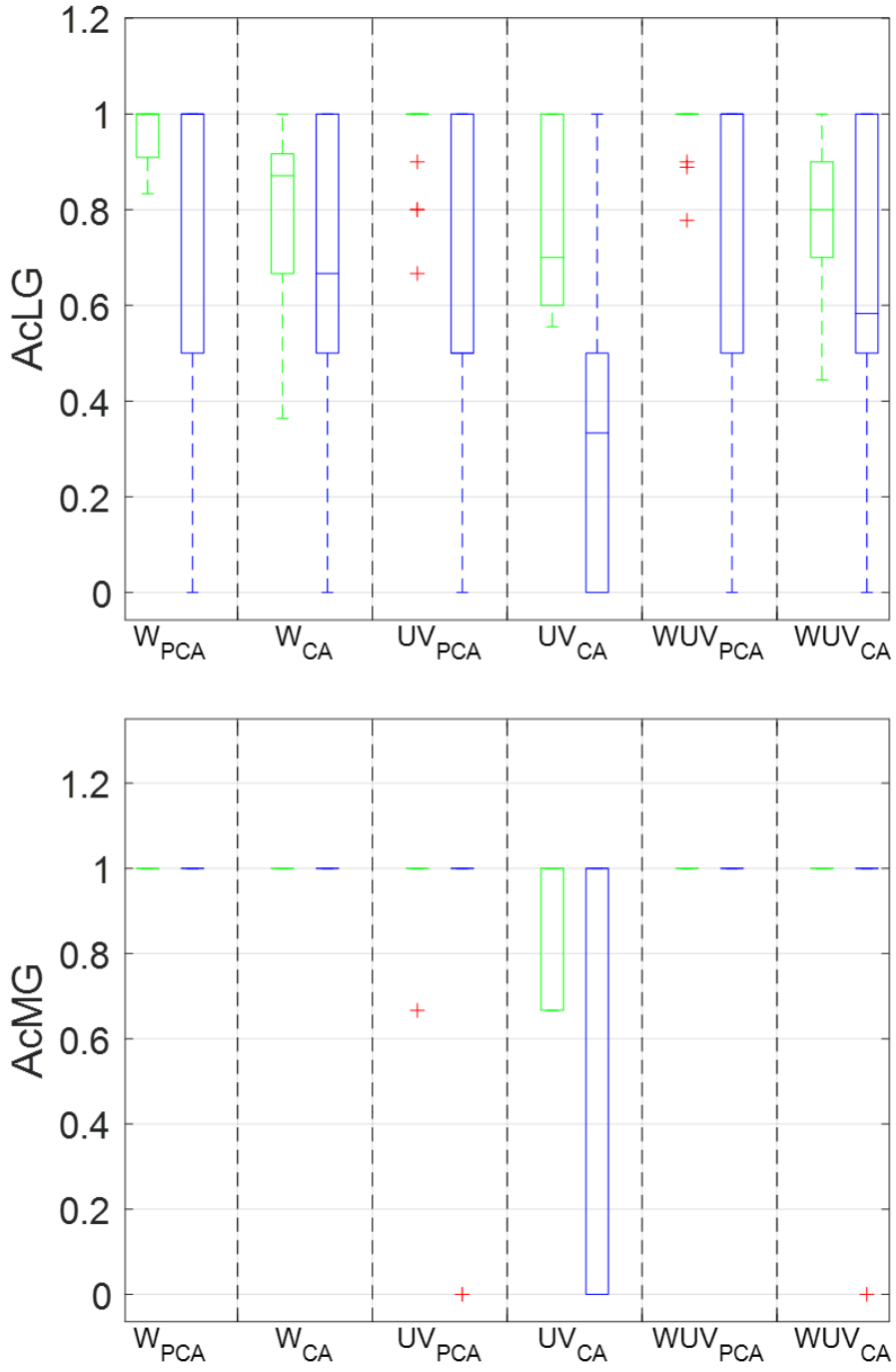


Figure 50. Summary of the prediction performance of AcLG and AcMG for the six scenarios considered after removing section #5; green boxes correspond to training and blue boxes to testing; refer to Table 29 for descriptions of the metrics.

In both regression and classification, PCA provides in general better results than CA in terms of mean and dispersion of the metrics considered. This is more evident for regression when the combination of colors from white and UV lights (WUV) is considered. These results indicate that CA can capture some significant optical information from televiewer scans, but probably ignores some supplementary information

that can contribute to characterize the fluorite grade, while the linear combinations of PCI in the principal components provide with a richer information on the pixel properties.

For regression testing sets, UV has worse R^2, VAF and RMSE than W light and slightly better MAPE. The mean and standard deviation of the RMSE for PCA are 5.15±1.32 % (UV_{PCA}) and 4.65±1.42% (W_{PCA}), with mean R^2 of 0.61 and 0.73, respectively. The best predictions according to the four metrics considered are obtained when both light sources are combined (WUV_{PCA}); the mean RMSE is then 3.32±0.90%, with a determination coefficient of 0.83±0.09.

The inputs from W_{PCA} provide the best classification accuracies for all rock classes: 0.84±0.12, 0.84±0.19, 0.77±0.30, and 1 for total, waste, low grade and medium grade for testing sets. For the waste, this means that 16% of the samples are classified wrongly as ore in average terms, while 23% of low-grade ore sections are misclassified. All medium grade ore sections are classified correctly for the 30 divisions of the dataset. Low grade ore is generally the category with a worse classification accuracy compared with waste and medium grade. These metrics involve that no further chemical assaying is required when a sample is classified as medium grade, while the classification into the other rock classes should be taken as first estimation of the ore grade, that needs to be confirmed through conventional assaying, especially for low-grade ore, to increase the reliability of the results.

4.7 Limitations

In future studies, there are still some shortcomings need to be addressed. At first, enlarging the database would likely improve the model performance and thus increase the significance of the results, allowing a further reduction in the amount of chemical analyses. Secondly, for case studies where it is not possible to enlarge the dataset available for the model development, repeated cross-validation is a good strategy to deal with relatively small databases as shown in this chapter. The repeated cross-validation can also be combined with other machine learning or deep learning techniques (Phoon and Zhang, 2023), such as convolutional neural networks (Zhang et al., 2021), random forest (Fernández et al., 2023; Liu et al., 2023), or recurrent neural networks (W. Zhang et al., 2022), all of them applied to improving the robustness of prediction models in geoscience or geotechnical issues. Finally, the proposed televiewer-based procedure can be applied on other types of minerals and various light sources are worthwhile to be investigated.

4.8 Conclusions

This chapter proposes a new fluorite grade prediction approach based on RGB values obtained from optical televiewer scanning and machine learning techniques in the Lújar underground mine located in Órgiva, Granada, Spain. RGB intensities of the pixels are procured from borehole walls by televiewer scanning with white and UV light illumination. The composition of the rock corresponding to the image logs was determined by chemical analysis of the drill cuttings sampled and used as the output. Percentiles of the color intensities (PCI) of borehole images are used as input parameters

for regression and classification issues of fluorite grade. Three types of color information have been tested, comprising PCI from W light scans, UV light scans and the combination of both (WUV). Two kinds of feature extraction techniques are employed for input selection: the first one is from the significantly correlated inputs with fluorite components; the second one is from PCA technique.

The support vector machine (SVM) is used to establish the prediction models. The hyperparameters of the SVM (C and γ) are optimized using a salp swarm metaheuristic algorithm. The results of the prediction models are assessed by repeated cross-validation and rated with classical statistical indicators. A "take one out" method is proposed for outlier data detection. One of the data sections was removed with this method, resulting in an improved prediction capacity.

In general, the combination of white light and UV light scans is more effective to predict fluorite grade from regression. If a single light source is used, the white light would be recommended. The average regression results for testing sets are $R^2 = 0.83$ and RMSE = 3.32% from WUV_{PCA} scenario. For classification, the best result is obtained with white light, W_{PCA}, with average classification accuracies, of 0.84 (total), 0.84 (for waste), 0.77 (for low grade ore) and 1 (for medium-grade ores).

The relatively low-cost and convenience of ore image procurement and processing makes this novel approach robust and easy to implement for fluorite grade prediction. Given the limited errors and acceptable prediction accuracies, the approach described here can be used as a first assessment of the fluorite grade, helping to save a fraction of laboratory analysis work. Additional work is needed in order to investigate the reliability e.g. with other fluorite ores or other minerals. The collection of larger datasets would improve the significance of the results, but, since the models are intrinsically site dependent, they can hardly be generalized to other operations. However, the method proposed can be adapted to other mining sites in order to develop an ore grade prediction model based on RGB intensities of the images of the blasthole walls. Such a methodology will reduce the time offset for grade control, especially when medium ore grade is observed, allowing to detect these areas from the first steps of the drilling and to take prompt decisions on mine development. The availability of a system that would automatically log the boreholes would boost this procedure towards a nearly online assessment of the ore grade of the deposit. The potential of UV light scans still needs to be explored, where different wavelengths are likely worthwhile investigating to improve the prediction accuracy of fluorite ore.

Chapter 5. Fluorite ore recognition using spectral clustering and smartphone digital images calibrated with a ColorChecker: A case study at the Lujar underground mine, Spain

Nomenclature

avB (Mean of blue pixel intensities)	P5 (Blue Flower)
avG (Mean of green pixel intensities)	P8 (Neutral 8)
avR (Mean of red pixel intensities)	P21 (Neutral 6.5)
Ac (Accuracy)	PCI (Pixel color intensity)
AHCF (Agglomerative Hierarchical Cluster Tree)	PC (Principal component)
BMU (Best matching unit)	PCA (Principal component analysis)
d (Euclidean distance)	Pr (Precision)
D (Dolomite)	Re (Recall)
FN (False Negative)	RGB (Red, green and blue)
FP (False Positive)	ROI (Region of Interest)
Fs (F1 score)	SC (Spectral clustering)
GLCM (Gray-level co-occurrence matrix)	SDE (Standard deviational ellipsoid)
GMM (Gaussian Mixture Model)	SOM (Self-Organizing Map)
HM (Medium-high or high ore grade)	Sp (Specificity)
KNN (K-nearest neighbor)	Tc (Texture-correlation)
L (Limestone)	TN (True Negative)
LLRBF (Local linear radial basis function)	TP (True Positive)
LO (Low ore grade)	U (Uncertainty)
O (Ore)	W (Waste)
P2 (Light skin)	XRF (X-Ray Fluorescence)

5.1 Introduction

Ore grade assessment and prediction play a crucial role in the mining operations and resources exploitation, their risk evaluation and environmental sustainability. Accurate estimation of ore grades allows to optimize mining processes, allocate resources effectively and contribute to sustainable mining resource utilization. Generally, the most accurate and direct way to determine the ore grade is by chemical component analysis, a time-consuming and costly process that is difficult to circumvent. A classical mathematical technique to limit the need of chemical data is kriging (Ali Akbar, 2012; Emery, 2006; Hekmatnejad et al., 2019). It assumes that variables have a certain degree of linear spatial correlation and by calculating the spatial correlation between sample points, an optimal spatial interpolation method is derived for estimating the numerical values of the unsampled area. These assumptions make kriging methods to be sensitive to the choice of the appropriate variogram function (Oliver and Webster, 2014; Van Groenigen, 2000) to interpret the interpolation results. In addition, the kriging method is also influenced by the sampling distribution and not applicable to nonlinear problems.

Since ore grade can be considered a variable related to spatial distribution, lithology, drilling parameters and some other rock characteristics, some scholars proposed to integrate these influential factors in the prediction of ore grade by machine learning techniques (Chatterjee et al., 2010a; Kaplan et al., 2021; Mahmoudabadi et al., 2009; Samanta et al., 2006; Samanta and Bandopadhyay, 2009). For instance, Samanta (2010) employed spatial coordinates of the offshore placer gold deposit to be inputs and utilized a radial basis function to predict the gold concentration. The predicted results showed that the radial basis function had better prediction performance than ordinary kriging method and multi-layer perceptron. Tsae et al. (2023) collected 14,294 drilling borehole datasets where lithology, alteration, deposit coordinates and drilling properties were used as inputs and the copper grade was the output. The artificial neural network presented better overall prediction results than some classical machine learning techniques such as random forest, linear regression and light gradient boosting machine. Jafrasteh and Fathianpour (2017) developed some optimized local linear radial basis function (LLRBF) neural networks to fit the skewed drill hole data from a phosphate deposit. The rock lithology and the deposit coordinates were considered as inputs and the phosphate grade value was the predicted target. In that case, the simultaneous perturbation artificial bee colony algorithm-based back-propagation approach presented the best optimization performance for LLRBF-based ore grade prediction model. Kaplan and Topal (2020) combined K-nearest neighbor (KNN) and multi-layer feed-forward neural network to forecast the grade distribution of a gold deposit where the KNN was used as an indirect approach to the determination of rock types and alterations for unsampled positions and neural network was used for developing the prediction model. It could be found that the participation of rock type and alteration improved the model effect compared with the individual usage of sample coordinates to be inputs. Though these studies provided valuable reference for the ore grade assessment, however, auxiliary measurements of drilling properties and rock characteristics also increase the workload.

In this study, a set of 494 pellet samples made from drilling chips are photographed by a smartphone to recognize fluorite grade. To avoid the impact of environmental factors on the image colors, a ColorChecker is used for color correction. Some significant color and

texture properties are processed by a principal component analysis for developing clustering models. As a result, different clustering can be labelled as waste or fluorite with different occurrence probabilities. This study provides a cheap and fast screening criterion for new pellets. Hence it would be assisted to recognize waste, lowering grading costs and increasing the availability of the equipment used for chemical analysis.

5.2 Literature review

Currently, the image properties have been widely used for ore grade assessment or ore recognition due to the fact that minerals may exhibit specific colors and textures (Chatterjee, 2013; Chatterjee et al., 2010b; Chatterjee and Bhattacherjee, 2011; Shatwell et al., 2023; Zhang et al., 2014). For instance, Liu et al. (2024) designed a novel mineral recognition framework named OreFormer where the modified convolutional neural networks were used. Liu et al. (2021a) applied four deep learning networks to conduct coal classification where anthracite, gas coal and coking coal with different density grades were predicted appropriately. Patel and Chatterjee (2016) designed a probabilistic neural network-based model on limestone samples. The features were extracted from the three-color channels, i.e., red, green and blue. The predicted results indicated that the proposed model could achieve misclassification lower than 6%. Qiu et al. (2021) utilized seven image features extracted from the gray histogram for ash content prediction of coal. These seven features included probability of each gray level, mean, variance, skewness, kurtosis, energy and entropy. The best prediction performance could be achieved by polynomial regression after feature selection with 4% relative error. Liu et al. (2021b) employed a typical convolutional neural network named VGG net to classify gas coal, coking coal and anthracite with different water content. The VGG net could achieve a prediction accuracy in excess of 94% for all three rock types.

Research gap: With the advancement of portability, connectivity and camera resolution of smartphones, they have been widely applied across numerous image analysis domains, such as the segmentation of images (Campos-Taberner et al., 2016), digital image colorimetry (Fan et al., 2021) and photomicrograph acquisition (Roy et al., 2014). However, it is rarely considered for procuring the color information from the tasks of ore grade estimation. This study proposes a smartphone as photography device to procure fluorite-based pellet images, to be analyzed for fluorite grade recognition and prediction. In order to overcome the variations in lighting conditions which may critically introduce noise or inconsistencies in image colors, this study proposes to employ a ColorChecker (Pascale, 2005) that allows correcting the colors of pellet images to obtain the genuine color of each pixel.

Significant color characteristics are extracted from the corrected pellet images and used for predicting the fluorite grade by a simple unsupervised machine learning technique. As a background for this approach, the authors demonstrate that color information of images of in-borehole walls drilled in the same mining site are sensitive to the fluorite content that it is predicted with reasonable accuracy (Li et al., 2023a).

5.3 Data collection and description

5.3.1 The pellets

Drilling chips of production blastholes and exploration boreholes from Lújar mine (F/Pb/Zn) were collected while drilling for grade control purposes; refer to Section 4.3. for more details of Lújar mine. The samples are quartered, dried, grinded to a size below 80 μm, and further quartered to obtain a 10 g sample (mixing of 95% sample with 5% Fluxana CEREOX wax as binding agent) from which pressed powder pellets of 32 mm of diameter and 5 mm height are prepared. The pellets were analyzed in the mine laboratory as was made with the drilling chips (see Section 4.3.1). This provides the amount of calcium fluoride, calcium carbonate, calcium magnesium carbonate (dolomite), silica, ferric oxide, aluminum oxide, zinc and lead.

A dataset composed by 494 pellets from Lujar underground mine is considered for this analysis. The samples were collected from levels 70 (198) and 345 (296) during 2020 and 2021. The distributions of the main chemical compounds of the pellets are shown in Figure 1; the kernel density estimate (Holger, 2015) of the probability density (grey patch) is overlapped with the box-and-whisker diagram. Fluorite (CaF_2), limestone (L) and dolomite (D) are predominant (left graph), while the oxides are less abundant (right graph, note the different Y-axis scale). The pellets are dolomite-rich in general, with a fluorite content below 20% in most of the cases. Pellets with fluorite as predominant compound are scarce (see blue dots in the fluorite distribution, left graph, Figure 51), while fluorite is less than 3.6% in 124 pellets (a 25% of the total) that are rich in either dolomite or limestone. This reflects the chemical composition of the deposit, with typical grades in the order of 15% in fluorite, complicating the in-situ ore recognition (Amor and Navarro, 2016). Silica is the predominant oxide in the pellets (see right graph, Figure 51), with a 75 percentile of 6.5%. The amount of the other oxides is in general smaller, though in some pellets they are atypically high which could affect to the color. There are also traces of zinc and lead, lower than 1.5% in 95% of the pellets; for pellets with a large lead (>30%) amount, the color becomes dark (i.e. greyish-black).

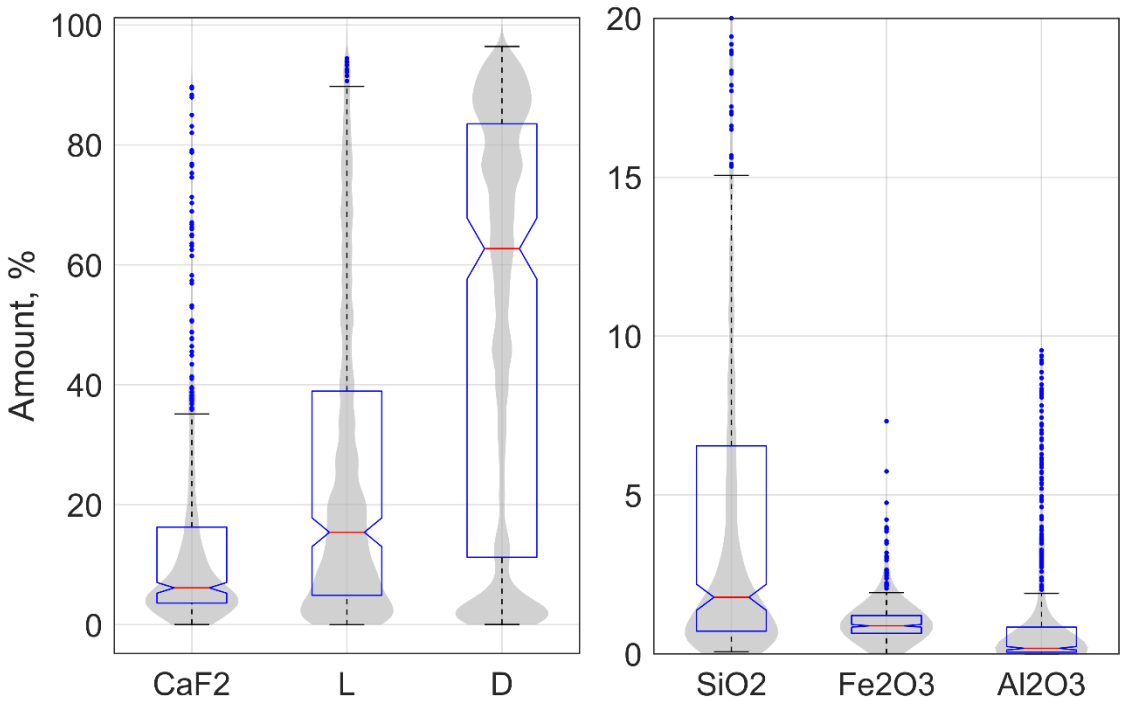


Figure 51. Violin and box plots of the main compounds of the pellets: major (left graph) and minor (right graph) compounds.

5.3.2 Experimental layout

The pellets were photographed with a smartphone Porsche design Huawei Mate 40 RS with LEICA OPTICS and a resolution up to 50 Mpx following the layout in Figure 52. Sunlight was used as the light source. The pellets were placed with the clean side (i.e. opposite to the damaged one by the X-Ray beam) towards the camera. A standardized television test pattern, i.e., PM5544-PAL test pattern (see Figure 52) is used as the background to provide more color information for color correction. Some camera settings like F-number (1.9) and the focal length (27 mm) are constant. The shutter speed or the ISO are automatically tuned by the smart phone. They are in the range 1/451 – 1/98 s for exposure time and 50–160, respectively. The images are recorded in "jpg" format. Variations in the camera settings and in the light intensity distort the pellet colors in the images. To assess these errors and correct them, a low-cost 24 patch ColorChecker Classic Mini target (63.5 x 109.0 mm) manufactured by "X-Rite" was placed on the top part of the image in the same relative position with respect to the pellet. The ground sampling distance is 0.0687 mm.

Figure 52. Experimental layout for the measurement of pellet colors.

5.4 Color evaluation and correction

A script was written in MATLAB (The MathWorks Inc, 2021) to crop automatically the pellets and to detect the ColorChecker. The pellet surface photographed includes an identification number and/or damaged parts; in order to leave these out, only a circular region of approximately 5.2 mm of diameter, comprising 4448 pixels, is considered as the Region of Interest (ROI) for the analysis. To avoid the number or damged parts, the selection of ROI part is not constant. The surface composition of the pellets is assumed to be uniform and representative.

An affine transformation of the measured colors (i.e. a constant value is added to the linear combination of the rows of the matrix color correction and the color channels of the image) is made with the MATLAB function *colorChecker*. The purpose is to obtain a color correction matrix $\boldsymbol{C}$, that leads to:

$$\boldsymbol{R} \approx \boldsymbol{M}\boldsymbol{C} \tag{5.1}$$

where $\boldsymbol{R}$ represents the corrected image matrix, an $n_P \times 3$ matrix with the red, green and blue intensities of the n_P patches of the ColorChecker and $\boldsymbol{M}$ is an $n_P \times 4$ matrix with the

red, green and blue intensities of each patch in the original images with a 4th column of ones.

The matrix $\boldsymbol{C}$ has 12 unknown elements, so if all the patches are considered, Eq. (1) is an overdetermined system of linear equations, which is solved using the left division as follows:

$$\boldsymbol{C} = (\boldsymbol{M}'\boldsymbol{M})\backslash\boldsymbol{M}'\boldsymbol{R} \tag{5.2}$$

Where $\boldsymbol{M}'$is the transposed of $\boldsymbol{M}$. For $\boldsymbol{M}$ square (n_P =4), Eq. (5.2) is equivalent to:

$$\boldsymbol{C} = \boldsymbol{M}^{-1}\boldsymbol{R} \tag{5.3}$$

The objective of this calculation is to obtain the correction matrix C that minimizes the difference between the nominal color intensities of the patches and their color intensities in the image. That correction matrix is then applied to the pellet area. This process is applied to every pellet.

The mean Euclidean distance (*d*) (Danielsson, 1980) is used to measure the distance between pellets and patches, explained in Eq. (5.4); where *x* and *y* represents the coordinates of pellets and patches, respectively, *n* denotes the number of pellets. Figure 53 (left graph) shows that the mean colors of each pellet in the red, green and blue (RGB) color space are concentrated around three patches. These are: Neutral 8 (P20; *d*=0.175), Neutral 6.5 (P21; *d*=0.197), and Light Skin (P2; *d*=0.229) where P represents the patch id shows in the right graph in Figure 3. The next closest color patch, the Blue Flower (P5), is at a distance of 0.365 to our pellets and is already outside the 99 % coverage of the colors of the pellets (blue parallelepiped). The longer distances for the rest of the patches, 0.402 – 0.902 suggest that they do not appear in our pellet colors.

$$d(x,y) = \frac{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}}{n} \tag{5.4}$$

Figure 54 shows as a reference, the errors in the original photos between the measured and the reference colors of the four closest patches to our pellets (P2, P5, P20 and P21). If the colors of the photos are transformed considering these four-color patches to calculate the matrix $\boldsymbol{C}$ in Eq. (5.3), the reproduction of these colors is excellent (high accuracy and small dispersion) in all the images. However, it means that there is a cost of higher errors for the rest of the color patches (the distribution of errors shifts towards higher values and becomes wider). This is not a problem as these colors are well separated with respect to the colors of our pellets as can be seen in Figure 53 (left graph). If the number of patches is increased to e.g. 14 (i.e. patches at mean distances below 0.61 are considered) to obtain the correction matrix **C**, the discrimination capacity is worse, and the rest of the colors are not better reproduced than in the original photos. Figure 55 shows as an example, the original and the transformed photos using Light Skin (P2), Blue Flower (P5), Neutral 8 (P20) and Neutral 6.5 (P21) patches for color correction; the change in the colors is apparent.

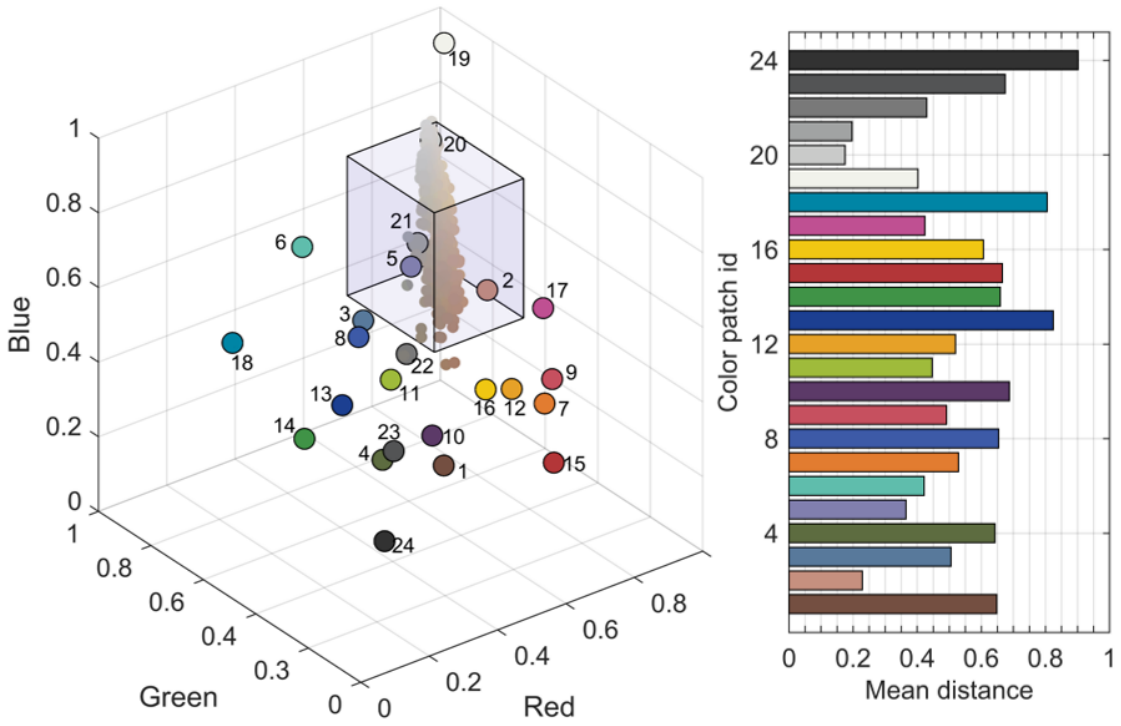


Figure 53. Position of the color coordinates of the ROIs of the pellets versus the patches of the ColorChecker. Left: Scatter plot of the mean pellet colors in RGB space (the parallelepiped shows the 99 % coverage region of the mean pellet colors). Right: Mean Euclidean distances between the colors of the pellets and of the patches of the ColorChecker.

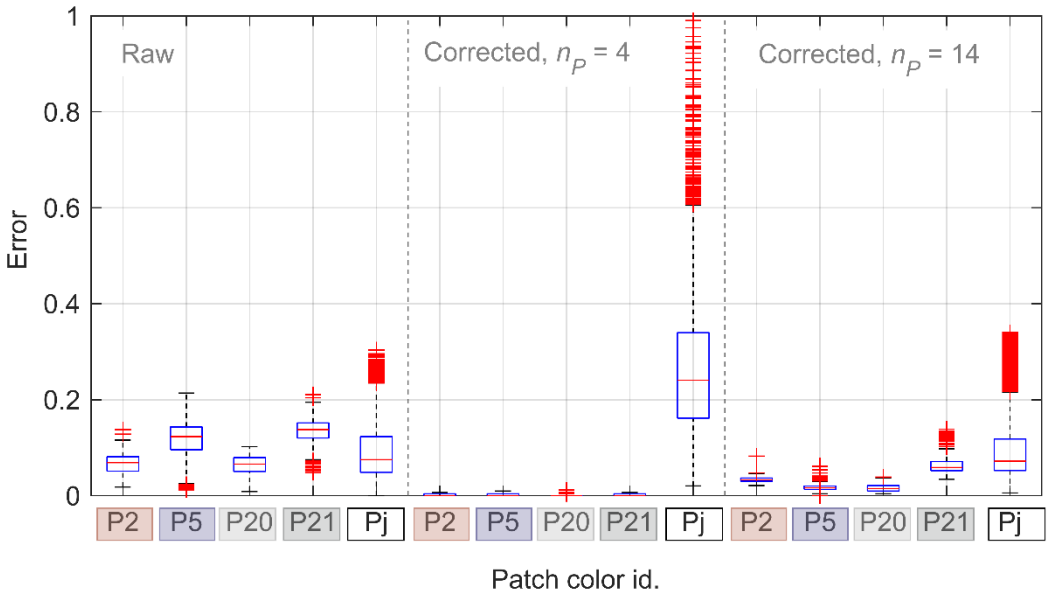


Figure 54. Euclidean color distance between measured and reference color intensities in the normalized RGB space for Light Skin (P2), Blue Flower (P5), Neutral 8 (P20), Neutral 6.5 (P21) and the rest of the color patches (j=1, 3, 4, 6–19, and 22–24) for 494 images.

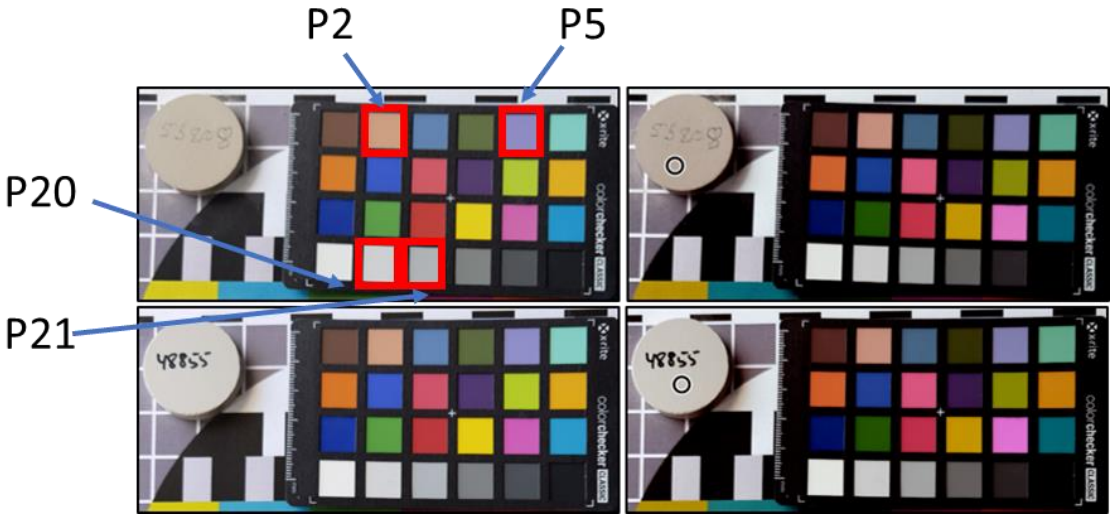


Figure 55. Original (left) and transformed (right) images for pellets 53208 (top) and 48855 (bottom); the black circle is the ROI considered for the analysis.

5.5 Model definition and development

5.5.1 Model definition

Various color parameters are calculated, however, only a few parameters have stronger correlation with the fluorite grade and thus selected to develop the model. The pixel color intensities (PCIs) of the corrected photos of the ROI of each pellet are described with the mean of the red, green and blue intensities of the pixels, that are abbreviated as avR, avG and avB, respectively. The resulting distributions are shown in Figure 56; mean red intensities range from 0.85 to 0.55 while blue intensities are the smaller ones, down to 0.3. A comparison of the median of each color channel before and after correction (green versus red horizonal lines, respectively), indicates the color shift towards slightly darker colors. The Blue Flower patch (P5) stays now inside the 99% coverage of the mean intensities together with the Light Skin (P2), Neutral 8 (P20), and Neutral 6.5 (P21) patches. Mean green intensities are strongly correlated with the mean of the red and blue intensities with Spearman coefficients of 0.88 and 0.98, respectively and are discarded as an input for predicting fluorite grade.

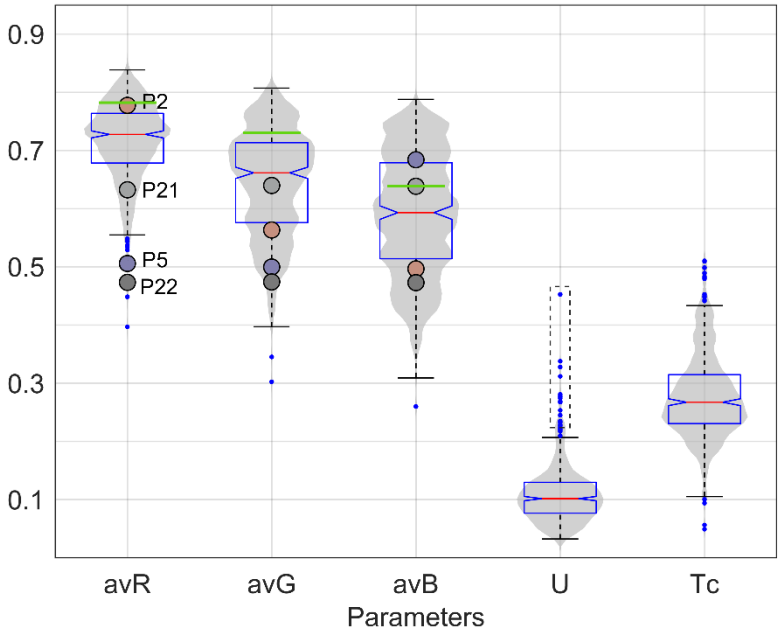


Figure 56. Violin and box plots of the descriptor parameters of the images of the ROI of the pellets: mean of red, green and blue intensities (avR, avG, avB, respectively), uncertainty (U) and texture-correlation (Tc). The circular markers show the colors of the patches lying within the 99 % coverage of the mean color intensities. The green horizontal lines are the median of the pixel intensities of ROI in raw pellet images.

The dispersion in the color clouds for each pellet in the RGB space is described with the standard deviational ellipsoid (SDE), which is extensively applied to investigate uncertainty of datasets in 2D and 3D spaces (Yang et al., 2020). The SDE is calculated with the covariance matrix of the PCIs under the assumption that the data comes from a normal distribution using the MATLAB function (*error_ellipse*) defined by Johnson (2023) (A.J., 2023). Its major axis indicates the direction of largest variability, and the scatter in PCIs is taken as the length of the major semi axis. Other directions don't present apparent variability. Therefore, only this single parameter describes the uncertainty (U) in the PCIs and is used instead of the standard deviation, or other dispersion measurements, of the color pixel intensities.

Figure 57 shows, as an example, the color intensities of a pellet in the RGB space and the SDE at a 95% confidence level. This ellipsoid is representative of the uncertainty observed in the rest of the pellets; it is elongated with the major axis significantly longer than the other two having an angle with the green axis of 46.3° sd 1.0°, and an inclination with respect to the blue axis of 54.2° sd 1.6° (mean and standard deviation from all the pellets). The major semi-axis with a length of 0.0895 is close to the median uncertainty of all pellets, and it lumps mainly dispersion in the red and blue colors. The minor semi axes are both about 0.0168.

Figure 56 shows that most of the pellets' images have uncertainties in the PCIs in the range 0.03-0.2 (see the extremes of the whiskers in the U series plot), while few have atypical high uncertainties, up to about 0.45, see the blue dots). Such large dispersion is due to the presence of black spots in the pellet surface that cannot be associated to

differences in the chemical composition and may be a consequence of the pellet degradation. This is apparent in the first four pellets shown in Figure 58, which uncertainties in the PCIs and the distance about the median normalized by the median absolute deviation about the median (Z_U) given in the top part decrease from left to right. These spots are scarce in the rightmost pellet of Figure 58, with Z_U below the threshold value of 3.819 that corresponds to a 99% coverage assuming a normal distribution. Pellets with $Z_U \geq 5$ are discarded (Miller et al., 2018) since they may bias the prediction of fluorite content. These are 15 pellets with similar ID number (hence preparation dates), which uncertainties are marked with a dashed rectangle in Figure 56. Hence 479 pellets are retained for the analysis. Note the green spots in the pellets' numbers in Figure 58; these are produced by using four color patches to correct the images, and they do not influence the analysis since they are outside the ROI.

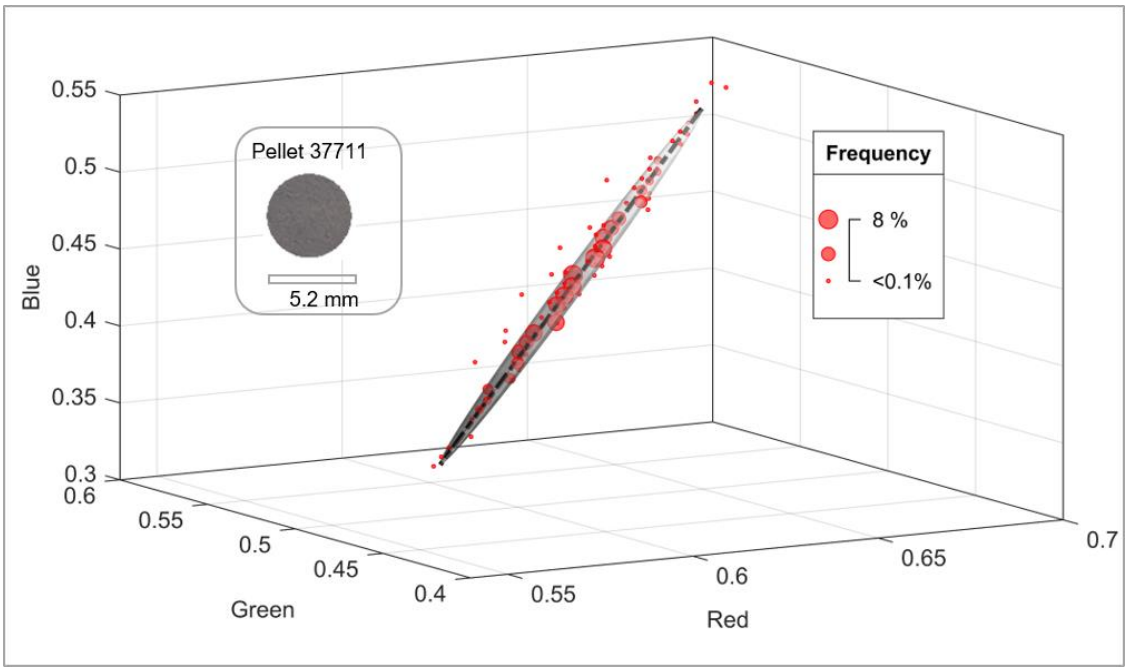


Figure 57. Color pixel intensities in the normalized RGB space of the ROI of the corrected image of pellet 37711 and its 95 % confidence SDE; the size of the points is proportional to their relative probability

Figure 58. Pellets (top) and their ROIs (bottom) with very high (left) to high (right) uncertainties in the PCIs that correspond to the upper and lower dots in boxplot of the U series in Figure 56

The texture analysis is carried out from the gray-level co-occurrence matrix (GLCM) of the ROI of the pellet (Haralick and Shanmugam, 1974). The GLCM is defined as the distribution of co-occurring grayscale values at a given offset in an image. First, the original image needs to be transformed into gray level. The GLCM determines how often pairs of pixels with particular intensity values and offset appear in the image; it can be calculated as follows:

$$C_{\Delta_a,\Delta_b}(x,y) = \sum_{a=1}^{n}\sum_{b=1}^{m}\begin{cases}1, & if\ I(a,b) = x \text{ and } I(a+\Delta_a, b+\Delta_b) = y \\ 0, & \text{otherwise}\end{cases} \quad (5.5)$$

where *x* and *y* are the pixel values of the original image *I*; *a* and *b* define the pixel positions; the offset (Δ_a, Δ_b) specifies the spatial relationship, and *I(a,b)* represents the pixel intensity located at *(a,b).* Then the four parameters can be output by the function "*graycoprops*" in Matlab. In this case, the pixels intensity of the grayscale images is divided into 20 equal width classes, so the intensity level is 1 to 20. The directions considered in the scaled GLCM to calculate how often a pixel with an *i* gray level occurs adjacent to pixels with the *j* value (i.e. an element of the GLCM matrix) are horizontal, vertical and main diagonals (bottom left to top right or 45° and top left to bottom right or 135°). Among the different metrics of the GCLM (Shu et al., 2017), the correlation feature (Tc) is selected. It measures the linear dependence between pixel values in different parts of an image, providing information on how well defined and oriented texture patterns are within the image.

The mean of the correlations for four GLCMs along the aforementioned four directions is considered as an omnidirectional descriptor of the texture of the pellets, which distribution is shown in Figure 56 (rightmost box). It indicates weak to medium linear dependency of grey scale intensities between adjacent pixels. Other metrics like contrast, energy or homogeneity do not show a significant correlation with the fluorite content and are discarded.

To determine the independence among these four inputs, the Pearson correlation was calculated as shown in Figure 59. It can be seen that avR and avB has the highest correlation equal to 0.79. Meanwhile, different input parameters also have some correlation. The lowest correlation is from avR and Tc. Although the number of variables that describe the properties of the ROIs images is not large, Principal Component Anlysis (PCA) (Li et al., 2021b, 2021a) can be applied. It is used to decrease the data complexity and to facilitate to plot the fluorite content as function of the image properties. The resultant principal components (PCs) represent new variables that are not correlated with each other and are formed by combining the original variables in a way that maximizes the variation among them.

The cumulative variance explained by each component is shown in Table 31; the first two PCs explain more 90 % of the total variance, and more than 95% if the third is included. The loadings of the parameters in each PC are also given in Table 31. For the first PC, the mean values of the blue and red pixel intensities have a positive effect and both the uncertainty in PCIs and texture have a negative influence. The second PC is clearly dominated by the texture, and the color intensities have a limited positive effect. For the third PC, the texture has a negligible influence and the largest loading with a positive influence corresponds to the mean of the red intensity of the pixels, while the mean of the blue intensity of the pixels has a negative effect, but with a lower contribution. So, finally, the three components are used for the development of fluorite grade prediction models.

