Determining the lowest quantity of variables required for geometallurgical machine learning-based modelling: A review

E.E. Mack, B.P. von der Heyden, M. Tadie, and T.M. Louw

Stellenbosch University, South Africa

An important aspect of geometallurgy is predicting the metallurgical performance of mineral deposits, often including extraction and processing characteristics. The use of machine learning is growing as an industry tool for geometallurgical prediction, with random forest being one of the most frequently used algorithms. However, using all the available geometallurgical variables for prediction can lead to overfitting and poor forecasts Therefore, determining the minimum quantity of variables required for accurate process modelling should be prioritised. The SHapley Additive exPlanations (SHAP) analysis is an innovative technique that provides explanations for individual predictions made by machine learning models that describe the process. It calculates the contribution of each variable to the prediction outcome, providing a ranking of feature (variable) importance. Based on SHAP analysis results, influential variables are retained to determine accuracy variance for the secondary model. The proposed method is hypothesised to produce acceptable performance of the secondary model compared to the primary model. This paper highlights the usefulness of SHAP analysis in determining the most important variables for prediction modelling. By using this technique, reduction in the number of variables required for accurate prediction can be utilised, resulting in more efficient and effective geometallurgy prediction models.

Keywords: Machine learning, random forests, dimensionality reduction, geometallurgy

INTRODUCTION

Machine learning is a field of study which can improve mineral resource extraction and beneficiation processes through regression and classification schemes. The combination of machine learning with geometallurgical modelling remains an important yet understudied topic. Geometallurgical machine learning-driven prediction models have gained traction in, for example, iron and gold mining sectors and there is scope to further apply these approaches to other commodities (e.g., Mn, Sn, etc.). Machine learning often produces overfitted results and therefore the optimisation of the computation process is critical to allow for efficiency without compromising forecast accuracy. There are two intersecting issues within the geometallurgical machine learning field, the abundance of data (or lack thereof) and the suitability of the algorithm. This paper focuses on the latter issue. An effective method to avoid overfitting is by removing variables that have little to no influence on the model; known as feature selection. This allows for influential variables to be retained to allow for accurate model reproduction, whereas non-influential variables are removed from the modelling process, as well as feature selection decision making on whether costly/difficult to measure data is required or not. By creating a framework for determining the process of removing the non-influential variables, the overall efficiency of machine learning approaches applied to geometallurgical problems may be streamlined.

Non-linear relationships between the ore's processing, geological and metallurgical variables create complex mathematical modelling challenges for which differing machine learning algorithms produce variable responses. The process of reducing the number of variables is often referred to as dimension reduction, which involves decreasing the number of variables relative to the number of samples. Using machine learning to correctly understand the relationships between variables is key to determining the minimum quantity of variables required for accurate geometallurgical modelling. During dimensionality reduction, variables cannot be randomly discarded since each variable has a level of influence on the output of the model, and the level of this influence may act individually on the model result, or through more complex interplay with other variables (e.g. derived variables). For example, upgrading of grade in a process is achieved by implementing specific metallurgical methods controlled by the ore type (e.g., magnetic separation, flotation, gravity separation etc.). Optimisation of the beneficiation process can occur when the metallurgical methods are correctly updated, models calibrated and specifically set according to continuously updated feed material, and relevant output material characteristics. Continuous, up-to-date data describing the input, processing and output properties of known ore allows for real-time updating of the model, allowing for effective fine-tuning of the geometallurgical processing stream. The vast quantity of data required for real-time updating makes human-based interpretation and modelling ineffective on the required timescale, therefore optimised machine learning algorithms applied to geometallurgical modelling are regarded as an advantageous step in maximising production yields.

A brief review of machine learning applied to geometallurgy

Decision trees, K-nearest neighbour, random forests, neural networks and support vector machines are commonly utilised machine learning algorithms in the geological and metallurgical fields (Dumakor-Dupey & Arya, 2021). These machine learning algorithms provide an alternative to fundamental and phenomenological models; however, they need to be calibrated using relevant data. Table I summarises several of these machine learning-based techniques, highlighting their relative advantages and disadvantages, and providing relevant references to their application in geological and geometallurgical problems. The remainder of this section interrogates a selection of case studies in which application of machine learning protocols demonstrably improves the geometallurgical workflows. The final part of this section synthesises the key take-home messages and learnings derived from the review.

