

Data-driven optimization of brittleness index for hydraulic fracturing

Lei Hou ^{a,*}, Jianhua Ren ^b, Yi Fang ^c, Yiyang Cheng ^b

^a School of Engineering, The University of Warwick, Coventry CV4 7AL, UK

^b Research Institute of Exploration & Development, East China Company of SINOPEC, Nanjing 210011, China

^c Department of Geology and Geological Engineering, South Dakota School of Mines and Technology, Rapid City 55701, USA

Abstract

Evaluation of brittleness index (BI) is a fundamental principle of a hydraulic fracturing design. A wide variety of BI calculations often baffle field engineers. The traditional value comparison may also not make the best of BI. Moreover, it is often mixed up with the fracability in field applications, thus causing concerns. We, therefore, redefine fracability as the fracturing pressure under certain rock mechanical (mainly brittleness), geological and injecting conditions to clarify the confusion. Then, we propose a data-driven workflow to optimize BIs by controlling the geological and injecting conditions. The machine learning (ML) workflow is employed to predict the fracability (fracturing pressure) based on field measurement. Three representative ML algorithms are applied to average the prediction, aiming to restrict the interference of algorithm performances. The contribution of brittleness on pressure/fracability prediction by error analysis (rather than the traditional method of BI-value comparison) is proposed as the new criterion for optimization. Six classic BI correlations (mineral-, logging- and elastic-based) are evaluated, three of which are optimized for the derivation of a new BI using the backward elimination strategy. The stress ratio (ratio of minimum and maximum horizontal principal stress), representing the geological feature, is introduced into the derived calculation based on the independent variable analysis. The reliability of the new BI is verified by error analyses using data of eight fracturing stages from seven different wells. Approximately 40%~50% of the errors are reduced based on the new BI. The differences among the performances of algorithms are also significantly restrained. The new brittleness index provides a more reliable option for evaluating the brittleness and fracability of the fracturing formation. The machine learning workflow also proposes a promising application scenario of the BI for hydraulic fracturing, which makes more efficient and broader usages of the BI compared with the traditional value comparison.

Keywords: brittleness index; fracability; hydraulic fracturing; machine learning; pressure prediction

* Corresponding author.

E-mail address: lei.hou@warwick.ac.uk (L. Hou).

1 Introduction

Brittleness index (BI), an essential parameter to evaluate the mineral, mechanical and failure characteristics of formations, is commonly applied in mining (e.g. rock cutting) and petroleum engineering (e.g. drilling and fracturing) (Kim, Walton, Larson, & Berry, 2021; Meng, Wong, & Zhou, 2020; Tarasov & Potvin, 2013). For hydraulic fracturing, it is an essential criterion for the optimization of “sweet point” that is easy to crack and tends to create complex fracture networks – namely the evaluation of fracability (how easy the reservoir can be cracked) (Arora & Mishra, 2015; Li, Xie, & Wang, 2020; Mullen & Enderlin, 2012). Initiating from the mineral component analysis, the BI calculations for hydraulic fracturing are involving more and more rock mechanical and engineering features (Iyare, Blake, & Ramsook, 2020; Zhang, Ranjith, & Perera, 2016).

Dozens of BI calculations are proposed for various formations with unique lithological and elastic properties, which could be roughly classified into (i) the mineral-based brittleness index (MBI) (Buller et al., 2010; Jarvie, Hill, Ruble, & Pollastro, 2007), (ii) the logging-based brittleness index (LBI) (Xiaochun Jin, Shah, Truax, & Roegiers, 2014), and (iii) the elastic-based brittleness index (EBI) (Rickman, Mullen, Petre, Grieser, & Kundert, 2008; Sharma & Chopra, 2015). Current efforts evaluate the BI calculations mainly based on numerical simulation (Xia et al., 2019), logging data (Mews, Alhubail, & Barati, 2019), and static tests (Feng, Zhang, Rezagholilou, Roshan, & Sarmadivaleh, 2019). Limited test at field-practical scales is presented, thus restricting the reliability and persuasion of these BI calculations. However, the successful applications of machine learning techniques in petroleum engineering have paved a potential data-driven approach for BI optimization at field-practical scales (Lin, Kang, Oh, & Canbulat, 2020; D. Sun et al., 2020).