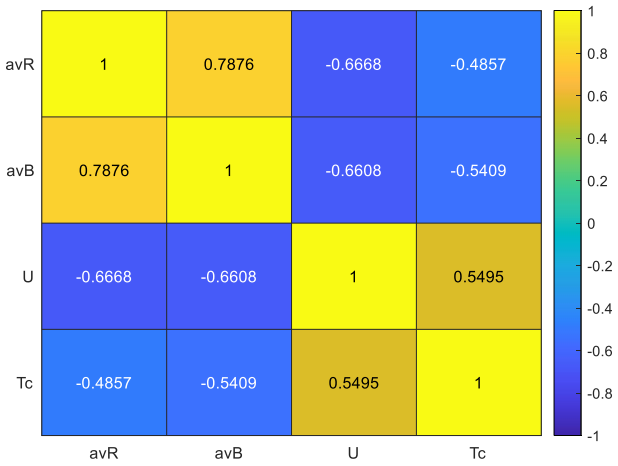


Figure 59. Correlation between inputs using Pearson method

Table 31. Summary of PCA.

	PC1	**PC2**	**PC3**
Percentage explained of cumulative variability in the data	75.97	90.69	96.69
Loading			
Mean of red PCIs, *avR*	+0.44	+0.20	+0.79
Mean of blue PCIs, *avB*	+0.77	+0.37	-0.52
Uncertainty in PCIs, *U*	-0.24	+0.06	-0.30
Texture-correlation, *Tc*	-0.40	+0.90	+0.06

5.5.2 Methodology

To predict the fluorite grade, five popular clustering methods were employed in this study in MATLAB R2021a environment, namely k-means Cluster, Agglomerative Hierarchical Cluster Tree, Gaussian Mixture Model, Self-Organizing Map and Spectral clustering. It is noted that, the main purpose of this study is to develop a new process to discriminate ore and waste, therefore, other more novel methods were not considered. For this, a set of 80% of the pellets (383 observations) is employed for generating the clustering model, while the rest (96 observations) will be used to test the identified patterns in the data; the distributions of fluorite grade and metallic oxides are similar for both data sets. The three principal components from smartphone image parameters would generate different clusters and thus the clustering results would be compared with ore/waste classification as function of the fluorite grade. In this study, three clusters were generated, while different number of clusters, like two, provide worse classification metrics and are not considered. The clusters could represent waste or ore depending on the values of the three components.

5.5.3 Model development

Figure 60 and Figure 61 shows the partition of the training set (circular markers) in three clusters for two of the methods considered, the k-means and GMM; pellets in each cluster are differentiated by different colors. The cluster centroids are marked with orange triangles that can be clearly seen for cluster C1, but are hidden below the pellet's clouds for the other two clusters; their PC coordinates are shown in Table 32. For all the methods, cluster C1 (dark grey circles) has its centroid at a value of the first principal component (PC1) near -0.160, the centroid for C2 (dark green circles) corresponds to a near-zero PC1. While for cluster C3 (dark blue circles), the PC1 coordinate of the centroid is over than 0.15.

Main differences between methods occur at the boundaries between clusters, so the classification of the furthest pellets from the centroids varies with the method. For instance, pellets with PC1 and PC3 around zero, above the main cloud, are assigned to cluster C2 and C3 by k-means (see green and blue circles inside the dashed ellipse in Figure 60), while they are classified into cluster C1 with the GMM (see the dark grey circles inside the ellipse). In addition, nearly all the pellets with PC1 in the range of 0.1–0.2 are assigned to cluster C3 in the k-means method while some of these pellets are labeled as cluster C2 for GMM.

The median silhouette coefficient combines the cohesion and separation of clustering to evaluate the appropriateness of assigning data points to different clusters. The closer the median silhouette coefficient is to 1, the larger is the difference between clusters. The resulting median silhouette coefficient of the data partition in three clusters are 0.3936 (k-means), 0.3713 (AHCF), 0.3084 (GMM), 0.3925 (SOM), and 0.3880 (SC). Although, these values do not suggest a strong structure with limited separation between clusters (Rousseeuw, 1987) as can be visually assessed from Figure 60, they provide a classification of the pellets in three groups based only on the color attributes of the images.

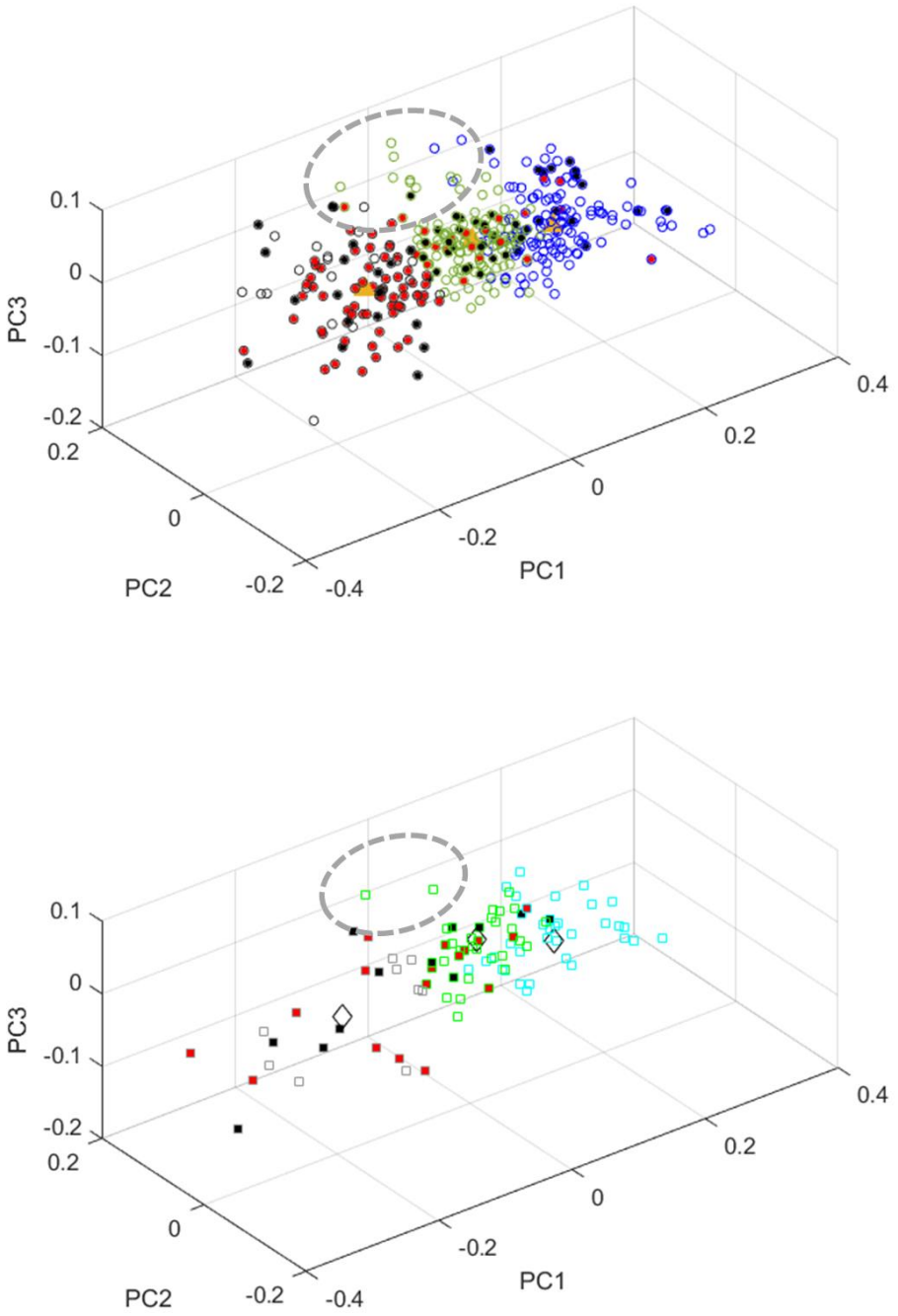


Figure 60. Partition of training and testing datasets in clusters by k-means: (a) training set (top); (b) testing set (bottom). Where C1 (dark grey circles: training set; light grey squares: testing set), C2 (dark green circles: training set; light green squares: testing set) and C3 (dark blue circles: training set; cyan squares: testing set). Orange triangles and black diamond are the cluster centers of training set and testing set, respectively (see their coordinates in Table 2). The filling color of the markers indicates ore grade class: void (CaF_2<10%), black (10% ≤CaF2<20%), and red (CaF_2≥20%).

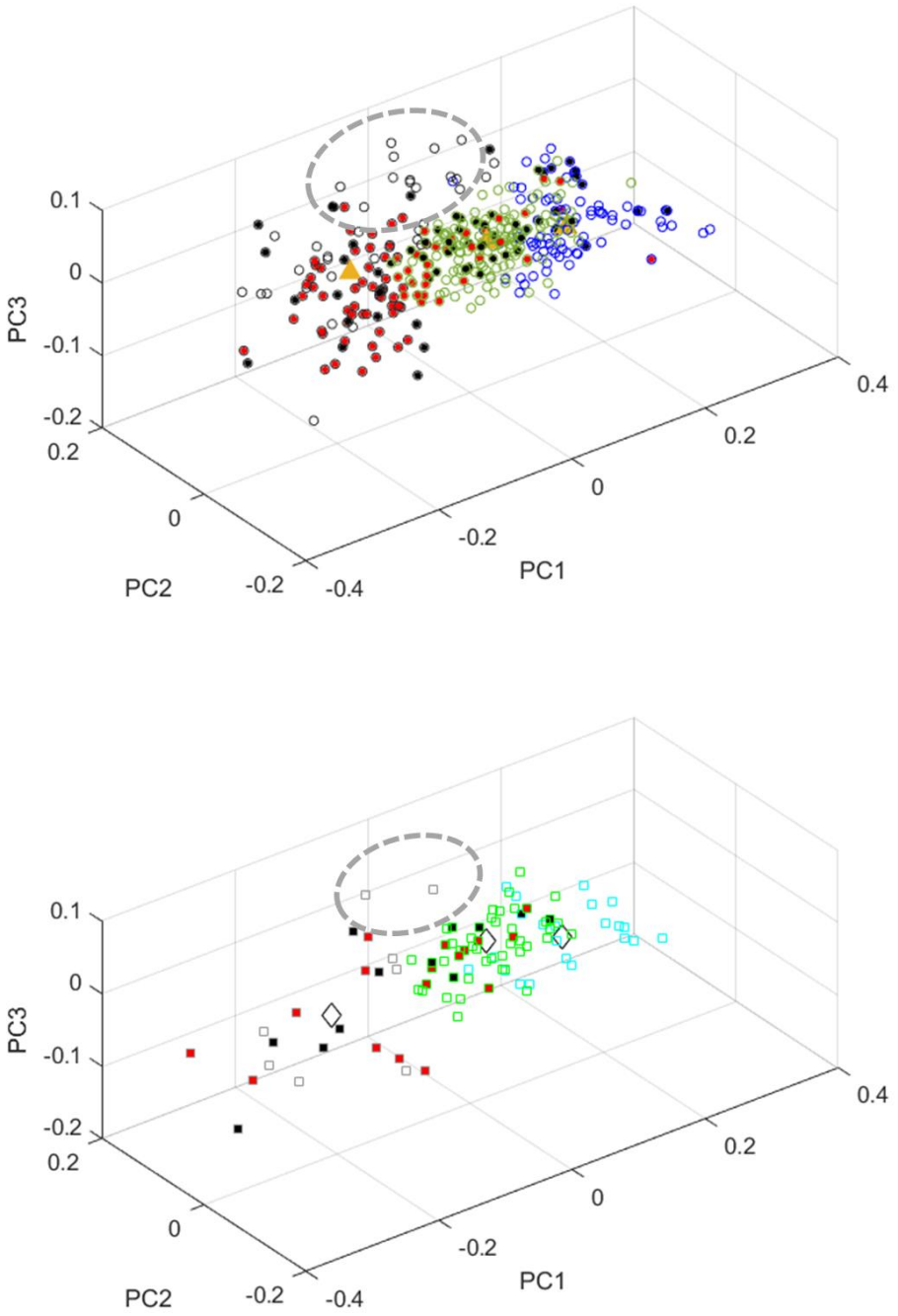


Figure 61. Partition of training and testing datasets in clusters by GMM: (a) training set (top); (b) testing set (bottom). Where C1 (dark grey circles: training set; light grey squares: testing set), C2 (dark green circles: training set; light green squares: testing set) and C3 (dark blue circles: training set; cyan squares: testing set). Orange triangles and black diamond are the cluster centers of training set and testing set, respectively (see their coordinates in Table 2). The filling color of the markers indicates ore grade class: void (CaF_2<10%), black (10% ≤CaF2<20%), and red (CaF_2≥20%).

To assess visually if the mineral composition affects the image properties of the ROIs of the pellets and thus their position in Figure 60, the markers are void for waste (CaF_2<10%), black-filled for low ore grade (10%≤CaF_2<20%), and red-filled for medium-high or high ore grade (CaF_2≥20%). These threshold values were defined according to the mine criterion (Li et al., 2023a). The distribution of fluorite in each cluster is shown Figure 62, and the percentage of pellets of each ore grade type in each cluster is given in Table 33; the cells are filled in dark green or dark yellow when ore or waste are dominant, respectively.

Table 32. PC coordinates of the cluster's centroids.

Method	**PC**	**Training data**			**Testing data**			**Distance**
		C1	C2	C3	C1	C2	C3	
K-means	PC1	-0.160	-0.007	0.127	-0.195	0.004	0.122	0.036
	PC2	-0.003	-0.014	0.002	-0.005	-0.008	-0.006	0.016
	PC3	0.000	0.019	-0.016	-0.010	0.028	-0.016	0.009
AHCF	PC1	-0.167	-0.019	0.117	-0.195	0.003	0.111	0.031
	PC2	-0.002	-0.010	-0.002	-0.005	-0.008	-0.004	0.024
	PC3	0.001	0.021	-0.019	-0.010	0.030	-0.015	0.007
GMM	PC1	-0.153	0.011	0.148	-0.197	0.018	0.138	0.052
	PC2	0.035	-0.019	0.003	0.014	-0.010	-0.001	0.015
	PC3	0.003	0.013	-0.025	-0.015	0.023	-0.019	0.012
SOM	PC1	-0.162	-0.006	0.128	-0.195	0.004	0.123	0.036
	PC2	-0.003	-0.012	0.002	-0.005	-0.007	-0.009	0.013
	PC3	0.001	0.015	-0.014	-0.010	0.023	-0.010	0.013
SC	PC1	-0.167	0.019	0.173	-0.163	0.031	0.154	0.010
	PC2	-0.002	-0.009	0.012	-0.006	-0.007	-0.006	0.017
	PC3	0.001	0.007	-0.010	-0.008	0.020	-0.013	0.026

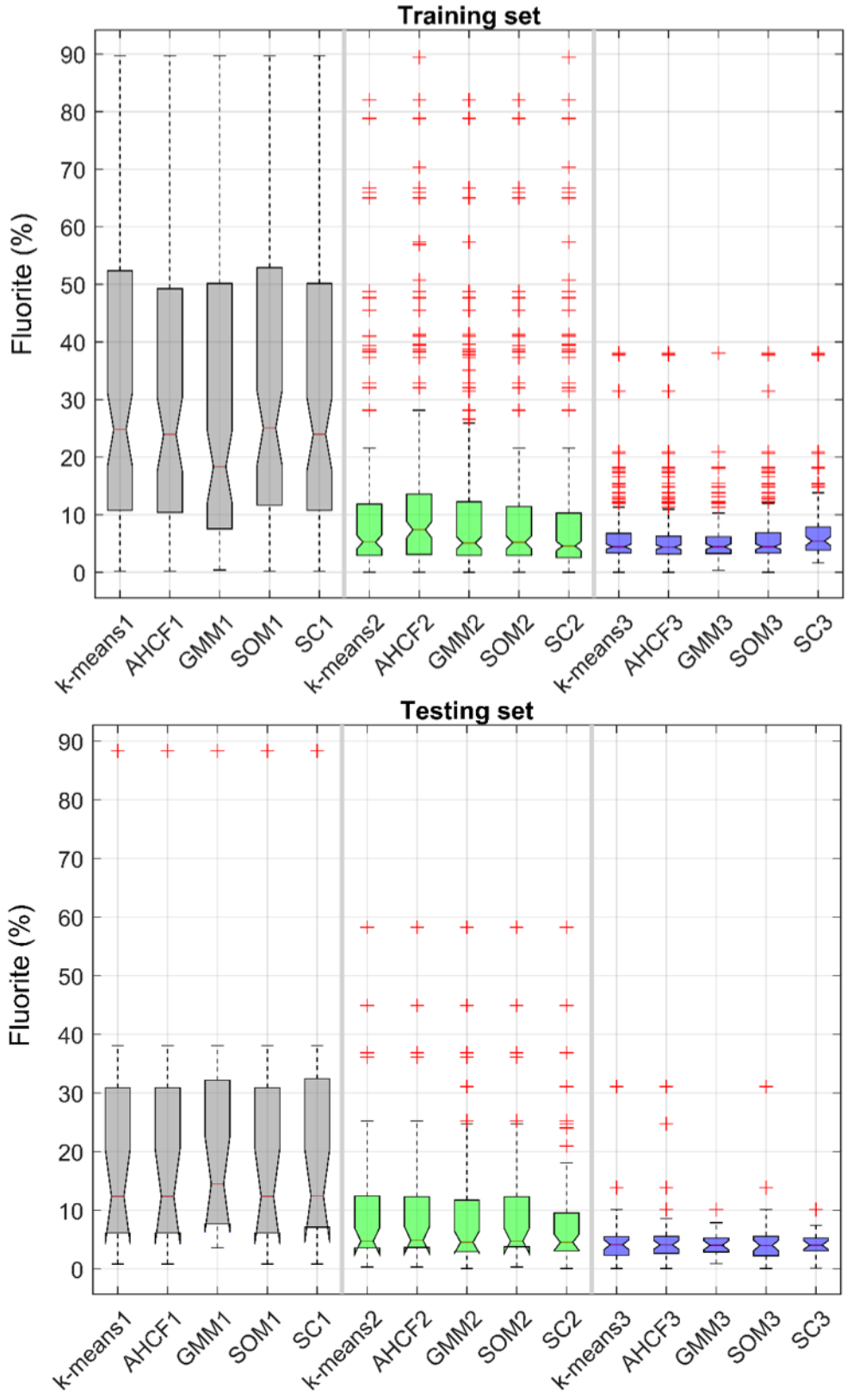


Figure 62. Distribution of fluorite for three clusters. Where 1, 2 and 3 represents the cluster C1 (grey color), C2 (green color) and C3 (blue color), respectively.

For the k-means, ore pellets (black or red filled markers) are dominant in cluster C1. The amount of ore pellets is rare in cluster C3 and the amount of ore pellets in cluster C2 is

between C1 and C3. Most of the pellets in cluster C3 belong to the waste grade, and few (15.83%) have a higher ore grade. The occurrence probability of medium or high-grade ore in C3 is low (3.6%). Pellets in cluster C2 have a similar median fluorite grade as C3, but the fluorite content distribution has a longer tail with a 75% percentile slightly above 10% and the maximum value (as defined by the upper whisker) over than the grade of 20% (see Figure 62). This involves that the percentage of waste pellets decreases from 84.17% (cluster C3) to 68.09% (cluster C2). While the percentage of ore pellets, mainly with high-medium ore grade, increases from 3.6% in C3 to 12.77% in C2 (see Table 33). Nearly 78% of the pellets in cluster C1 are original from low or medium-high grade (19.42% plus 58.25% of the pellets, respectively). Meanwhile, the distribution of fluorite grade shifts to higher values with interquartile range about 10 to 50% (see Figure 62). The SOM method provides similar clustering division as the k-means. For the AHCF, the numbers of pellets in cluster C1 decreases compared with k-means and SOM, while cluster C3 contains more pellets, i.e., nearly 40% of the data. The number of pellets with high-medium ore grade in cluster C2 increases to 18.98% (see the maximum fluorite value in Figure 62 that is near 30%). For the GMM, up to about 50% pellets belong to the cluster C2, where 66.67% of the pellets in this cluster are waste. The percentage of waste pellets is larger in cluster C3 than in cluster C2, similar to those observed by k-means, AHCF and SOM. The median fluorite in cluster C1 decreases compared with the other four methods to around 19% involving that cluster C1 has more waste pellets. For the SC, much more pellets are assigned to the cluster C2 compared with the other four clustering models.

Table 33. Membership and ore grade recognition of the clusters defined from the principal components (PC) of the image properties.

Methods		**Training set**			**Testing set**		
		C1	C2	C3	C1	C2	C3
K-means	Number of pellets	103	141	139	23	38	35
	Percentage of pellets, %	26.89	36.81	36.29	23.96	39.58	36.46
	HM, %	**58.25**	12.77	3.60	34.78	21.05	2.86
	LO, %	19.42	19.15	12.23	26.09	10.53	5.71
	Waste, %	22.33	**68.09**	**84.17**	**39.13**	**68.42**	**91.43**
AHCF	Number of pellets, %	93	137	153	23	35	38
	Percentage of pellets, %	24.28	35.77	39.95	23.96	36.46	39.58
	HM, %	**55.91**	18.98	3.27	34.78	20.00	5.26
	LO, %	21.51	19.71	11.11	26.09	11.43	5.26
	Waste, %	22.58	**61.31**	**85.62**	**39.13**	**68.57**	**89.47**
GMM	Number of pellets	111	183	89	22	48	26
	Percentage of pellets, %	28.98	47.78	23.24	22.92	50.00	27.08
	HM, %	**47.75**	15.30	2.25	**36.36**	18.75	0.00
	LO, %	18.02	18.03	12.36	27.27	8.33	7.69
	Waste, %	34.23	**66.67**	**85.39**	**36.36**	**72.92**	**92.31**
SOM	Number of pellets	102	147	134	23	38	35
	Percentage of pellets, %	26.63	38.38	34.99	23.96	39.58	36.46
	HM, %	**58.82**	12.24	3.73	34.78	21.05	2.86
	LO, %	19.61	18.37	12.69	26.09	10.53	5.71
	Waste, %	21.57	**69.39**	**83.58**	**39.13**	**68.42**	**91.43**
SC	Number of pellets	95	220	68	25	52	19
	Percentage of pellets, %	24.80	57.44	17.75	26.04	54.17	19.79
	HM, %	**56.84**	11.36	5.88	**36.00**	15.38	0.0
	LO, %	21.05	14.55	17.65	28.00	7.69	5.26

	Waste, %	22.11	**74.09**	**76.47**	**36.00**	**76.92**	**94.74**

Note: Dominant fluorite grade in each cluster is in bold; HM represents medium-high or high ore grade, LO represents the low ore grade.

Meanwhile, it can be found that the three clusters have different distributions of ferric and aluminum oxides with all clustering methods (see Figure 63); most of the pellets in C3 have the smaller amount of oxides, generally below 1.5 %; pellets in C2 have in general a higher grade of metallic oxides than in C3, up to 4%; and pellets in the ore class (cluster C1) have the largest amount of these oxides, up to 10 %. In this regard, cluster C2 can be considered a transition class between ore and waste clusters, where waste is still predominant but has a lower probability of occurrence.

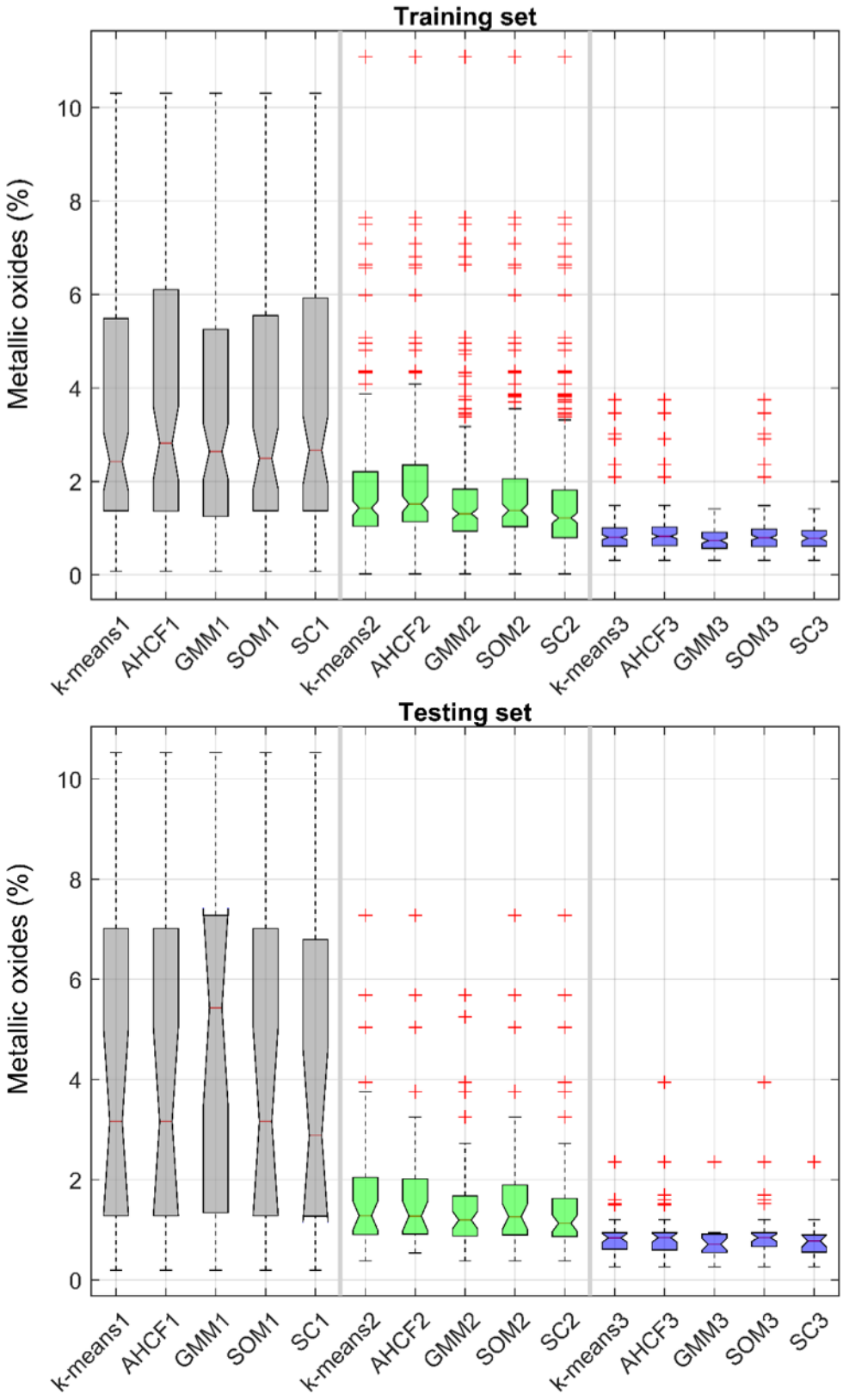


Figure 63. Distribution of metallic oxides for three clusters, where 1, 2 and 3 represents the cluster C1 (grey color), C2 (green color) and C3 (blue color), respectively.

5.5.4 Models evaluation

In this study, it can be found that the occurrence probability of fluorite content ≥10% from cluster C1 is much higher than the other clusters. Then it can be assumed that cluster C1 is labelled as ore and clusters C2 and C3 can be described as waste. Regarding this, the low ore grade pellets and medium-high or high ore grade pellets are combined. Then the percentage of pellets correctly assigned in one cluster can be considered as precision (Pr) for ore or waste and Table 34 can be obtained.

$$Precision\ (Pr) = TP/(TP+FP) \tag{5.6}$$

Where: True Positive (TP): the number of instances that are correctly predicted as positive; False Positive (FP): the number of instances that are incorrectly predicted as positive. Table 34 shows the precision in ore for cluster C1, and the precision in waste for clusters C2 and C3; green and yellow denotes ore and waste, respectively. It should be noted that when we calculate the Pr for ore, then the ore would be the positive case and in the case of metric calculation of waste, on the contrary, the waste would be positive case. For instance, the physical meaning of the Pr for the GMM method of the training set is as follows. There is a 65.77% (47.75%+18.02%) probability that pellet samples would be recognized to be ore samples if their PCs are located in cluster C1, and there is 66.67% or 85.39% probability that a pellet would be predicted as a waste sample if they are in cluster C2 or C3, respectively. According to this, it can be seen that the best precision for ore pellets of the training set is from C1 with the usage of SOM, i.e., 78.43% (58.82%+19.61%) and the best precision for waste pellets of the training set is from C3 with the usage of AHCF, i.e., 85.62%.

Table 34. Precision for ore (green) and waste (yellow) in each cluster of different clustering methods

Methods	**Training set**			**Testing set**		
	C1	C2	C3	C1	C2	C3
k-means	77.67%	68.09%	84.17%	60.87%	68.42%	91.43%
AHCF	77.42%	61.31%	85.62%	60.87%	68.57%	89.47%
GMM	65.77%	66.67%	85.39%	63.64%	72.92%	92.31%
SOM	78.43%	69.39%	83.58%	60.87%	68.42%	91.43%
SC	77.89%	74.09%	76.47%	64.00%	76.92%	94.74%

In the next step, to measure the overall classification performance of the clustering models for waste/ore, the waste clusters are integrated into a single one. By doing this, the confusion matrix (Liu et al., 2024, 2023; Wang et al., 2023) can be obtained as shown in Figure 64. Some significant classification metrics can be obtained from the confusion matrix and their definitions can be seen as follows:

$$Accuracy\ (Ac) = (TP+TN)/(TP+TN+FP+FN) \tag{5.7}$$

$$Recall\ (Re) = TP/(TP+FN) \tag{5.8}$$

$$Specificity\ (Sp) = TN/(TN+FP) \tag{5.9}$$

$$F1\ Score\ (Fs) = 2\times(Precision\times Recall)/(Precision+Recall) \tag{5.10}$$

where, True Negative (TN): the number of instances that are correctly predicted as negative; False Negative (FN): the number of instances that are incorrectly predicted as negative. Their meanings are demonstrated in Figure 64. Take the k-means results for example, the upper two figures represent the pellet assignment in each cluster and corresponding Pr. By the accumulation of the number with the same filling color, then we can get the confusion matrix for the training set and testing set.

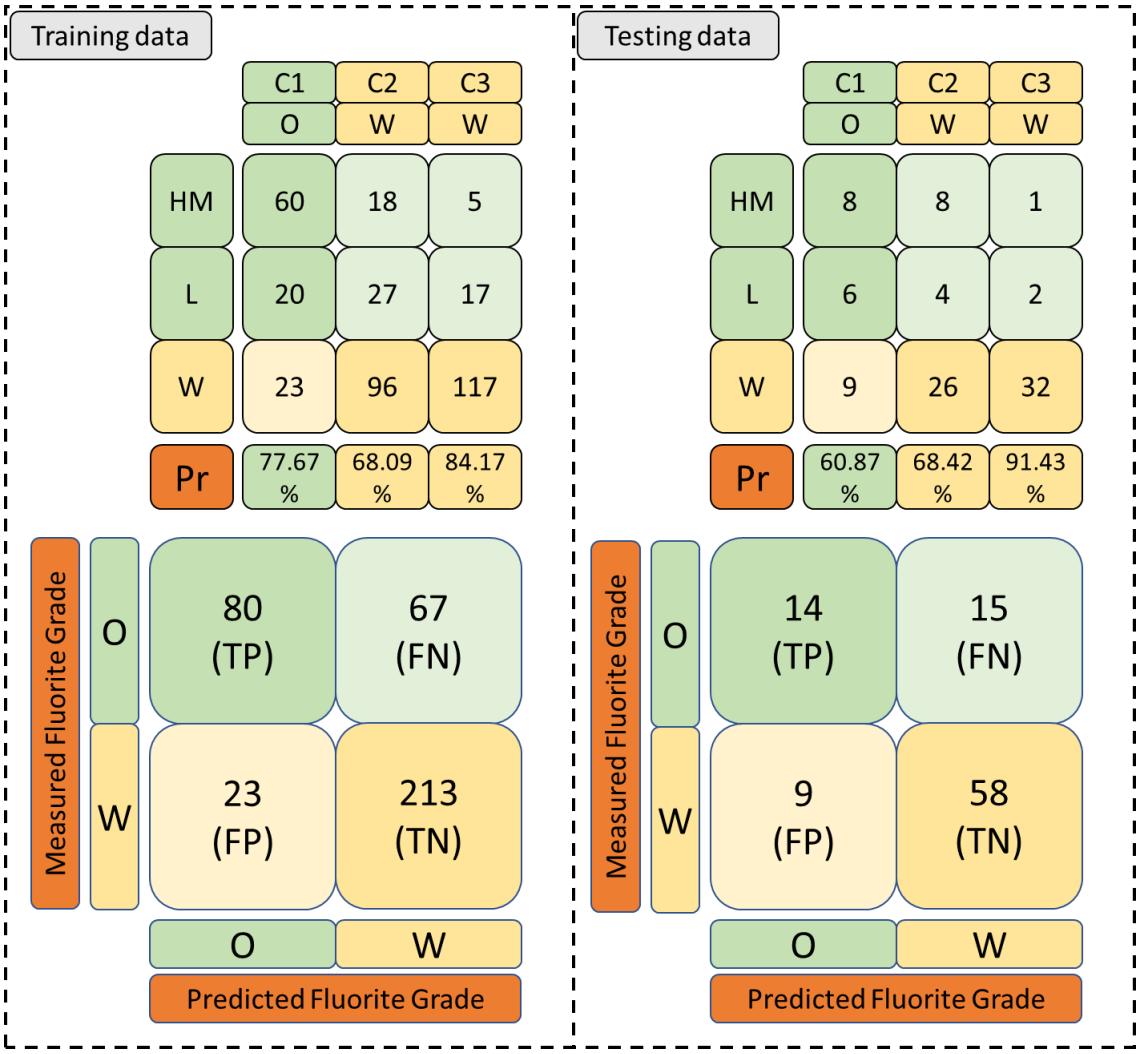


Figure 64. The confusion matrix from k-means method for the training set (left graph) and testing set (right graph); O and W represent the ore and waste, respectively. HM represents medium-high or high ore grade, L represents the low ore grade

The random division of training and testing set has been implemented. There is no obvious prediction variability for different random division. The SOM provides the best metrics with the exception of the recall for waste (ReW) for the training set (Figure 65, top graph), however, its ReW is also competitive. The GMM leads to a significant worst performance with an overall accuracy (Ac) of 0.7, and smallest Pr for the ore of 0.65. Compared to the other clustering methods used in this study, GMM exhibited poorer performance in C1. It can be attributed to its reliance on the assumption that the data are

generated from a mixture of Gaussian distributions. The actual data in this study presented non-Gaussian cluster structures, which made GMM less effective. In general, all the methods, lead to a low recall (Re) for the ore. On the contrary, waste has a better Re (up to 0.9) and a Pr near 0.8 (note that this value is smaller than in Table 4 due to lumping C2 and C3).

The patterns observed from the training set are validated with the pellets of the testing group; these are assigned to the cluster with the nearest centroid using the L1 distance (orange triangles in Figure 60) according to the three PC coordinates for k-means, AHCF and SC methods. For GMM and SOM, the developed clustering model based on the training set would be applied to the testing data. The distribution of the pellets in different clusters for the testing set has been compared with the training data in Figure 60. It can be seen that they have the similar clustering performance. The detailed coordinates of the clusters' centroids for the testing set can be seen in Table 32. In each method, three distances from up to down, represent the distance between the centroids of clusters C1, C2 and C3 of the training set and testing set. This suggests a limit separation between them. The trends observed in the distribution of fluorite and metallic oxides for each cluster from the training set are similar (see Figure 62 and Figure 63, bottom graphs).

The results from the testing set can be seen in Table 33 and Table 34. According to Table 34, it can be found that some methods produce excellent Pr for the waste in cluster C3, for instance, the SC method can reach about 95%. This means that if the samples' PCs are located in cluster C3, there is an about 95% probability among the pellets of that belong to waste. The other clustering methods present a slightly smaller Pr for the waste in cluster C3, from 89.47% (AHCF) to 92.31% (GMM). Finally, according to the Pr for each cluster, mainly due to its higher precision of waste in cluster C2 (see Table 34), and the better overall classification metrics for the testing set (see Appendix 4, Table 55), the SC would be recommended for recognizing new fluorite-based pellets.

For new pellets, the same procedure could be followed (described in Figure 66) to calculate the color attributes and their position in the PC coordinate system, and hence its classification. An increase in the size of the datasets would be expected to increase the significance of the color parameter-based fluorite grade classification.

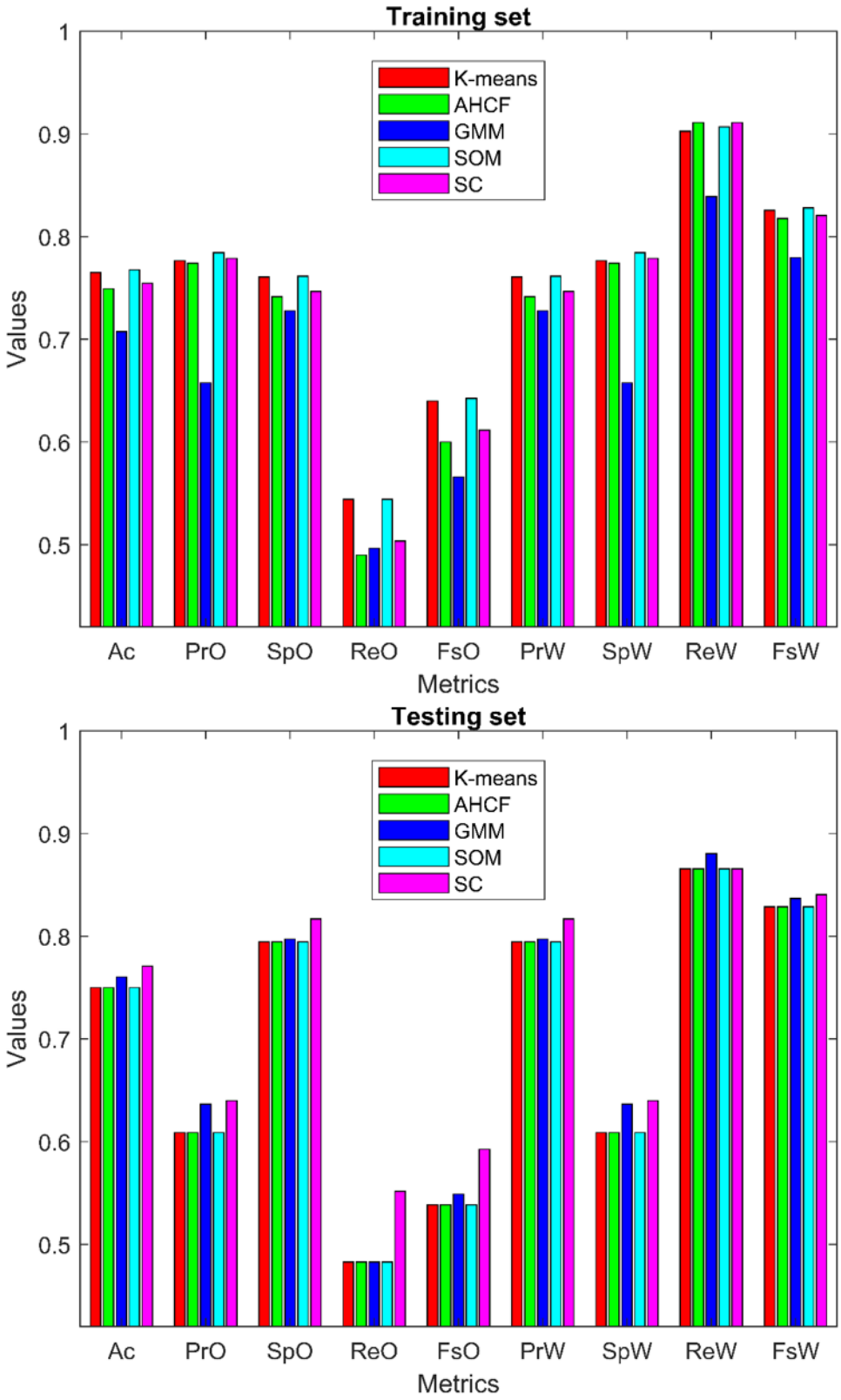


Figure 65. The classification metrics of different clustering methods; where O and W are ore and waste, respectively.

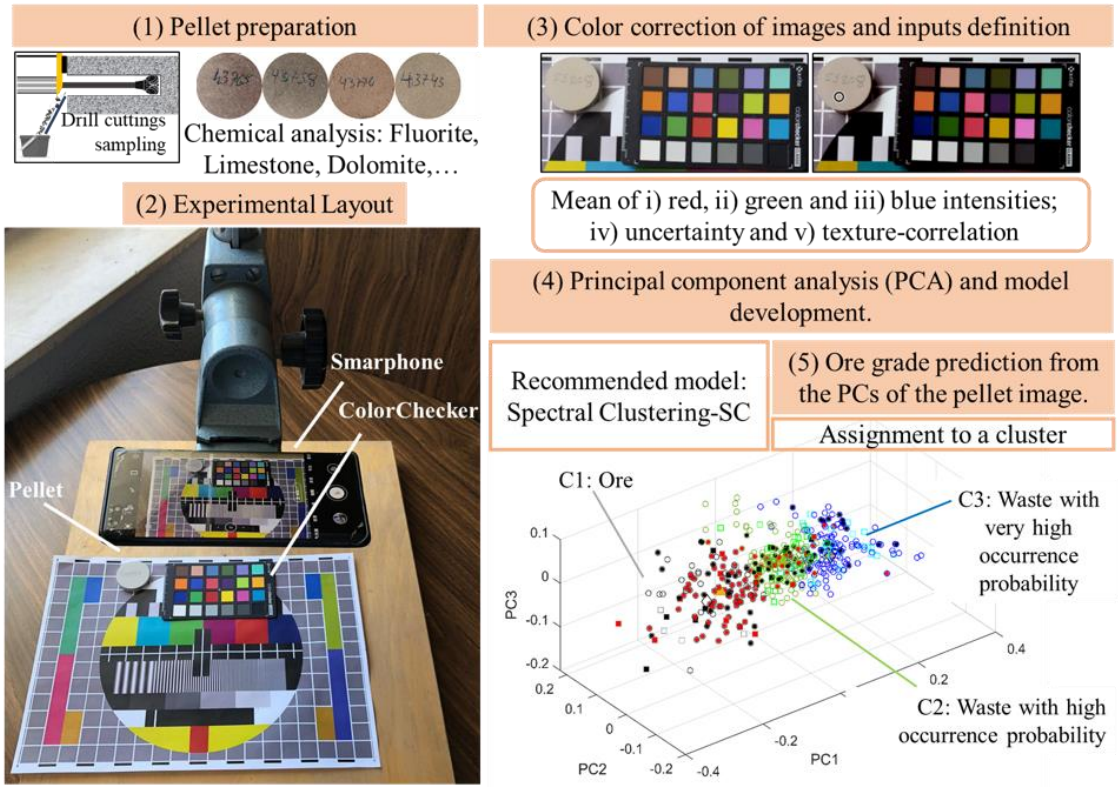


Figure 66. Flowsheet for fluorite grade recognition based on pellet images properties

5.6 Discussion

This study attempts to find specific color attributes from different fluorite grade pellets to distinguish waste and fluorite ore. The fluorite pellets are made from drilling chips and a smartphone is used to take the pellet photos. A ColorChecker is used to avoid the lightness interference and to get the genuine colors of pellet images. The PCA results from mean value of Red and Blue channels, uncertainty in PCIs and a texture parameter-correlation are used to develop five clustering models, namely k-means, AHCF, GMM, SOM and SC. The same training sets and testing sets are employed for all the methods; however, different cluster assignments are procured. For a new pellet, these steps would be easy to implement and the codes would automatically provide the pellet position in the PC coordinate system considered, and thus its classification in any of the three clusters employed.