Table I. Most commonly used machine learning algorithms for geometallurgical datasets

Name	Description	Advantages	Disadvantages	References
Decision Trees	Popular technique used in geometallurgy. Tree-like structure representing decisions and possible consequences. Divide data into smaller subsets based on data attributes.	Easy to interpret, handle categorical and numerical data and also used to model complex systems. Tree-based algorithms are strong nonlinear feature handlers.	Often overfit data, decision boundaries often unstable, imbalanced worksheets produce performance issues.	• (Nobahar *et al.*, 2022) – Fleet selection in a Kaolin Mine • (Rodriguez-Galiano *et a*l., 2015) – Gold prospectivity modelling.
Random Forests	Combines multiple decision trees to make more accurate predictions. Constructs multiple decision trees and combines predictions.	Handles categorical and numerical data, handles missing and outlying data, used for both	Computationally heavy, difficult to interpret, can overfit noisy datasets.	• (Sheng *et al.*, 2015) – Iron ore classification • (O'Brien *et al.*, 2014) – Polymetallic gahnite

	Used for both regression and classification.	classification and regression. Tree-based algorithms are strong nonlinear feature handlers.		classification • (Bhuiyan *et al.*, 2019) – Gold geometallurgical relationship models.
Support Vector Machines (SVMs)	Type of supervised learning algorithm used for regression and classification analysis. Find the hyperplane that maximises the margin between two classes.	Can handle both nonlinear and linear datasets, works well on small datasets, both regression and classification.	Computationally heavy, choice of kernel function subjective, sensitive to choice of hyperparameters. Are feature space-geometry aware, therefore complications arise with untransformed data.	• (Ahmad *et al.*, 2018) – Intrusion detection • (Chatterjee & Bandopadhyay, 2011) – Platinum grade estimation • (Taboada *et al.*, 2007) – Slate quality evaluation.
Artificial Neural Networks (ANNs)	Use layers of interconnected nodes to learn patterns in data. Used for classification and regression problems.	Handle both linear and non-linear datasets, learn complex patterns, both classification and regression.	Computationally heavy, prone to overfitting, architecture and hyperparameters difficult to optimise. Commonly overfit data due to predominantly being stuck in local minima. The approximation ability also is dependent on the activation function type.	• (Wu & Zhou, 1993) – Reserve estimation • ((Al-Alawi & Tawo, 1998) – Bauxite resource evaluation • (Kaplan & Topal, 2020) – Gold grade prediction.
K-Nearest Neighbours (KNN)	Type of supervised algorithm used for classification and regression. Finds K closest data points to test point using labels to predict label of test point. KNN is non-parametric, meaning it doesn't make assumptions about distribution of data.	Simple, classification and regression, doesn't make assumptions about distributions.	Computationally heavy for large datasets, sensitive to choice of K and distance metric, requires large amount of memory to store training data set. The algorithm is feature space-geometry aware, causing complications	• (Kaplan & Topal, 2020) – Gold grade prediction • (Koch *et al.*, 2019) – Porphyry copper drill core mineralogical characterisation.

			with untransformed data.	
Naïve Bayes	Probabilistic algorithm for classification problems. Computes probability of each class using input data and selecting class with highest probability. Based on Bayes theorem, assumes features are independent.	Simple, both categorical and numerical data, works well on high-dimensionality datasets.	Assumes features are independent, does not perform well on datasets with strong dependencies between features, requires large amount of training data to estimate probabilities accurately.	• (Köse *et al.*, 2012) – Ore mineral segmentation and quantification.
Gradient Boosting	Used for regression and classification problems, working by iteratively adding decision trees to model and adjusting weights of data points to improve predictions.	Both categorical and numerical data, can handle missing and outlying data, both regression and classification.	Computationally expensive, overfit on noisy datasets, choice of hyperparameters can be difficult. The algorithm is feature space-geometry aware, causing complications with untransformed data.	• (Prades & Deutsch, 2016) – Acid consumption, recovery and impurity of copper ore processing.

Evaluation of Neural Networks, Random Forest, Regression Trees and Support Vector Machines (Rodriguez-Galiano et al.*, 2015)*

Rodriguez-Galiano *et al.,* (2015) used artificial neural networks, regression trees, random forests and support vector machines to predict mineral prospectivity and produce possible target locations for likely mineral deposits. The four different algorithms used epithermal gold data comprised of 46 gold occurrence locations, where the input data (variables) were composed of known mineralised structures, geochemical surveys, exploited deposits, gravity and magnetic surveys as well as geological information. Although this study focuses on mineral prospectivity, geometallurgical responses and properties can be taken into account at a later stage to allow for feasibility studies and possible mine planning once specific outcomes such as production targets have been identified. The performance measure used to compare the four algorithms is the Kappa value. Kappa values are a statistical measure that indicates how much agreement there is between the predicted and observed classifications when judgments or evaluations are made, taking into account what would be expected by chance alone. Random forests and support vector machines produced the largest Kappa values of 0.92 and 0.87 respectively, whereas artificial neural networks and regression trees produced Kappa values of 0.77 and 0.66 respectively. The four algorithms were compared using the accuracy of delineation of prospective areas, sensitivity to estimation of hyper-parameters, sensitivity to input data size and interpretability of model parameters. The results generated from each algorithm indicated that random forests displayed higher stability and robustness (ability to maintain accuracy regardless of the level of data noise and variability) while also outperforming the other algorithms due to having the highest measured accuracy. A conclusion by Rodriguez-Galiano *et al.*, (2015) was that the decision tree-based algorithms like regression trees and random forests involve a less complicated training process compared to

artificial neural networks and support vector machines, concluding that the random forest algorithm produced the most accurate results while being easier to train. By reducing the number of input variables according to their relevance and impact on the output, a faster model production turnaround time can be achieved.

Performance Improvements during Mineral Processing Using Material Fingerprints Derived from Machine Learning – A Conceptual Framework (van Duijvenbode* et al.*, 2020)
Van Duijvenbode *et al.*, (2020) defined a data-driven link between an ore's characteristics and its processing behaviour as a "material fingerprint". Material fingerprints are machine learning-based classification schemes utilising material attributes as a source of ore differentiation within the range of attributes (chemical composition, texture and mineralogy) of a mine's mineral reserves. By utilising geometallurgical ore attributes as a means of ore characterisation, the improved classification of ore types may improve processing behaviour forecasts. Van Duijvenbode *et al.*, (2020) state that a larger dataset of ore characteristics upgrades the fingerprint classification while improving confidence and providing higher fidelity in the characterisation process. However, larger datasets increase processing time while also increasing costs related to the data acquisition process. The study further stated that new data improves confidence and fidelity of the fingerprints. Therefore, determining which variables impact the machine learning output (i.e., fingerprints) allows a faster computational processing time and provides insights into which variables should be prioritised in the data acquisition process.

Integrating Geometallurgical ball mill throughput predictions into short-term stochastic production scheduling in mining complexes (Both & Dimitrakopoulos (2022))
Both & Dimitrakopoulos (2022) presented a process that integrated a throughput prediction model for a ball mill into a short-term stochastic production schedule. The aim of the study was to increase revenue for the mining operation by decreasing mining costs, processing costs and rehandling costs while increasing the metal recovery rate. The geometallurgical data used for the multiple linear regression prediction model were comprised of geological domains, material types, blast hole drilling data, rock density and mill throughput rates providing a large dataset with numerous material and processing characteristics. The primary challenge for linking the simulated ore characteristics to the observed production data output at the processing plant related to tracking of material throughout the processing facilities. The aim of characterising ore based on its attributes allows it to be matched with the later processing response. Integration of a throughput prediction model into a stochastic short-term mine production scheduling model involves not only the ore characteristics as mentioned, but also comminution and scheduling data. The vast amount of data to model a scheduling plan can create significant challenges within the modelling process. By incorporating all variables in each modelling stage, non-influential data can result in skewed outputs with increased uncertainty in the ore characterisation and comminution stages.

RECOMMENDATIONS

From the three case studies presented above, and from those summarised in Table 1, it is clear that machine learning models may be hindered by the vast amounts of input data that enables overfitting . This may slow down the prediction and characterisation processes while also increasing data acquisition costs. Utilisation of machine learning for geometallurgical data is increasing, although there are few optimisation methods being applied to these algorithms and associated geometallurgical data. This highlights the need to process data during the pre- and syn-modelling phases. Here we present SHAP analysis as a tool to optimise machine learning-based geometallurgical workflows.

SHapley Additive exPlanations Analysis

SHapley Additive exPlanations Analysis (SHAP) analysis ranks features (also referred to as variables) according to their influence on any utilised machine learning model (Lundberg & Lee, 2017). Kumar *et al.,* (2020) utilised SHAP analysis to reduce dimensionality where a shapley value is assigned to each variable once the model has run. Values assigned to each variable are based on influence and impact on the model, where variables with least impact can be filtered out. SHAP analysis aims to decompose predictions from the machine learning model into the contribution of each feature. By breaking down

the predictions into the contributions, an understanding of how the model reaches the predicted output can be achieved. SHAP analysis does not remove variables with least influence, but rather provides a way to quantify contributions of each feature regardless of the feature's level of influence.