Based on the highest Pr for waste pellets and the overall classification performance for the testing set (see Figure 65, low graph), the Spectral clustering (SC) is recommended. Clusters C2 and C3 might be labelled waste, since waste pellets are predominant in these two clusters with precision of 76.92% and 94.74%, respectively (see Table 34 for SC method), and cluster C1 might be labelled as ore with a precision of 64%. In cluster C3, the number of high-medium fluorite pellets is 0 (see Table 33) and the number of low-grade fluorite pellets is also limited, a 5.3% of the pellets in that cluster. Under this situation, if color attributes of a new pellet are located in this, it has a high probability of being waste and thus is eligible to save chemical analysis cost.

The lower Pr for ore in cluster C1 indicates that chemical analysis is advisable if a new pellet falls in this cluster, as the percentage of waste samples in this cluster is up to 36% (see results for SC method for the testing set in Table 34). The chemical analysis can be also considered with new pellets in cluster C2, where about 23% of the pellets in that cluster are owed to a wrong prediction.

The interaction between different minerals may result in various color characteristics. In this study, the ore pellets do not show a color gradient from low to high grade, thus increasing the difficulty of the work. The existence of ferric oxide in some samples may further alter significantly the color. Nevertheless, the recognition rate (precision) for waste in cluster C3 can reach a higher level for all the methods. The clustering results shown in Figure 63 suggest the influence of these minerals in the resulting cluster definition, while also indicating that it is possible to discriminate fluorite by color attributes and the clustering method, especially with SC method.

5.7 Limitations

Although this study has contributed to validate the use of images for fluorite recognition from color properties there are still limitations need to considered. At first, the potential of classification or the regression techniques for fluorite grade prediction is worthwhile to be investigated, such as extreme learning machine (Zhang et al., 2023a), extreme gradient boosting (Zhang et al., 2023b), LightGBM (Zhang et al., 2023c) and deep learning (Phoon and Zhang, 2023; Wu et al., 2023). Secondly, more pellet images and their corresponding chemical composition should be collected to explore the stability of clustering models. Thirdly, the proposed procedure provides excellent waste classification only for one of the clusters, while the results for the other two are not good enough to eliminate chemical analysis for pellets classified in these clusters. The working procedure shown in Figure 66 would be advisable to other deposits which has more specific color characteristics, e.g., hematite (reddish), chalcopyrite (brass color or golden) or dioptase (emerald green or blue-green).

5.8 Conclusion

This study presents a promising avenue for image-based ore grade recognition, on the premise that ores sometimes exhibit distinct colors corresponding to variations in their mineral composition and concentrations. A smartphone with a high-resolution camera is employed to procure photographs of pellets made from finely ground and compacted drilling chips samples. The photos are corrected by a ColorChecker to obtain the non-lighting-interfered ore colors. Image-based analysis is integrated with five clustering algorithms, i.e., k-means, Agglomerative Hierarchical Cluster Tree, Gaussian Mixture Model, Self-Organizing Map and Spectral clustering. Three clusters are generated according to the three principal components obtained from mean of red and blue PCIs, uncertainty in PCIs, and texture-correlation. The precision for each cluster is calculated and it can be found that Spectral clustering presents the best precision for waste with 94.74% in cluster C3 and the best overall classification performance for the testing set. Cluster C2 could be assigned to waste with high occurrence probability and cluster C1 to

ore. The lower precision from clusters C2 and C3 makes them necessary to consider chemical analysis if the pellets are clustered in these two clusters. Besides on the differences in fluorite grade distributions, all clusters have different distributions of ferric and aluminum oxides. They generally increase as the first principal component (PC1) decreases with below 1.5 % for waste cluster C3 and up to 10 % to for ore cluster C1. As for other clustering methods, they also generate desirable recognition performance for waste in cluster C3. But the precision is lower, in the range of 89.47%-92.31%.

The use of color properties of pellets images acquired with a smartphone provides a fast and cheap screening criterion of pellets. It could be deemed waste if the color attributes correspond to cluster C3, hence lowering grading costs and increasing equipment availability. The equipment required in this study is low cost and does not need a rigid operational environment. The potential for utilizing smartphone-captured images and their color properties as a predictive tool for ore grading provides a new insight for the mining industry. It also represents a cost-effective method that can be easily applied regardless of the geographic location or resource limitations, and does not require high professional skills. Though not complete substituting geochemistry analysis, it can reduce the time cost dedicated by operators to this laborious task. In addition, it can reduce maintenance costs by extending the average lifespan of the X-ray tube, since it reduces the number of analyses to be performed.

The assessment of the image-based recognition process developed here is probably something worth trying with other types of minerals, e.g. iron and copper ores. They are expected to have higher color properties correlation with ore grades. Besides the clustering approach used here, perhaps a variety of machine learning methods could also be applied.

Chapter 6. Lithology identification using borehole images by contrast-limited adaptive histogram equalization (CLAHE) and machine learning models

Nomenclature

Ac (Accuracy)	NIQE (Natural Image Quality Evaluator)
AHE (Adaptive histogram equalization)	PCC (Pearson Correlation Coefficient)
ASTER (Advanced spaceborne thermal emission and reflection radiometer)	PIQE (Perception based Image Quality Evaluator)
B (Blue)	Pr (Precision)
BL (Brecciated limestone)	Pr0 (Precision for massive limestone)
BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator)	Pr1 (Precision for brecciated limestone)
BT (Bagged Tree)	Pr2 (Precision for high amount of clay)
BY (Bayesian)	Q1B (25th percentile of blue pixels)
CC (Color counting)	Q1G (25th percentile of green pixels)
CLAHE (Contrast-Limited Adaptive Histogram Equalization)	Q1R (25th percentile of red pixels)
CNN (Convolutional neural network)	Q2B (median of blue pixels)
CR (Correlation)	Q2G (median of green pixels)
CT (Contrast)	Q2R (median of red pixels)
EN (Energy)	Q3B (75th percentile of blue pixels)
ET (Entropy)	Q3G (75th percentile of green pixels)
FN (False negatives)	Q3R (75th percentile of red pixels)
FP (False positives)	R (Red)
Fs (F1 score)	Re (Recall)
Fs0 (F1 score for massive limestone)	Re0 (Recall for massive limestone)
Fs1 (F1 score for brecciated limestone)	Re1 (Recall for brecciated limestone)
Fs2 (F1 score for high amount of clay)	Re2 (Recall for high amount of clay)
G (Green)	RF (Random Forest)
GBM (Gradient Boost Machine)	Sp (Specificity)
GLCM (Gray level co-occurrence matrix)	Sp0 (Specificity for massive limestone)
HC (High amount of clay)	Sp1 (Specificity for brecciated limestone)
HG (Homogeneity)	Sp2 (Specificity for high amount of clay)
LGBM (Light Gradient Boosting Machine)	SVC (Support Vector Classification)
MFFNN (Multilayer feed-forward neural network)	S2 (Sentinel-2)
ML (Massive limestone)	TN (True negatives)
N (North)	TP (True positives)
NE (Northeast)	XGB (Extreme Gradient Boosting)

6.1 Introduction

Lithology identification plays a crucial role in the assessment of the rock mass mineral and structural characteristics, critical for the design and operational routines especially during the initial construction phase, but it also serves as a valuable guide for addressing unfavorable geological conditions encountered during the whole operation (Bosch et al., 2002; Guzzetti et al., 1996) in e.g. mining (Chen et al., 2021), tunnel excavation (Xu et al., 2023), geotechnical hazard evaluation (Tang et al., 2011), water resource management (Nickolas et al., 2017), environment protection (Adithya et al., 2021) and others. One of the traditional and reliable methods is to recognize the lithology from drilling cores by naked eyes (K. Li et al., 2021). However, it demands practitioners with exceptional professional expertise and extensive fieldwork experience, the task being not only physically demanding and time-intensive but also highly subjective. Some other traditional methods involve interpretation of lithology from X-ray, CT scan, electron microscope, isotope detection by mass spectroscopy (Chawshin et al., 2021; Mauriohooho et al., 2016; Tipper et al., 2008; Van Hoesen and Orndorff, 2004) and other. They can provide valuable and comprehensive insights for rock internal structure but the production of samples and analytical methods are expensive.

Some indirect ways have been proposed to distinguish rock lithologies from physical properties of rocks (Konaté et al., 2017; Mishra et al., 2022; Sebtosheikh and Salehi, 2015; Wang and Zhang, 2008). For instance, Iserhien-Emekeme et al. (2017) utilized the Allied Ohmega Terrameter to supervise seventeen vertical electrical soundings and two sediments could be inferred, i.e., unsaturated and saturated according to different resistivity values. Zhang et al. (2018) proposed to predict the lithology log utilizing post-stack seismic data and deep learning techniques. To facilitate the calculation of deep learning, the extracted seismic log as well as its nearby logs within a specified time window were used as input. The continuous wavelet transform was performed to get seismic spectra by a Ricker wavelet function and a two-dimensional input could be obtained. As a result, binary classification tasks were processed, i.e., shale and sand. Spectra-based model presented better prediction accuracy than the original post-stack seismic data-based model. Corina and Hovda (2018) proposed an automatic procedure to predict shale and not-shale formations. Gamma-ray from wireline logging was used to distinguish different lithologies and the classification model was developed based on kernel density estimator. Hossain et al. (2021) developed an intelligent lithology prediction model by rough set theory in which more than 5000 samples were used from 11 well logs. Ten attributes were considered such as gamma ray log, porosity or neutron log and density log. The predicted results indicated that rough set theory-based model presented higher prediction accuracy than support vector machine, artificial neural network and linear discriminant analysis. To overcome the difficulties of exploration of coal from carbonaceous and non-combustible beds, Kumar et al. (2022) applied five supervised learning techniques by integrating Gamma-ray, density and resistivity logs. The five machine learning methods proposed could achieve more than 80% accuracy for new datasets. Liu et al. (2021) applied global-attention-mechanism-based long-short-term-memory (LSTM) to predict five lithologies based on 12 significant tunneling parameters. It could be found that global-attention-mechanism-based LSTM performed better than non-optimized LSTM and some other machine learning techniques.

The aforementioned methods, while effective, lack of direct observation and thus may require more interpretation and explanation by geological expertise or complicated

models. In addition, many non-vision-based methods require expensive instruments and stringent environment for measurements and data collection.

6.2 Literature review

To overcome the shortcomings of traditional and indirect methods, there is a growing interest in harnessing the power of machine learning or deep learning techniques and computer vision to enhance the efficiency and accuracy of geological assessment (Alzubaidi et al., 2021; Ergün Hatir and İnce, 2021; Faria et al., 2022; Xu et al., 2021). On the one hand, computer vision can directly observe and describe the appearance characteristics of rock samples. On the other hand, the artificial intelligence (AI) techniques can model complicated relationship between image information and lithology types. Moreover, the aid of AI provides faster prediction and enhance the efficiency.

For instance, (Alzubaidi et al., 2021) developed three convolutional neural network (CNN) models to automatically recognize sandstone, limestone and shale by collecting the images from drilling core trays in hydrocarbon reservoir. 1 cm step log could be predicted with 93% accuracy based on the ResNeXt-50 model optimized by a probability smoothing function. Faria et al. (2022) extracted the features from Brazilian pre-salt images based on 570 thin sections. These features involved border and texture characteristics. By a process of feature compression techniques named Autoencoder, the original feature number was reduced from 104 to 50 and used to be inputs by deep neural network. Four lithologies were predicted, i.e., stromatolite, spherulite, laminate and grainstone and the overall prediction accuracy achieved 83%. Xu et al. (2021) employed the Faster R-CNN prediction model and residual learning technique to extract the contour features, texture features and color feature from rock samples. 30 types of lithologies were predicted, with the faster R-CNN model performing better than YOLO v4 method. Xu et al. (2022) took images of thin slices of 30 types of rock samples with a polarizing microscope, a total 14950 microscopic images were obtained and used for establishing the dataset. Seven deep learning techniques were employed. Transfer learning and data augmentation techniques were employed to optimize the model training network. It was found that the Xception-based approach achieved the highest prediction accuracy with 98.65%, together with a rapid model training speed. Galdames et al. (2019) combined the characteristics from color, range and hyperspectral images of rock samples to classify 13 lithologies. Linear and nonlinear support vector machine was used for conducting classification task; the features provided by hyperspectral-based images achieved the best classification accuracy at 99.95%. More recently, Shirmard et al. (2022) investigated and compared the potential of three multispectral remote-sensing data to classify lithologies, namely land imager, advanced spaceborne thermal emission and reflection radiometer (ASTER), and Sentinel-2 (S2). The integration of CNN and ASTER parameters generated the best performance. Bahrami et al. (2024) used the spectral features derived from ASTER as inputs. The input importance analysis was applied to remove the less important imagery properties. Several machine learning methods were tested to map the lithology distribution, with an accuracy in excess of 80%. While all these works can effectively conduct lithology classification or recognition, the requirement of sufficient number of rock samples or comparably costly equipment tends to hinder their application and generalization.

Endoscopes, equipped with high-resolution cameras, have become alternative ways for capturing detailed images of rock properties in remote or challenging environments, to be

used for rock mass quality evaluation in underground mines (Majcherczyk et al., 2005; Malkowski et al., 2008), evaluation of geological conditions ahead of a tunnel face (Hohashi et al., 2019), inspection of jetted holes (Li et al., 2022; Reinsch et al., 2018), and rock mass structural recognition (Fernández et al., 2023). These applications indicate that the endoscope may be a good tool for exploring lithologies.

Research gap: However, the investigation of the integration of AI techniques and images from endoscopes to predict lithology is rare. To fill the gap, this chapter explores the application of machine learning techniques to predict lithology types from borehole images by an endoscope. Compared with previous equipment, an endoscope is low-cost, flexible and its implementation doesn't need too much professional training.

In this work, lithology classification is made according to the clay content into three classes, i.e., massive limestone, brecciated limestone and high amount of clay. The images are automatically extracted from videos to reflect the lithology condition with the corresponding depth. The gray pixel intensity threshold and three none-reference image quality metrics, namely Perception based Image Quality Evaluator, Natural Image Quality Evaluator and Blind/Referenceless Image Spatial Quality Evaluator are used for the determination of image quality. Contrast-limited Adaptive Histogram Equalization (CLAHE) technique is adopted to improve the image quality. Ten color characteristics involving three percentiles of red, green and blue pixel values as well as color counting and five texture characteristics including correlation, entropy, homogeneity, contrast and energy are used as inputs. Feature importance scores were calculated from the original and CLAHE-enhanced scenarios. Compared with previous image-based lithology prediction methodologies, this study provides a new intelligent way for geologists and researchers to approach lithological characterization, aiming to provide a better knowledge of the rock mass for blast design and explosive loading of the boreholes, helping to prevent undesirable events such as fines generation, and, importantly, flyrock from explosive accumulation in underground cavities which could pose a safety concern. This assessment can be expedited and reliance on human interpretation reduced with the method proposed.

6.3 Site description and research problem

The studied quarry is located in the municipality of Valdilecha, in the province of Madrid, Spain. Geologically, the quarry is located within the Meso-Tertiary Tagus Basin or Madrid Basin, in the transition zone from intermediate to central basin facies. The area is predominantly made up of Neogene materials (Upper Miocene and Pliocene) deposited in intramountainous continental basins, presenting a great variety of facies. These materials are sub-horizontal or slightly inclined at the edges. The stratigraphic column comprises different levels of materials of detrital, chemical or mixed origin, ordered from the oldest to the most recent as follows: firstly, conglomerates and sandstones of the intra-Miocene fluvial network, separated by a clear sedimentary break (marls). Secondly, limestone formations of the mooreland that crown the Miocene series. They cover the largest surface and have been the object of the existing exploitations. Finally, quaternary deposits belong to the glacis and terraces of the Anchuelo and Las Morenas streams (Alonso Zarza et al., 1993; De Vicente and Muñoz-Martín, 2013).

Generally, the lithologies can be categorized in three classes according to the amount of clay: massive limestone (ML), brecciated limestone (BL) and high amount of clay (HC). Massive limestone (first row of the picture shown in Figure 67) has an overall grayish color and uniform structure with relatively homogeneous appearances. Brecciated limestone (second row of the picture shown in Figure 67) is characterized by fragmented surfaces that could be angular or rounded and is made up of pre-existing limestone, clay or other rocks. High amount of clay presents a much more reddish color (third row of the picture shown in Figure 67). Among the mineralogy presented in the site, there are sandstones, sands and clays with abundant feldspars and variable proportion of metamorphic elements, such as iron, which constitute the red series of the deposit and that is why the characteristic color of these clays is reddish to brownish. Among the mechanical properties, it can be highlighted that clay has a very high plasticity, which increases as water absorption increases (Al-Shayea, 2001). In addition, as the clay content increases, the rock quality is lower. In some cases, these clays are found within joints.

Clay minerals can be characterized for being a material with a very fine granulometry, mostly 80% of the material under 4 microns. In addition, some are called expansive clays, which are capable of absorbing water from a wet environment, trapping water molecules in the vacancies between silicate layers, and might complicate blasting and other mining operations. Especially in the rainy season, such clay may block the bypass of the primary crusher and contaminates the fine product as it is adhered to larger fragments that are fed into the primary crusher. Therefore, the recognition of the clay amount is of importance for blast design and geotechnical engineering. Appropriate evaluation of the amount of clay would allow a correct borehole distribution (Chen et al., 2021), optimizing the use of explosives (e.g., the density of the explosive can be reduced in areas with high clay content or ultimately replaced by airbags or stemming) (Chen et al., 2021), overall improving blasting fragmentation and saving operational costs. Fast and accurate recognition of lithology would hence be the main goal of this study.

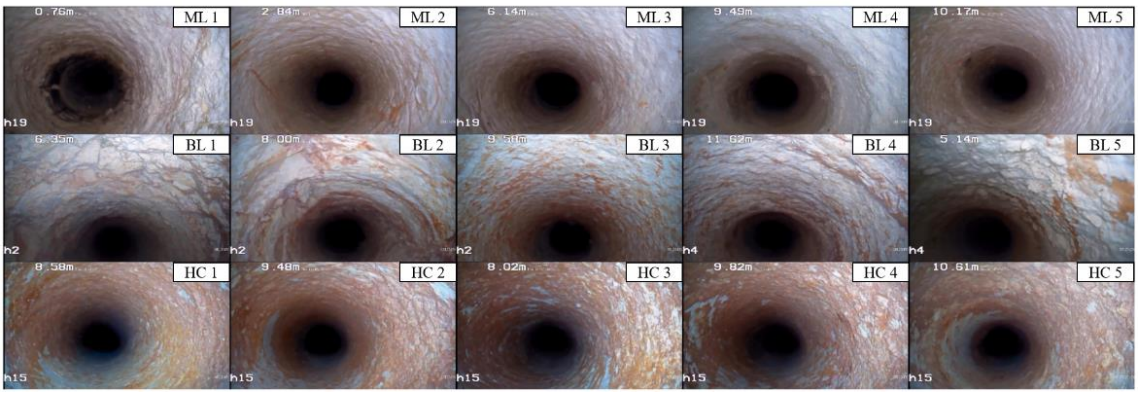


Figure 67. Typical lithologies observed from the drilling boreholes.

6.4 Data procurement, processing and description

6.4.1 Endoscope measurements

An endoscope was used for recording the borehole wall videos. It is manually introduced through the borehole and supported by a roller system, see Figure 68. The device employed in this study is produced by the “Forthaus Tech” company and includes a 23mm

diameter camera linked to a recording unit via a 20m insertion cable. The resolution of the camera is 752×582 pixels (photo resolution 1920×1080 pixels). Positioned beneath the unit, an encoder keeps track of the cable's coiling with a depth resolution of 1cm, while depth measurements, hole identification, and recording duration are superimposed onto the captured video frames. This functionality facilitates the identification of features (such as cavities, joints, faults, different lithologies, etc.) observed on the internal walls of the borehole. The camera is equipped with a centralizer, made with plastic straps, to ensure a good visualization of the hole walls. Before conducting the endoscope measurement, the borehole identification can be entered with the computer and the depth set to 0m. The measurement process goes on until the bottom of the borehole is observed.

Figure 68. Endoscope in the field.

6.4.2 Video and Image Procurement

For the procurement of the borehole wall images, the videos were read with a Matlab R2021a script and the frames were cropped every two seconds. Next, the numbers which represent the borehole wall depth need be recognized (outlined by a red rectangle, see Figure 69) to connect the borehole depth and corresponding lithologies. For doing this, three groups of numbers from 0 to 9 were cropped from different photos to recognize the digit in the first integer, decimal and centesimal, respectively. The digit would then be compared with the numbers by means of correlation function *corr2* in Matlab. The number with the highest correlation reflects the most appropriate recognition number. This process is automatically done with the designed codes that, among other, had to overcome difficulties such as non-uniform digit background and the presence of small characters with the date and time of recording superimposed to the depth. When the depth is over 9.99m, the recognition of the leftmost digit is omitted and automatically generated as 1. After this procedure, each image captured can be named with the borehole identification and depth. Three geologists determine the borehole wall lithology visually and the recognized results are considered as the ground truth data in this study.

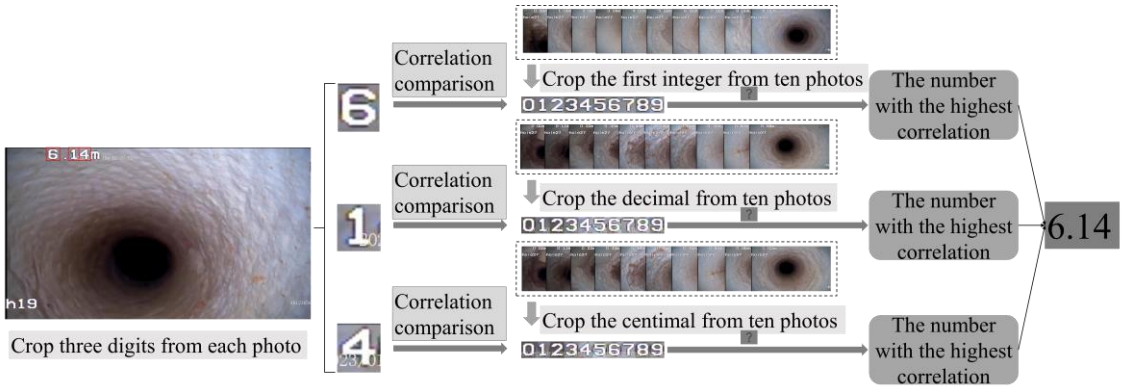


Figure 69. The process of the recognition of the borehole depth.

Before employing borehole images to do classification tasks, some filtering need be done to procure the desirable images. For this, all images are transformed into grey levels by the function *rgb2gray* in Matlab. Then the number of pixel values above 230 (which represent the over-exposed area) and below 20 (which represent the completely dark area) are calculated in each image. Images with frequency of pixel intensity higher than 230 or lower than 20 in excess of 0.14 are discarded. This is done to avoid images with a very high number of dark pixels, where only a small area can be distinguished, and images where the borehole wall is too bright, often due to a poor centralizing that results in images that do not capture all borehole wall information or are over-exposed. Some photos discarded are shown, with the offending pixel frequency, in Figure 70.

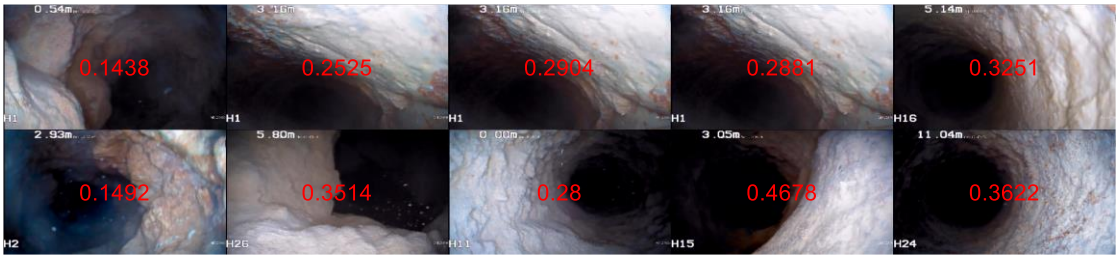


Figure 70. Examples of undesirable images and corresponding pixel frequency.

In the process of recording the borehole walls, sometimes it is inevitable to stop the camera due to obstructions by irregular borehole walls. In addition, influenced by the precision of the encoder, sometimes one depth may produce more than one borehole images. For selecting the best image from the same depth, three image quality assessment metrics were considered, i.e., Perception based Image Quality Evaluator (PIQE) (Venkatanath N et al., 2015), Natural Image Quality Evaluator (NIQE) (Mittal et al., 2013) and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) (Mittal et al., 2012). These three metrics belong to no-reference image quality metrics, working with statistical features of global or local image characteristics. A brief introduction and explanation of these metrics can be seen in Table 35.

Table 35. Description of three image quality assessment metrics.

Metrics	Description	Characteristics	Standard
BRISQUE	Trained based on images with known distortions.	Limited by the same type of distortion.	Generally, in the range [0, 100]. Lower values of score reflect better perceptual quality of image.
NIQE	Trained based on pristine images.	Not limited by the distortion.	Non-negative, the lower the score value, the better the perceived quality of image.
PIQE	Does not need a trained model.	Not limited by the image distortion. Evaluate the image quality based on the block-wise distortion and local variance.	Excellent: [0, 20]; Good: [21, 35]; Fair: [36, 50]; Poor: [51, 80]; Bad: [81,100].

The calculated results by the three image quality evaluation metrics for all images are shown in Figure 71 in the form of frequency histograms. Almost all BRISQUE values are over 50 and most BRISQUE values are in the range of [57, 62], all PIQE values are over than 58.5 and NIQE values are in the range [3.8, 5.6]. According to the image quality assessment standard listed in Table 35, it can be said that all borehole wall images obtained by the endoscope don't have a high image quality, mainly due to the difficulty in achieving good focusing of the endoscope lens during the movement of the camera. For the purpose of comparing the quality of images with the same depth, the image with the lowest sum of NIQE, PIQE and BRISQUE values is be retained. As an example, Figure 72 shows several image pairs with the same depth in the same borehole. The images with lower clarity have higher sum values of NIQE, PIQE and BRISQUE as shown in Table 36 and thus they are eliminated.

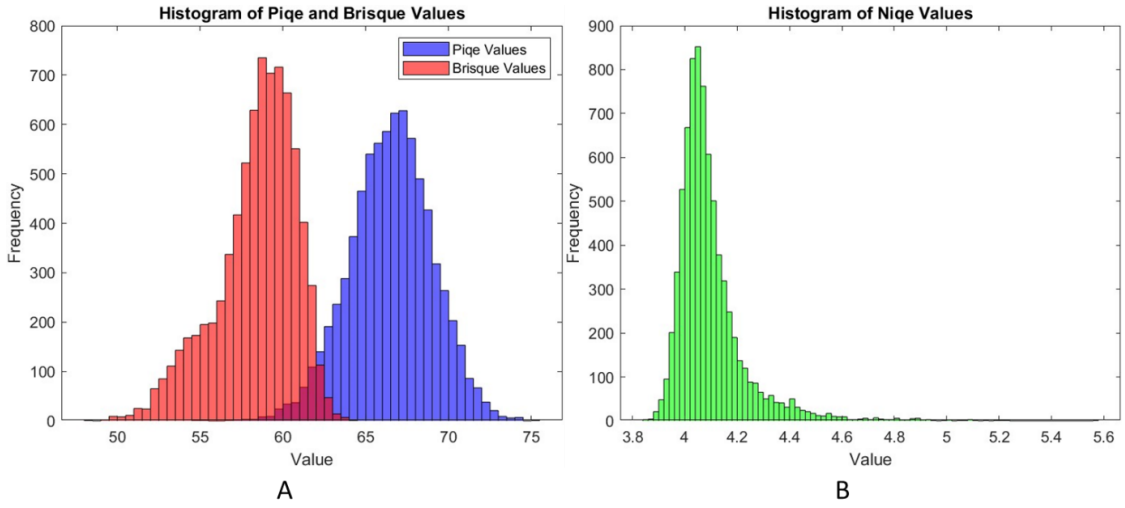


Figure 71. Histograms of image quality evaluation metrics for all images: (A), PIQE and BRISQUE, (B): NIQE.

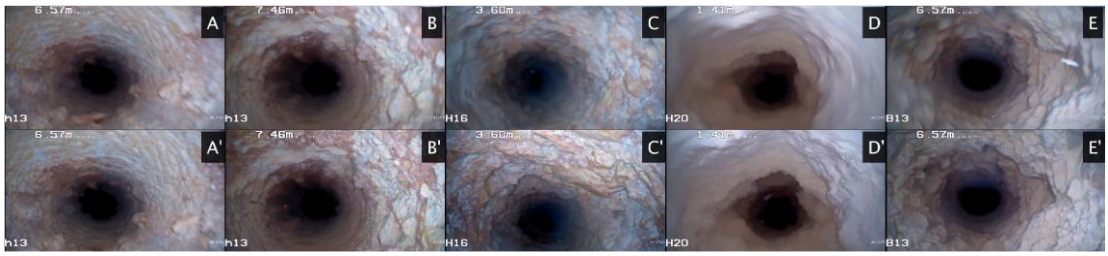


Figure 72. Several cases of photo selection: discarded photos (first row), retained photos (second row).

Table 36. The quality comparison of images with the same position by BRIQUE, NIQE and PIQE.

	Photo names	BRISQUE	NIQE	PIQE	Total score
Discarded	A	62.09	4.17	73.59	139.86
Retained	A’	60.35	4.13	71.42	135.90
Discarded	B	59.54	4.06	71.87	135.47
Retained	B’	56.43	4.15	66.43	127.01
Discarded	C	61.71	3.99	74.94	140.64
Retained	C’	57.55	4.08	74.72	136.36
Discarded	D	54.02	4.31	73.23	131.57
Retained	D’	54.36	4.20	65.97	124.53
Discarded	E	60.56	4.19	74.51	139.26
Retained	E’	59.38	4.11	73.92	137.41

6.4.3 Data description

The boreholes monitored were located in two different zones of the quarry, North (N) and Northeast (NE), and two benches. Table 37 shows the blast events and related borehole information. It should be noted that in the case of blasts 230530 (B11), only one borehole was considered because more data was needed for the HC class while the prevailing lithologies in those blasts were classes ML and BL. As a result, a total of 79 boreholes and 7583 images were used for the development of the lithology prediction model.

6.4.4 Contrast-limited adaptive histogram equalization (CLAHE) algorithm

Adaptive histogram equalization (AHE) is an image pre-processing technique used to improve contrast in images (Pizer et al., 1987). It calculates multiple histograms, each associated with a specific portion of the image, and utilizes them to redistribute the luminance values across the image. Consequently, it is effective in enhancing local contrast and refining edge definitions within local parts of an image, tough with a drawback of magnifying noise for relatively homogeneous regions. Contrast-Limited Adaptive Histogram Equalization (CLAHE) (Reza, 2004) addresses this issue by restricting the extent of local amplification. At first, CLAHE divides an image into sub-blocks, each of which is considered a local area, and the histogram equalization is performed on each sub-block to enhance local contrast. In addition, CLAHE introduces a contrast limitation after histogram equalization. Within each small block, it truncates the pixels in the histogram to ensure that the contrast enhancement is not excessive, this way limiting the possibility of amplifying noise in relatively uniform areas. In this study, the CLAHE technique is employed to increase the contrast of grey colors and reddish colors so as to decrease the negative influence of illumination. Some examples of original and CLAHE-processed images can be seen in Figure 73. In comparison, the processed images exhibit enhanced image characteristics, particularly in regions with shadowing. The reddish areas, typically corresponding to clay, and the gray regions, generally representing massive limestone, are more distinctly delineated. Given that the classification of lithologies in this study primarily relies on the ratio of clay to massive limestone, these improvements are expected to provide better differentiation of lithological units.

Table 37. Borehole data description for the study.

Blast code	No. of boreholes considered	Number of available images	Number of lithologies		Location (quarry area)	Bench number
230120 (B1)	10	1081	ML	131	NE	3
			BL	815		
			HC	135		
230131 (B2)	15	920	ML	67	NE	3
			BL	837		
			HC	16		
230213 (B3)	13	1413	ML	1397	N	1
			BL	16		
			HC	0		
230308 (B5)	10	898	ML	36	NE	3
			BL	818		
			HC	44		
230316 (B6)	30	3211	ML	2998	N	1
			BL	213		
			HC	0		
230530 (B11)	1	60	ML	0	N	1
			BL	35		
			HC	25		

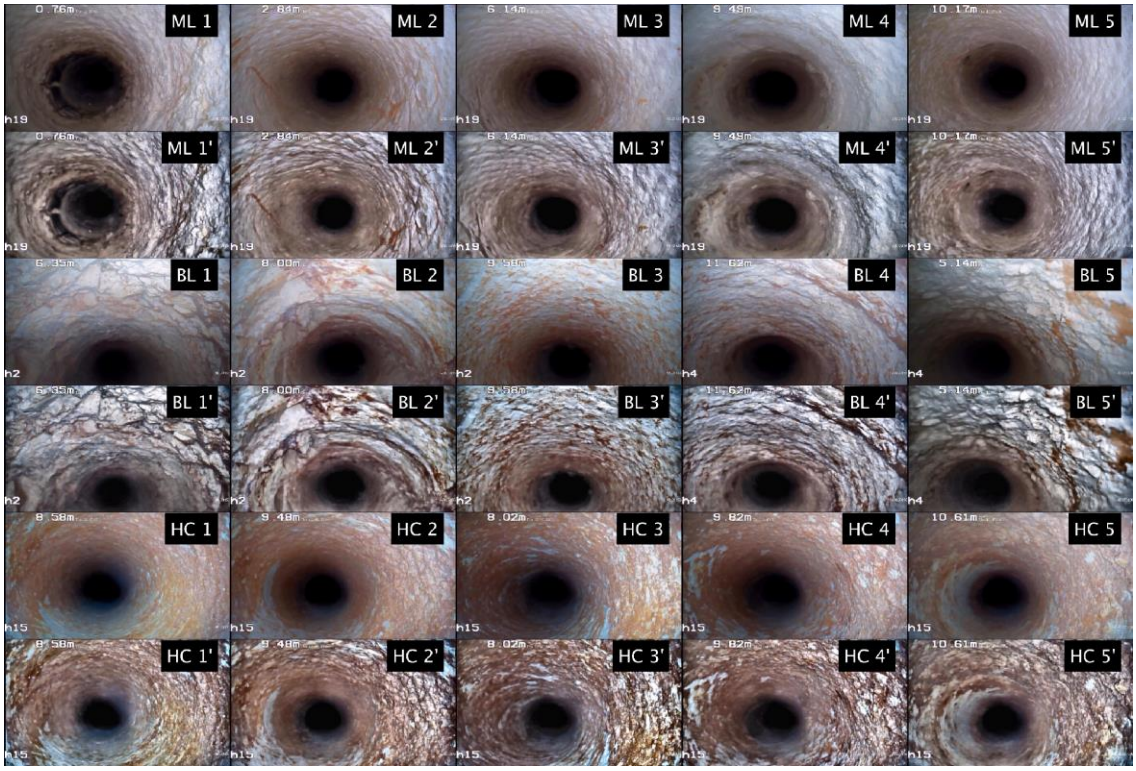


Figure 73. Comparison between original (cardinal row) and processed (even row) images using CLAHE technique.

6.5 Model development and evaluation

6.5.1 Determination of input parameters

Although some undesirable photos with a high amount of very bright or very dark pixels have been discarded by the process described in section 3.2, there may still be dark or over-exposed areas in the images retained that cannot be used for feature extraction. In general, red (R), green (G) and blue (B) pixel values over 235 or less than 30 are over-exposed or too dark. Therefore, all image intensity properties are extracted on the R, G and B range of [30, 235]. A pre-defined parameter named "color counting" (CC) is applied where the number of unique colors can be counted in an image. First, this function identifies the unique colors along with their number of pixels, then removing colors with a pixel count below a certain threshold, to reduce the number of colors considered. Herein, this threshold is set to be 5%. The main reason for using this function is that images from class ML generally show a narrower color distribution than classes BL and HC. In addition, some regular RGB intensity properties are also calculated, including the 25th percentile, the median and the 75th percentile of R, G and B pixels, abbreviated Q1R, Q1G, Q1B, Q2R, Q2G, Q2B, Q3R, Q3G and Q3B, respectively. These ten parameters are used to describe the pixel distribution. Since the patterns of classes ML and HC are comparatively monotonous and the class BL is comparatively intertwined, texture features are also considered. For characterizing the texture properties, 5 parameters are calculated, i.e., contrast (CT), energy (EN), correlation (CR), homogeneity (HG) and entropy (ET) (Aouat et al., 2021; Tsai et al., 2008). To obtain the four parameters CT,

EN, CR and HG, the gray level co-occurrence matrix (GLCM) needs to be calculated (Aouat et al., 2021). It is defined in Eq. (5.5).

The five texture parameters are calculated as follows:

$$CT = \sum_{i=1}^{N} \sum_{j=1}^{N} P(i,j) \cdot (i-j)^2 \tag{6.1}$$

$$EN = \sum_{i=1}^{N} \sum_{j=1}^{N} P(i-j)^2 \tag{6.2}$$

$$CR = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (i \cdot j \cdot P(i,j)) - \mu_x \cdot \mu_y}{\sigma_x \cdot \sigma_y} \tag{6.3}$$

$$HG = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{P(i,j)}{1+|i-j|} \tag{6.4}$$

$$ET = -\sum_{i=1}^{N} \sum_{j=1}^{N} P(i,j) \cdot \log_2(P(i,j)) \tag{6.5}$$

where *i* and *j* define the row and column index in the GLCM, $i, j \in \{1,2,\cdots,N\}$, *N* being the image gray levels. *P(i,j)* indicates the co-occurrence probability of gray levels *i* and *j*. $P(i,j)^2$ denotes the square of each value in the GLCM, reflecting the concentration of the distribution. σ_x and σ_y are the standard deviation of row and column gray values. μ_x and μ_y are the product of the mean values, used to adjust the offset. $|i-j|$ indicates the absolute differences in gray levels. $\log_2(P(i,j))$ symbolizes the logarithm of the distribution, which is used as a measure of the uncertainty.

CT measures the variation in intensity or color between neighboring pixels in an image; higher CT values indicate a greater difference in pixel values, resulting in more pronounced texture patterns. EN quantifies the randomness or uncertainty in the distribution of pixel values within an image; images with more complex or varied textures tend to have higher EN values, while uniform or less textured images have lower EN values. CR measures the linear dependence between pixel values in different parts of an image, providing information on how well defined and oriented texture patterns are within the image. HG measures the similarity or uniformity of pixel values in an image; higher HG values indicate that the texture is more consistent and uniform, while lower HG values suggest variations in the texture (Mutlag et al., 2020). ET is a statistical measure of randomness, typically used to assess the uncertainty or complexity of a system. In image processing, the higher the ET, the more complex the texture of the image, and the lower the ET, the simpler or more uniform the texture of the image. In total, 15 image properties are used as inputs for the lithology classification models. A flow chart showing how the inputs are determined from the images is given in Figure 74.

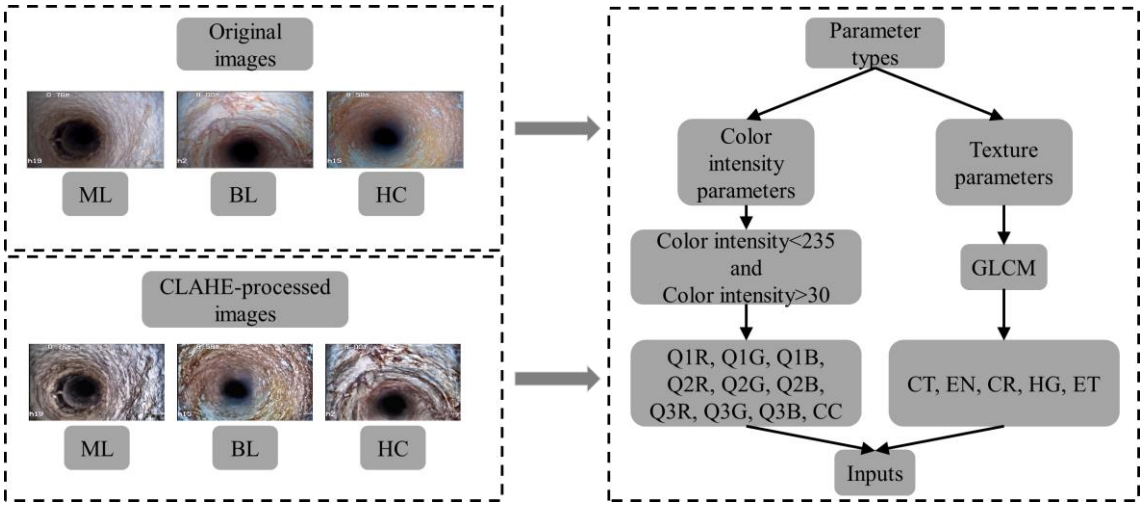


Figure 74. General flow to extract inputs for lithology prediction.

The distribution of inputs from the original and CLAHE-processed images is presented in Figure 75 and Figure 76, respectively. For the original inputs, CC is clearly distinct for different classes based on the range from lower to upper quartiles. This indicates that this image characteristic might play a significant role in discriminating the lithology classes. The upper and lower quartiles of inputs Q1R, Q1G, Q1B, Q2R, Q2G, Q2B and CT present gradually increasing trend from class ML to HC. The distribution of HG indicates a gradually decreasing trend from class ML to HC. For Q3R, EN and CR from classes BL and HC, they have similar range of upper and lower quartiles. For the CLAHE-processed inputs, a similar CC distribution can be observed as from the original inputs. However, the gradually decreasing trend from class ML to HC is observed from inputs Q1R, Q1G, Q2R, Q2G, Q3R, Q3G, HG and CR. The rising trend is only seen in CT. Q3B present similar data distribution for classes ML and BL.