Kumar *et al*,. (2020) recommends that the model should be rebuilt using a new dataset comprised of only the influential variables in order to calculate the accuracy of the model where the number of variables could decrease by 10-20%, dependant on the dataset. The accuracy of the new model should not vary greatly from the original model, where the performance of both predictions should be measured by utilising performance measures such as error rate, sensitivity, precision, accuracy, standard deviation, and other mathematic performance measures dependant on the type of model. The sum of the contributions of all features within the final prediction of the test model can be expressed as:

$$f(x) = g(z') = \varphi_0 + \sum_{i=1}^{K} \varphi_i z'$$

where $f(x)$ is the predicted value of the model and g is the explanatory model (training model), and $z' \in \{0,1\}^k$, K is the dimension of the input feature, φ_0 is the average predicted value of the model and φ_i is the contribution of the i-th feature (Shapley value) (Liu & Liu, 2022). It is generally recommended to scale the features before computing SHAP values. Scaling the features ensures that the contributions of the features are on a comparable scale and prevents the influence of features with larger numerical ranges from dominating the results. SHAP values do not indicate feature importance, rather quantify influence of the feature on the model's output. Therefore, features with low importance by themselves may have a high influence on the model when paired with other features. This underlines the importance of SHAP values and why influence on the output of a model is critical to understand. Once features are ranked using SHAP analysis, SHAP-based feature selection is used to preferentially select the influential variables according to the desired criteria of selection which is moderated by known fundamental and/or phenomenological relationships governing the specific section of the mining operations under investigation. Once the features have been selected, a new model must be trained to compare performance after cross-validation techniques have been applied to the new model.

An advantage of SHAP analysis is that the process is not restricted to only one type of machine learning algorithm, but instead can be applied to virtually any model that can produce a prediction using a set of input features. However, the effectiveness of SHAP analysis as a tool to determine the influential variables is shouldn't be viewed as dependent on the model. This is an important feature of SHAP analysis, as the machine learning algorithm can be tailored to the field of study and use SHAP analysis post-prediction.

Implications and advantages

The reduction of variables based on influence on the machine learning model within a geometallurgical dataset has two main advantages: model overfitting and data acquisition cost decreases. These can be summarised as:

Model overfitting

By eliminating unnecessary input variables in a machine learning algorithm, we can mitigate the risk of overfitting. Rather than focusing on computational cost, an alternative strategy involves optimising the machine learning process by carefully selecting the most influential features. By reducing the number of input features, the model becomes more adept at identifying patterns and avoids allocating excessive attention to variables that offer little or no impact on the modelled task. Consequently, this proactive reduction in the model's complexity reduces the risk of overfitting, ensuring a more efficient and effective learning process.

Data acquisition costs

By strategically selecting and minimising the input variables used in data collection and analysis processes, resource allocation can be optimised. This approach enables a more focused and efficient data acquisition process, where only the most relevant and influential variables are considered.

Consequently, unnecessary expenses related to acquiring and processing large amounts of data that may have limited impact on the desired outcomes can be avoided. By streamlining the data acquisition process through variable reduction, cost savings can be achieved while still obtaining valuable insights for geological exploration, metallurgical analysis, and geometallurgical modelling.

CONCLUSION

Incorporating variable reduction techniques, such as those based on SHAP analysis, is important for improving and enhancing machine learning-based geometallurgical modelling. From the aforementioned studies, random forests display consistent accuracy in the modelling process and therefore pose a suggestion of which machine learning algorithm should be applied with geometallurgical data. By identifying and focusing on the most influential features through SHAP analysis, data collection and analysis can be streamlined, thereby reducing costs. This approach ensures that resources are allocated efficiently, targeting only the variables that have a significant impact on the desired outcomes. Through effective feature selection, mining operations can optimise their decision-making processes, leading to improved profitability and more accurate geometallurgical models. Ultimately, the integration of variable reduction techniques based on SHAP analysis empowers mining companies to make informed decisions, capitalise on valuable insights, and drive greater efficiency and profitability in their operations.

REFERENCES

Ahmad, I., Basheri, M., Iqbal, M. J., & Rahim, A. (2018). Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. *IEEE Access*, *6*, 33789–33795. https://doi.org/10.1109/ACCESS.2018.2841987

Al-Alawi, S. M., & Tawo, E. E. (1998). Application of Artificial Neural Networks in Mineral Resource Evaluation. *Journal of King Saud University - Engineering Sciences*, *10*(1), 127–138. https://doi.org/10.1016/S1018-3639(18)30692-5

Bhuiyan, M., Esmaieli, K., & Ordóñez-Calderón, J. C. (2019). Application of data analytics techniques to establish geometallurgical relationships to bond work index at the Paracutu Mine, Minas Gerais, Brazil. *Minerals*, *9*(5). https://doi.org/10.3390/min9050302