For the determination of interdependence between inputs and lithology classes, the Pearson Correlation Coefficient (PCC) is calculated, see Table 38. For this, the lithology classes are assigned a number 0, 1 and 2. The highest PCC is produced by CC for two types of inputs. Most original inputs have positive PCC with lithology types while 10 inputs show negative PCC with lithologies for the CLAHE-processed inputs. Two low PCCs are obtained from CLAHE-processed inputs, Q2B and Q3B. Compared with the PCCs from the original images, Q1R, Q2R, Q2G, Q3R, Q3G, CC and five texture parameters produce higher absolute PCCs while the PCCs of Q1G, Q1B, Q2B and Q3B are lower.

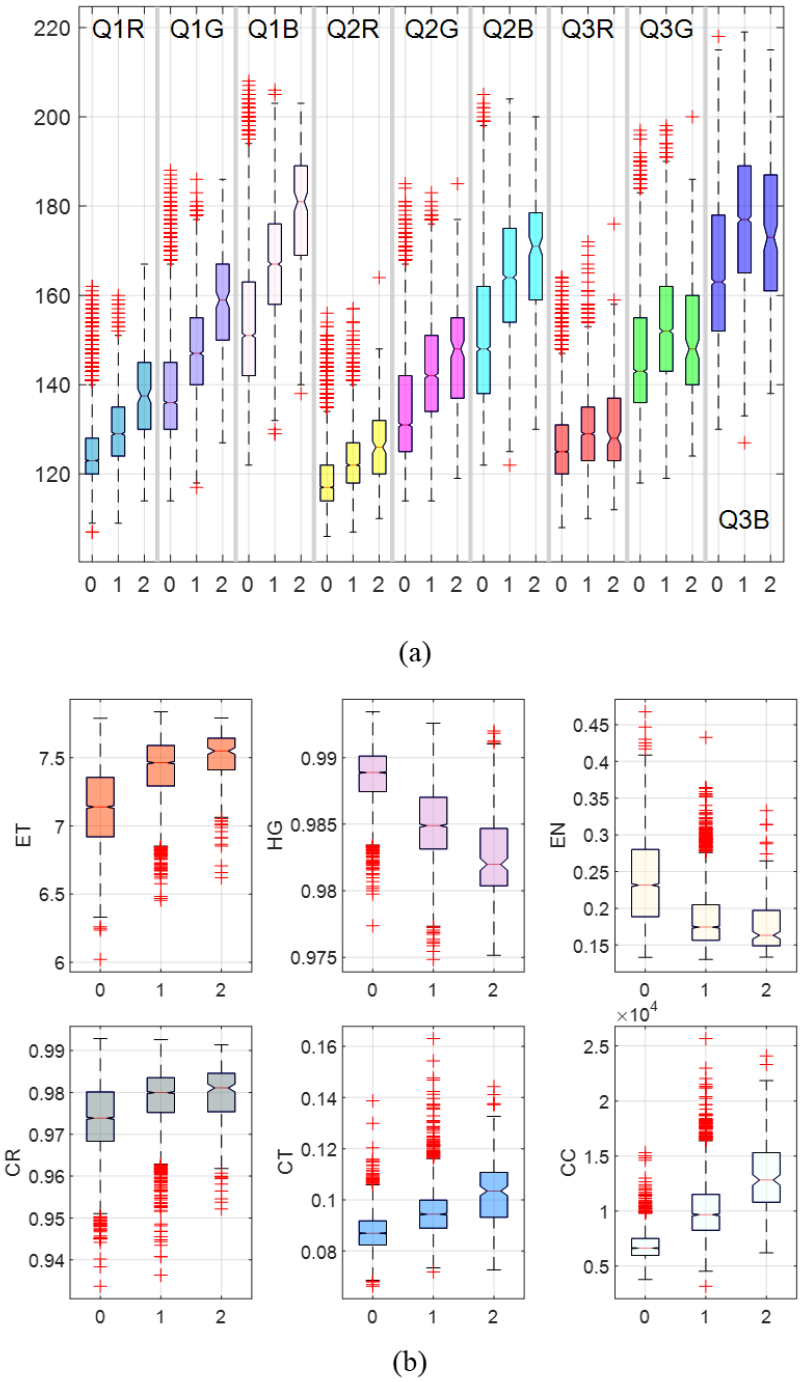


Figure 75. Distributions of the original inputs by box plots: (a) color properties (except from CC); (b) CC and five texture properties.

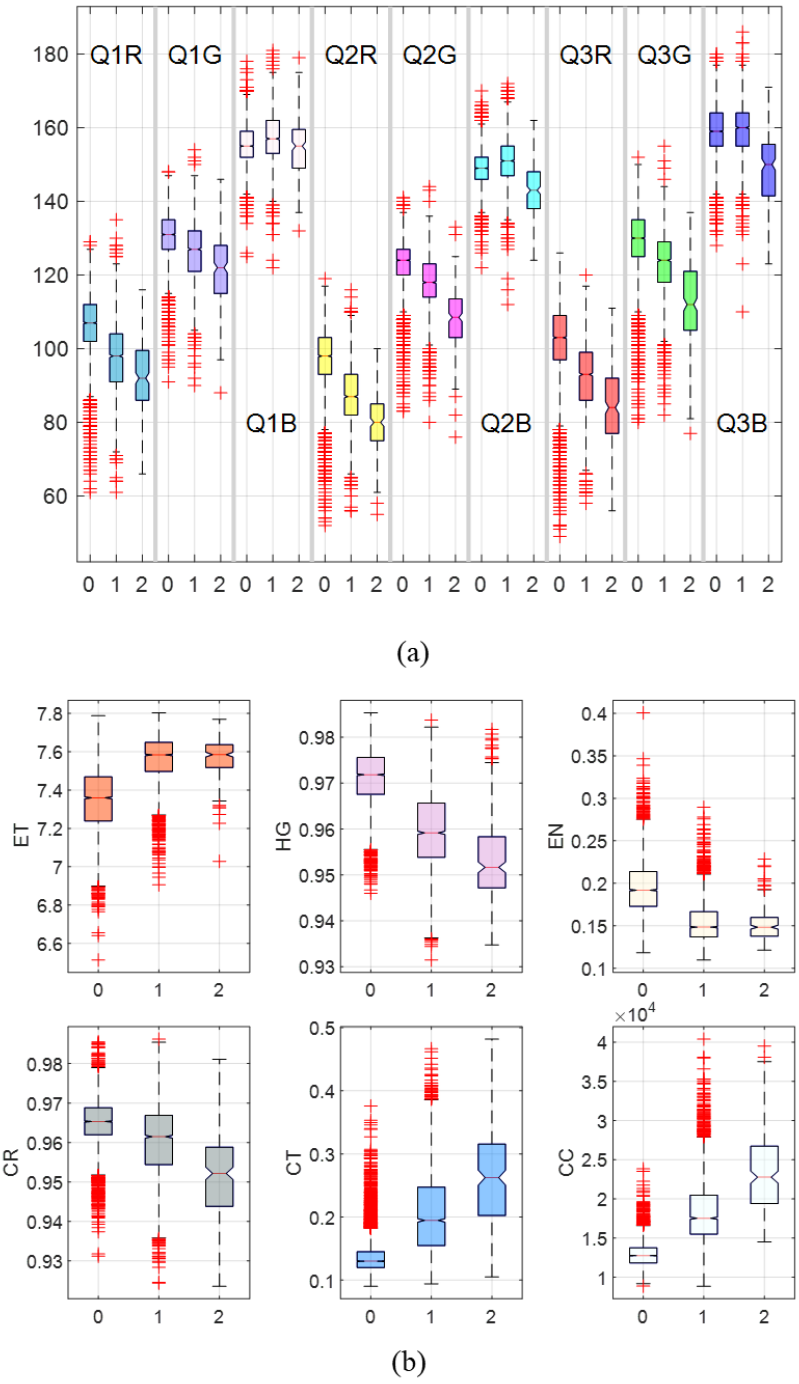


Figure 76. Distributions of the CLAHE-processed inputs by box plots: (a) color properties (except from CC); (b) CC and five texture properties

Table 38. PCC between original as well as CLAHE-processed inputs and lithology types.

Inputs	Q1R	Q1G	Q1B	Q2R	Q2G
Original	0.35	0.40	0.44	0.28	0.32
CLAHE-processed	-0.44	-0.31	0.14	-0.52	-0.41
Inputs	Q2B	Q3R	Q3G	Q3B	CC
Original	0.37	0.16	0.21	0.29	0.67
CLAHE-processed	0.05	-0.45	-0.38	-0.08	0.71
Inputs	CT	CR	EN	HG	ET
Original	0.46	0.28	-0.43	-0.61	0.46
CLAHE-processed	0.60	-0.38	-0.56	-0.62	0.55

6.5.2 Modelling methodology

Several classical and comparatively novel classification techniques are employed, namely Support Vector Classification (SVC), Random Forest (RF), Bagged Tree (BT), Gradient Boost Machine (GBM), Extreme Gradient Boosting (XGB) and Light Gradient Boosting Machine (LGBM).

Bayesian (BY) optimization algorithm is an optimization technique used to find the minimum or maximum of an objective function that is expensive to evaluate (Zhou et al., 2021). In this study, it is used for tuning the hyper-parameters in the aforementioned classification models. It builds a probabilistic model, typically using a Gaussian Process, to predict the function's behavior and guide the search towards promising areas (Zhou et al., 2021).

6.5.3 Normalization and cross validation

Since the inputs have different magnitude, normalization is used to scale features to a standard range [0, 1] as follows:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{6.6}$$

where x, x_{min}, x_{max} and $x_{normalized}$ represents, the original, minimum, maximum and normalized variable, respectively. This ensures that all features contribute equally to the model, preventing features with larger scales from dominating the learning process. Five-fold cross validation is employed to reduce overfitting where the training dataset is divided into five equal parts (Fushiki, 2011). The model is trained on four parts and validated on the remaining one. This process is repeated five times, with each part used

exactly once as the test set. The results are then averaged to provide an overall performance metric, reducing the bias and variance of the model evaluation.

6.5.4 Evaluation of classifier performance

To evaluate the overall classification efficiency of the model, five classical classification indicators were calculated and compared, i.e., accuracy (Ac; Eq. (5.7)), precision (Pr; Eq. (5.6)), recall (Re; Eq. (5.8)), specificity (Sp; Eq. (5.9)) and F1 score (Fs; Eq. (5.10)). They have been described in Chapter 5.

6.5.5 Model development

70% of total images (5308) are used for developing the classification model and 30% (2275) are used for testing the model. To ensure the consistency of the lithology distribution, a stratified sampling is used and a constant random seed is applied to all lithology classification models. As a result, 3241 images of class ML, 1914 images of class BL and 154 images of class HC are extracted to establish the training set and 1388 images of class 0, 820 images of class 1 and 66 images of class 2 are retained to develop the testing set. The accuracy from five-fold cross validation is used as the score to tune the hyper-parameters in six classification models. The number of iterations is set to 50 since more iterations would consume more computation time and less iterations might cause under-fitting. Generally, the CLAHE-processed scenario produces a better score than the original one. The optimization progress can be seen in Figure 77; it converges after the 25th iteration, approximately. The optimized hyper-parameters, optimization bound and modelling time can be seen in Table 39. These optimized hyper-parameters would be used for the development of lithology classification models based on the Python 3.7 environment. The specifications of the used device include: Processor: Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz, RAM: 16.0GB and Windows 10 system.

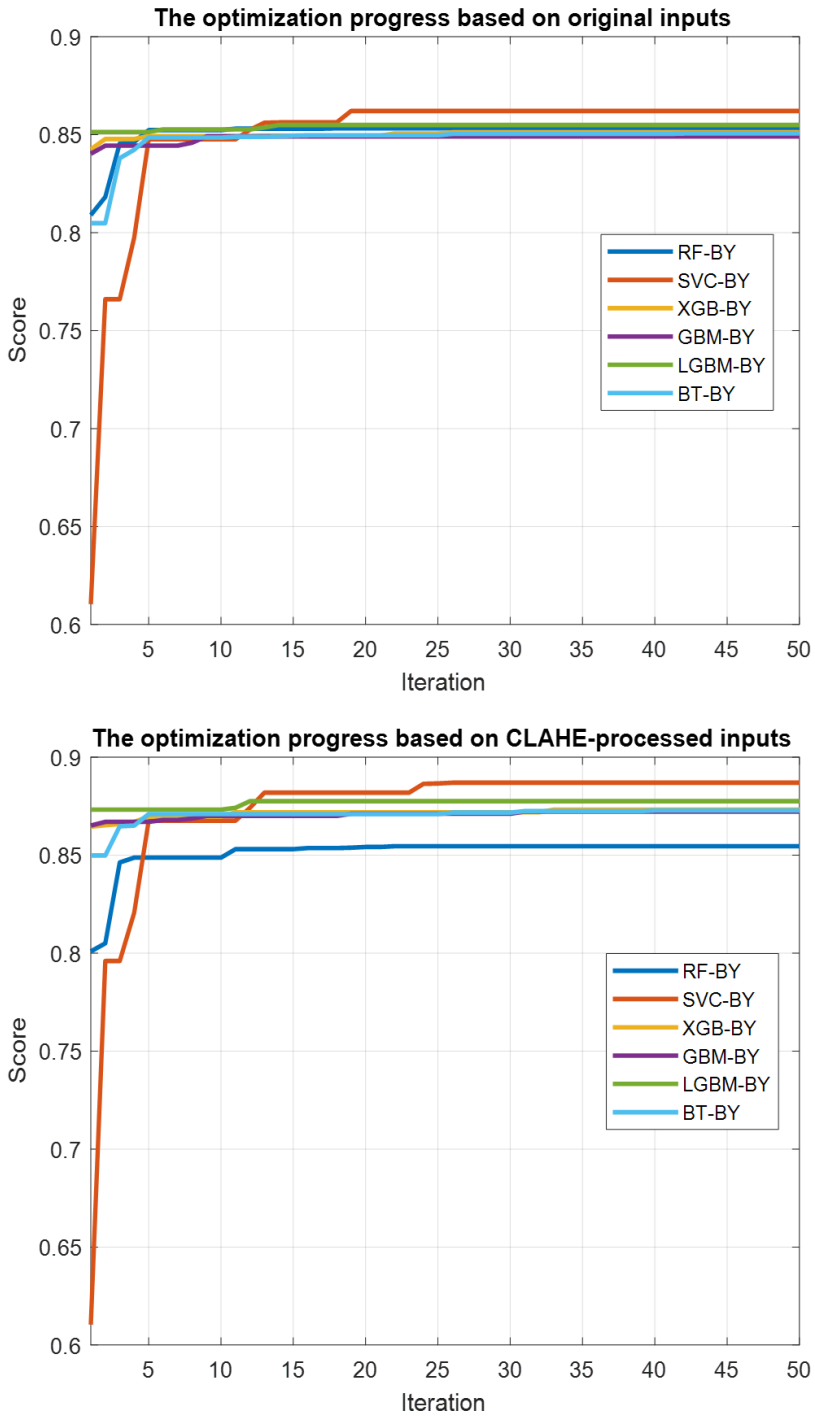


Figure 77. Optimization process based on five-fold cross validation, six optimized classification models and two types of inputs.

Table 39. Optimized hyper-parameters, bounds and corresponding values.

Method	Optimized hyper-parameters	Original scenario		CLAHE-processed scenario		Lower bound	Upper bound
		Optimized values	Modelling time (s)	Optimized values	Modelling time (s)		
SVC	C	16.828	74.95	62.486	70.24	0.001	100
	g	4.292		1.769		0.001	100
RF	n_estimators	28	32.73	41	30.12	1	100
	max_depth	17		16		1	20
XGB	learning_rate	0.382	22.72	0.376	21.85	0.01	0.5
	min_child_weight	15		17		1	20
LGBM	learning_rate	0.037	99.75	0.058	98.21	0.01	0.5
	min_data_in_leaf	29		96		10	100
GBM	learning_rate	0.158	957.38	0.117	940.85	0.01	0.5
	min_samples_split	7		13		2	20
BT	n_estimators	64	409.32	49	386.94	1	100
	max_depth	10		16		1	20

6.6 Results

The confusion matrices are shown in Figure 78 and Figure 79. For both training and testing sets, the class ML is predicted correctly in general. Most ML wrong prediction cases are predicted as BL and only one case is predicted to be HC. Some class BL cases are classified to be ML while the correct prediction is still predominant; only a few class BL cases are identified as class HC. Therefore, it can be concluded that the recognition of ML and BL is favorable.

Wrong predictions of class HC generally go to BL and only a few cases are classified as ML. For the training set, the best prediction performance of class HC is from the RF-BY where 147 cases are correctly predicted out of 154 cases for two input scenarios. For the testing set, the discrimination between classes BL and HC is relatively vague. The reason for this could be the subjective assessment of the amount of clay, sometimes causing inaccurate classification by geologists. The best identification of class HC is achieved by the model GBM-BY and LGBM-BY for the original and CLAHE-processed inputs, respectively, where 23 and 29 cases are predicted correctly out of 66 cases. In general, CLAHE-processed inputs have demonstrated similar prediction capacity for class ML and better recognition for classes BL and HC, compared with the unprocessed inputs.

By means of the confusion matrices, the five classification metrics are calculated, shown in Figure 99 to Figure 104 and Table 56 to Table 57 in the Appendix 5. The classification metric for each class is noted with “0”, “1” and “2” for ML, BL and HC, respectively. Both input scenarios present favorable performance for both training and testing sets, except for Fs2 and Re2. The overall prediction performance from the testing set is lower than the training set. For testing set using original inputs, Fs2 and Re2 are quite low. However, Pr2 is much higher, up to 73.91% by BT-BY model. This means that, among the images predicted to be class HC, there is a percentage of 73.91% that actually are class 2. The highest Re2, 34.85% (which indicates that about 65% of class 2 images are recognized to be class 0 or class 1) is produced by RF-BY. Since the Fs is a balanced metric between Re and Pr, it doesn’t show a satisfactory performance for class 2, either. Sp for class 2 is high which indicates that the models have a low inclination of predicting classes 0 and 1 to be class 2. The highest Ac is 86.33% generated by SVC-BY, while most of this is contributed by classes 0 and 1 because they occupy a larger proportion in the dataset. Compared with the prediction results from the original inputs, most evaluation metrics have been enhanced by the CLAHE. The highest Re2 improves from 34.85% to 43.94% by LGBM-BY. Pr2 increases from 73.91% to 90.00% by XGB-BY so the model provides a higher prediction confidence for class 2. As for other classification metrics, they also improve. For instance, the highest Ac increases from 86.33% to 89.10% by SVC-BY. The best Pr0 is 91.46% by SVC-BY compared with 87.99% from the original input variables.

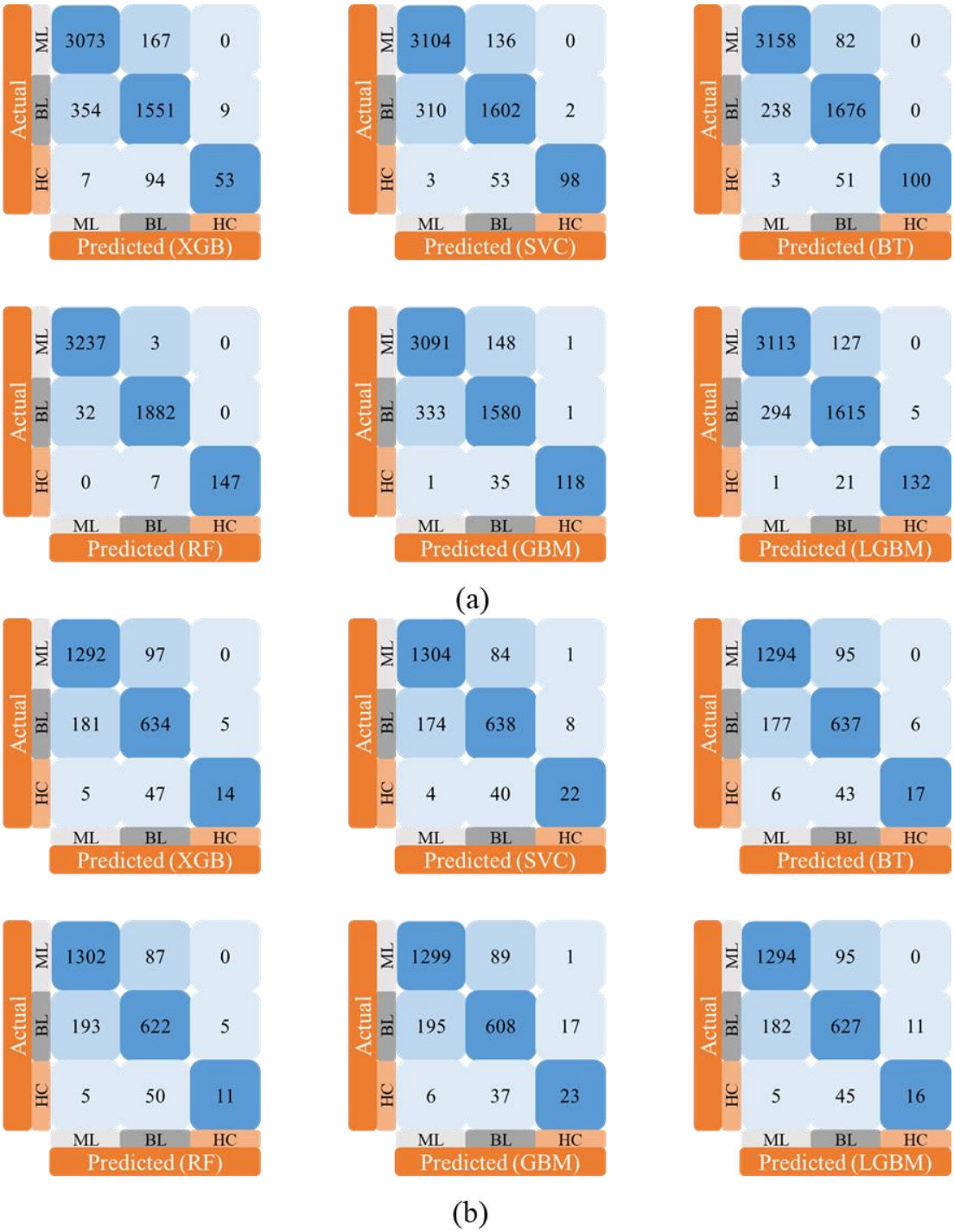


Figure 78. Confusion matrices based on the original inputs and six optimized classification models: (a) training set, (b) testing set.

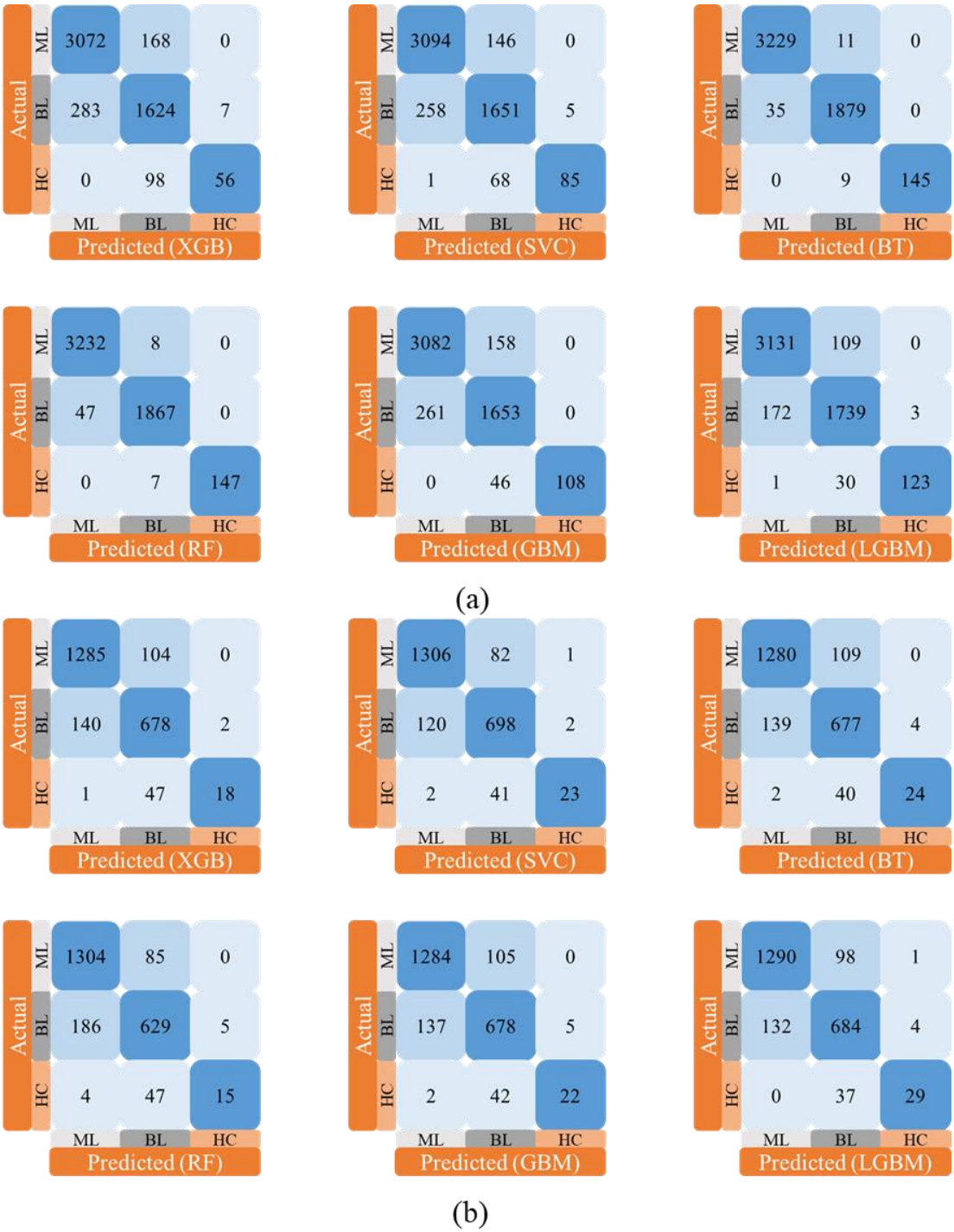


Figure 79. Confusion matrices based on the CLAHE-processed inputs and six optimized classification models: (a) training set, (b) testing set.

The overall evaluation system proposed by (Zorlu et al., 2008) is adopted to determine the best classification model. In it, each of the six models is graded 1 to 6 for each metric, the better getting a higher mark. The best classification model is the one that has the highest addition of marks. Adding the total marks from the training and testing sets, the best classification model using the original and CLAHE-processed inputs is, respectively, BT-BY and LGBM-BY with total marks 116 (59+57) and 113 (51+62), see Table 56 and Table 57 in the Appendix 5. By comparing the metrics of both scenarios, LGBM-BY

model combined with CLAHE technique has a better overall classification performance, so this is the one recommended for lithology recognition.

Although the distinction between BL and HC is vague in some cases, the Pr of the three classes is relatively high, from 83.52% to 90.72%. The Re of 92.87% and 83.41% for ML and BL indicate that the prediction for these two classes is quite accurate. However, for class HC, it is unsatisfactory, i.e., 43.94%, although the discrimination between ML and HC is high as shown by the results of the confusion matrix.

A sample comparison of measured and predicted lithology is represented in Figure 80 for borehole B11 in blast 230530 (North pit) (collar coordinates X, Y, Z 474535, 4463159, 863 UTM 30 North and WGS84 System), for the two best scenarios, BT-BY and LGBM-BY. Lithologies are shown with red (ML), yellow (BL) and green (HC). Some sections, shown in black, were not measured, or were discarded due to poor image quality (this sometimes happens in the collar zone of the borehole due to drilling fine material or debris falling from the surface, see the black zone in the lower part of the bars), or were not chosen in the training or the testing sets. Two lithologies, BL and HC, were present (Measured-TR and TS bars) in that borehole. In the training set (TR), all HC (green) sections are predicted correctly by LGBM-BY while four HC sections are predicted incorrectly by BT-BY. LGBM-BY yields three wrong predictions for BL and BT-BY four. In the testing set (TS), LGBM-BY still performs better and six sections are predicted incorrectly, compared with seven by BT-BY. Other cases of lithology distribution and prediction are shown in the Appendix 5, Figure 105 to Figure 107, for different blocks of the quarry site; only results with the best model, LGBM-BY with CLAHE-process inputs, are presented there.

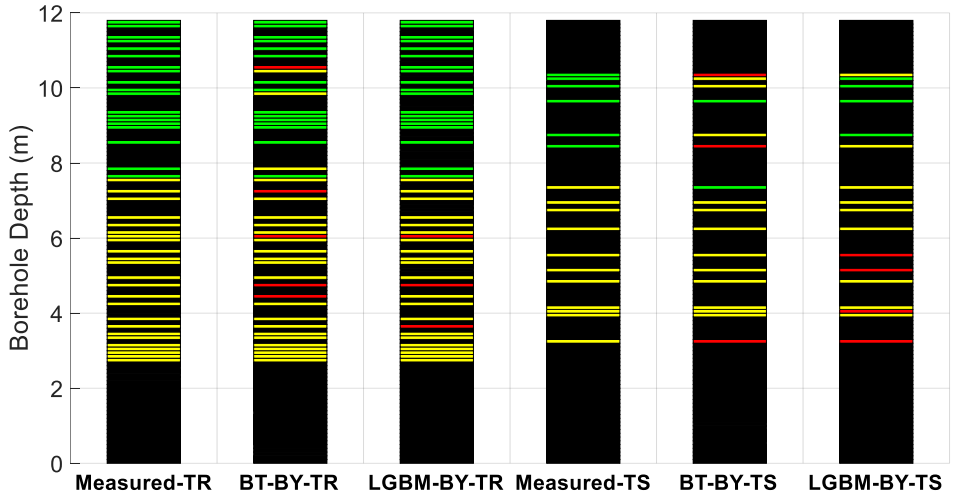


Figure 80. Example of distributions of measured and predicted lithologies in a borehole (B11, blast 230530, North pit) by BT-BY and LGBM-BY. Red: ML, yellow: BL, green: HC, black: sections without data or where data are not used in the training (TR) or the testing (TS) sets.

6.7 Discussion

6.7.1 Feature importance scores

To determine the influence of the different borehole image properties on the models, scores of feature importance are calculated for BT-BY and LGBM-BY, see Figure 81. CC is the most significant attribute for lithology classification among all image characteristics. For CLAHE-processed inputs, the top five ranking include CC, ET, HG, CR and Q1R. For original inputs, the top five ranking is CC, HG, CT, Q1R and CR, which suggests that that it would be possible to discriminate other types of lithologies when they feature differential distributions of these image characteristics. Specifically, the outstanding relevance of CC means that different types of rock can be distinguished if they have different diversities of colors. This probably obvious fact provides physical soundness to the method described in this work and the results obtained.

6.7.2 Comparison with previous studies

To assess the significance of recognizing lithology using borehole images from an endoscope and machine learning models, some recent works employing AI techniques are summarized in Table 40. Several studies have applied deep learning-based networks, such as CNN, ResNeXt-50 and Xception, with a robust predictive accuracy. Due to practical considerations, deep learning usually requires higher computational power and provides less interpretability. Despite that, we have investigated the direct use of deep learning techniques for our problem without success. Reasons for that may be the sometimes vague distinction between the patterns of BL and HC and the undesirable illumination in some sections of the footages. From an equipment perspective, obtaining endoscope images from boreholes is an easy way to capture the underground lithology information, compared with core sampling, thin sections or televiewer. The endoscope can be deployed in remote or hard-to-access areas without a complex setup, with a lower cost than advanced radar or ASTER systems, and with a higher resolution; radar is sensitive to surface roughness and structure, making it suitable for distinguishing lithologies in exposed areas, but less effective for subtle changes under the surface.

6.7.3 Model stability

Since the random division of the training and testing sets would probably influence the model robustness and generalization, checking the model stability is required. Repeated random splits allow for a statistical analysis of the performance metrics (Li et al., 2023a). To this end, the whole lithology dataset was randomly divided 30 times generating 30 groups of training and testing sets, for which optimized model hyper-parameters and prediction performance were generated. The significance of the performance is assessed by comparing the lithology forecasting results from the single random division in Section 5 with the statistics produced by the 30 random divisions.

Take the LGBM-BY model on the CLAHE-processed images for example, with the same iteration number (50) and score (i.e., accuracy from five-fold cross validation) as explained in Section 4.4. The 30 groups of optimal hyper-parameters are plotted in Figure 82. The accuracy for the cross-validation sets is in excess of 87%, even if the hyper-parameters show some variation. Applying these hyper-parameters to build the prediction models, their performance for the 30 training and testing sets is given as boxplots of the

performance metrics in Figure 83; some descriptive values are given in Table 58 in the Appendix 5.

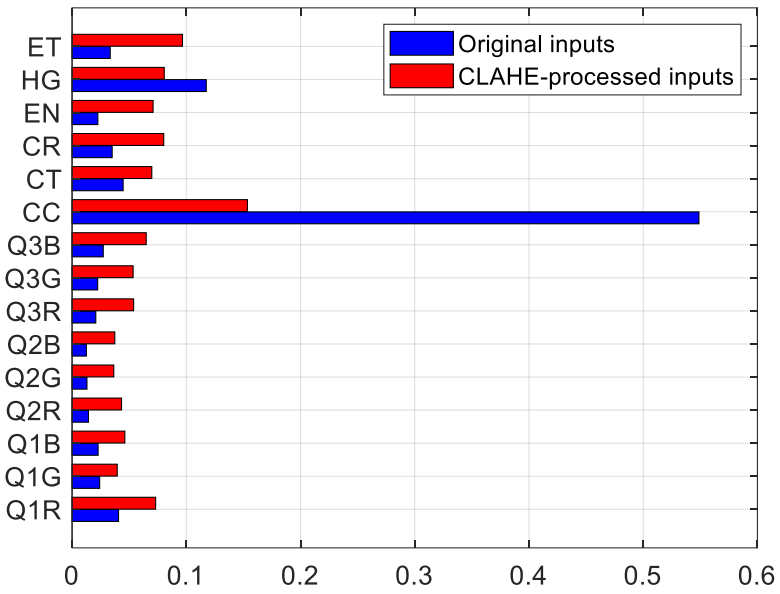


Figure 81. Feature importance scores for original inputs with BT-BY and for CLAHE-processed inputs with LGBM-BY.

The interquartile ranges of the evaluation metrics produced by the 30 random divisions are relatively narrow, especially for the testing sets, see also the small standard deviations in Table 58. The metrics from the training set of the single division (section 5) do not fall in the inter-quartile range except for Sp0, though they are generally not extremes nor outliers, see also Table 58. The above suggests a stable behavior of the model LGBM-BY with CLAHE- processed images.

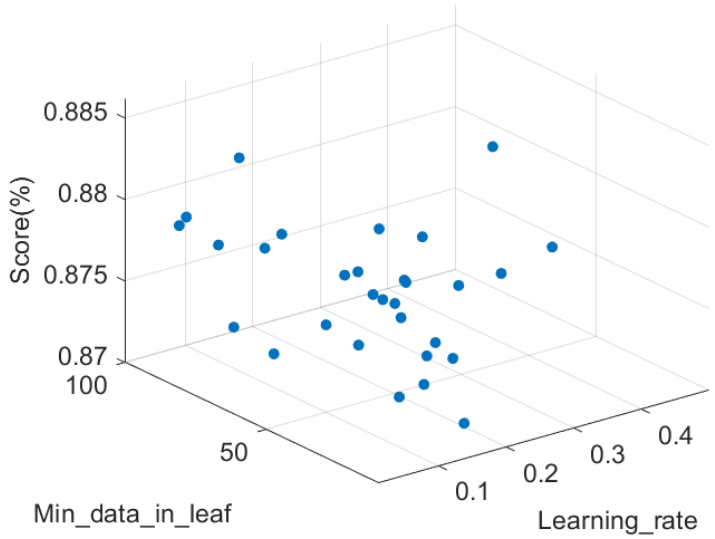


Figure 82. Distribution of 30 groups of optimal hyper-parameters using LGBM-BY and CLAHE-processed images, and their resulting scores.

Table 40. Recent works using AI techniques and imagery to predict or recognize lithology.

Reference	Image types or inputs	AI methods	Data size	Ac
Zhang et al. (2017)	Borehole images from televiewer	CNN	1500	95%
Nanjo and Tanaka (2019)	Thin sections	CNN	306	83.90%
Dos Anjos et al. (2021)	Cropped microtomographic images	CNN	6000	81.33%
Xu et al. (2021)	Rock images	ResNeXt-50	8974	99.19%
Shayeganpour et al. (2021)	ASTER + simulated panchromatic S2	RF	1950	98%
Alzubaidi et al. (2021)	Cropped core tray images	ResNeXt-50	76500	92.00%
Xu et al. (2022)	Rock microscopic images	Xception	14950	98.65%
Fu et al. (2022)	Cropped core tray images	ResNeXt-50	15000	99.60%
Faria et al. (2022)	Thin sections	MFFNN	570	83%
Manap and San (2022)	Optical images, radar images and digital elevation model data	NN	11402	97.36%
Shirmard et al. (2022)	ASTER	CNN	-	99%-100%
Bahrami et al. (2024)	ASTER	SVC	20966	85%

Note: CNN - conventional neural network; MFFNN - multilayer feed-forward neural network; NN – neural network; RF - random forest; SVC - support vector classification.

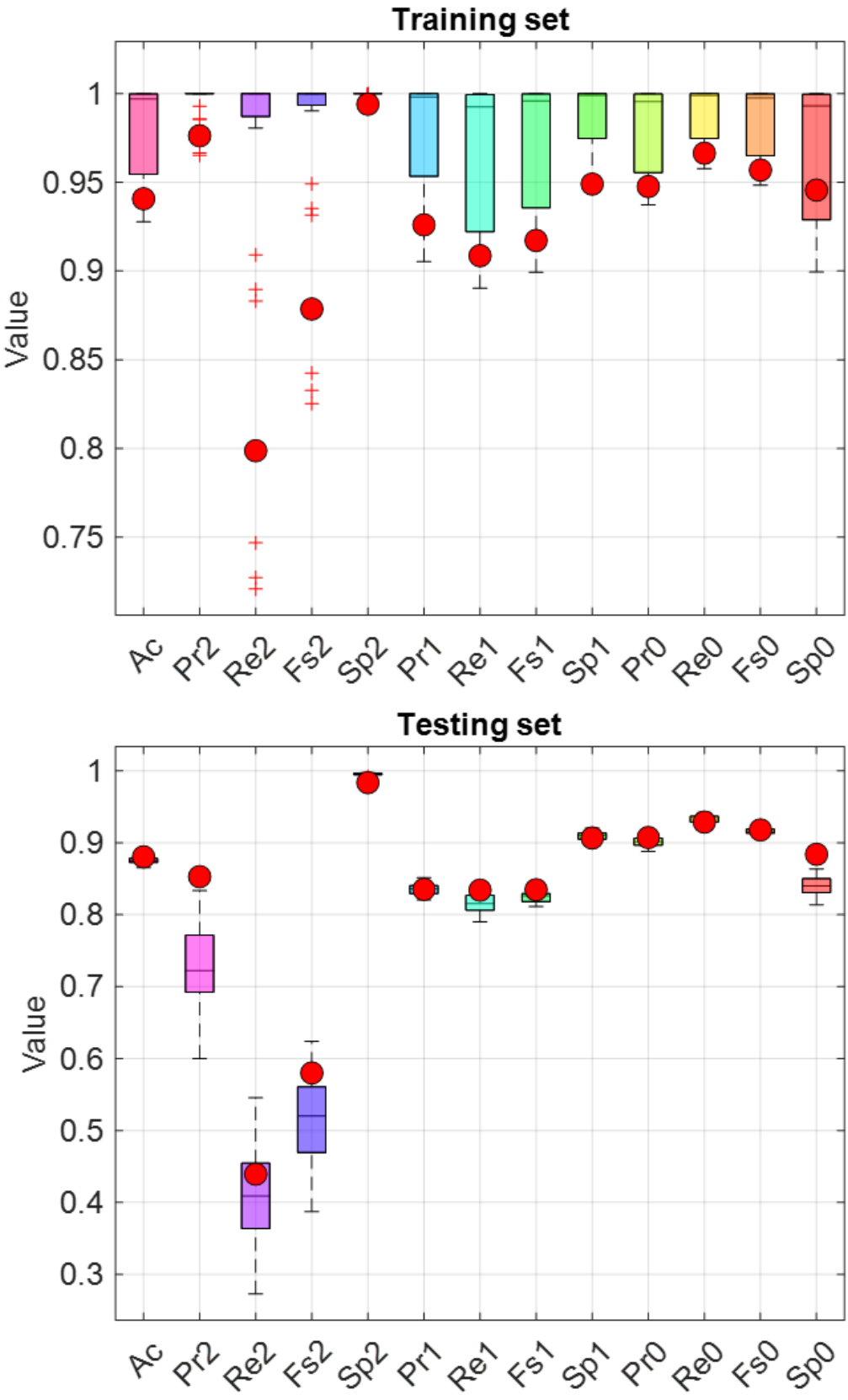


Figure 83. Lithology prediction performance metrics statistics based on 30 random divisions using LGBM-BY and CLAHE-processed images: red dots: performance of the single division of LGBM-BY, see Section 5.

6.7.4 Practical applications

The borehole lithology recognition method proposed in this study works well with data monitored with a simple equipment in the usually hard conditions of a drilling and blasting environment. It enables fast lithological classification, reducing reliance on costly and time-consuming core sampling, and providing continuous profiling of boreholes. The decision-making during drilling, optimizing costs and operational efficiency can be enhanced and thus assist in the explosives loading phase. Together with that, it also aids in stratigraphic correlation, geotechnical analysis and in assessing rock mass stability, offering cost-effective subsurface insights.

For an optimum application, proper borehole lighting should be ensured. The endoscope should be driven into the borehole using a controlled mechanism with a stable and reasonably slow speed to guarantee minimal distortion and consistent image quality. Images should be captured from the videos at specific intervals, e.g., two seconds in our case. Shadows or overexposures should be avoided, as they could affect analysis. To assign the captured frames a depth, the digits representing the depth are cropped and correlated with specially prepared digit images. Image cleaning discards photos with poor illumination and low quality, after which some key visual features such as color intensity and texture can be extracted to be inputs of the trained LGBM-BY model, from which the lithology labels are obtained as outputs. By integrating the depth metadata with the predicted lithologies, spatial lithology distribution plots can be visualized. Figure 84 shows a flow chart of the method.

A similar procedure could be considered for application with other lithology types if they feature differential color intensity or texture image attributes, albeit different image characteristics might be considered in other mining or geology sites. The interdependence analysis is a sound means to determine the inputs for the lithology prediction.

6.8 Limitations

Some aspects might be considered in order to improve the performance at identifying lithology types, namely: (1) a more stable centralizer, to ensure a good alignment of the camera in order to provide uniform illumination and more consistent image layout, see an example of this in Figure 85 (left); (2) a higher resolution of the endoscope camera should procure more precise texture parameters; (3) stable and slow movement of the endoscope camera should reduce the number of blurred images and the sliding of surrounding rock debris, as shown in Figure 85 (right); and (4) some subjectivity in the lithology identification (i.e. the amount of clay in our case) from visual inspection is inevitable, leading to some confusion of the classification model. This is difficult to overcome, but validation on some other types of lithologies would be of interest in order to confirm a broader reliability of the procedure developed here.

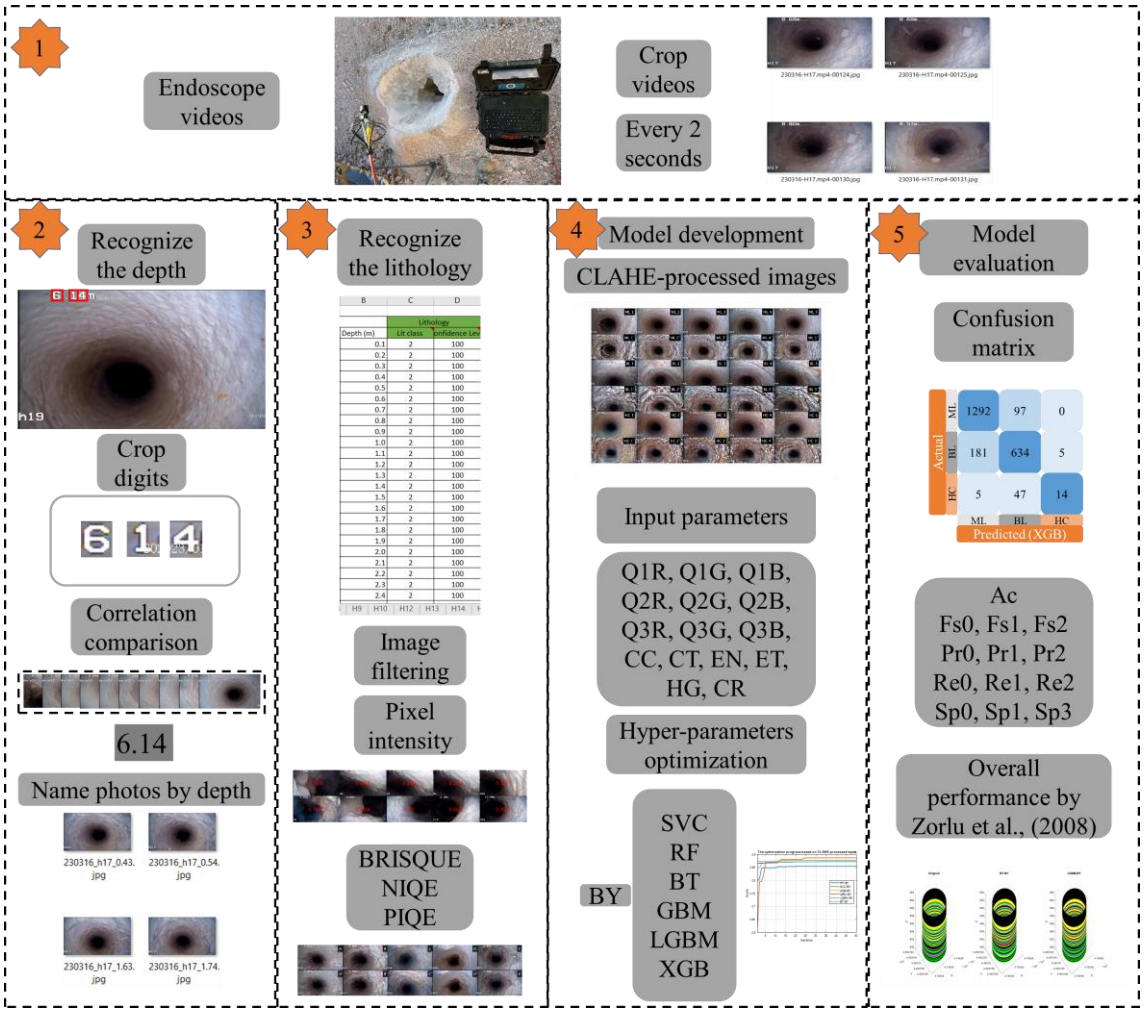


Figure 84. General flow chart of proposed method to recognize lithology.

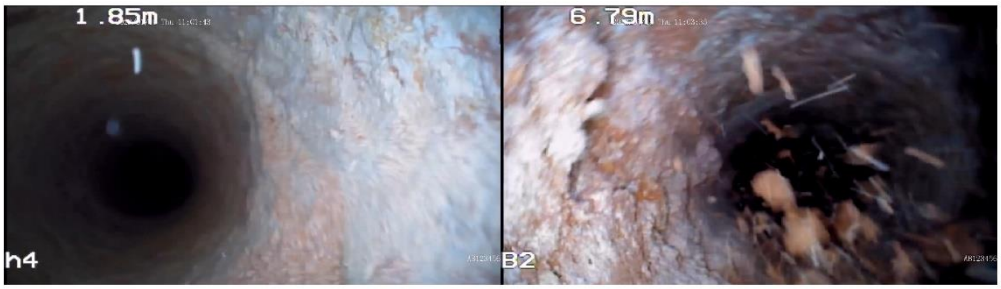


Figure 85. Detrimental cases of images: (left) uneven illumination; (right) unstable movement of the endoscope camera.

6.9 Conclusions

A new lithology identification and prediction methodology is proposed from borehole wall images captured from endoscope video footages, for the assessment of sections with massive limestone, brecciated limestone and high amount of clay, in a limestone quarry environment. The different amount of clay of the three lithologies is key as they present

different color distribution and patterns. Images are automatically captured from the endoscope videos and their corresponding depth is recognized by pixel correlation of depth labels with numerals patterns. The grey pixel range and three no-reference image quality assessment metrics (PIQE, NIQE and BRISQUE) are used to discard unfavorable images. As a result, 7583 images are obtained from six blasting blocks and 79 boreholes. The Contrast-limited Adaptive Histogram Equalization (CLAHE) technique is adopted to improve image quality. 15 inputs are extracted involving ten RGB distribution properties and five texture parameters. A Pearson Correlation and the distributions of the inputs for different classes showed that there is some relationship between these inputs and the lithologies.

Six classification models are established and optimized by Bayesian optimization. Optimal hyper-parameters are used for the development of the classification models. With the original image inputs, the best overall classification performance is achieved by Bagged Tree-Bayesian optimization, with an accuracy of 85.63% and a precision for massive limestone, brecciated limestone and high amount of clay of 87.61%, 82.19% and 73.91%, respectively. CLAHE-handled input images may improve the overall performance with a Light Gradient Boosting Machine (LGBM) technique, the accuracy reaching 88.04% and the precision 90.72%, 83.52% and 85.29% for the same classes above, respectively. The color counting is identified, according to the feature importance ranking, as the most important factor for the classification model. This suggests that the process could be successfully used for other cases where the lithologies involved present different color or texture properties. The performance of 30 repeated random divisions of training and testing sets indicates that the LGBM with Bayesian optimization and CLAHE-strengthened images is stable.

The use of image properties from endoscope footages and classification models offers a fast and cheap approach to the identification of the three lithologies sought with a high recognition rate for the new images. The equipment required is inexpensive and can be used in a harsh mining environment. Data processing, once the method has been optimized, is fast, allowing for a readily available information on the quality of rock to be mined.

Chapter 7. General conclusions and future work

7.1 General conclusions

This thesis provides an approach to predict different aspects of mining, like rock fragmentation from blasting, gas relative permeability in reservoirs, ore/waste discrimination in a fluorite mine and lithology recognition in a limestone deposit using machine learning techniques. Some of these topics can be tackled with analytical methods, as occurs in rock fragmentation from blasting and gas relative permeability in reservoirs. In other cases, the relations between influential parameters and predicted targets are so complex that they cannot be integrated in specific, analytical functions. Machine learning constitutes an excellent solution in these cases.

The thesis comprises some published data and new measurements with state-of-the-art systems. A total of twelve supervised or unsupervised learning techniques and eleven optimization algorithms have been employed and their performance has been compared. No overall conclusion about a particular method that provides a good performance in all the cases can be drawn, but results obtained suggest that different methods can be a good choice for similar problems. In this regard, this thesis provides a working procedure valid for different sets of data to predict relevant parameters of the mining operations like the median size of the run-of-mine or the ore grade.

The main achievements are:

- Application of v-support vector regression to predict the median size of fragments in blasting considering a dataset from 76 blasts. This technique, rarely used in mining issues, explains nearly 99% of the variability in fragmentation. Such prediction performance is better than that obtained with traditional data analysis-based methods. The cost, however, is that the effect of the blasting parameters in new predictions is unknown.
- Development of a prediction model of the gas relative permeability in reservoirs by optimized kernel extreme learning machine models. The data (1024 measurements from published works) is from several different reservoirs and thus the developed models have good prediction potential for other similar gas relative permeability situations.

- Two image analysis-based methods are proposed to predict the fluorite grade. At first, the inputs are extracted from the televiewer images. A desirable differentiation can be achieved by support vector machine, i.e., waste, low grade and medium grade. In the next step, a faster and lower-cost scenario is applied, i.e., the procurement of fluorite images from pellets by a smartphone. Spectral clustering establishes three clusters and achieve good discrimination of waste from ore.
- Based on the above outcomes, an endoscope is applied to collect lithology images from drilling boreholes. The Contrast-limited adaptive histogram equalization is employed to enhance the feature performance to distinguish three lithologies. As a result, the Light Gradient Boosting Machine with Bayesian optimization can achieve the best overall prediction performance.

The novelty and relevance of this thesis is supported by three papers published in high impact journals and two more under review at the moment when this thesis is submitted.

7.2 Practical implications

This dissertation is not only a combination of application cases of machine learning techniques in mining, but also has its practical implications for the achievement of sustainable mining:

- Using machine learning to predict blasting fragmentation helps optimize blast designs, improving operational efficiency and reducing equipment wear by ensuring the right fragment sizes. This leads to lower costs by minimizing explosive use and optimizing downstream processes like hauling and crushing.
- Accurate predictions of the gas relative permeability of reservoirs enable better control of gas migration and leakage in storage sites, reducing the risk of emissions into the atmosphere. This contributes to mitigating environmental pollution, particularly greenhouse gas emissions, and ensures the safe management of underground gas storage. By optimizing the design and monitoring of gas reservoirs, it helps in reducing energy consumption and operational costs, while enhancing the sustainability and safety of mining operations.
- The combination of machine learning techniques and the image characteristics of the blasthole walls from the televiewer enables real-time, non-destructive analysis of ore quality, improving the efficiency and accuracy of ore classification. This approach reduces reliance on traditional, time-consuming sampling methods and enhances decision-making for mine planning and resource extraction. In addition, it optimizes resource utilization by identifying high-grade ores more precisely, reducing waste and lowering operational costs.
- Using smartphone photography of pellets instead of a televiewer is another alternative to assess ore/waste and define ore/waste boundaries in the deposit and ore blending strategies. Smartphones are more accessible, portable, and cost-effective, allowing for quicker and more flexible data collection in the field. This makes it easier to capture images in various conditions without the need for specialized equipment. Smartphones integrated with machine learning algorithms

for real-time analysis thus enable faster decision-making and improve operational efficiency.

- Using an endoscope, as an alternative to televiewer, to capture in-hole images from boreholes and applying machine learning to predict lithology is another method to investigate rock mass without the need for extensive drilling or sampling. It allows for real-time, in-situ lithology classification, improving the accuracy of geological models and reducing operational costs. By automating rock type identification, it enhances decision-making for drilling, blasting, and resource management, ultimately leading to more efficient and sustainable mining operations. This technique can be also used to calibrate models based on Measurement-While-Drilling signals.
- Industry advancements: the methodologies and models presented in this thesis have the potential to develop some user-friendly interfaces or software to optimize the mining operations.

7.3 Future directions

Despite this doctoral dissertation has introduced models and methodologies to improve mining operations, how to apply these techniques in reality is still worthwhile to be researched. The following ideas can be investigated:

- Database development: although the proposed prediction models can achieve the desirable performance, the limited number of current databases may obscure the application of models in reality. Future development directions include leveraging data augmentation techniques, transfer learning, and synthetic data generation to enhance the model's robustness. Collaborations to share domain-specific datasets may also offer solutions by allowing models to be trained on distributed data. These approaches will help overcome the challenges of limited data and improve prediction accuracy.
- Model optimization: exploring a broader range of algorithms and model architectures, such as ensemble learning and hybrid models that combine the strengths of multiple techniques. Additionally, advancements in deep learning, and reinforcement learning can open up new possibilities for more complex and accurate predictions.
- The ore grade discrimination models have been proved to be effective for fluorite-based minerals. Future development directions should focus on integrating advanced computer vision techniques and deep learning models to enhance image analysis accuracy. Implementing multi-modal data fusion, which combines information from different imaging sources and geological data, can improve prediction robustness. Collaborating with domain experts to refine feature extraction and continuously updating models with new data will further optimize the prediction process and ensure adaptability to varying geological conditions.
- Building on the effectiveness of using endoscope images for lithology prediction, future research should integrate multi-sensor data, such as incorporating geophysical measurements and core sample data. This can provide a more comprehensive understanding of subsurface geology.

- Finally, implementing real-time analysis capabilities and developing user-friendly interfaces will also enhance the accessibility and practicality of these predictive tools in mining and geological exploration.

REFERENCES

Abad ARB, Ghorbani H, Mohamadian N, Davoodi S, Mehrad M, Aghdam SK, et al. Robust hybrid machine learning algorithms for gas flow rates prediction through wellhead chokes in gas condensate fields. Fuel 2022;308:121872. https://doi.org/10.1016/j.fuel.2021.121872.

Adithya VSP, Chidambaram S, Prasanna MV, Venkatramanan S, Tirumalesh K, Thivya C, et al. Health Risk Implication and Spatial Distribution of Radon in Groundwater Along the Lithological Contact in South India. Arch Environ Contam Toxicol 2021;80:308–18. https://doi.org/10.1007/s00244-020-00798-9.

Ahmadi MA. Connectionist approach estimates gas–oil relative permeability in petroleum reservoirs: Application to reservoir simulation. Fuel 2015;140:429–39. https://doi.org/10.1016/j.fuel.2014.09.058.

Ahmadi M-A, Ahmadi MR, Hosseini SM, Ebadi M. Connectionist model predicts the porosity and permeability of petroleum reservoirs by means of petro-physical logs: Application of artificial intelligence. Journal of Petroleum Science and Engineering 2014;123:183–200. https://doi.org/10.1016/j.petrol.2014.08.026.

Ahmadi MA, Chen Z. Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petro-physical logs. Petroleum 2019;5:271–84. https://doi.org/10.1016/j.petlm.2018.06.002.

Ahmed U, Crary SF, Coates GR. Permeability Estimation: The Various Sources and Their Interrelationships. Journal of Petroleum Technology 1991;43:578–87. https://doi.org/10.2118/19604-PA.

Aigbedion I. A case study of permeability modeling and reservoir performance in the absence of core data in the Niger Delta, Nigeria. Journal of Applied Sciences 2007;7:772–6.

A.J. J. Error_ellipse 2023. https://www.mathworks.com/matlabcentral/fileexchange/4705-error_ellipse.

Ali Akbar D. Reserve estimation of central part of Choghart north anomaly iron ore deposit through ordinary kriging method. International Journal of Mining Science and Technology 2012;22:573–7. https://doi.org/10.1016/j.ijmst.2012.01.022.

Alonso Zarza AM, Calvo JP, García Del Cura MA. Palaeogeomorphological Controls on the Distribution and Sedimentary Styles of Alluvial Systems, Neogene of the NE of the

Madrid Basin (Central Spain). In: Marzo M, Puigdefábregas C, editors. Alluvial Sedimentation. 1st ed., Wiley; 1993, p. 277–92. https://doi.org/10.1002/9781444303995.ch19.

Al-Shayea NA. The combined effect of clay and moisture content on the behavior of remolded unsaturated soils. Engineering Geology 2001;62:319–42. https://doi.org/10.1016/S0013-7952(01)00032-1.

Alzubaidi F, Mostaghimi P, Swietojanski P, Clark SR, Armstrong RT. Automated lithology classification from drill core images using convolutional neural networks. Journal of Petroleum Science and Engineering 2021;197:107933. https://doi.org/10.1016/j.petrol.2020.107933.

Amor C, Navarro R. Minería del flúor en Sierra de Lújar. Rocas y minerales: Técnicas y procesos de minas y canteras 2016;11:46–58.

Anifowose FA, Labadin J, Abdulraheem A. Ensemble machine learning: An untapped modeling paradigm for petroleum reservoir characterization. Journal of Petroleum Science and Engineering 2017;151:480–7. https://doi.org/10.1016/j.petrol.2017.01.024.

Aouat S, Ait-hammi I, Hamouchene I. A new approach for texture segmentation based on the Gray Level Co-occurrence Matrix. Multimedia Tools and Applications 2021;80:24027–52. https://doi.org/10.1007/s11042-021-10634-4.

Armaghani DJ, Hajihassani M, Mohamad ET, Marto A, Noorani SA. Blasting-induced flyrock and ground vibration prediction through an expert artificial neural network based on particle swarm optimization. Arabian Journal of Geosciences 2014;7:5383–96. https://doi.org/10.1007/s12517-013-1174-0.

Arora S, Singh S. Butterfly optimization algorithm: a novel approach for global optimization. Soft Computing 2019;23:715–34. https://doi.org/10.1007/s00500-018-3102-4.

Asl PF, Monjezi M, Hamidi JK, Armaghani DJ. Optimization of flyrock and rock fragmentation in the Tajareh limestone mine using metaheuristics method of firefly algorithm. Engineering with Computers 2018;34:241–51. https://doi.org/10.1007/s00366-017-0535-9.

Auger-Méthé M, Clair CCS, Lewis MA, Derocher AE. Sampling rate and misidentification of Lévy and non-Lévy movement paths: comment. Ecology 2011;92:1699–701. https://doi.org/10.1890/10-1704.1.

Bahrami A, Monjezi M, Goshtasbi K, Ghazvinian A. Prediction of rock fragmentation due to blasting using artificial neural network. Engineering with Computers 2011;27:177–81. https://doi.org/10.1007/s00366-010-0187-5.

Baykan NA, Yılmaz N. Mineral identification using color spaces and artificial neural networks. Computers & Geosciences 2010;36:91–7. https://doi.org/10.1016/j.cageo.2009.04.009.

Berrezueta E, Ordóñez-Casado B, Bonilla W, Banda R, Castroviejo R, Carrión P, et al. Ore Petrography Using Optical Image Analysis: Application to Zaruma-Portovelo Deposit (Ecuador). Geosciences 2016;6:30. https://doi.org/10.3390/geosciences6020030.

Boesch DF, Rabalais NN, editors. Long-term Environmental Effects of Offshore Oil and Gas Development. CRC Press; 1987. https://doi.org/10.4324/9780203497777.

Bosch M, Zamora M, Utama W. Lithology discrimination from physical rock properties. GEOPHYSICS 2002;67:573–81. https://doi.org/10.1190/1.1468618.

Breiman L. Random Forests. Machine Learning 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

Bro R, Smilde AK. Principal component analysis. Analytical Methods 2014;6:2812–31. https://doi.org/10.1039/C3AY41907J.

Brooks R, Corey A. Hydraulic properties of porous media. Hydrology Papers, Colorado State University 1964.

Campos-Taberner M, García-Haro FJ, Camps-Valls G, Grau-Muedra G, Nutini F, Crema A, et al. Multitemporal and multiresolution leaf area index retrieval for operational local rice crop monitoring. Remote Sensing of Environment 2016;187:102–18. https://doi.org/10.1016/j.rse.2016.10.009.

Chang C-C, Lin C-J. Training *v* -Support Vector Regression: Theory and Algorithms. Neural Computation 2002;14:1959–77. https://doi.org/10.1162/089976602760128081.

Chatterjee S. Vision-based rock-type classification of limestone using multi-class support vector machine. Applied Intelligence 2013;39:14–27. https://doi.org/10.1007/s10489-012-0391-7.

Chatterjee S, Bandopadhyay S, Machuca D. Ore Grade Prediction Using a Genetic Algorithm and Clustering Based Ensemble Neural Network Model. Mathematical Geosciences 2010a;42:309–26. https://doi.org/10.1007/s11004-010-9264-y.

Chatterjee S, Bhattacherjee A. Genetic algorithms for feature selection of image analysis-based quality monitoring model: An application to an iron mine. Engineering Applications of Artificial Intelligence 2011;24:786–95. https://doi.org/10.1016/j.engappai.2010.11.009.

Chatterjee S, Bhattacherjee A, Samanta B, Pal SK. Image-based quality monitoring system of limestone ore grades. Computers in Industry 2010b;61:391–408. https://doi.org/10.1016/j.compind.2009.10.003.

Chen Y, Chen J, Wang P, Zhou M, Yang H, Li J. Design method of blasthole charge structure based on lithology distribution. Scientific Reports 2021;11:24247. https://doi.org/10.1038/s41598-021-03758-y.

Cherkassky V, Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks 2004;17:113–26. https://doi.org/10.1016/S0893-6080(03)00169-2.

Chopra N, Mohsin Ansari M. Golden jackal optimization: A novel nature-inspired optimizer for engineering applications. Expert Systems with Applications 2022;198:116924. https://doi.org/10.1016/j.eswa.2022.116924.

Chung SH, Katsabanis PD. Fragmentation prediction using improved engineering formulae. Fragblast 2000;4:198–207. https://doi.org/10.1076/frag.4.3.198.7392.

Coates GR, Dumanoir JL. A new approach to improved log-derived permeability. SPWLA 14th Annual Logging Symposium 1973, 1973.

Corina AN, Hovda S. Automatic lithology prediction from well logging using kernel density estimation. Journal of Petroleum Science and Engineering 2018;170:664–74. https://doi.org/10.1016/j.petrol.2018.06.012.

Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20:273–97. https://doi.org/10.1007/BF00994018.

Cover TM, Thomas JA. Elements of Information Theory. Wiley; 2005. https://doi.org/10.1002/047174882X.

Cunningham C. Fragmentation Estimations and the Kuz-Ram Model—Four Years on, Keystone, Colorado, USA: 1987, p. 475–87.

Danielsson P-E. Euclidean distance mapping. Computer Graphics and Image Processing 1980;14:227–48. https://doi.org/10.1016/0146-664X(80)90054-4.

Day WHE, Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods. Journal of Classification 1984;1:7–24. https://doi.org/10.1007/BF01890115.

De Vicente G, Muñoz-Martín A. The Madrid Basin and the Central System: A tectonostratigraphic analysis from 2D seismic lines. Tectonophysics 2013;602:259–85. https://doi.org/10.1016/j.tecto.2012.04.003.

Desta FS, Buxton MWN. The use of RGB Imaging and FTIR Sensors for mineral mapping in the Reiche Zeche underground test mine, Freiberg. Proceedings of Real Time Mining, Amsterdam, The Netherlands: 2017.

Dimitraki L, Christaras B, Marinos V, Vlahavas I, Arampelos N. Predicting the average size of blasted rocks in aggregate quarries using artificial neural networks. Bulletin of Engineering Geology and the Environment 2019;78:2717–29. https://doi.org/10.1007/s10064-018-1270-1.

Donskoi E, Manuel JR, Austin P, Poliakov A, Peterson MJ, Hapugoda S. Comparative study of iron ore characterisation using a scanning electron microscope and optical image analysis. Applied Earth Science 2013;122:217–29. https://doi.org/10.1179/1743275814Y.0000000042.

Donskoi E, Poliakov A, Manuel JR, Peterson M, Hapugoda S. Novel developments in optical image analysis for iron ore, sinter and coke characterisation. Applied Earth Science 2015;124:227–44. https://doi.org/10.1179/1743275815Y.0000000013.

Donskoi E, Suthers SP, Fradd SB, Young JM, Campbell JJ, Raynlyn TD, et al. Utilization of optical image analysis and automatic texture classification for iron ore particle characterisation. Minerals Engineering 2007;20:461–71. https://doi.org/10.1016/j.mineng.2006.12.005.

Dumakor-Dupey NK, Arya S. Machine Learning—A Review of Applications in Mineral Resource Estimation. Energies 2021;14:4079. https://doi.org/10.3390/en14144079.

Ebrahimi E, Monjezi M, Khalesi MR, Armaghani DJ. Prediction and optimization of back-break and rock fragmentation using an artificial neural network and a bee colony algorithm. Bulletin of Engineering Geology and the Environment 2016;75:27–36. https://doi.org/10.1007/s10064-015-0720-2.

Elkatatny S, Mahmoud M, Tariq Z, Abdulraheem A. New insights into the prediction of heterogeneous carbonate reservoir permeability from well logs using artificial

intelligence network. Neural Computing and Applications 2018;30:2673–83. https://doi.org/10.1007/s00521-017-2850-x.

Emery X. Two Ordinary Kriging Approaches to Predicting Block Grade Distributions. Mathematical Geology 2006;38:801–19. https://doi.org/10.1007/s11004-006-9048-6.

Enayatollahi I, Aghajani Bazzazi A, Asadi A. Comparison Between Neural Networks and Multiple Regression Analysis to Predict Rock Fragmentation in Open-Pit Mines. Rock Mechanics and Rock Engineering 2014;47:799–807. https://doi.org/10.1007/s00603-013-0415-6.

Ergün Hatir M, İnce İ. Lithology mapping of stone heritage via state-of-the-art computer vision. Journal of Building Engineering 2021;34:101921. https://doi.org/10.1016/j.jobe.2020.101921.

Erofeev A, Orlov D, Ryzhov A, Koroteev D. Prediction of Porosity and Permeability Alteration Based on Machine Learning Algorithms. Transport in Porous Media 2019;128:677–700. https://doi.org/10.1007/s11242-019-01265-3.

Esen S, Onederra I, Bilgin HA. Modelling the size of the crushed zone around a blasthole. International Journal of Rock Mechanics and Mining Sciences 2003;40:485–95. https://doi.org/10.1016/S1365-1609(03)00018-2.

Esmaeili M, Osanloo M, Rashidinejad F, Aghajani Bazzazi A, Taji M. Multiple regression, ANN and ANFIS models for prediction of backbreak in the open pit blasting. Engineering with Computers 2014;30:549–58. https://doi.org/10.1007/s00366-012-0298-2.

Esmaeili M, Salimi A, Drebenstedt C, Abbaszadeh M, Aghajani Bazzazi A. Application of PCA, SVR, and ANFIS for modeling of rock fragmentation. Arabian Journal of Geosciences 2015;8:6881–93. https://doi.org/10.1007/s12517-014-1677-3.

Esmaeili S, Sarma H, Harding T, Maini B. Review of the effect of temperature on oil-water relative permeability in porous rocks of oil reservoirs. Fuel 2019;237:91–116. https://doi.org/10.1016/j.fuel.2018.09.100.

Fan J, Ma X, Wu L, Zhang F, Yu X, Zeng W. Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agricultural Water Management 2019;225:105758. https://doi.org/10.1016/j.agwat.2019.105758.

Fan Y, Li J, Guo Y, Xie L, Zhang G. Digital image colorimetry on smartphone for chemical analysis: A review. Measurement 2021;171:108829. https://doi.org/10.1016/j.measurement.2020.108829.

Fang Q, Nguyen H, Bui X-N, Nguyen-Thoi T, Zhou J. Modeling of rock fragmentation by firefly optimization algorithm and boosted generalized additive model. Neural Computing and Applications 2021;33:3503–19. https://doi.org/10.1007/s00521-020-05197-8.

Faria EL, Coelho JulianaM, Matos TF, Santos BCC, Trevizan WA, Gonzalez JL, et al. Lithology identification in carbonate thin section images of the Brazilian pre-salt reservoirs by the computational vision and deep learning. Computational Geosciences 2022;26:1537–47. https://doi.org/10.1007/s10596-022-10168-0.

Farsi M, Mohamadian N, Ghorbani H, Wood DA, Davoodi S, Moghadasi J, et al. Predicting Formation Pore-Pressure from Well-Log Data with Hybrid Machine-Learning

Optimization Algorithms. Natural Resources Research 2021;30:3455–81. https://doi.org/10.1007/s11053-021-09852-2.

Fernández A, Sanchidrián JA, Segarra P, Gómez S, Li E, Navarro R. Rock mass structural recognition from drill monitoring technology in underground mining using discontinuity index and machine learning techniques. International Journal of Mining Science and Technology 2023;33:555–71. https://doi.org/10.1016/j.ijmst.2023.02.004.

Fu D, Su C, Wang W, Yuan R. Deep learning based lithology classification of drill core images. PLoS ONE 2022;17:e0270826. https://doi.org/10.1371/journal.pone.0270826.

Fushiki T. Estimation of prediction error by using K-fold cross-validation. Statistics and Computing 2011;21:137–46. https://doi.org/10.1007/s11222-009-9153-8.

Galdames FJ, Perez CA, Estévez PA, Adams M. Rock lithological classification by hyperspectral, range 3D and color images. Chemometrics and Intelligent Laboratory Systems 2019;189:138–48. https://doi.org/10.1016/j.chemolab.2019.04.006.

Galdames FJ, Perez CA, Estévez PA, Adams M. Classification of rock lithology by laser range 3D and color images. International Journal of Mineral Processing 2017;160:47–57. https://doi.org/10.1016/j.minpro.2017.01.008.

Gao W, Karbasi M, Hasanipanah M, Zhang X, Guo J. Developing GPR model for forecasting the rock fragmentation in surface mines. Engineering with Computers 2018;34:339–45. https://doi.org/10.1007/s00366-017-0544-8.

Ghaeini N, Mousakhani M, Amnieh HB, Jafari A. Prediction of blasting-induced fragmentation in Meydook copper mine using empirical, statistical, and mutual information models. Arabian Journal of Geosciences 2017;10:409. https://doi.org/10.1007/s12517-017-3189-4.

Ghassemi A. A Review of Some Rock Mechanics Issues in Geothermal Reservoir Development. Geotechnical and Geological Engineering 2012;30:647–64. https://doi.org/10.1007/s10706-012-9508-3.

Gheibie S, Aghababaei H, Hoseinie SH, Pourrahimian Y. Modified Kuz—Ram fragmentation model and its use at the Sungun Copper Mine. International Journal of Rock Mechanics and Mining Sciences 2009;46:967–73. https://doi.org/10.1016/j.ijrmms.2009.05.003.

Gholami R, Shahraki AR, Jamali Paghaleh M. Prediction of Hydrocarbon Reservoirs Permeability Using Support Vector Machine. Mathematical Problems in Engineering 2012;2012:1–18. https://doi.org/10.1155/2012/670723.

Gordan B, Jahed Armaghani D, Hajihassani M, Monjezi M. Prediction of seismic slope stability through combination of particle swarm optimization and neural network. Engineering with Computers 2016;32:85–97. https://doi.org/10.1007/s00366-015-0400-7.

Guzzetti F, Cardinali M, Reichenbach P. The Influence of Structural Setting and Lithology on Landslide Type and Pattern. Environmental & Engineering Geoscience 1996;II:531–55. https://doi.org/10.2113/gseegeosci.II.4.531.

Hajihassani M, Jahed Armaghani D, Marto A, Tonnizam Mohamad E. Ground vibration prediction in quarry blasting through an artificial neural network optimized by imperialist

competitive algorithm. Bulletin of Engineering Geology and the Environment 2015;74:873–86. https://doi.org/10.1007/s10064-014-0657-x.

Haralick RM, Shanmugam KS. Combined spectral and spatial processing of ERTS imagery data. Remote Sensing of Environment 1974;3:3–13. https://doi.org/10.1016/0034-4257(74)90033-9.

Hasanipanah M, Amnieh HB, Arab H, Zamzam MS. Feasibility of PSO–ANFIS model to estimate rock fragmentation produced by mine blasting. Neural Computing and Applications 2018;30:1015–24. https://doi.org/10.1007/s00521-016-2746-1.

Heidari AA, Mirjalili S, Faris H, Aljarah I, Mafarja M, Chen H. Harris hawks optimization: Algorithm and applications. Future Generation Computer Systems 2019;97:849–72. https://doi.org/10.1016/j.future.2019.02.028.

Hekmatnejad A, Emery X, Alipour-Shahsavari M. Comparing linear and non-linear kriging for grade prediction and ore/waste classification in mineral deposits. International Journal of Mining, Reclamation and Environment 2019;33:247–64. https://doi.org/10.1080/17480930.2017.1386430.

Hohashi R, Sekine I, Kobayashi Y, Ishigaki K. Visualization of Ground Ahead of Tunnel Face by Industrial Endoscope. 5th ISRM Young Scholars' Symposium on Rock Mechanics and International Symposium on Rock Engineering for Innovative Future Okinawa, Japan: 2019, p. 049.

Holger H. violin.m - Simple violin plot using matlab default kernel density estimation 2015. https://www.mathworks.com/matlabcentral/fileexchange/45134-violin-plot.

Holland JH. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. The MIT Press; 1992. https://doi.org/10.7551/mitpress/1090.001.0001.

Honarpour M, Koederitz L, Harvey AH. Relative Permeability of Petroleum Reservoirs. CRC Press; 2018. https://doi.org/10.1201/9781351076326.

Honarpour M, Mahmood SM. Relative-Permeability Measurements: An Overview. Journal of Petroleum Technology 1988;40:963–6. https://doi.org/10.2118/18565-PA.

Hossain TM, Watada J, Aziz IA, Hermana M, Meraj ST, Sakai H. Lithology Prediction Using Well Logs: A Granular Computing Approach. International journal of innovative computing, information & control 2021. https://doi.org/10.24507/ijicic.17.01.225.

Hu H, Lu W, Yan P, Chen M, Gao Q, Yang Z. A new horizontal rock dam foundation blasting technique with a shock-reflection device arranged at the bottom of vertical borehole. European Journal of Environmental and Civil Engineering 2020;24:481–99. https://doi.org/10.1080/19648189.2017.1399168.

Hu X, Huang S. Physical Properties of Reservoir Rocks. In: Hu X, Hu S, Jin F, Huang S, editors. Physics of Petroleum Reservoirs, Berlin, Heidelberg: Springer Berlin Heidelberg; 2017, p. 7–164. https://doi.org/10.1007/978-3-662-53284-3_2.

Huang G Bin, Zhu QY, Siew CK. Extreme learning machine: Theory and applications. Neurocomputing 2006;70:489–501. https://doi.org/10.1016/j.neucom.2005.12.126.

Huang G, Zhou H, Ding X, Zhang R. Extreme Learning Machine for Regression and Multiclass Classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 2012;42:513–29. https://doi.org/10.1109/TSMCB.2011.2168604.

Huang J, Asteris PG, Manafi Khajeh Pasha S, Mohammed AS, Hasanipanah M. A new auto-tuning model for predicting the rock fragmentation: a cat swarm optimization algorithm. Engineering with Computers 2022;38:2209–20. https://doi.org/10.1007/s00366-020-01207-4.

Ilin A, Velasco F, Navarro R, Tornos F. New data on Alpine type fluorite deposits: Case of Lújar mine in Betic Cordillera (SE Spain). MACLA: Revista española de la Sociedad de Mineralogía 2019;24:63–4.

Iserhien-Emekeme R, Ofomola M, Bawallah M, Anomohanran O. Lithological Identification and Underground Water Conditions in Jeddo Using Geophysical and Geochemical Methods. Hydrology 2017;4:42. https://doi.org/10.3390/hydrology4030042.

Jack Feng C-X, Yu Z-GS, Kingi U, Pervaiz Baig M. Threefold vs. fivefold cross validation in one-hidden-layer and two-hidden-layer predictive neural network modeling of machining surface roughness data. Journal of Manufacturing Systems 2005;24:93–107. https://doi.org/10.1016/S0278-6125(05)80010-X.

Jafrasteh B, Fathianpour N. A hybrid simultaneous perturbation artificial bee colony and back-propagation algorithm for training a local linear radial basis neural network on ore grade estimation. Neurocomputing 2017;235:217–27. https://doi.org/10.1016/j.neucom.2017.01.016.

Jooshaki M, Nad A, Michaux S. A Systematic Review on the Application of Machine Learning in Exploiting Mineralogical Data in Mining and Mineral Industry. Minerals 2021;11:816. https://doi.org/10.3390/min11080816.

Jorgensen DG. Estimating Geohydrologic Properties from Borehole-Geophysical Logs. Groundwater Monitoring & Remediation 1991;11:123–9. https://doi.org/10.1111/j.1745-6592.1991.tb00388.x.

Jug J, Strelec S, Gazdek M, Kavur B. Fragment Size Distribution of Blasted Rock Mass. IOP Conference Series: Earth and Environmental Science, 2017;95:042013. https://doi.org/10.1088/1755-1315/95/4/042013.

Kaplan UE, Dagasan Y, Topal E. Mineral grade estimation using gradient boosting regression trees. International Journal of Mining, Reclamation and Environment 2021;35:728–42. https://doi.org/10.1080/17480930.2021.1949863.

Kaplan UE, Topal E. A New Ore Grade Estimation Using Combine Machine Learning Algorithms. Minerals 2020;10:847. https://doi.org/10.3390/min10100847.

Kaur S, Awasthi LK, Sangal AL, Dhiman G. Tunicate Swarm Algorithm: A new bio-inspired based metaheuristic paradigm for global optimization. Engineering Applications of Artificial Intelligence 2020;90:103541. https://doi.org/10.1016/j.engappai.2020.103541.

Kennedy J, Eberhart R. Particle swarm optimization. Proceedings of ICNN'95 - International Conference on Neural Networks, vol. 4, Perth, WA, Australia: IEEE; 1995, p. 1942–8. https://doi.org/10.1109/ICNN.1995.488968.

Khandelwal M, Monjezi M. Prediction of Backbreak in Open-Pit Blasting Operations Using the Machine Learning Method. Rock Mechanics and Rock Engineering 2013;46:389–96. https://doi.org/10.1007/s00603-012-0269-3.

Konaté AA, Ma H, Pan H, Qin Z, Ahmed HA, Dembele NDJ. Lithology and mineralogy recognition from geochemical logging tool data using multivariate statistical analysis. Applied Radiation and Isotopes 2017;128:55–67. https://doi.org/10.1016/j.apradiso.2017.06.041.

Koopialipoor M, Jahed Armaghani D, Hedayat A, Marto A, Gordan B. Applying various hybrid intelligent systems to evaluate and predict slope stability under static and dynamic conditions. Soft Computing 2019;23:5913–29. https://doi.org/10.1007/s00500-018-3253-3.

Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Physical Review E 2004;69:066138. https://doi.org/10.1103/PhysRevE.69.066138.

Kulatilake PHSW, Hudaverdi T, Wu Q. New Prediction Models for Mean Particle Size in Rock Blast Fragmentation. Geotechnical and Geological Engineering 2012;30:665–84. https://doi.org/10.1007/s10706-012-9496-3.

Kulatilake PHSW, Qiong W, Hudaverdi T, Kuzu C. Mean particle size prediction in rock blast fragmentation using neural networks. Engineering Geology 2010;114:298–311. https://doi.org/10.1016/j.enggeo.2010.05.008.

Kumar S, Gautam S, Varun R, Khan MA. A Robust Machine Learning Model to Predict the Relative Permeability of an Oil Porous Medium at Elevated Temperatures. Day 3 Thu, April 28, 2022, SPE; 2022. https://doi.org/10.2118/209313-MS.

Kumar Sharma S, Rai P. Establishment of blasting design parameters influencing mean fragment size using state-of-art statistical tools and techniques. Measurement 2017;96:34–51. https://doi.org/10.1016/j.measurement.2016.10.047.

Kumar T, Seelam NK, Rao GS. Lithology prediction from well log data using machine learning techniques: A case study from Talcher coalfield, Eastern India. Journal of Applied Geophysics 2022;199:104605. https://doi.org/10.1016/j.jappgeo.2022.104605.

Lane GR, Martin C, Pirard E. Techniques and applications for predictive metallurgy and ore characterization using optical image analysis. Minerals Engineering 2008;21:568–77. https://doi.org/10.1016/j.mineng.2007.11.009.

Li E, Segarra P, Sanchidrián JA, Gómez S, Fernández A, Navarro R, et al. Application of percentile color intensities of borehole images for automatic fluorite grade assessment. Ore Geology Reviews 2023a;163:105790. https://doi.org/10.1016/j.oregeorev.2023.105790.

Li E, Yang F, Ren M, Zhang X, Zhou J, Khandelwal M. Prediction of blasting mean fragment size using support vector regression combined with five optimization algorithms. Journal of Rock Mechanics and Geotechnical Engineering 2021a;13:1380–97. https://doi.org/10.1016/j.jrmge.2021.07.013.

Li E, Zhang N, Xi B, Zhou J, Gao X. Compressive strength prediction and optimization design of sustainable concrete based on squirrel search algorithm-extreme gradient boosting technique. Frontiers of Structural and Civil Engineering 2023b;17:1310–25. https://doi.org/10.1007/s11709-023-0997-3.

Li E, Zhou J, Shi X, Jahed Armaghani D, Yu Z, Chen X, et al. Developing a hybrid model of salp swarm algorithm-based support vector machine to predict the strength of fiber-reinforced cemented paste backfill. Engineering with Computers 2021b;37:3519–40. https://doi.org/10.1007/s00366-020-01014-x.

Li H, Wang H. A bi-level training approach based on extreme learning machine autoencoder for data classification. Proceedings - 2022 18th International Conference on Computational Intelligence and Security, CIS 2022 2022:171–5. https://doi.org/10.1109/CIS58238.2022.00043.

Li J, Huang Zhe, Li G, Huang Zhongwei, Dai J, Cheng K. Field test of radial jet drilling technology in a surface formation. Journal of Petroleum Science and Engineering 2022;218:110928. https://doi.org/10.1016/j.petrol.2022.110928.

Li N, Hao H, Gu Q, Wang D, Hu X. A transfer learning method for automatic identification of sandstone microscopic images. Computers & Geosciences 2017;103:111–21. https://doi.org/10.1016/j.cageo.2017.03.007.

LI W, MU L, ZHAO L, LI J, WANG S, FAN Z, et al. Pore-throat structure characteristics and its impact on the porosity and permeability relationship of Carboniferous carbonate reservoirs in eastern edge of Pre-Caspian Basin. Petroleum Exploration and Development 2020;47:1027–41. https://doi.org/10.1016/S1876-3804(20)60114-8.

Liao K, Wu Y, Miao F, Li L, Xue Y. Using a kernel extreme learning machine with grey wolf optimization to predict the displacement of step-like landslide. Bulletin of Engineering Geology and the Environment 2020;79:673–85. https://doi.org/10.1007/s10064-019-01598-9.

Liu C, Li M, Zhang Y, Han S, Zhu Y. An Enhanced Rock Mineral Recognition Method Integrating a Deep Learning Model and Clustering Algorithm. Minerals 2019;9:516. https://doi.org/10.3390/min9090516.

Liu S, Wang L, Zhang W, Sun W, Fu J, Xiao T, et al. A physics-informed data-driven model for landslide susceptibility assessment in the Three Gorges Reservoir area. Geoscience Frontiers 2023;14:101621. https://doi.org/10.1016/j.gsf.2023.101621.

Liu Y, Wang X, Zhang Z, Deng F. OreFormer: Ore Sorting Transformer Based on ConvNet and Visual Attention. Natural Resources Research 2024;33:521–38. https://doi.org/10.1007/s11053-023-10298-x.

Liu Y, Zhang Z, Liu X, Wang L, Xia X. Deep learning-based image classification for online multi-coal and multi-class sorting. Computers & Geosciences 2021a;157:104922. https://doi.org/10.1016/j.cageo.2021.104922.