Both, C., & Dimitrakopoulos, R. (2022). Integrating geometallurgical ball mill throughput predictions into short-term stochastic production scheduling in mining complexes. *International Journal of Mining Science and Technology*, *33*(2), 185–199. https://doi.org/10.1016/j.ijmst.2022.10.001

Chatterjee, S., & Bandopadhyay, S. (2011). Goodnews Bay Platinum Resource Estimation Using Least Squares Support Vector Regression with Selection of Input Space Dimension and Hyperparameters. *Natural Resources Research*, *20*(2), 117–129. https://doi.org/10.1007/s11053-011-9140-6

Dumakor-Dupey, N. K., & Arya, S. (2021). Machine learning—a review of applications in mineral resource estimation. *Energies*, *14*(14). https://doi.org/10.3390/en14144079

Kaplan, U. E., & Topal, E. (2020). A new ore grade estimation using combine machine learning algorithms. *Minerals*, *10*(10), 1–17. https://doi.org/10.3390/min10100847

Koch, P. H., Lund, C., & Rosenkranz, J. (2019). Automated drill core mineralogical characterization method for texture classification and modal mineralogy estimation for geometallurgy. *Minerals Engineering*, *136*(May 2017), 99–109. https://doi.org/10.1016/j.mineng.2019.03.008

Köse, C., Alp, I., & Ikibaş, C. (2012). Statistical methods for segmentation and quantification of minerals in ore microscopy. *Minerals Engineering*, *30*, 19–32. https://doi.org/10.1016/j.mineng.2012.01.008

Kumar, C. S., Choudary, M. N. S., Bommineni, V. B., Tarun, G., & Anjali, T. (2020). Dimensionality Reduction based on SHAP Analysis: A Simple and Trustworthy Approach. *Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020*, 558–560. https://doi.org/10.1109/ICCSP48568.2020.9182109

Liu, J. J., & Liu, J. C. (2022). Permeability Predictions for Tight Sandstone Reservoir Using Explainable Machine Learning and Particle Swarm Optimization. *Geofluids*, *2022*(2). https://doi.org/10.1155/2022/2263329

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-Decem*(Section 2), 4766–4775.

Nobahar, P., Pourrahimian, Y., & Mollaei Koshki, F. (2022). Optimum Fleet Selection Using Machine Learning Algorithms—Case Study: Zenouz Kaolin Mine. *Mining*, *2*(3), 528–541. https://doi.org/10.3390/mining2030028

O'Brien, J. J., Spry, P. G., Nettleton, D., Xu, R., & Teale, G. S. (2014). Using Random Forests to distinguish gahnite compositions as an exploration guide to Broken Hill-type Pb-Zn-Ag deposits in the Broken Hill domain, Australia. *Journal of Geochemical Exploration*, *149*, 74–86. https://doi.org/10.1016/j.gexplo.2014.11.010

Prades, C. F., & Deutsch, C. V. (2016). *Comparison of Machine Learning Techniques for Predicting and Learning from Geometallurgical Multivariate Databases*. 1–20. http://www.ccgalberta.com

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Sheng, L., Zhang, T., Niu, G., Wang, K., Tang, H., Duan, Y., & Li, H. (2015). Classification of iron ores by laser-induced breakdown spectroscopy (LIBS) combined with random forest (RF). *Journal of Analytical Atomic Spectrometry*, *30*(2), 453–458. https://doi.org/10.1039/c4ja00352g

Taboada, J., Matías, J. M., Ordóñez, C., & García, P. J. (2007). Creating a quality map of a slate deposit using support vector machines. *Journal of Computational and Applied Mathematics*, *204*(1), 84–94. https://doi.org/10.1016/j.cam.2006.04.030

Van Duijvenbode, J. R., Buxton, M. W. N., & Soleymani Shishvan, M. (2020). Performance improvements during mineral processing using material fingerprints derived from machine learning—A conceptual framework. *Minerals*, *10*(4). https://doi.org/10.3390/min10040366

Wu, X., & Zhou, Y. (1993). Reserve estimation using neural network techniques. *Computers and Geosciences*, *19*(4), 567–575. https://doi.org/10.1016/0098-3004(93)90082-G

Ethan Mack

Stellenbosch University

My name is Ethan Eric Mack and am currently 23 years old, originally from northern Kwazulu-Natal and am now currently completing my MSc Earth Science degree focusing on geometallurgy at Stellenbosch University. I am currently focusing on applying machine learning to improve manganese ore characterization and subsequent mine planning through the African Rainbow Minerals Chair at Stellenbosch University.