Liu Y, Zhang Z, Liu X, Wang L, Xia X. Performance evaluation of a deep learning based wet coal image classification. Minerals Engineering 2021b;171:107126. https://doi.org/10.1016/j.mineng.2021.107126.

Liu Z, Li L, Fang X, Qi W, Shen J, Zhou H, et al. Hard-rock tunnel lithology prediction with TBM construction big data using a global-attention-mechanism-based LSTM network. Automation in Construction 2021;125:103647. https://doi.org/10.1016/j.autcon.2021.103647.

Lloyd S. Least squares quantization in PCM. IEEE Transactions on Information Theory 1982;28:129–37. https://doi.org/10.1109/TIT.1982.1056489.

Mahdaviara M, Menad NA, Ghazanfari MH, Hemmati-Sarapardeh A. Modeling relative permeability of gas condensate reservoirs: Advanced computational frameworks. Journal of Petroleum Science and Engineering 2020;189:106929. https://doi.org/10.1016/j.petrol.2020.106929.

Mahmoudabadi H, Izadi M, Menhaj MB. A hybrid method for grade estimation using genetic algorithm and neural networks. Computational Geosciences 2009;13:91–101. https://doi.org/10.1007/s10596-008-9107-9.

Majcherczyk T, Malkowski P, Niedbalski Z. Describing Quality of Rocks Around Underground Headings: Endoscopic Observations of Fractures, ISRM International Symposium - EUROCK, Brno, Czech Republic: 2005, p. 058.

Malkowski P, Niedbalski Z, Majcherczyk T. Endoscopic method of rock mass quality evaluation-new experiences. The 42nd U.S. Rock Mechanics Symposium (USRMS), vol. 08, San Francisco, California: 2008, p. 237.

Marathe R, Turner ML, Fogden A. Pore-Scale Distribution of Crude Oil Wettability in Carbonate Rocks. Energy & Fuels 2012;26:6268–81. https://doi.org/10.1021/ef301088j.

Marcot BG, Hanea AM. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? Computational Statistics 2021;36:2009–31. https://doi.org/10.1007/s00180-020-00999-9.

Marschallinger R. Automatic mineral classification in the macroscopic scale. Computers & Geosciences 1997;23:119–26. https://doi.org/10.1016/S0098-3004(96)00074-X.

Matinkia M, Hashami R, Mehrad M, Hajsaeedi MR, Velayati A. Prediction of permeability from well logs using a new hybrid machine learning algorithm. Petroleum 2023;9:108–23. https://doi.org/10.1016/j.petlm.2022.03.003.

Merris R. Laplacian matrices of graphs: a survey. Linear Algebra and Its Applications 1994;197–198:143–76. https://doi.org/10.1016/0024-3795(94)90486-3.

Mery N, Marcotte D. Quantifying Mineral Resources and Their Uncertainty Using Two Existing Machine Learning Methods. Mathematical Geosciences 2022;54:363–87. https://doi.org/10.1007/s11004-021-09971-9.

Miller JN, Miller JC, Miller RD. Statistics and chemometrics for analytical chemistry. Seventh edition. Harlow, United Kingdom: Pearson Education Limited; 2018.

Mirjalili S, Gandomi AH, Mirjalili SZ, Saremi S, Faris H, Mirjalili SM. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. Advances in Engineering Software 2017;114:163–91. https://doi.org/10.1016/j.advengsoft.2017.07.002.

Mirjalili S, Mirjalili SM, Hatamlou A. Multi-Verse Optimizer: a nature-inspired algorithm for global optimization. Neural Computing and Applications 2016;27:495–513. https://doi.org/10.1007/s00521-015-1870-7.

Mirjalili S, Mirjalili SM, Lewis A. Grey Wolf Optimizer. Advances in Engineering Software 2014;69:46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007.

Mishra A, Sharma A, Patidar AK. Evaluation and Development of a Predictive Model for Geophysical Well Log Data Analysis and Reservoir Characterization: Machine Learning Applications to Lithology Prediction. Natural Resources Research 2022;31:3195–222. https://doi.org/10.1007/s11053-022-10121-z.

Mittal A, Moorthy AK, Bovik AC. No-Reference Image Quality Assessment in the Spatial Domain. IEEE Trans on Image Process 2012;21:4695–708. https://doi.org/10.1109/TIP.2012.2214050.

Mittal A, Soundararajan R, Bovik AC. Making a “Completely Blind” Image Quality Analyzer. IEEE Signal Processing Letters 2013;20:209–12. https://doi.org/10.1109/LSP.2012.2227726.

Moayedi H, Jahed Armaghani D. Optimizing an ANN model with ICA for estimating bearing capacity of driven pile in cohesionless soil. Engineering with Computers 2018;34:347–56. https://doi.org/10.1007/s00366-017-0545-7.

Mojtahedi SFF, Ebtehaj I, Hasanipanah M, Bonakdari H, Amnieh HB. Proposing a novel hybrid intelligent model for the simulation of particle size distribution resulting from blasting. Engineering with Computers 2019;35:47–56. https://doi.org/10.1007/s00366-018-0582-x.

Monjezi M, Amini Khoshalan H, Yazdian Varjani A. Prediction of flyrock and backbreak in open pit blasting operation: a neuro-genetic approach. Arabian Journal of Geosciences 2012;5:441–8. https://doi.org/10.1007/s12517-010-0185-3.

Monjezi M, Amiri H, Farrokhi A, Goshtasbi K. Prediction of Rock Fragmentation Due to Blasting in Sarcheshmeh Copper Mine Using Artificial Neural Networks. Geotechnical and Geological Engineering 2010a;28:423–30. https://doi.org/10.1007/s10706-010-9302-z.

Monjezi M, Bahrami A, Yazdian Varjani A. Simultaneous prediction of fragmentation and flyrock in blasting operation using artificial neural networks. International Journal of Rock Mechanics and Mining Sciences 2010b;47:476–80. https://doi.org/10.1016/j.ijrmms.2009.09.008.

Monjezi M, Mohamadi HA, Barati B, Khandelwal M. Application of soft computing in predicting rock fragmentation to reduce environmental blasting side effects. Arabian Journal of Geosciences 2014;7:505–11. https://doi.org/10.1007/s12517-012-0770-8.

Monjezi M, Rezaei M, Yazdian Varjani A. Prediction of rock fragmentation due to blasting in Gol-E-Gohar iron mine using fuzzy logic. International Journal of Rock Mechanics and Mining Sciences 2009;46:1273–80. https://doi.org/10.1016/j.ijrmms.2009.05.005.

Moon TK. The expectation-maximization algorithm. IEEE Signal Processing Magazine 1996;13:47–60. https://doi.org/10.1109/79.543975.

Moussa T, Elkatatny S, Mahmoud M, Abdulraheem A. Development of New Permeability Formulation From Well Log Data Using Artificial Intelligence Approaches. Journal of Energy Resources Technology 2018;140. https://doi.org/10.1115/1.4039270.

Mutlag WK, Ali SK, Aydam ZM, Taher BH. Feature Extraction Methods: A Review. Journal of Physics: Conference Series, 2020;1591:012028. https://doi.org/10.1088/1742-6596/1591/1/012028.

Nainggolan DR, Sitorus R, Eveny ON, Sariandi F. Correlation between uniaxial compressive strength (UCS) and blasting geometry on rock excavation at PT Agincourt Resources. IOP Conference Series: Earth and Environmental Science, 2018;212:012065. https://doi.org/10.1088/1755-1315/212/1/012065.

Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics 2013;7. https://doi.org/10.3389/fnbot.2013.00021.

Nazari H, Hajizadeh F. Estimation of permeability from a hydrocarbon reservoir located in southwestern Iran using well-logging data and a new intelligent combined method. Carbonates and Evaporites 2023;38:20. https://doi.org/10.1007/s13146-022-00840-y.

Nickolas LB, Segura C, Brooks JR. The influence of lithology on surface water sources. Hydrological Processes 2017;31:1913–25. https://doi.org/10.1002/hyp.11156.

Okada N, Maekawa Y, Owada N, Haga K, Shibayama A, Kawamura Y. Automated Identification of Mineral Types and Grain Size Using Hyperspectral Imaging and Deep Learning for Mineral Processing. Minerals 2020;10:809. https://doi.org/10.3390/min10090809.

Okon AN, Adewole SE, Uguma EM. Artificial neural network model for reservoir petrophysical properties: porosity, permeability and water saturation prediction. Modeling Earth Systems and Environment 2021;7:2373–90. https://doi.org/10.1007/s40808-020-01012-4.

Oliver MA, Webster R. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. CATENA 2014;113:56–69. https://doi.org/10.1016/j.catena.2013.09.006.

Ouchterlony F. The Case for the Median Fragment Size as a Better Fragment Size Descriptor than the Mean. Rock Mechanics and Rock Engineering 2016;49:143–64. https://doi.org/10.1007/s00603-015-0722-1.

Ouchterlony F, Sanchidrián JA. The Fragmentation-Energy Fan Concept and the Swebrec Function in Modeling Drop Weight Testing. Rock Mech Rock Eng 2018;51:3129–56. https://doi.org/10.1007/s00603-018-1458-5.

Pascale D. RGB coordinates of the Macbeth ColorChecker 2005.

Patel AK, Chatterjee S. Computer vision-based limestone rock-type classification using probabilistic neural network. Geoscience Frontiers 2016;7:53–60. https://doi.org/10.1016/j.gsf.2014.10.005.

Patel AK, Chatterjee S, Gorai AK. Development of a machine vision system using the support vector machine regression (SVR) algorithm for the online prediction of iron ore grades. Earth Science Informatics 2019;12:197–210. https://doi.org/10.1007/s12145-018-0370-6.

Peña A, Caja MA, Campos JR, Santos C, Pérez JL, Fernández PR, et al. Application of machine learning models in thin sections image of drill cuttings: lithology classification and quantification (Algeria tight reservoirs). EAGE/ALNAFT Geoscience Workshop, Algiers, Africa,: European Association of Geoscientists & Engineers; 2019, p. 1–5. https://doi.org/10.3997/2214-4609.2019X60047117.

Perez CA, Estévez PA, Vera PA, Castillo LE, Aravena CM, Schulz DA, et al. Ore grade estimation by feature selection and voting using boundary detection in digital image analysis. International Journal of Mineral Processing 2011;101:28–36. https://doi.org/10.1016/j.minpro.2011.07.008.

Phoon K-K, Zhang W. Future of machine learning in geotechnics. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 2023;17:7–22. https://doi.org/10.1080/17499518.2022.2087884.

Pizer SM, Amburn EP, Austin JD, Cromartie R, Geselowitz A, Greer T, et al. Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing 1987;39:355–68. https://doi.org/10.1016/S0734-189X(87)80186-X.

Poli R, Kennedy J, Blackwell T. Particle swarm optimization: An overview. Soft Computing 2007;1:33–57. https://doi.org/10.1007/s11721-007-0002-0.

Qi C, Tang X. A hybrid ensemble method for improved prediction of slope stability. International Journal for Numerical and Analytical Methods in Geomechanics 2018;42:1823–39. https://doi.org/10.1002/nag.2834.

Qiao J, Zeng J, Chen D, Cai J, Jiang S, Xiao E, et al. Permeability estimation of tight sandstone from pore structure characterization. Marine and Petroleum Geology 2022;135:105382. https://doi.org/10.1016/j.marpetgeo.2021.105382.

Qiu Y, Zhou J, Khandelwal M, Yang H, Yang P, Li C. Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. Engineering with Computers 2022;38:4145–62. https://doi.org/10.1007/s00366-021-01393-9.

Qiu Z, Dou D, Zhou D, Yang J. On-line prediction of clean coal ash content based on image analysis. Measurement 2021;173:108663. https://doi.org/10.1016/j.measurement.2020.108663.

Quan Q, Hao Z, Xifeng H, Jingchun L. Research on water temperature prediction based on improved support vector regression. Neural Computing and Applications 2022;34:8501–10. https://doi.org/10.1007/s00521-020-04836-4.

Ramil A, López AJ, Pozo-Antonio JS, Rivas T. A computer vision system for identification of granite-forming minerals based on RGB data and artificial neural networks. Measurement 2018;117:90–5. https://doi.org/10.1016/j.measurement.2017.12.006.

Regnet JB, David C, Robion P, Menéndez B. Microstructures and physical properties in carbonate rocks: A comprehensive review. Marine and Petroleum Geology 2019;103:366–76. https://doi.org/10.1016/j.marpetgeo.2019.02.022.

Reinsch T, Paap B, Hahn S, Wittig V, Van Den Berg S. Insights into the radial water jet drilling technology – Application in a quarry. Journal of Rock Mechanics and Geotechnical Engineering 2018;10:236–48. https://doi.org/10.1016/j.jrmge.2018.02.001.

Reza AM. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement. The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology 2004;38:35–44. https://doi.org/10.1023/B:VLSI.0000028532.53893.82.

Rodriguez JD, Perez A, Lozano JA. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2010;32:569–75. https://doi.org/10.1109/TPAMI.2009.187.

Roy S, Pantanowitz L, Amin M, Seethala RR, Ishtiaque A, Yousem SA, et al. Smartphone adapters for digital photomicrography. Journal of Pathology Informatics 2014;5:24. https://doi.org/10.4103/2153-3539.137728.

Samanta B. Radial Basis Function Network for Ore Grade Estimation. Natural Resources Research 2010;19:91–102. https://doi.org/10.1007/s11053-010-9115-z.

Samanta B, Bandopadhyay S. Construction of a radial basis function network using an evolutionary algorithm for grade estimation in a placer gold deposit. Computers & Geosciences 2009;35:1592–602. https://doi.org/10.1016/j.cageo.2009.01.006.

Samanta B, Bandopadhyay S, Ganguli R. Comparative Evaluation of Neural Network Learning Algorithms for Ore Grade Estimation. Mathematical Geology 2006;38:175–97. https://doi.org/10.1007/s11004-005-9010-z.

Sanchidrián JA, Ouchterlony F. Blast-Fragmentation Prediction Derived From the Fragment Size-Energy Fan Concept. Rock Mechanics and Rock Engineering 2023;56:8869–89. https://doi.org/10.1007/s00603-023-03496-9.

Sanchidrián JA, Segarra P, López LM. Energy components in rock blasting. International Journal of Rock Mechanics and Mining Sciences 2007;44:130–47. https://doi.org/10.1016/j.ijrmms.2006.05.002.

Sander R, Pan Z, Connell LD. Laboratory measurement of low permeability unconventional gas reservoir rocks: A review of experimental methods. Journal of Natural Gas Science and Engineering 2017;37:248–79. https://doi.org/10.1016/j.jngse.2016.11.041.

Sayadi A, Monjezi M, Talebi N, Khandelwal M. A comparative study on the application of various artificial neural networks to simultaneous prediction of rock fragmentation and backbreak. Journal of Rock Mechanics and Geotechnical Engineering 2013;5:318–24. https://doi.org/10.1016/j.jrmge.2013.05.007.

Sayevand K, Arab H, Golzar SB. Development of imperialist competitive algorithm in predicting the particle size distribution after mine blasting. Engineering with Computers 2018;34:329–38. https://doi.org/10.1007/s00366-017-0543-9.

Schölkopf B, Smola AJ, Williamson RC, Bartlett PL. New Support Vector Algorithms. Neural Computation 2000;12:1207–45. https://doi.org/10.1162/089976600300015565.

Sebtosheikh MA, Salehi A. Lithology prediction by support vector classifiers using inverted seismic attributes data and petrophysical logs as a new approach and investigation of training data set size effect on its performance in a heterogeneous carbonate reservoir. Journal of Petroleum Science and Engineering 2015;134:143–9. https://doi.org/10.1016/j.petrol.2015.08.001.

Segarra P, Sanchidrián JA, Navarro J, Castedo R. The Fragmentation Energy-Fan Model in Quarry Blasts. Rock Mechanics and Rock Engineering 2018;51:2175–90. https://doi.org/10.1007/s00603-018-1470-9.

Seyyedattar M, Zendehboudi S, Butt S. Relative Permeability Modeling Using Extra Trees, ANFIS, and Hybrid LSSVM–CSA Methods. Natural Resources Research 2022;31:571–600. https://doi.org/10.1007/s11053-021-09950-1.

Shams S, Monjezi M, Majd VJ, Armaghani DJ. Application of fuzzy inference system for prediction of rock fragmentation induced by blasting. Arabian Journal of Geosciences 2015;8:10819–32. https://doi.org/10.1007/s12517-015-1952-y.

Shatwell DG, Murray V, Barton A. Real-time ore sorting using color and texture analysis. International Journal of Mining Science and Technology 2023;33:659–74. https://doi.org/10.1016/j.ijmst.2023.03.004.

Sheykhinasab A, Mohseni AA, Barahooie Bahari A, Naruei E, Davoodi S, Aghaz A, et al. Prediction of permeability of highly heterogeneous hydrocarbon reservoir from conventional petrophysical logs using optimized data-driven algorithms. Journal of Petroleum Exploration and Production Technology 2023;13:661–89. https://doi.org/10.1007/s13202-022-01593-z.

Shi X, Zhou J, Wu B, Huang D, Wei W. Support vector machines approach to mean particle size of rock fragmentation due to bench blasting prediction. Transactions of Nonferrous Metals Society of China 2012;22:432–41. https://doi.org/10.1016/S1003-6326(11)61195-3.

Shi Y, Eberhart R. A modified particle swarm optimizer. 1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360), Anchorage, AK, USA: IEEE; 1998, p. 69–73. https://doi.org/10.1109/ICEC.1998.699146.

Shokooh Saljooghi B, Hezarkhani A. A new approach to improve permeability prediction of petroleum reservoirs using neural network adaptive wavelet (wavenet). Journal of Petroleum Science and Engineering 2015;133:851–61. https://doi.org/10.1016/j.petrol.2015.04.002.

Shu L, McIsaac K, Osinski GR, Francis R. Unsupervised feature learning for autonomous rock image classification. Computers & Geosciences 2017;106:10–7. https://doi.org/10.1016/j.cageo.2017.05.010.

Singh V, Rao SM. Application of image processing and radial basis neural network techniques for ore sorting and ore classification. Minerals Engineering 2005;18:1412–20. https://doi.org/10.1016/j.mineng.2005.03.003.

Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and Computing 2004;14:199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

Song H, Liu C, Lao J, Wang J, Du S, Yu M. Intelligent Microfluidics Research on Relative Permeability Measurement and Prediction of Two-Phase Flow in Micropores. Geofluids 2021;2021:1–12. https://doi.org/10.1155/2021/1194186.

Spathis AT. A Correction Relating to the Analysis of the Original Kuz-Ram Model. Fragblast 2004;8:201–5. https://doi.org/10.1080/13855140500041697.

Starr RC, Ingleton RA. A New Method for Collecting Core Samples Without a Drilling Rig. Groundwater Monitoring & Remediation 1992;12:91–5. https://doi.org/10.1111/j.1745-6592.1992.tb00413.x.

Subasi A, El-Amin MF, Darwich T, Dossary M. Permeability prediction of petroleum reservoirs using stochastic gradient boosting regression. Journal of Ambient Intelligence and Humanized Computing 2022;13:3555–64. https://doi.org/10.1007/s12652-020-01986-0.

Sun T, Chen F, Zhong L, Liu W, Wang Y. GIS-based mineral prospectivity mapping using machine learning methods: A case study from Tongling ore district, eastern China. Ore Geology Reviews 2019;109:26–49. https://doi.org/10.1016/j.oregeorev.2019.04.003.

Sutton CD. Classification and Regression Trees, Bagging, and Boosting. Handbook of Statistics, vol. 24, Elsevier; 2005, p. 303–29. https://doi.org/10.1016/S0169-7161(04)24011-1.

Székely GJ, Rizzo ML. Partial distance correlation with methods for dissimilarities. The Annals of Statistics 2014;42. https://doi.org/10.1214/14-AOS1255.

Székely GJ, Rizzo ML. Brownian distance covariance. The Annals of Applied Statistics 2009;3. https://doi.org/10.1214/09-AOAS312.

Tanaka S, Tsuru H, Someno K, Yamaguchi Y. Identification of Alteration Minerals from Unstable Reflectance Spectra Using a Deep Learning Method. Geosciences 2019;9:195. https://doi.org/10.3390/geosciences9050195.

Tang C, Zhu J, Qi X. Landslide hazard assessment of the 2008 Wenchuan earthquake: a case study in Beichuan area. Canadian Geotechnical Journal 2011;48:128–45. https://doi.org/10.1139/T10-059.

The MathWorks Inc. 2021.

Thomas S, Pillai GN, Pal K. Prediction of peak ground acceleration using ϵ-SVR, ν-SVR and Ls-SVR algorithm. Geomatics, Natural Hazards and Risk 2017;8:177–93. https://doi.org/10.1080/19475705.2016.1176604.

Thompson S, Fueten F, Bockus D. Mineral identification using artificial neural networks and the rotating polarizer stage. Computers & Geosciences 2001;27:1081–9. https://doi.org/10.1016/S0098-3004(00)00153-9.

Thornton D, Kanchibotla SS, Brunton I. Modelling the Impact of Rockmass and Blast Design Variation on Blast Fragmentation. Fragblast 2002;6:169–88. https://doi.org/10.1076/frag.6.2.169.8663.

Tian J, Qi C, Sun Y, Yaseen ZM, Pham BT. Permeability prediction of porous media using a combination of computational fluid dynamics and hybrid machine learning methods. Engineering with Computers 2021;37:3455–71. https://doi.org/10.1007/s00366-020-01012-z.

Timur A. An Investigation Of Permeability, Porosity, & Residual Water Saturation Relationships For Sandstone Reservoirs. The Log Analyst 1968.

Tsae NB, Adachi T, Kawamura Y. Application of Artificial Neural Network for the Prediction of Copper Ore Grade. Minerals 2023;13:658. https://doi.org/10.3390/min13050658.

Tsai D-Y, Lee Y, Matsuyama E. Information Entropy Measure for Evaluation of Image Quality. Journal of Digital Imaging 2008;21:338–47. https://doi.org/10.1007/s10278-007-9044-5.

Urang JG, Ebong ED, Akpan AE, Akaerue EI. A new approach for porosity and permeability prediction from well logs using artificial neural network and curve fitting techniques: A case study of Niger Delta, Nigeria. Journal of Applied Geophysics 2020;183:104207. https://doi.org/10.1016/j.jappgeo.2020.104207.

Valentín MB, Bom CR, Coelho JM, Correia MD, De Albuquerque Márcio P., De Albuquerque Marcelo P., et al. A deep residual convolutional neural network for automatic lithological facies identification in Brazilian pre-salt oilfield wellbore image logs. Journal of Petroleum Science and Engineering 2019;179:474–503. https://doi.org/10.1016/j.petrol.2019.04.030.

Van Groenigen JW. The influence of variogram parameters on optimal sampling schemes for mapping by kriging. Geoderma 2000;97:223–36. https://doi.org/10.1016/S0016-7061(00)00040-9.

Vapnik VN. The Nature of Statistical Learning Theory. New York, NY: Springer New York; 2000. https://doi.org/10.1007/978-1-4757-3264-1.

Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, Channappayya SS, Medasani SS. Blind image quality evaluation using perception based features. 2015 Twenty First National Conference on Communications (NCC), Mumbai, India: IEEE; 2015, p. 1–6. https://doi.org/10.1109/NCC.2015.7084843.

Vesanto J, Alhoniemi E. Clustering of the self-organizing map. IEEE Transactions on Neural Networks 2000;11:586–600. https://doi.org/10.1109/72.846731.

Von Luxburg U. A tutorial on spectral clustering. Statistics and Computing 2007;17:395–416. https://doi.org/10.1007/s11222-007-9033-z.

Wang K, Zhang L. Predicting formation lithology from log data by using a neural network. Petroleum Science 2008;5:242–6. https://doi.org/10.1007/s12182-008-0038-9.

Wang M, Chen H, Li H, Cai Z, Zhao X, Tong C, et al. Grey wolf optimization evolving kernel extreme learning machine: Application to bankruptcy prediction. Engineering Applications of Artificial Intelligence 2017;63:54–68. https://doi.org/10.1016/j.engappai.2017.05.003.

Wang M, Shi X, Zhou J. Charge design scheme optimization for ring blasting based on the developed Scaled Heelan model. International Journal of Rock Mechanics and Mining Sciences 2018a;110:199–209. https://doi.org/10.1016/j.ijrmms.2018.08.004.

Wang M, Shi X, Zhou J, Qiu X. Multi-planar detection optimization algorithm for the interval charging structure of large-diameter longhole blasting design based on rock fragmentation aspects. Engineering Optimization 2018b;50:2177–91. https://doi.org/10.1080/0305215X.2018.1439943.

Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and Intelligent Laboratory Systems 1987;2:37–52. https://doi.org/10.1016/0169-7439(87)80084-9.

Wu C, Hong L, Wang L, Zhang R, Pijush S, Zhang W. Prediction of wall deflection induced by braced excavation in spatially variable soils via convolutional neural network. Gondwana Research 2023;123:184–97. https://doi.org/10.1016/j.gr.2022.06.011.

Xi B, Li E, Fissha Y, Zhou J, Segarra P. LGBM-based modeling scenarios to compressive strength of recycled aggregate concrete with SHAP analysis. Mechanics of Advanced Materials and Structures 2024a;31:5999–6014. https://doi.org/10.1080/15376494.2023.2224782.

Xi B, Zhang N, Li E, Li J, Zhou J, Segarra P. A comprehensive comparison of different regression techniques and nature-inspired optimization algorithms to predict carbonation depth of recycled aggregate concrete. Frontiers of Structural and Civil Engineering 2024b;18:30–50. https://doi.org/10.1007/s11709-024-1041-y.

Xie C, Nguyen H, Bui X-N, Nguyen V-T, Zhou J. Predicting roof displacement of roadways in underground coal mines using adaptive neuro-fuzzy inference system optimized by various physics-based optimization algorithms. Journal of Rock Mechanics

and Geotechnical Engineering 2021;13:1452–65. https://doi.org/10.1016/j.jrmge.2021.07.005.

Xu S, Yang Z, Wu S, Wang L, Wei W, Yang F, et al. Fractal Analysis of Pore Structure Differences Between Shale and Sandstone Based on the Nitrogen Adsorption Method. Natural Resources Research 2022;31:1759–73. https://doi.org/10.1007/s11053-022-10056-5.

Xu Z, Ma W, Lin P, Hua Y. Deep learning of rock microscopic images for intelligent lithology identification: Neural network comparison and selection. Journal of Rock Mechanics and Geotechnical Engineering 2022;14:1140–52. https://doi.org/10.1016/j.jrmge.2022.05.009.

Xu Z, Ma W, Lin P, Shi H, Pan D, Liu T. Deep learning of rock images for intelligent lithology identification. Computers & Geosciences 2021;154:104799. https://doi.org/10.1016/j.cageo.2021.104799.

Xu Z, Yu T, Lin P, Li S. Adverse Geology Identification Through Mineral Anomaly Analysis During Tunneling: Methodology and Case Study. Engineering 2023:S2095809923000061. https://doi.org/10.1016/j.eng.2022.09.013.

Yadav S, Shukla S. Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, India: IEEE; 2016, p. 78–83. https://doi.org/10.1109/IACC.2016.25.

Yang M-S, Lai C-Y, Lin C-Y. A robust EM clustering algorithm for Gaussian mixture models. Pattern Recognition 2012;45:3950–61. https://doi.org/10.1016/j.patcog.2012.04.031.

Yang W, Zhang X, Ma H, Zhang G. Laser Beams-Based Localization Methods for Boom-Type Roadheader Using Underground Camera Non-Uniform Blur Model. IEEE Access 2020;8:190327–41. https://doi.org/10.1109/ACCESS.2020.3032368.

Yang Y, Zhang Q. A hierarchical analysis for rock engineering using artificial neural networks. Rock Mechanics and Rock Engineering 1997;30:207–22. https://doi.org/10.1007/BF01045717.

Yu Z, Shi X, Miao X, Zhou J, Khandelwal M, Chen X, et al. Intelligent modeling of blast-induced rock movement prediction using dimensional analysis and optimized artificial neural network technique. International Journal of Rock Mechanics and Mining Sciences 2021a;143:104794. https://doi.org/10.1016/j.ijrmms.2021.104794.

Yu Z, Shi X, Zhou J, Chen X, Miao X, Teng B, et al. Prediction of Blast-Induced Rock Movement During Bench Blasting: Use of Gray Wolf Optimizer and Support Vector Regression. Natural Resources Research 2020a;29:843–65. https://doi.org/10.1007/s11053-019-09593-3.

Yu Z, Shi X, Zhou J, Chen X, Qiu X. Effective Assessment of Blast-Induced Ground Vibration Using an Optimized Random Forest Model Based on a Harris Hawks Optimization Algorithm. Applied Sciences 2020b;10:1403. https://doi.org/10.3390/app10041403.

Yu Z, Shi X, Zhou J, Gou Y, Rao D, Huo X. Machine-Learning-Aided Determination of Post-blast Ore Boundary for Controlling Ore Loss and Dilution. Natural Resources Research 2021b;30:4063–78. https://doi.org/10.1007/s11053-021-09914-5.

Yu Z, Shi X, Zhou J, Rao D, Chen X, Dong W, et al. Feasibility of the indirect determination of blast-induced rock movement based on three new hybrid intelligent models. Engineering with Computers 2021c;37:991–1006. https://doi.org/10.1007/s00366-019-00868-0.

Zhang G, Wang Z, Chen Y. Deep learning for Seismic Lithology Prediction. Geophysical Journal International 2018. https://doi.org/10.1093/gji/ggy344.

Zhang K, Zhang Ke, Bao R, Liu X. A framework for predicting the carbonation depth of concrete incorporating fly ash based on a least squares support vector machine and metaheuristic algorithms. Journal of Building Engineering 2023. https://doi.org/10.1016/j.jobe.2022.105772.

Zhang N, Xi B, Li J, Liu L, Song G. Utilization of CO2 into recycled construction materials: A systematic literature review. Journal of Material Cycles and Waste Management 2022;24:2108–25. https://doi.org/10.1007/s10163-022-01489-4.

Zhang S, Bui X-N, Trung N-T, Nguyen H, Bui H-B. Prediction of Rock Size Distribution in Mine Bench Blasting Using a Novel Ant Colony Optimization-Based Boosted Regression Tree Technique. Natural Resources Research 2020;29:867–86. https://doi.org/10.1007/s11053-019-09603-4.

Zhang W, Gu X, Hong L, Han L, Wang L. Comprehensive review of machine learning in geotechnical reliability analysis: Algorithms, applications and further challenges. Applied Soft Computing 2023a;136:110066. https://doi.org/10.1016/j.asoc.2023.110066.

Zhang W, Gu X, Tang L, Yin Y, Liu D, Zhang Y. Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge. Gondwana Research 2022;109:1–17. https://doi.org/10.1016/j.gr.2022.03.015.

Zhang W, He Y, Wang L, Liu S, Meng X. Landslide Susceptibility mapping using random forest and extreme gradient boosting: A case study of Fengjie, Chongqing. Geological Journal 2023b;58:2372–87. https://doi.org/10.1002/gj.4683.

Zhang W, Li H, Li Y, Liu H, Chen Y, Ding X. Application of deep learning algorithms in geotechnical engineering: a short critical review. Artificial Intelligence Review 2021;54:5633–73. https://doi.org/10.1007/s10462-021-09967-1.

Zhang W, Wu C, Tang L, Gu X, Wang L. Efficient time-variant reliability analysis of Bazimen landslide in the Three Gorges Reservoir Area using XGBoost and LightGBM algorithms. Gondwana Research 2023c;123:41–53. https://doi.org/10.1016/j.gr.2022.10.004.

Zhang Y, Chen H, Yang B, Fu S, Yu J, Wang Z. Prediction of phosphate concentrate grade based on artificial neural network modeling. Results in Physics 2018;11:625–8. https://doi.org/10.1016/j.rinp.2018.10.011.

Zhang Z, Yang J, Wang Y, Dou D, Xia W. Ash content prediction of coarse coal by image analysis and GA-SVM. Powder Technology 2014;268:429–35. https://doi.org/10.1016/j.powtec.2014.08.044.

Zhao D, Wang H, He W, Ding K, Zhang H. Research on rock sample lithology identification algorithm based on ResNet self-supervised learning. In: Zhang L, Chen S, AlShawabkeh M, editors. International Conference on Algorithms, High Performance

Computing, and Artificial Intelligence (AHPCAI 2021), Sanya, China: SPIE; 2021, p. 54. https://doi.org/10.1117/12.2626531.

Zhou J, Asteris PG, Armaghani DJ, Pham BT. Prediction of ground vibration induced by blasting operations through the use of the Bayesian Network and random forest models. Soil Dynamics and Earthquake Engineering 2020;139:106390. https://doi.org/10.1016/j.soildyn.2020.106390.

Zhou J, Chen C, Du K, Jahed Armaghani D, Li C. A new hybrid model of information entropy and unascertained measurement with different membership functions for evaluating destressability in burst-prone underground mines. Engineering with Computers 2022a;38:381–99. https://doi.org/10.1007/s00366-020-01151-3.

Zhou J, Chen C, Khandelwal M, Tao M, Li C. Novel approach to evaluate rock mass fragmentation in block caving using unascertained measurement model and information entropy with flexible credible identification criterion. Engineering with Computers 2022b;38:3789–809. https://doi.org/10.1007/s00366-020-01230-5.

Zhou J, Dai Y, Du K, Khandelwal M, Li C, Qiu Y. COSMA-RF: New intelligent model based on chaos optimized slime mould algorithm and random forest for estimating the peak cutting force of conical picks. Transportation Geotechnics 2022c;36:100806. https://doi.org/10.1016/j.trgeo.2022.100806.

Zhou J, Li C, Arslan CA, Hasanipanah M, Bakhshandeh Amnieh H. Performance evaluation of hybrid FFA-ANFIS and GA-ANFIS models to predict particle size distribution of a muck-pile after blasting. Engineering with Computers 2021a;37:265–74. https://doi.org/10.1007/s00366-019-00822-0.

Zhou J, Li X, Hani S. M, Wang S, Wei W. Identification of large-scale goaf instability in underground mine using particle swarm optimization and support vector machine. International Journal of Mining Science and Technology 2013;23:701–7. https://doi.org/10.1016/j.ijmst.2013.08.014.

Zhou J, Li X, Mitri HS. Classification of Rockburst in Underground Projects: Comparison of Ten Supervised Learning Methods. Journal of Computing in Civil Engineering 2016;30:04016003. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000553.

Zhou J, Li X, Shi X. Long-term prediction model of rockburst in underground openings using heuristic algorithms and support vector machines. Safety Science 2012;50:629–44. https://doi.org/10.1016/j.ssci.2011.08.065.

Zhou J, Qiu Y, Armaghani DJ, Zhang W, Li C, Zhu S, et al. Predicting TBM penetration rate in hard rock condition: A comparative study among six XGB-based metaheuristic techniques. Geoscience Frontiers 2021b;12:101091. https://doi.org/10.1016/j.gsf.2020.09.020.

Zhou J, Qiu Y, Khandelwal M, Zhu S, Zhang X. Developing a hybrid model of Jaya algorithm-based extreme gradient boosting machine to estimate blast-induced ground vibrations. International Journal of Rock Mechanics and Mining Sciences 2021c;145:104856. https://doi.org/10.1016/j.ijrmms.2021.104856.

Zhou J, Qiu Y, Zhu S, Armaghani DJ, Khandelwal M, Mohamad ET. Estimation of the TBM advance rate under hard rock conditions using XGBoost and Bayesian optimization. Underground Space 2021d;6:506–15. https://doi.org/10.1016/j.undsp.2020.05.008.

Zhou J, Qiu Y, Zhu S, Armaghani DJ, Li C, Nguyen H, et al. Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate. Engineering Applications of Artificial Intelligence 2021e;97:104015. https://doi.org/10.1016/j.engappai.2020.104015.

Zhou J, Shen X, Qiu Y, Li E, Rao D, Shi X. Improving the efficiency of microseismic source locating using a heuristic algorithm-based virtual field optimization method. Geomech Geophys Geo-Energ Geo-Resour 2021f;7:89. https://doi.org/10.1007/s40948-021-00285-y.

Zhou J, Zhang R, Qiu Y, Khandelwal M. A true triaxial strength criterion for rocks by gene expression programming. Journal of Rock Mechanics and Geotechnical Engineering 2023;15:2508–20. https://doi.org/10.1016/j.jrmge.2023.03.004.

Zorlu K, Gokceoglu C, Ocakoglu F, Nefeslioglu HA, Acikalin S. Prediction of uniaxial compressive strength of sandstones using petrography-based models. Engineering Geology 2008;96:141–58. https://doi.org/10.1016/j.enggeo.2007.10.009.

Appendix 1. Cumulative scores for different models with different population sizes

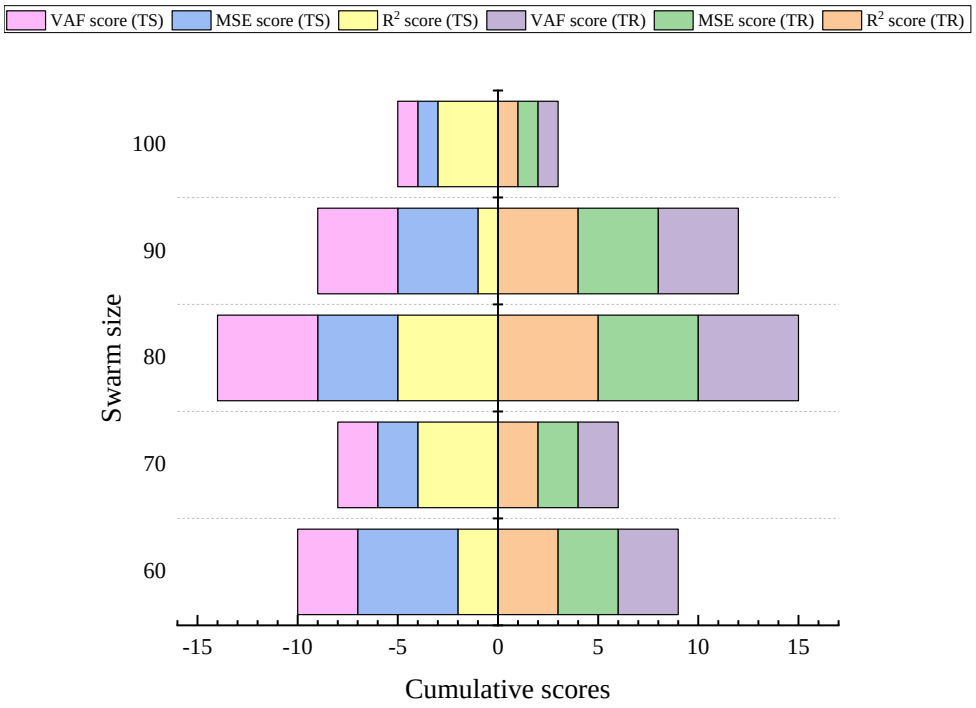


Figure 86. Cumulative scores with different population sizes for PSO-e-SVR.

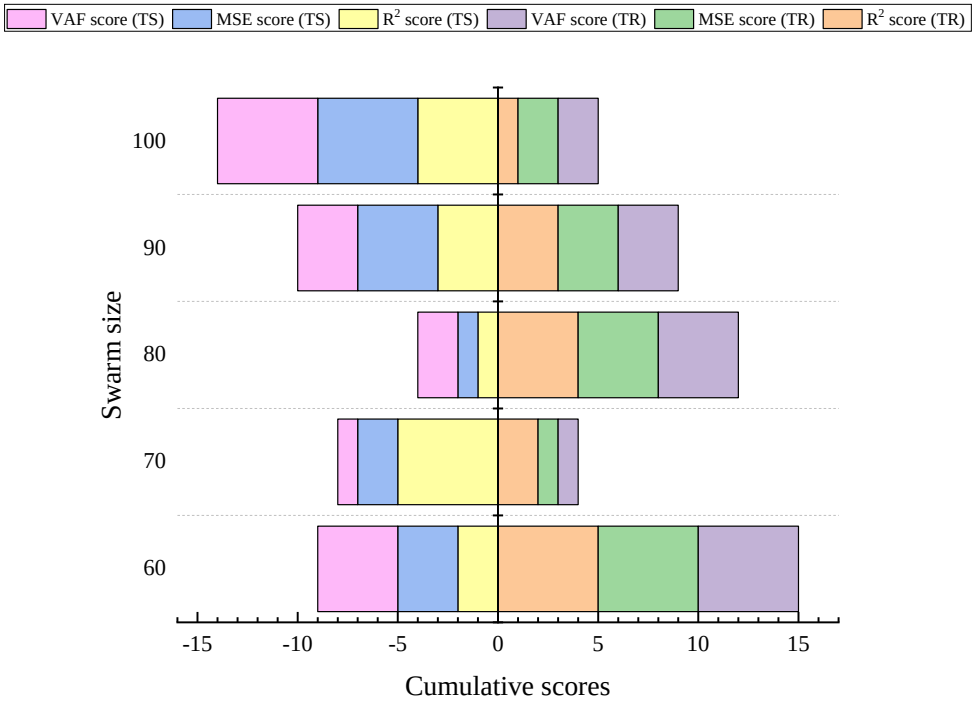


Figure 87. Cumulative scores with different population sizes for PSO-v-SVR.

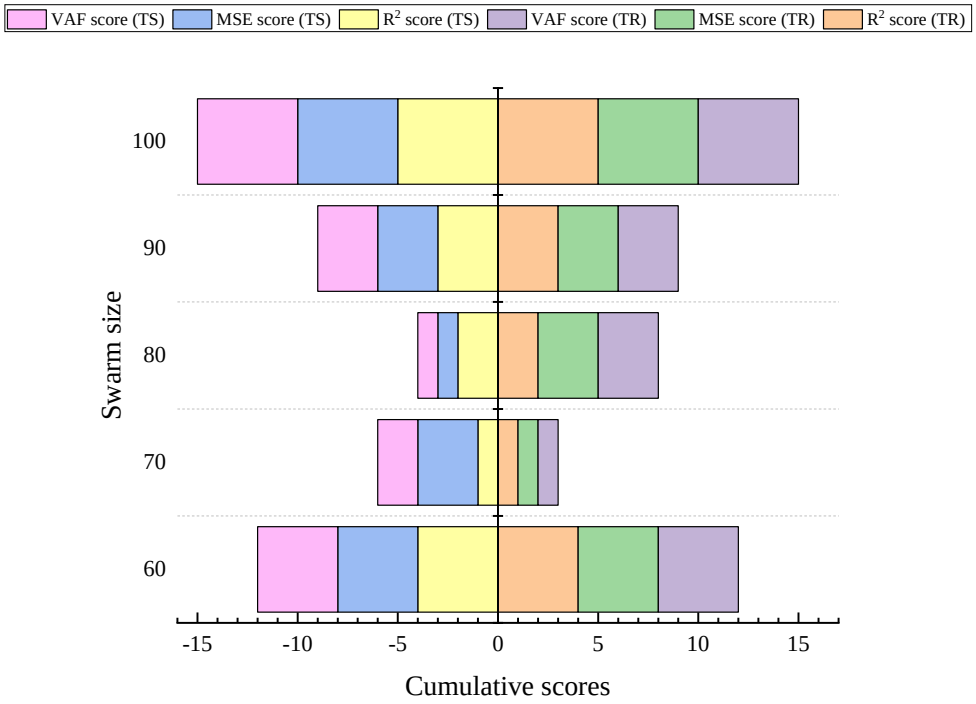


Figure 88. Cumulative scores with different population sizes for GA-e-SVR.

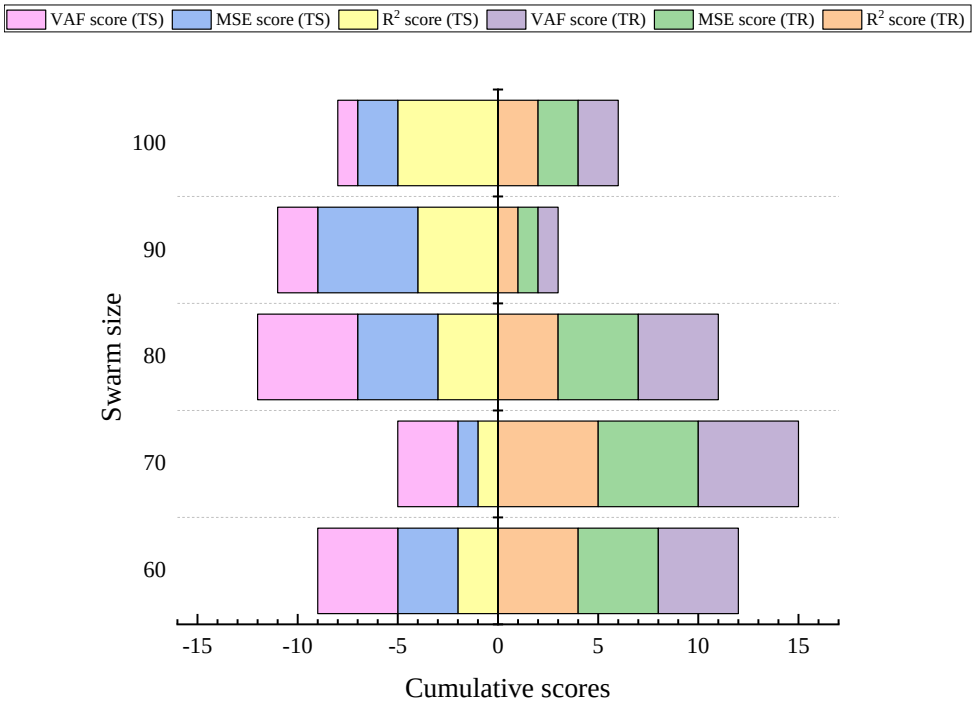


Figure 89. Cumulative scores with different population sizes for GA-v-SVR.

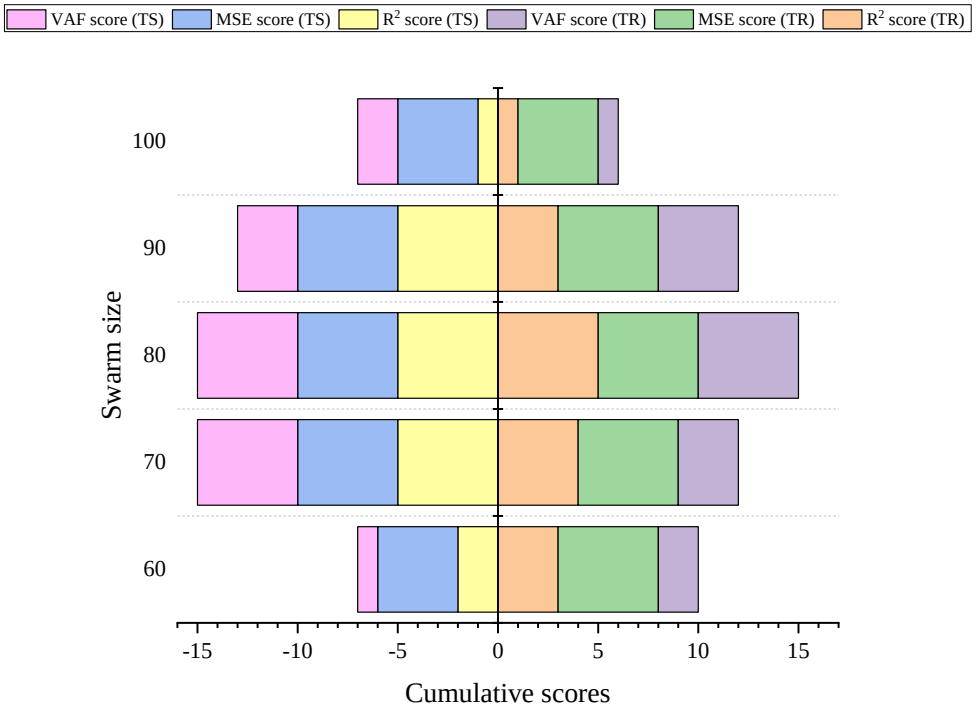


Figure 90. Cumulative scores with different population sizes for SSA-e-SVR.

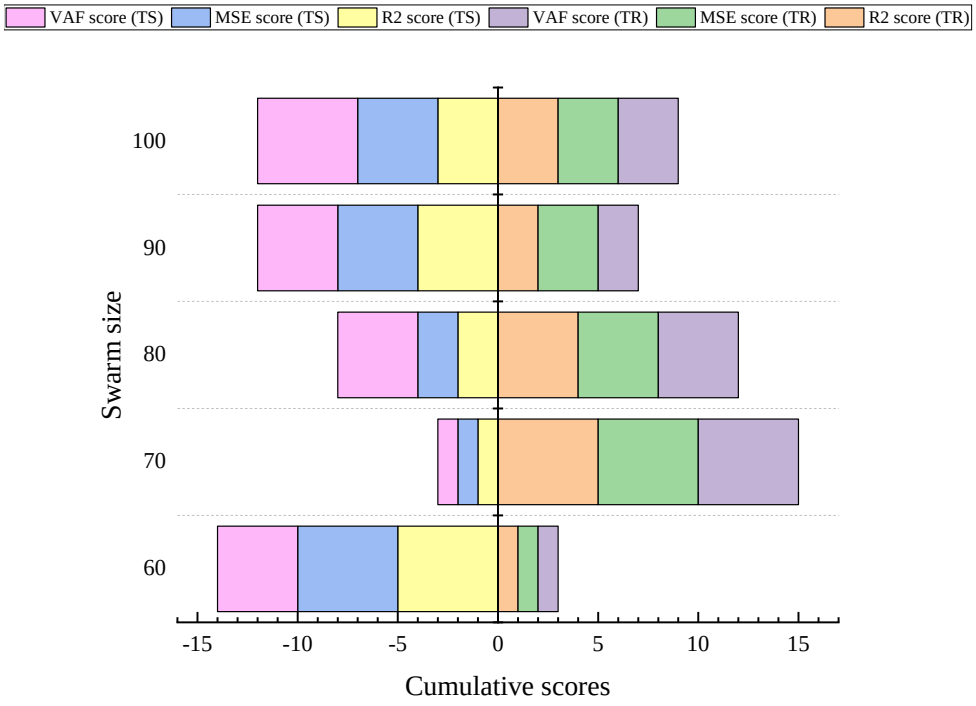


Figure 91. Cumulative scores with different population sizes for SSA-v-SVR.

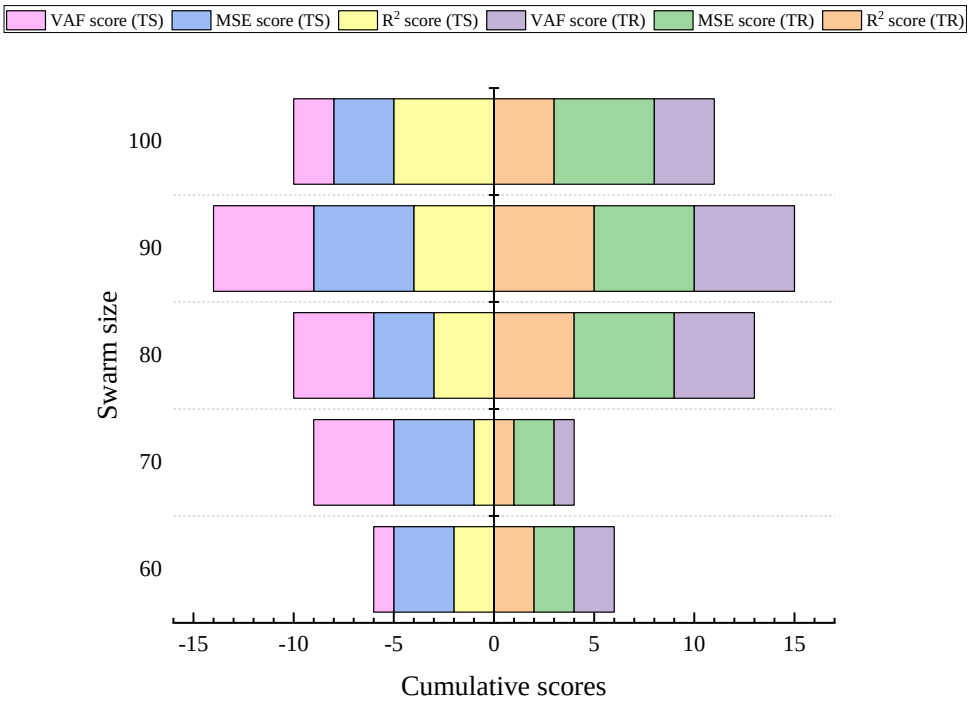


Figure 92. Cumulative scores with different population sizes for GWO-e-SVR.

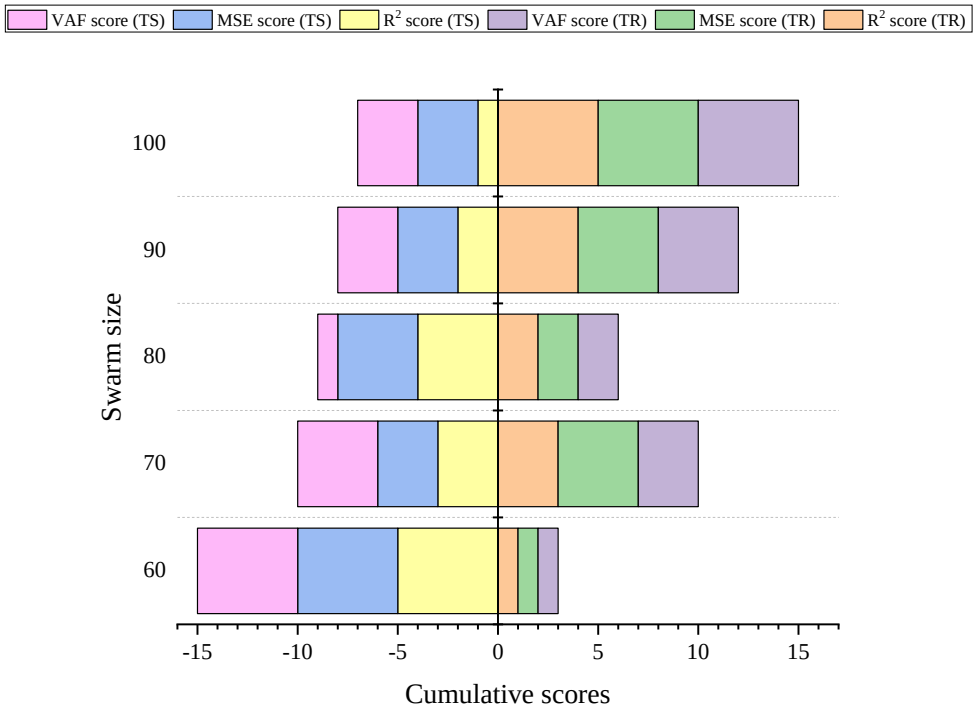


Figure 93. Cumulative scores with different population sizes for GWO-v-SVR.

Appendix 2. Prediction performance of five optimized KELM models for the gas relative permeability in reservoir

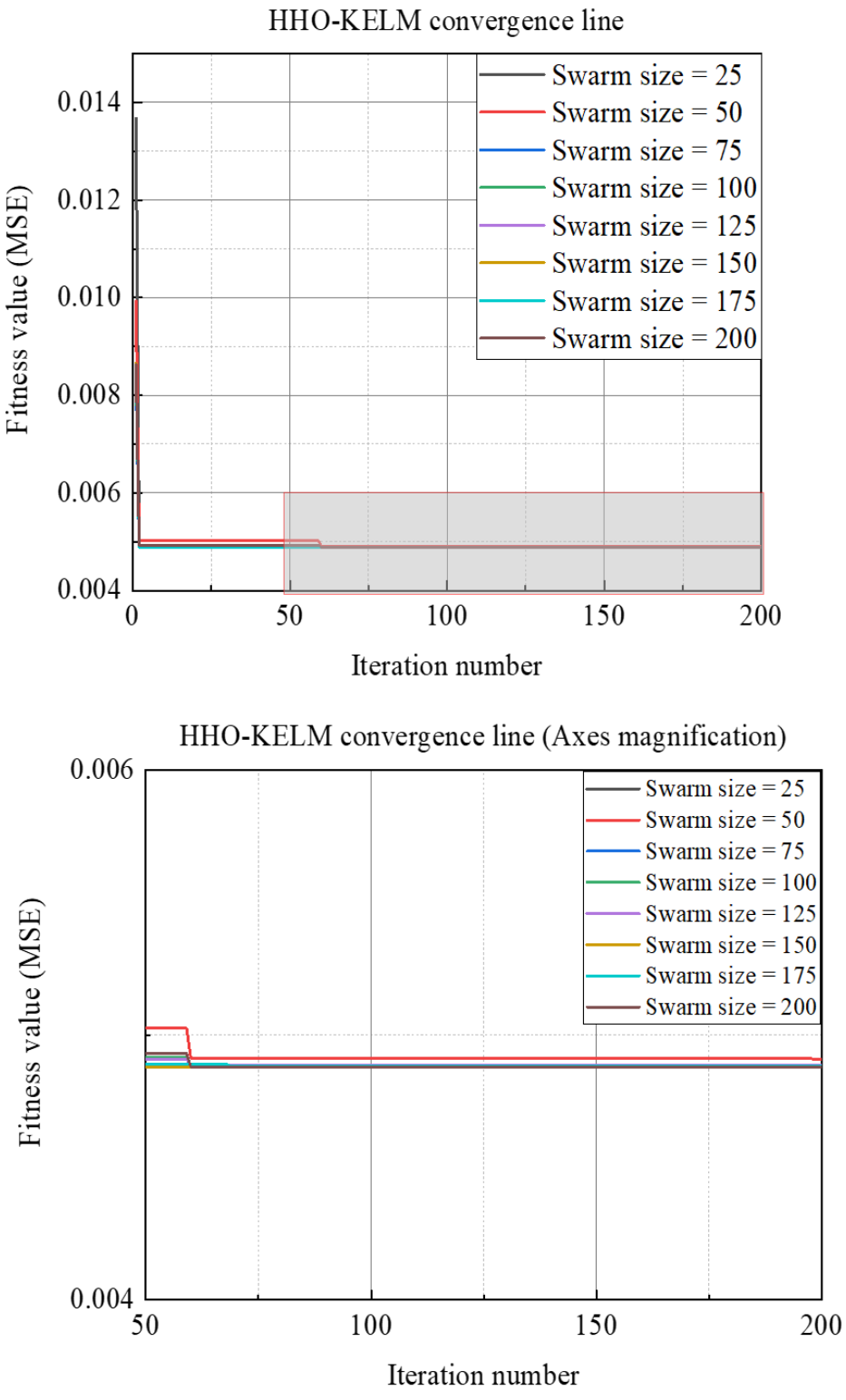


Figure 94. Optimization process of HHO-KELM.

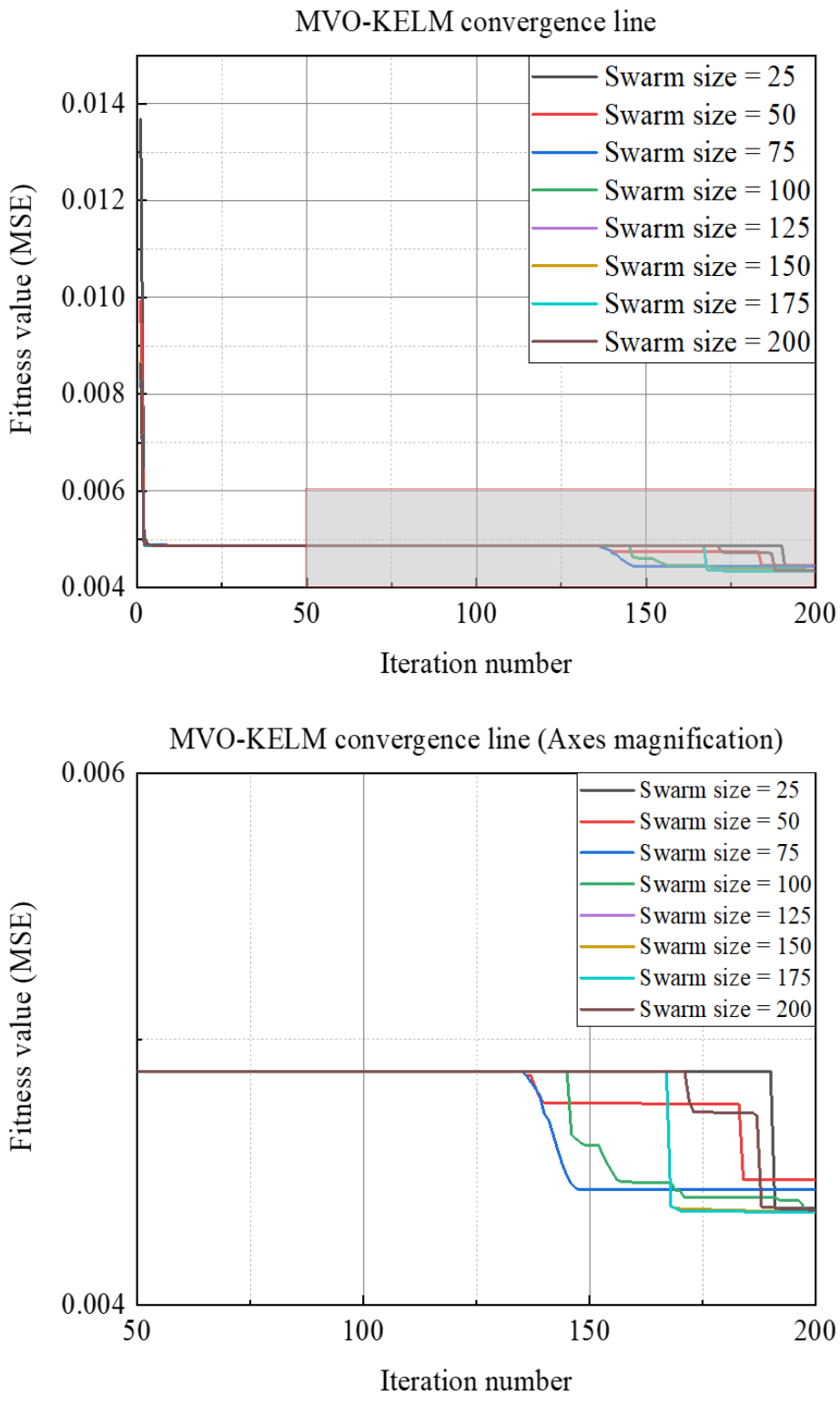


Figure 95. Optimization process of MVO-KELM.

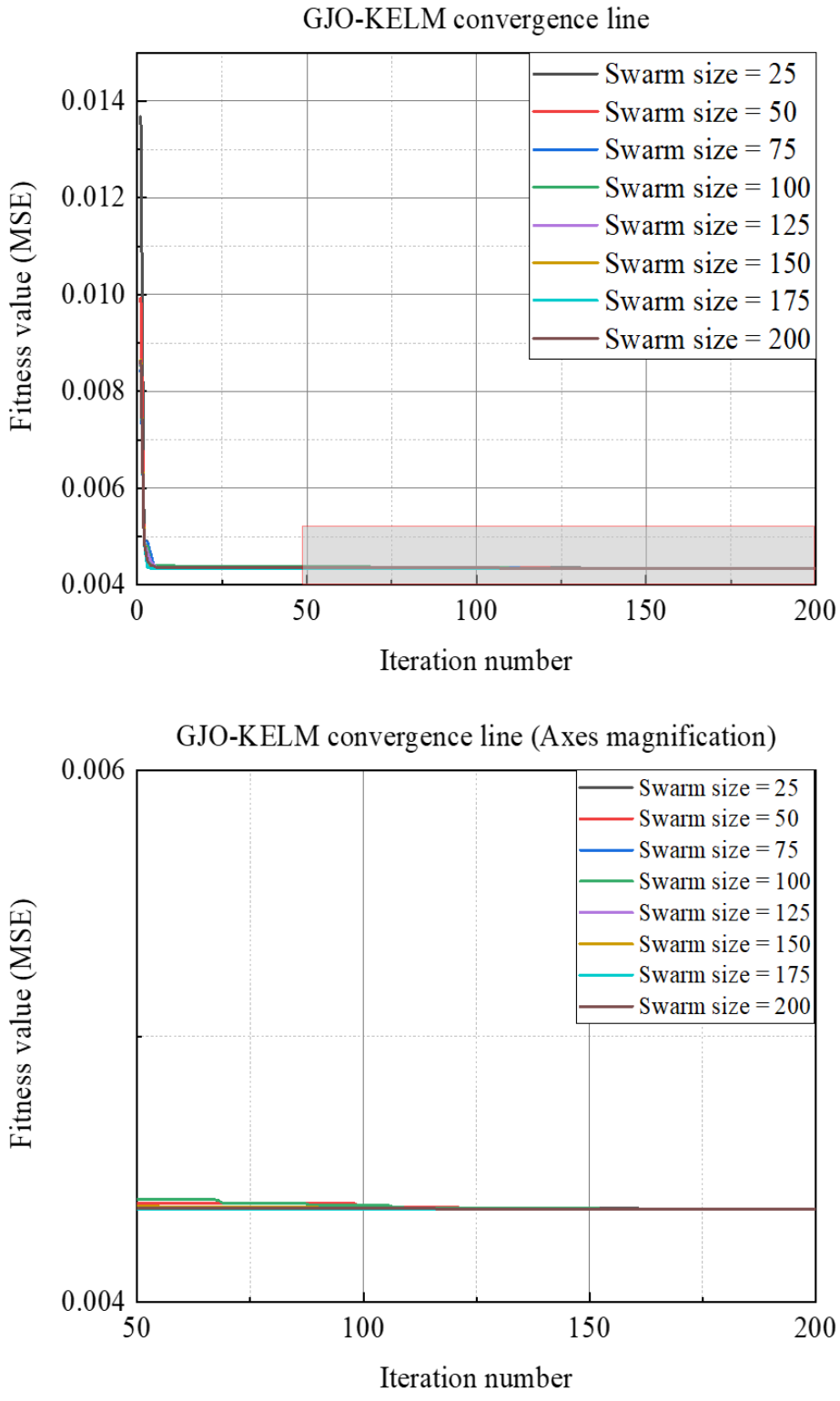


Figure 96. Optimization process of GJO-KELM.

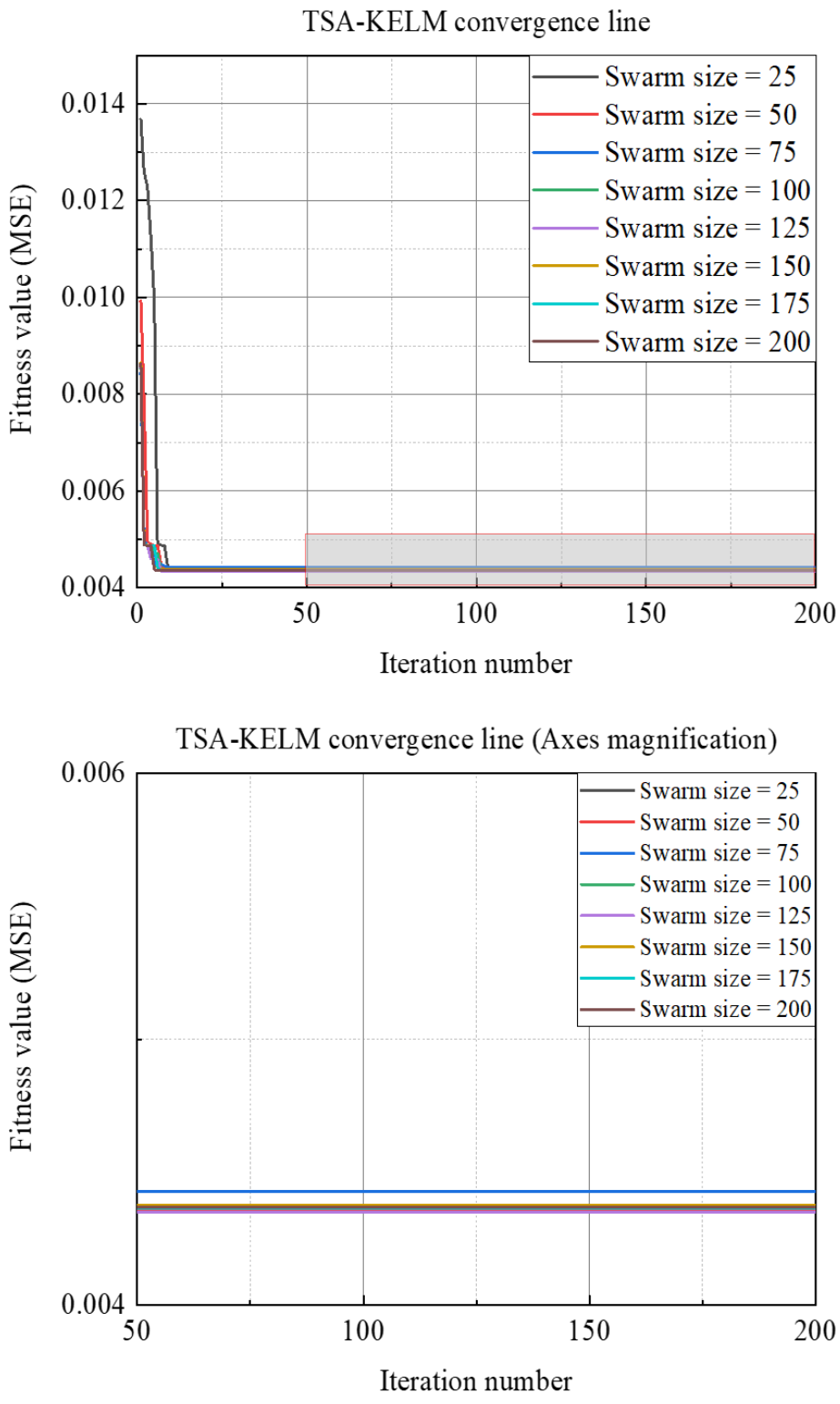


Figure 97. Optimization process of TSA-KELM.

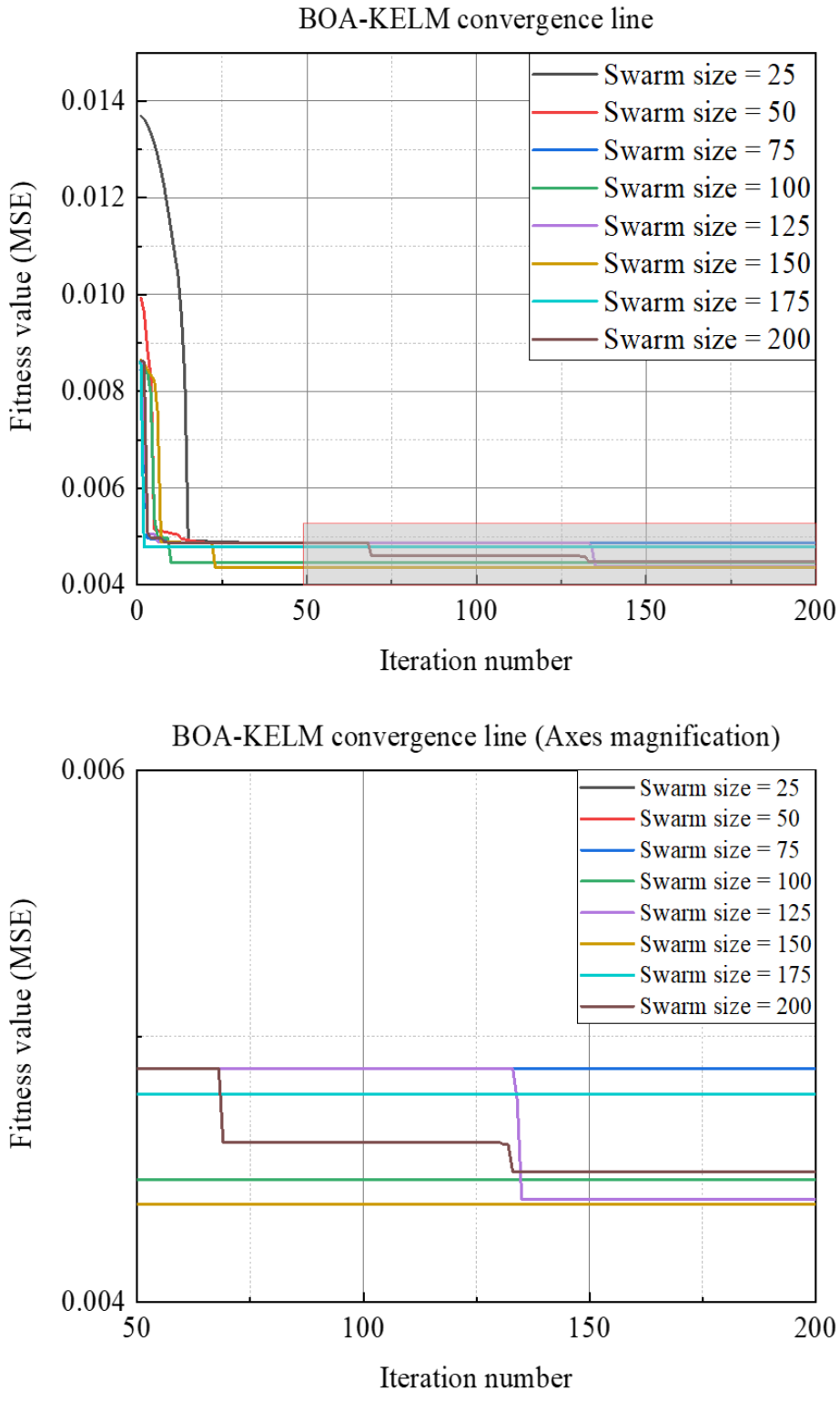


Figure 98 Optimization process of BOA-KELM.

Table 41. Prediction performance of HHO-KELM based prediction models (training set).

Swarm size	Training set					Hyper-parameters	
	R^2	RMSE	VAF	MAE	GI	C	γ
25	0.9918	0.685	99.1786	9.7961	0.1893	91.6281	0.1
50	0.9932	0.6249	99.3166	8.8178	0.2104	140.899	0.1
75	0.9914	0.7001	99.1421	10.0416	0.1846	82.6342	0.1
100	0.9918	0.6846	99.1796	9.7893	0.1894	91.8905	0.1
125	0.9911	0.7115	99.1139	10.226	0.1812	76.5107	0.1
150	0.9918	0.6848	99.179	9.7933	0.1893	91.7363	0.1
175	0.9914	0.7012	99.1394	10.0591	0.1843	82.0304	0.1
200	0.9918	0.6837	99.1818	9.7743	0.1897	92.4802	0.1

Table 42. Prediction performance of HHO-KELM based prediction models (testing set).

Swarm size	Testing set					Hyper-parameters	
	R^2	RMSE	VAF	MAE	GI	C	γ
25	0.975	0.5783	97.5029	4.332	0.3971	91.6281	0.1
50	0.9745	0.5841	97.4529	4.3362	0.3961	140.899	0.1
75	0.9751	0.5777	97.5078	4.3423	0.3964	82.6342	0.1
100	0.975	0.5783	97.5028	4.3317	0.3972	91.8905	0.1
125	0.9751	0.5776	97.5093	4.3513	0.3957	76.5107	0.1
150	0.975	0.5783	97.5029	4.3318	0.3971	91.7363	0.1
175	0.9751	0.5777	97.508	4.3431	0.3963	82.0304	0.1
200	0.975	0.5784	97.5024	4.331	0.3972	92.4802	0.1

Table 43. Prediction performance of BOA-KELM based prediction models (training set).

Swarm size	Training set					Hyper-parameters	
	R^2	RMSE	VAF	MAE	GI	C	γ
25	0.9918	0.6826	99.1844	9.7568	0.19	93.1724	0.1
50	0.9918	0.6839	99.1814	9.7775	0.1896	92.3569	0.1
75	0.9918	0.6844	99.18	9.7867	0.1894	91.9935	0.1
100	0.9854	0.9145	98.5365	13.4226	0.1375	95.2028	0.1682
125	0.9859	0.8985	98.5872	13.1284	0.1406	80.0979	0.1538
150	0.9902	0.7477	99.0218	10.6636	0.1736	151.978	0.1343
175	0.9944	0.5646	99.4422	7.7975	0.2378	374.111	0.1153
200	0.9853	0.917	98.5279	13.4733	0.1369	43.3536	0.1302

Table 44. Prediction performance of BOA-KELM based prediction models (testing set).

Swarm size	Testing set					Hyper-parameters	
	R^2	RMSE	VAF	MAE	GI	C	γ
25	0.975	0.5784	97.5019	4.3303	0.3973	93.1724	0.1
50	0.975	0.5784	97.5024	4.3312	0.3972	92.3569	0.1
75	0.975	0.5783	97.5027	4.3316	0.3972	91.9935	0.1
100	0.9785	0.5372	97.846	4.4518	0.3922	95.2028	0.1682
125	0.9782	0.5408	97.817	4.4256	0.3939	80.0979	0.1538
150	0.9789	0.5314	97.8917	4.1706	0.4164	151.978	0.1343
175	0.9757	0.5704	97.5716	4.2705	0.4031	374.111	0.1153
200	0.9761	0.5656	97.6122	4.5621	0.3807	43.3536	0.1302

Table 45. Prediction performance of TSA-KELM based prediction models (training set).

Swarm size	Training set					Hyper-parameters	
	R^2	RMSE	VAF	MAE	GI	C	γ
25	0.9875	0.8438	98.754	12.178	0.1517	87.672	0.14033
50	0.9878	0.8337	98.7837	12.006	0.1539	107.925	0.14646
75	0.9835	0.9695	98.3548	14.4171	0.1278	46.409	0.15009
100	0.9865	0.8782	98.6502	12.7648	0.1446	73.2649	0.14293
125	0.9886	0.8068	98.8608	11.5702	0.1597	121.278	0.14305
150	0.9865	0.8784	98.6495	12.769	0.1446	72.8539	0.14274
175	0.987	0.8626	98.6978	12.4958	0.1478	83.9841	0.14431
200	0.9869	0.8644	98.6922	12.5284	0.1474	79.5647	0.1424

Table 46. Prediction performance of TSA-KELM based prediction models (testing set).

Swarm size	Testing set					Hyper-parameters	
	R^2	RMSE	VAF	MAE	GI	C	γ
25	0.9784	0.5375	97.8431	4.3211	0.4028	87.672	0.14033
50	0.979	0.5306	97.8984	4.2764	0.4073	107.925	0.14646
75	0.9761	0.5652	97.6159	4.676	0.3725	46.409	0.15009
100	0.978	0.5433	97.7963	4.3944	0.3961	73.2649	0.14293
125	0.9791	0.5288	97.9124	4.2289	0.4116	121.278	0.14305
150	0.9779	0.5435	97.7946	4.3955	0.396	72.8539	0.14274
175	0.9784	0.5384	97.8361	4.3509	0.4002	83.9841	0.14431
200	0.9782	0.5404	97.8202	4.3608	0.3992	79.5647	0.1424

Table 47. Prediction performance of MVO-KELM based prediction models (training set).

Swarm size	Training set					Hyper-parameters	
	R^2	RMSE	VAF	MAE	GI	C	γ
25	0.987	0.8615	98.7011	12.478	0.148	92.3295	0.14834
50	0.9815	1.0282	98.1495	15.5281	0.1186	35.0492	0.15489
75	0.9837	0.9649	98.3704	14.3303	0.1286	44.3439	0.14618
100	0.9878	0.8338	98.7833	12.009	0.1538	102.822	0.14433
125	0.9882	0.8211	98.8201	11.7998	0.1566	114.582	0.14511
150	0.988	0.8283	98.7993	11.9212	0.155	100.91	0.14175
175	0.9885	0.8118	98.8466	11.6476	0.1587	125.572	0.14618
200	0.988	0.8279	98.8003	11.9168	0.155	99.3892	0.14098

Table 48. Prediction performance of MVO-KELM based prediction models (testing set).

Swarm size	Testing set					Hyper-parameters	
	R^2	RMSE	VAF	MAE	GI	C	γ
25	0.9786	0.5351	97.8628	4.3364	0.4018	92.3295	0.14834
50	0.9747	0.5817	97.4749	4.8531	0.3587	35.0492	0.15489
75	0.976	0.5664	97.6053	4.6708	0.3727	44.3439	0.14618
100	0.9788	0.5322	97.8851	4.2841	0.4065	102.822	0.14433
125	0.9791	0.5294	97.9079	4.2523	0.4095	114.582	0.14511
150	0.9788	0.5333	97.8766	4.2803	0.4067	100.91	0.14175
175	0.9793	0.5271	97.9258	4.228	0.4119	125.572	0.14618
200	0.9787	0.5339	97.8719	4.2823	0.4064	99.3892	0.14098

Table 49. Prediction performance of GJO-KELM based prediction models (training set).

Swarm size	Training set					Hyper-parameters	
	R^2	RMSE	VAF	MAE	GI	C	γ
25	0.9889	0.796	98.8911	11.3974	0.1622	131.162	0.14304
50	0.9889	0.7954	98.8929	11.3874	0.1623	130.864	0.14273
75	0.9889	0.7962	98.8907	11.4001	0.1622	130.537	0.14288
100	0.9889	0.7955	98.8926	11.3892	0.1623	131.251	0.1429
125	0.9889	0.7961	98.8908	11.3993	0.1622	130.741	0.14293
150	0.9889	0.7955	98.8925	11.3896	0.1623	131.037	0.14283
175	0.9889	0.7957	98.8919	11.3931	0.1623	130.95	0.14288
200	0.9889	0.7953	98.8931	11.3863	0.1624	131.403	0.14289

Table 50. Prediction performance of GJO-KELM based prediction models (testing set).

Swarm size	Testing set					Hyper-parameters	
	R^2	RMSE	VAF	MAE	GI	C	γ
25	0.9792	0.5274	97.9233	4.2088	0.4135	131.162	0.14304
50	0.9792	0.5276	97.922	4.2089	0.4135	130.864	0.14273
75	0.9792	0.5275	97.9222	4.2097	0.4134	130.537	0.14288
100	0.9792	0.5274	97.9229	4.2084	0.4135	131.251	0.1429
125	0.9792	0.5275	97.9225	4.2094	0.4134	130.741	0.14293
150	0.9792	0.5275	97.9225	4.2087	0.4135	131.037	0.14283
175	0.9792	0.5275	97.9225	4.209	0.4135	130.95	0.14288
200	0.9792	0.5274	97.923	4.2082	0.4136	131.403	0.14289

Appendix 3. Input analysis and prediction performance

Table 51. PCA results after removing section #5.

Light source	1	2	3	4	5	6	7	8	9
W[a]	63.06	93.26	97.41	-	-	-	-	-	-
UV	60.41	71.19	81.1	86.78	90.63	92.84	94.55	95.62	-
WUV[a]	37.15	67.29	83.16	86.95	90.43	92.19	93.74	94.98	96.02

[a]Percentiles with constant intensity are discarded for the analysis

Table 52. Correlation coefficients between PCI and fluorite content after removing section #5.

Percentile	White light			UV light		
	Red	Green	Blue	Red	Green	Blue
10	0.12	0.3	**0.35**	0.13	**0.48**	0.07
20	0.08	0.28	**0.34**	0.23	0.19	0.24
30	0.09	0.32	**0.36**	0.3	0.2	0.34
40	0.08	0.31	**0.34**	0.16	**0.37**	0.24
50	0.02	0.27	**0.37**	0.09	**0.38**	0.15
60	-0.05	0.26	**0.41**	0.07	0.23	0.16
70	-0.09	0.32	**0.43**	0.14	0.14	0.22
80	-0.14	**0.37**	**0.44**	0.05	0.2	0.27
90	-0.13	**0.35**	**0.4**	0.08	0.31	0.23
100	-	-	0.15	**0.59**	**0.5**	0.13

Note: Bold numbers are $|r| \geq 0.3$ and p-value≤ 0.05.

Table 53. Regression results.

REGRESSION		TR				TS			
		R^2	VAF	MAPE	RMSE	R^2	VAF	MAPE	RMSE
W_{PCA}	Mean	0.9	90.3	0.26	2.68	0.73	76.93	0.47	4.65
	Min	0.84	84.28	0.04	0.34	0.33	37.39	0.2	3.08
	Max	1	99.83	0.37	3.4	0.9	92.78	0.76	9.38
	Std.	0.03	3.21	0.08	0.59	0.13	11.78	0.13	1.42
W_{CA}	Mean	0.78	78.01	0.39	4.01	0.7	71.87	0.5	4.82
	Min	0.67	68.25	0.12	2.07	0.46	50.51	0.33	3.49
	Max	0.94	94.24	0.5	4.83	0.88	88.68	0.63	6.24
	Std.	0.09	8.47	0.13	0.92	0.1	9.31	0.08	0.67
UV_{PCA}	Mean	0.99	99.08	0.05	0.48	0.61	65.01	0.48	5.15
	Min	0.88	87.92	0.03	0.28	0.24	26.45	0.23	2.6
	Max	1	99.87	0.19	2.58	0.9	89.6	0.94	7.36
	Std.	0.03	2.51	0.04	0.55	0.19	17.99	0.15	1.32
UV_{CA}	Mean	0.75	76.04	0.3	3.77	0.6	66.9	0.47	5.3
	Min	0.65	67.76	0.12	1.78	-0.03	11.88	0.29	3.02
	Max	0.95	94.96	0.37	4.6	0.87	87.26	0.91	7.93
	Std.	0.07	6.47	0.05	0.6	0.22	20.59	0.13	1.37
WUV_{PCA}	Mean	0.99	99.35	0.04	0.44	0.83	84.72	0.33	3.32
	Min	0.93	93.02	0.03	0.27	0.64	67.34	0.14	1.8
	Max	1	99.88	0.15	1.99	0.95	97.55	0.56	5.54
	Std.	0.02	1.6	0.03	0.45	0.09	7.71	0.11	0.9
WUV_{CA}	Mean	0.89	89.41	0.21	2.48	0.77	79.81	0.39	4.03
	Min	0.84	84.33	0.03	0.29	0.48	60.49	0.25	2.63
	Max	1	99.85	0.3	3.12	0.91	91.59	0.58	6.07
	Std.	0.04	3.46	0.06	0.55	0.1	9.52	0.09	0.85

Table 54. Classification results.

CLASSIFICATION		TR				TS			
		AcT	AcW	AcLG	AcMG	AcT	AcW	AcLG	AcMG
W_{PCA}	Mean	0.95	0.92	0.96	1	0.84	0.84	0.77	1
	Min	0.83	0.77	0.83	1	0.5	0.5	0	1
	Max	1	1	1	1	1	1	1	1
	Std.	0.05	0.07	0.06	0	0.12	0.19	0.3	0
W_{CA}	Mean	0.88	0.9	0.82	1	0.76	0.76	0.67	1
	Min	0.69	0.69	0.36	1	0.5	0.33	0	1
	Max	1	1	1	1	1	1	1	1
	Std.	0.09	0.09	0.18	0	0.11	0.24	0.3	0
UV_{PCA}	Mean	0.96	0.94	0.97	0.99	0.71	0.76	0.62	0.93
	Min	0.68	0.42	0.67	0.67	0.33	0	0	0
	Max	1	1	1	1	1	1	1	1
	Std.	0.08	0.14	0.08	0.06	0.15	0.29	0.33	0.25
UV_{CA}	Mean	0.85	0.9	0.77	0.88	0.58	0.71	0.38	0.73
	Min	0.72	0.75	0.56	0.67	0.17	0.33	0	0
	Max	1	1	1	1	0.83	1	1	1
	Std.	0.11	0.07	0.17	0.16	0.17	0.25	0.36	0.45
WUV_{PCA}	Mean	0.99	0.98	0.99	1	0.81	0.79	0.74	1
	Min	0.88	0.92	0.78	1	0.5	0	0	1
	Max	1	1	1	1	1	1	1	1
	Std.	0.03	0.03	0.05	0	0.19	0.27	0.31	0
WUV_{CA}	Mean	0.91	0.96	0.81	1	0.79	0.84	0.66	0.97
	Min	0.8	0.75	0.44	1	0.5	0.33	0	0
	Max	1	1	1	1	1	1	1	1
	Std.	0.07	0.07	0.14	0	0.12	0.2	0.29	0.18

Appendix 4. Classification performance for different clustering models

Table 55. Overall classification performance from different clustering models.

Metrics (training set)	Methods				
	K-means	AHCF	GMM	SOM	SC
Accuracy, Ac	0.765	0.7493	0.7076	0.7676	0.7546
Precision, PrO	0.7767	0.7742	0.6577	0.7843	0.7789
Specificity, SpO	0.7607	0.7414	0.7279	0.7616	0.7465
Recall, ReO	0.5442	0.4898	0.4966	0.5442	0.5034
F1 score, FsO	0.64	0.6	0.5659	0.6426	0.6116
Precision, PrW	0.7607	0.7414	0.7279	0.7616	0.7465
Specificity, SpW	0.7767	0.7742	0.6577	0.7843	0.7789
Recall, ReW	0.9025	0.911	0.839	0.9068	0.911
F1 score, FsW	0.8256	0.8175	0.7795	0.8279	0.8206
Metrics (testing set)					
Accuracy, Ac	0.75	0.75	0.7604	0.75	0.7708
Precision, PrO	0.6087	0.6087	0.6364	0.6087	0.64
Specificity, SpO	0.7945	0.7945	0.7973	0.7945	0.8169
Recall, ReO	0.4828	0.4828	0.4828	0.4828	0.5517
F1 score, FsO	0.5385	0.5385	0.549	0.5385	0.5926
Precision, PrW	0.7945	0.7945	0.7973	0.7945	0.8169
Specificity, SpW	0.6087	0.6087	0.6364	0.6087	0.64
Recall, ReW	0.8657	0.8657	0.8806	0.8657	0.8657
F1 score, FsW	0.8286	0.8286	0.8369	0.8286	0.8406

Appendix 5. Measured and predicted borehole lithologies as well as prediction performance

Table 56. Overall classification performance using original images.

Training set	Ac	Fs0	Fs1	Fs2	Re0	Re1	Re2	Sp0	Sp1	Sp2	Pr0	Pr1	Pr2	Score sum
RF-BY	99.216	99.466	98.906	97.676	99.916	98.336	95.456	99.856	99.066	99.866	99.026	99.476	100.005	77
SVC-BY	90.503	93.263	86.483	77.172	95.803	83.703	63.642	92.813	91.133	98.922	90.843	89.452	98.003	35
LGBM-BY	91.564	93.654	87.844	90.725	96.084	84.384	85.715	93.324	91.574	99.575	91.344	91.614	96.352	53
XGB-BY	88.111	92.091	83.251	49.071	94.851	81.031	34.421	91.091	89.621	98.071	89.491	85.601	85.481	13
BT-BY	92.955	95.135	90.035	78.743	97.475	87.575	64.943	95.705	93.205	98.963	92.915	92.655	100.005	59
GBM-BY	90.222	92.752	85.942	86.134	95.402	82.552	76.624	92.092	90.582	99.314	90.252	89.623	98.334	35
Testing set	Ac	Fs0	Fs1	Fs2	Re0	Re1	Re2	Sp0	Sp1	Sp2	Pr0	Pr1	Pr2	Score sum
RF-BY	85.052	90.133	78.782	26.831	93.744	75.852	16.671	88.775	86.942	97.571	86.802	81.953	68.753	31
SVC-BY	86.336	90.846	80.666	45.366	93.885	77.806	33.335	89.286	87.976	98.045	87.996	83.736	70.974	73
LGBM-BY	85.143	90.174	79.023	34.413	93.162	76.463	24.243	88.042	87.203	97.783	87.373	81.752	59.262	36
XGB-BY	85.274	90.132	79.354	32.942	93.021	77.324	21.212	87.831	87.584	97.702	87.424	81.491	73.685	36
BT-BY	85.635	90.305	79.875	38.204	93.162	77.685	25.764	88.103	87.805	97.824	87.615	82.194	73.916	57
GBM-BY	84.841	89.931	78.251	42.995	93.523	74.151	34.856	88.394	86.241	98.086	86.601	82.835	56.101	36

Table 57. Overall classification performance using CLAHE-enhanced images.

Training set	Ac	Fs0	Fs1	Fs2	Re0	Re1	Re2	Sp0	Sp1	Sp2	Pr0	Pr1	Pr2	Score sum
RF-BY	98.835	99.165	98.375	97.676	99.756	97.545	95.456	99.616	98.635	99.866	98.575	99.206	100.004	70
SVC-BY	90.992	93.863	87.382	69.672	95.493	86.262	55.192	92.533	92.362	98.682	92.283	88.532	94.442	30
LGBM-BY	94.074	95.694	91.724	87.864	96.644	90.864	79.874	94.564	94.904	99.404	94.764	92.604	97.623	51
XGB-BY	89.531	93.161	85.381	51.611	94.811	84.851	36.361	91.401	91.521	98.131	91.561	85.931	88.891	13
BT-BY	98.966	99.296	98.566	96.995	99.665	98.176	94.165	99.465	98.976	99.835	98.936	98.955	100.004	70
GBM-BY	91.243	93.642	87.673	82.443	95.122	86.363	70.133	91.962	92.443	99.123	92.192	89.013	100.004	36
Testing set	Ac	Fs0	Fs1	Fs2	Re0	Re1	Re2	Sp0	Sp1	Sp2	Pr0	Pr1	Pr2	Score sum
RF-BY	85.631	90.461	79.571	34.881	93.885	76.711	22.731	89.125	87.381	97.741	87.281	82.654	75.001	24
SVC-BY	89.105	92.726	85.076	50.004	94.026	85.125	34.854	90.206	91.616	98.094	91.466	85.026	88.465	69
LGBM-BY	88.044	91.785	83.475	58.006	92.874	83.414	43.946	88.394	90.665	98.356	90.725	83.525	85.293	62
XGB-BY	87.082	91.303	82.232	41.862	92.513	82.683	27.272	87.753	90.183	97.872	90.113	81.791	90.006	35
BT-BY	87.082	91.102	82.263	51.065	92.151	82.562	36.365	87.241	90.132	98.135	90.082	81.962	85.714	36
GBM-BY	87.213	91.324	82.434	47.313	92.442	82.683	33.333	87.682	90.214	98.043	90.234	82.183	81.482	40

Table 58. Comparison of performance indicators based on 30 random splits and on single split using LGBM-BY and CLAHE-processed images.

Indicator	Ac	Pr2	Re2	Fs2	Sp2	Pr1	Re1	Fs1	Sp1	Pr0	Re0	Fs0	Sp0
Training set													
Min	0.93	0.97	0.72	0.83	1.00	0.91	0.89	0.90	0.95	0.94	0.96	0.95	0.90
Max	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Mean	0.98	1.00	0.96	0.98	1.00	0.98	0.96	0.97	0.99	0.98	0.99	0.98	0.97
Std	0.03	0.01	0.08	0.05	0.00	0.03	0.04	0.04	0.02	0.02	0.01	0.02	0.04
Single-division	0.94	0.98	0.80	0.88	0.99	0.93	0.91	0.92	0.95	0.95	0.97	0.96	0.95
Testing set													
Min	0.87	0.60	0.27	0.39	0.99	0.82	0.79	0.81	0.90	0.89	0.92	0.91	0.81
Max	0.88	0.83	0.55	0.62	1.00	0.85	0.84	0.84	0.92	0.91	0.94	0.92	0.86
Mean	0.87	0.73	0.41	0.52	1.00	0.84	0.81	0.82	0.91	0.90	0.93	0.92	0.84
Std	0.00	0.06	0.06	0.06	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01
Single-division	0.88	0.85	0.44	0.58	0.98	0.84	0.83	0.83	0.91	0.91	0.93	0.92	0.88

Note: Min, Max, Mean, Std represents the minimum, maximum, average and standard deviation statistics from 30 random splits.

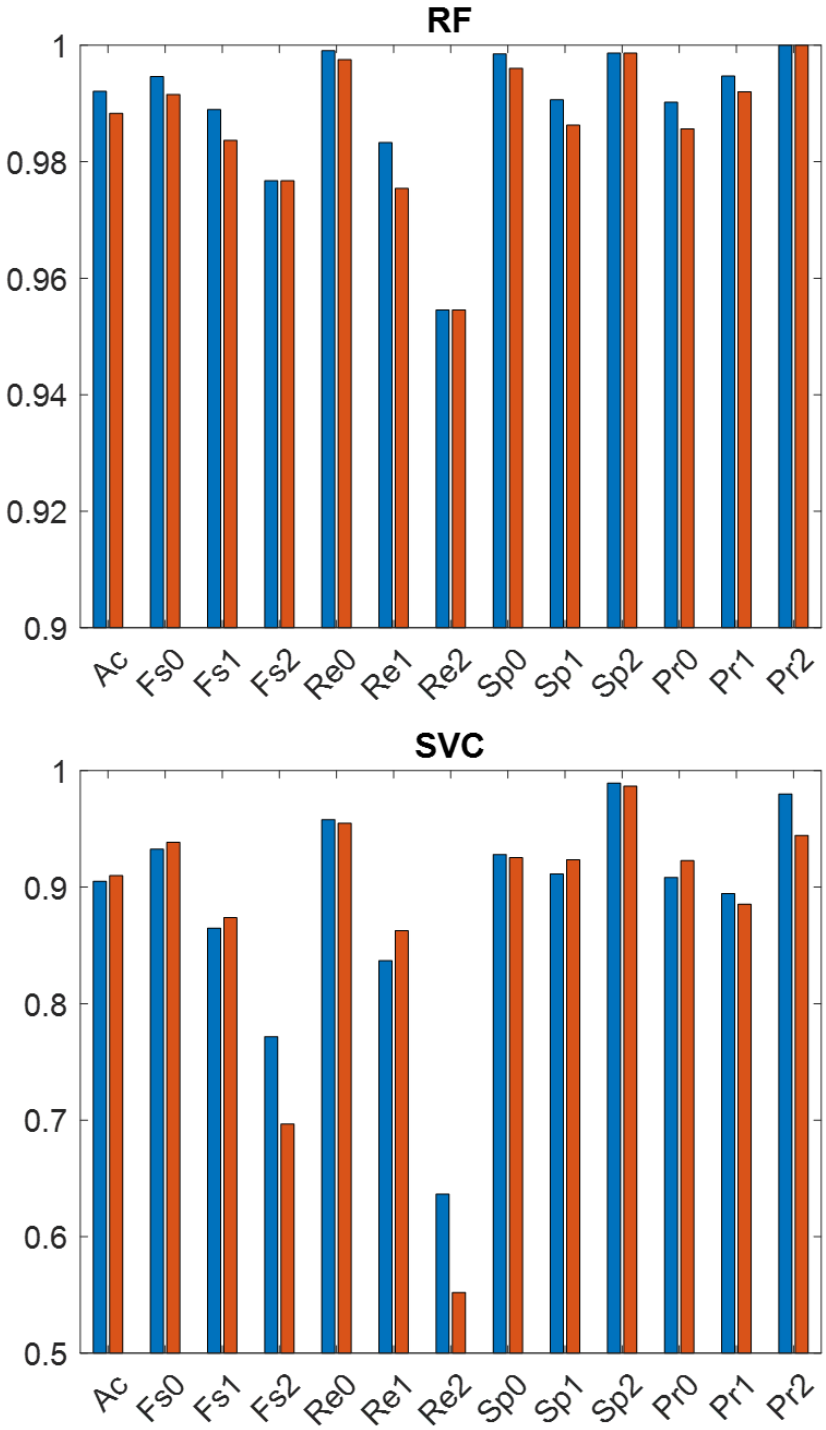


Figure 99. Training set performance of RF and SVC with original and CLAHE-processed inputs (represented by blue and orange, respectively), where 0, 1 and 2, represents the ML, BL and HC, respectively.

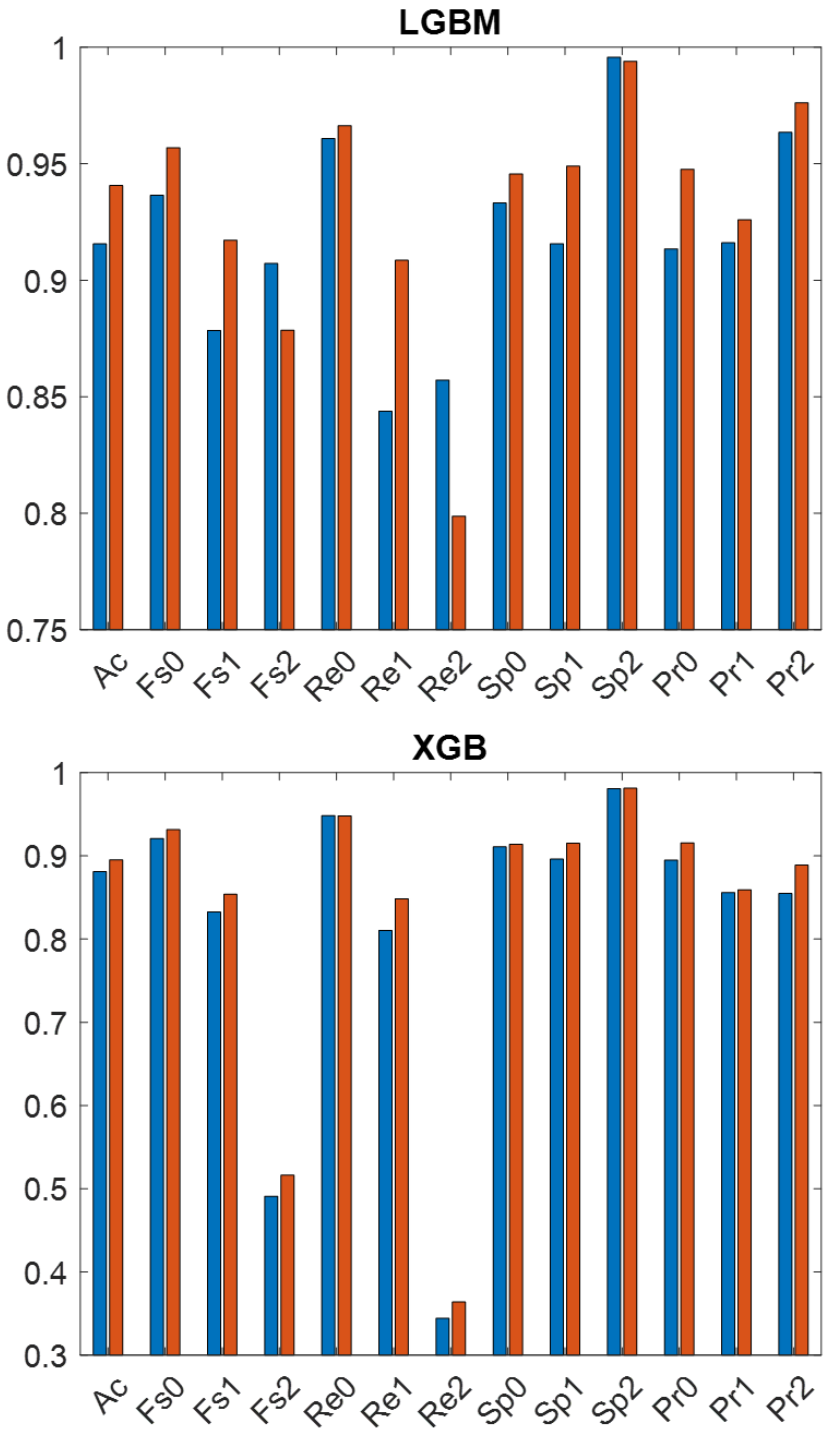


Figure 100. Training set performance of LGBM and XGB with original and CLAHE-processed inputs (represented by blue and orange, respectively), where 0, 1 and 2, represents the ML, BL and HC, respectively.

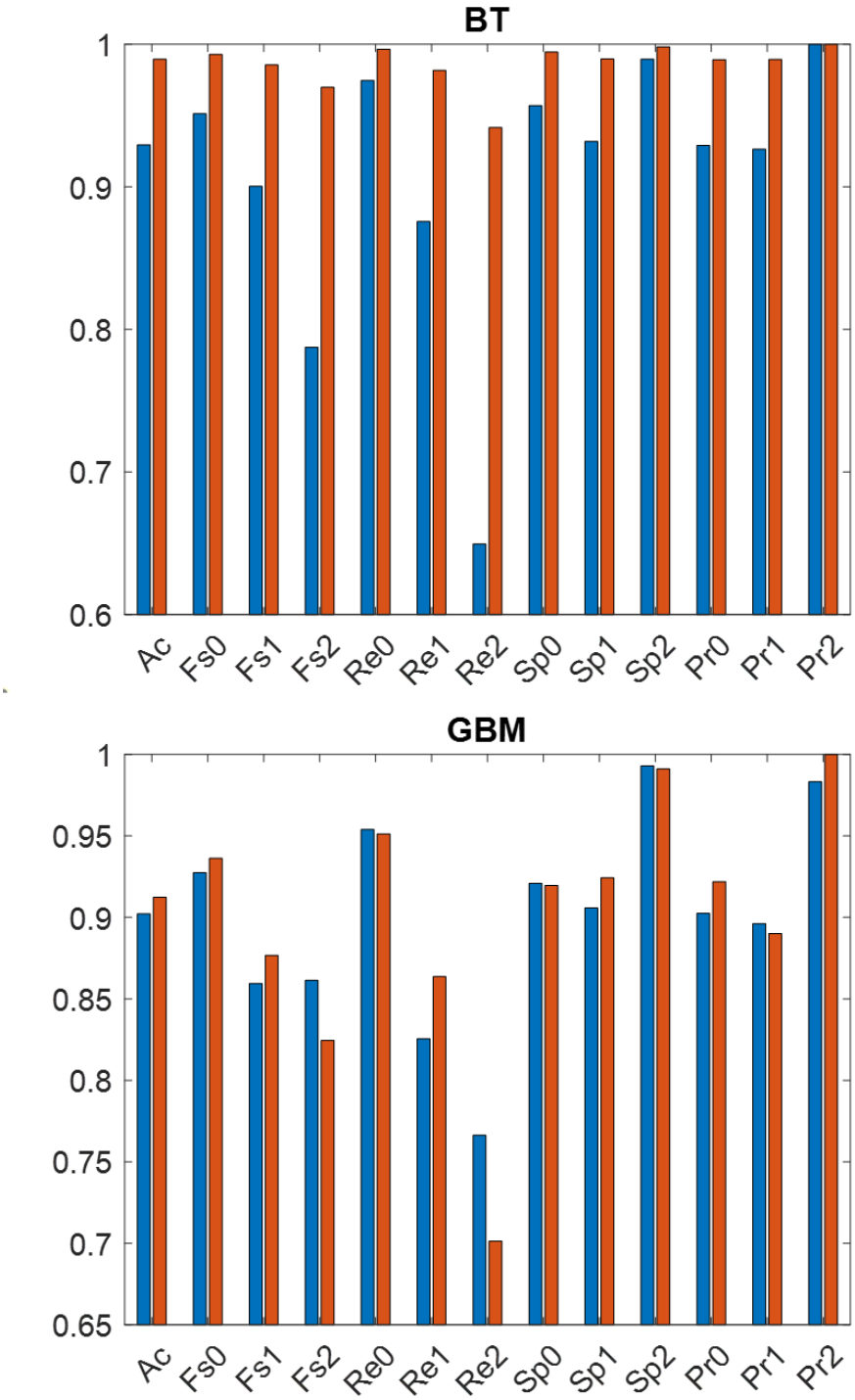


Figure 101. Training set performance of LGBM and XGB with original and CLAHE-processed inputs (represented by blue and orange, respectively), where 0, 1 and 2, represents the ML, BL and HC, respectively.

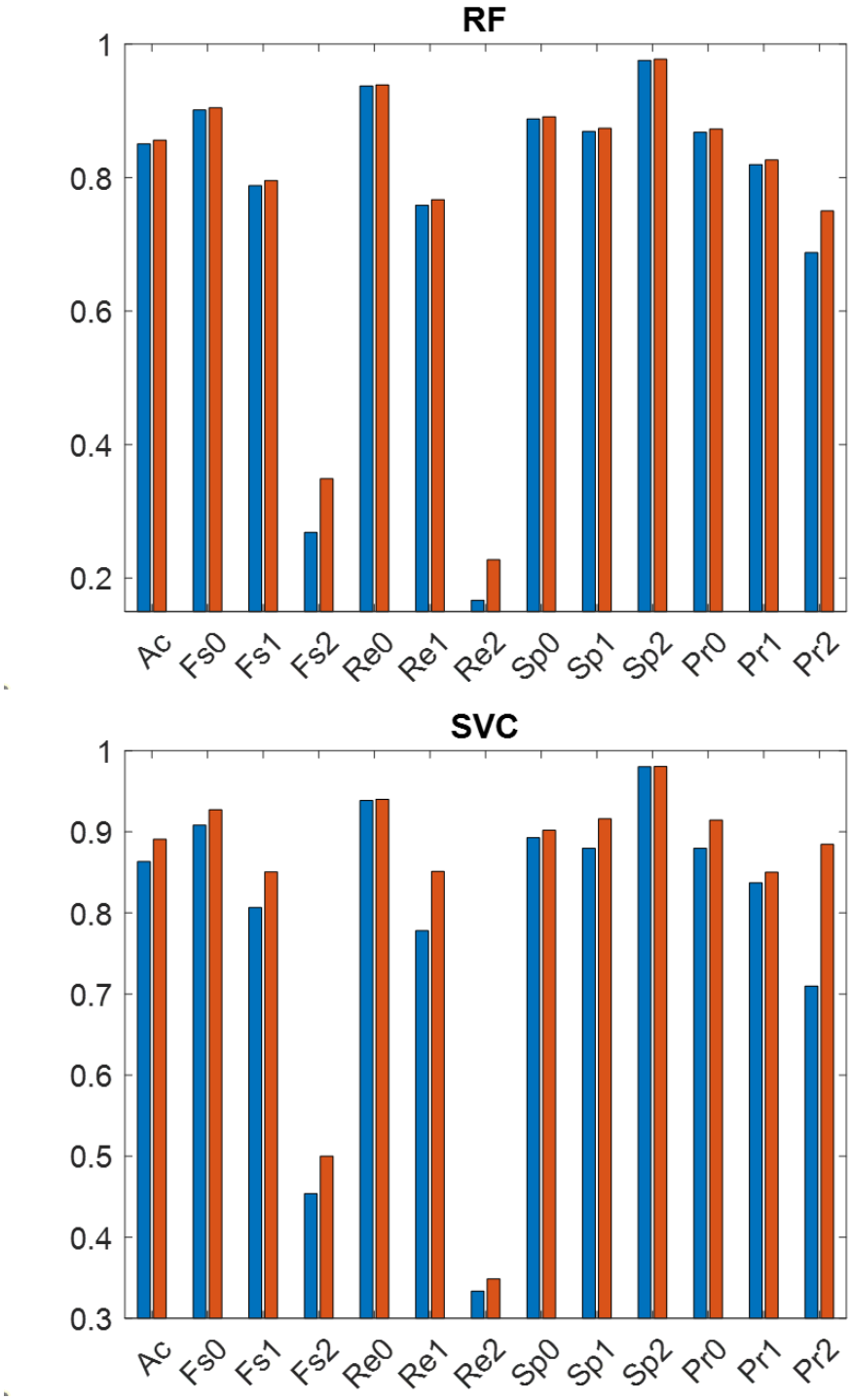


Figure 102. Testing set performance of RF and SVC with original and CLAHE-processed inputs (represented by blue and orange, respectively), where 0, 1 and 2, represents the ML, BL and HC, respectively.

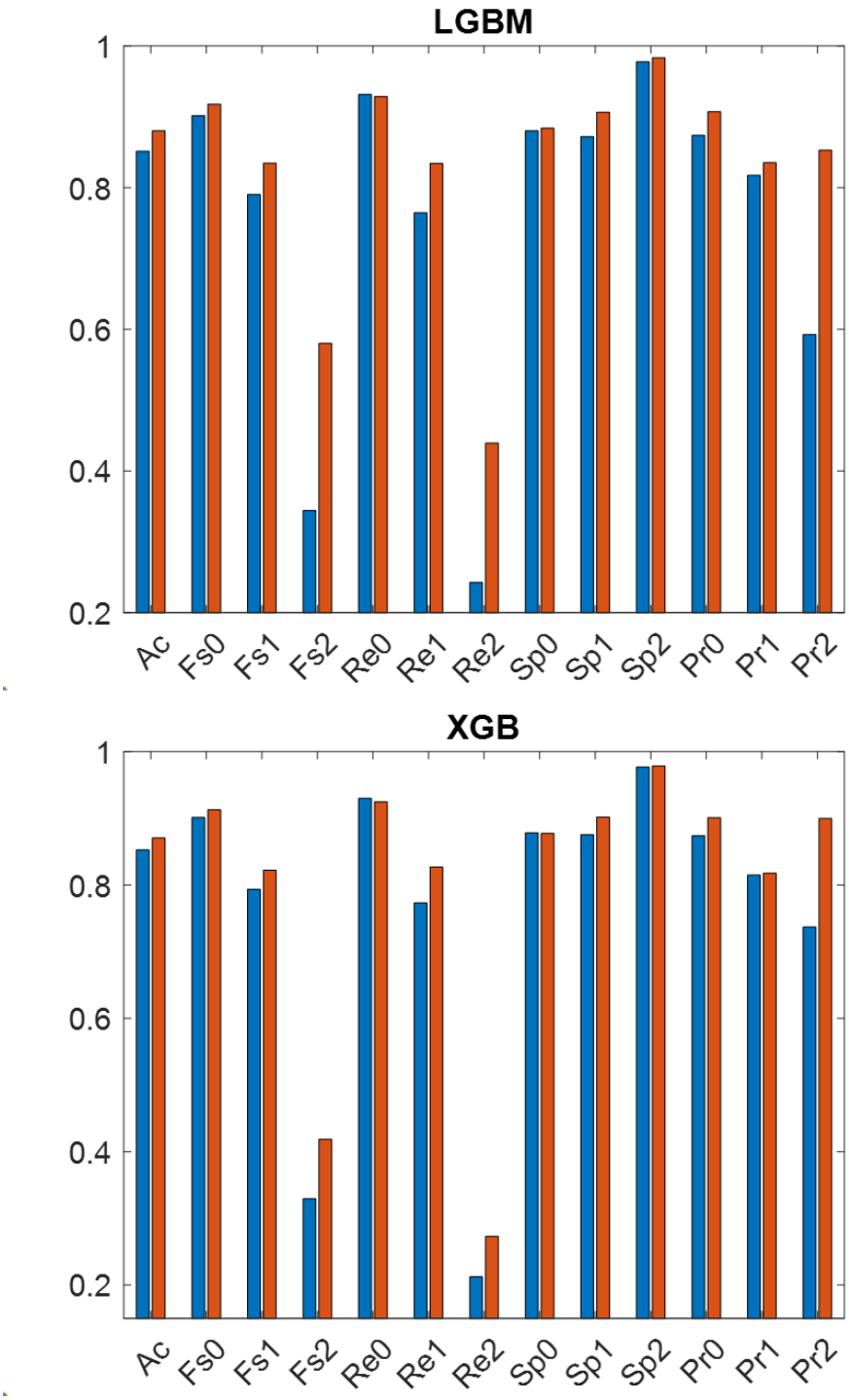


Figure 103. The testing set performance of LGBM and XGB with original and CLAHE-processed inputs (represented by blue and orange, respectively), where 0, 1 and 2, represents the ML, BL and HC, respectively.

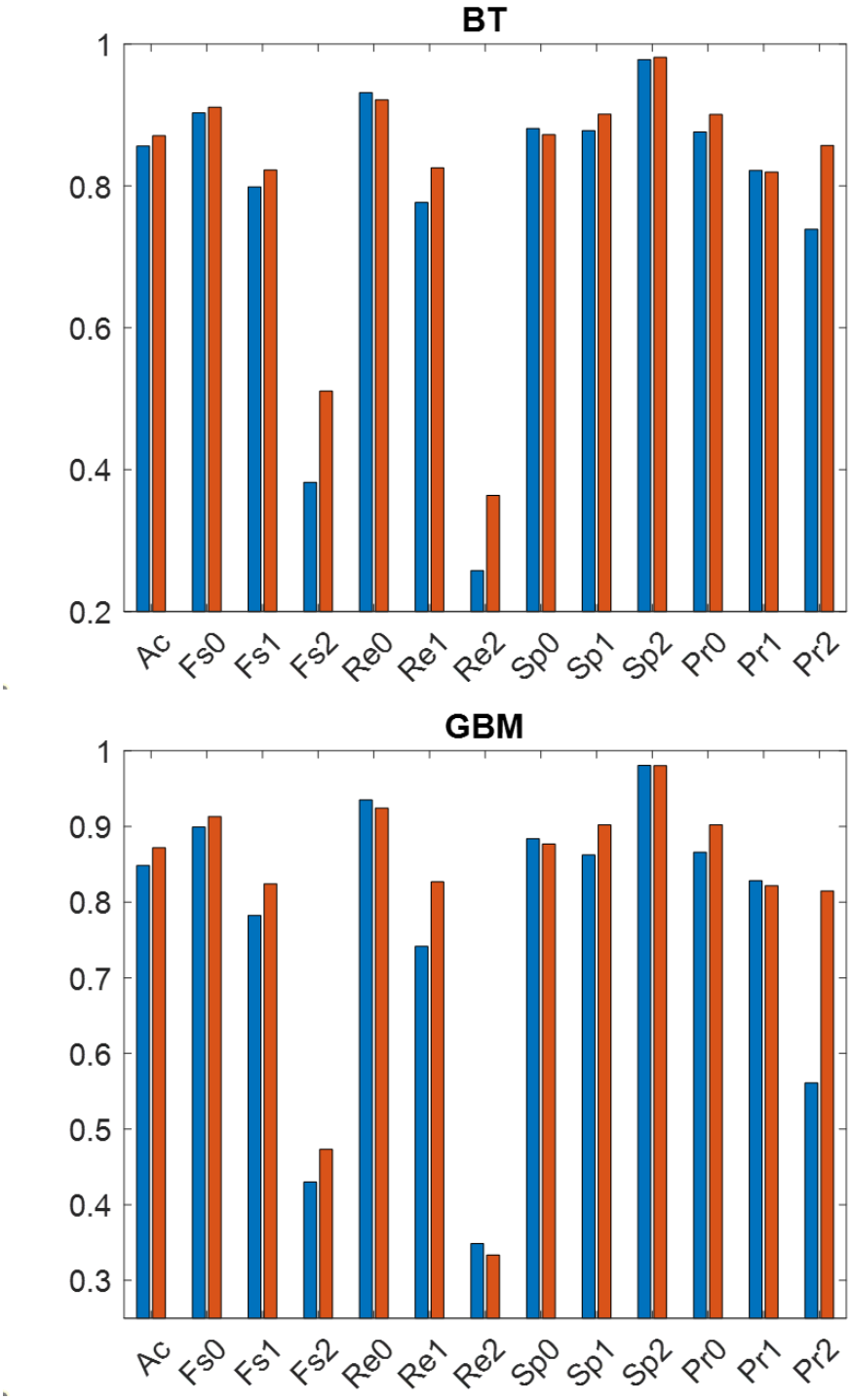


Figure 104. The testing set performance of BT and GBM with original and CLAHE-processed inputs (represented by blue and orange, respectively), where 0, 1 and 2, represents the ML, BL and HC, respectively.

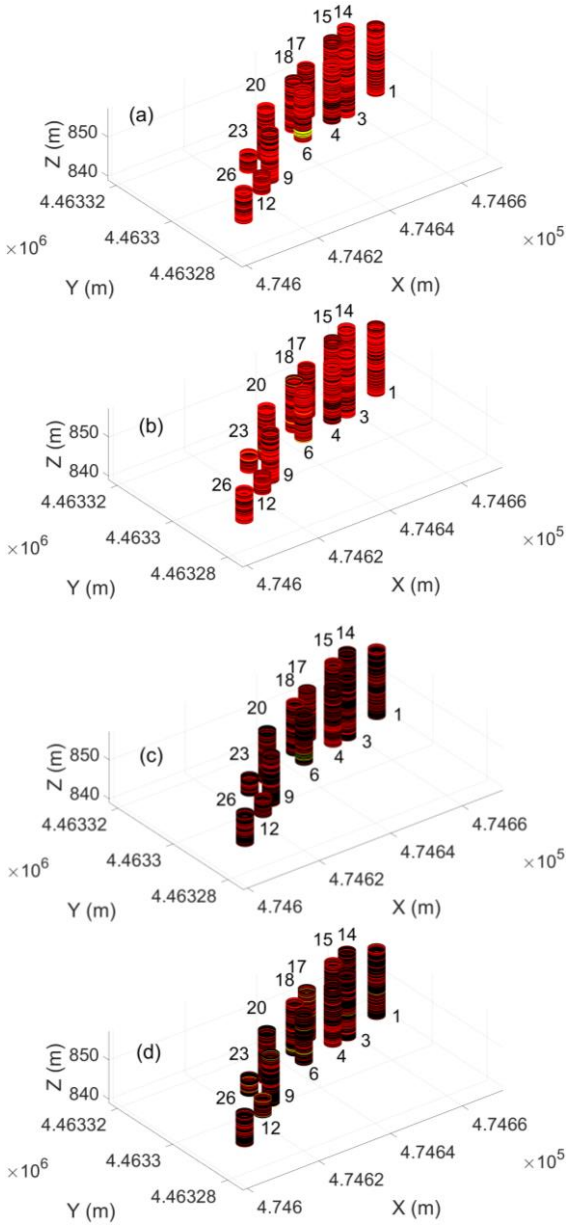


Figure 105. Comparison between measured and predicted lithology for B3, North pit, with CLAHE-processed inputs and LGBM-BY model: (a) training set, measured; (b) training set, predicted; (c) testing set, measured; (d) testing set, predicted. Red: ML; yellow: BL; green: HC; black: sections with no data or not chosen in the training or testing sets.

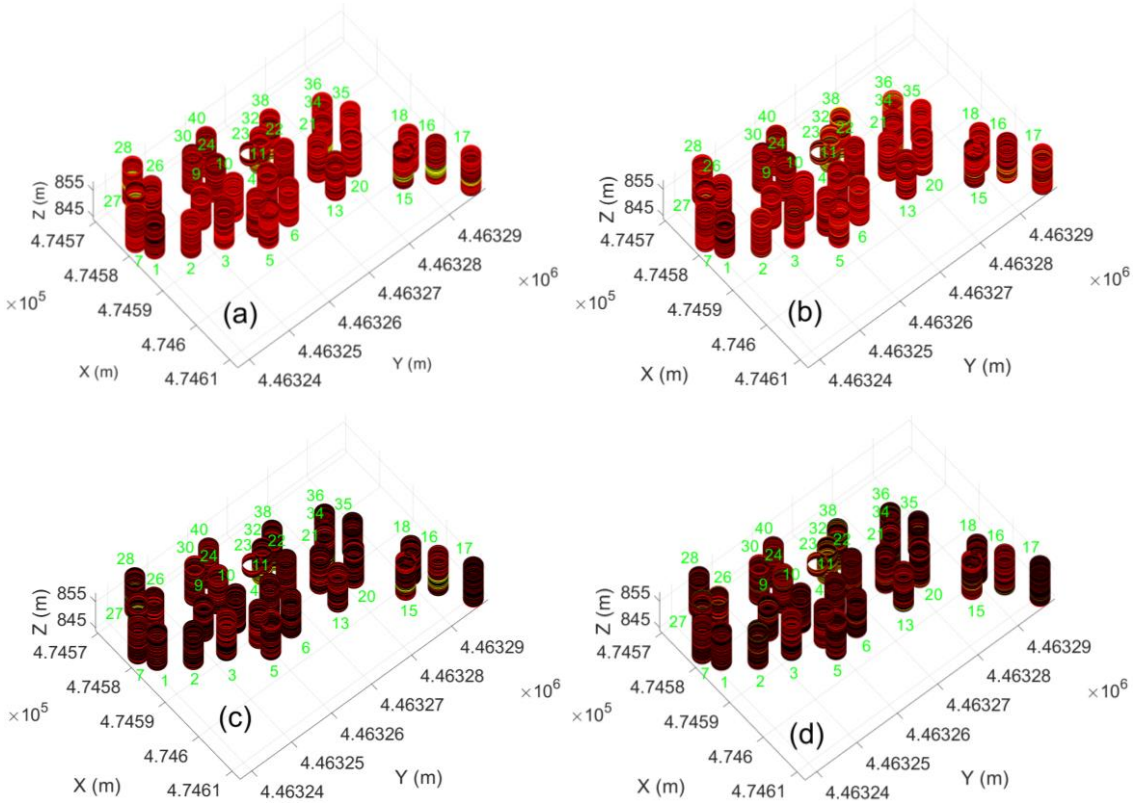


Figure 106. Comparison between measured and predicted lithology for B6, North pit, with CLAHE-processed inputs and LGBM-BY model: (a) training set, measured; (b) training set, predicted; (c) testing set, measured; (d) testing set, predicted. Red: ML; yellow: BL; green: HC; black: sections with no data or not chosen in the training or testing sets.

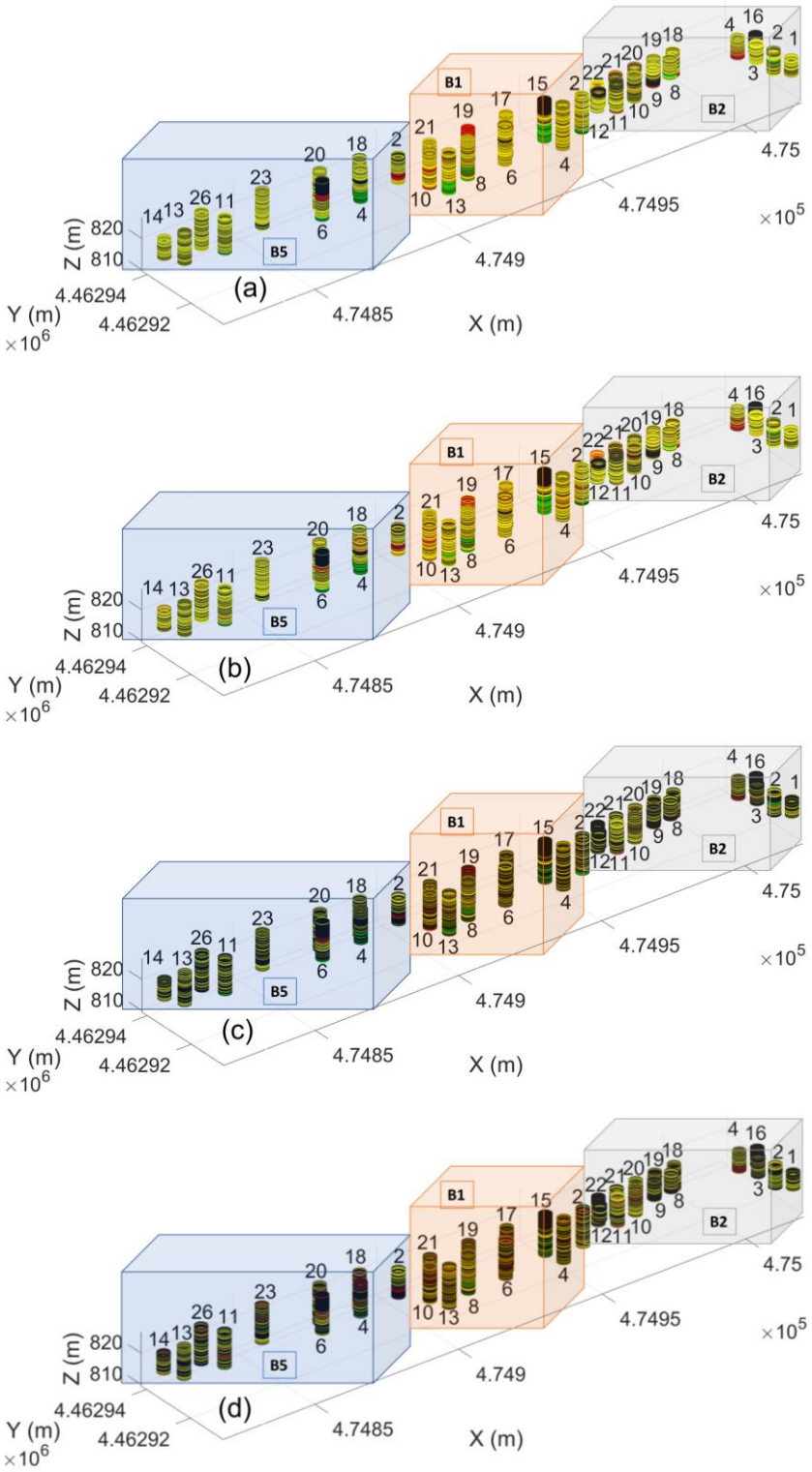


Figure 107. Comparison between measured and predicted lithology for the blocks of the NE pit with CLAHE-processed inputs and LGBM-BY model: (a) training set, measured; (b) training set, predicted; (c) testing set, measured; (d) testing set, predicted. Red: ML; yellow: BL; green: HC; black: sections with no data or not chosen in the training or testing sets.

Appendix 6. Paper A

Appendix 7. Paper B

Appendix 8. Paper C