The primary application of BI is to evaluate the fracability for hydraulic fracturing. However, the BI is often mixed up with the fracability in many fracturing applications, which has caused concerns about misleading the brittleness and ductility recognition of the targeting formation (Bai, 2016; Herwanger, Bottrill, & Mildren, 2015; Mathia, Ratcliffe, & Wright, 2016). Some researchers believe that the BI is more like a rock type indicator. The geological properties should also play an important role in the evaluation of fracability, for instance, the in-situ confining stress, pore pressure, natural

fractures and faults (Zhiqi Guo, Li, Liu, Feng, & Shen, 2013; Heidari, Khanlari, Torabi-Kaveh, Kargarian, & Saneie, 2013; Ye, Tang, & Xi, 2020). To distinguish fracability from BI, we propose that fracability should be an even broader concept that involves the injection parameters (pump rate, fluid and proppant properties) as well.

Therefore, we redefine fracability as the fracturing pressure, a more comprehensive response of the cracked reservoir under certain geological, rock mechanical and hydraulic injecting conditions. Then, a machine learning (ML) workflow is proposed for optimizing BIs by predicting the fracability/pressure based on field measurements. The prediction errors are analyzed to evaluate the contribution of each BI via a control variate method (controlling the interferences of geological, hydraulic and ML algorithm conditions). A new BI is derived by integrating the geological feature based on the optimization, variable analysis and backward elimination strategy, which significantly improves the predictions of fracability/pressure compared with classic BIs. Our optimization work clarifies the confusion between BI and fracability, promotes the calculation of BI and provides a more reliable option of BI for the recognition of targeting formation and design of fracturing schedule. It also demonstrates a promising application scene of BI based on the data-driven approach, which takes better advantage of BI than the traditional value comparisons.

2 Methodology

The field measurements of hydraulic fracturing are collected and used as the Basic dataset for comparisons. By appending the brittleness index to the Basic dataset, the BI datasets are built for data processing. Three classic machine learning algorithms, Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Artificial Neural Network (ANN) are used for data processing to mitigate the interference of algorithm performances on the predicted fracability. The mean absolute error (MAE) and root mean squared error (RMSE) are deployed for the prediction analyses and BI optimization. The backward elimination strategy and variable analysis are applied for the derivation of the new BI.

2.1 Data preparation

We collect 76 stages of shale gas fracturing and logging measurements (11860 groups of data in minute-level) in 7 wells (denoted as W1 to W7) from the Sichuan Basin (Longmaxi shale gas formation), China, as shown in Table 1. Each group of data involves parameters of fluid types (hydrochloric acid – HCL, slickwater, and gel), pump rate, proppant concentration, proppant mesh, and fracturing pressure. The logging interpretations for calculating BIs contain mineral components (Clay, Quartz, Carbonate and total organic carbon – TOC), Young’s modulus, Poisson ratio, density, porosity, wave velocities and compressional slowness (DTC).

Table 1 Division of training and predicting stages for data processing workflow

Well No.	W1	W2	W3	W4	W5	W6	W7
For Training / Number of stages	14	17	17	20	/	/	/
For Testing / Stage no.	Stage 3	Stage 5	Stage 7	Stage 9	Stage 11	Stage 18	Stages 10 & 15
Notes	/	/	/	/	Neighbour well	Distant well	Different formation

The training and predicting stages are specially designed (Table 1). There are 68 stages from wells W1~W4 used for algorithm training. The first four predicting stages (Stage 3, 5, 7 and 9) are selected successively from the same wells (W1~W4), aiming to control the uncertainties from geological heterogeneity in the shale formation. As a result, the optimization of BI based on the prediction is more reliable by minimizing the interference of geological uncertainties. Four more predicting stages are randomly selected from the first and second halves of the fracturing stages in two new wells, as shown in Table 1. W5 and W3 are neighbour wells. W6 is a distant well far from all others. W7 drills through different subdivided formations from all other wells. We choose representative stages from the first and second halves of a fracturing operation to restrict the potential influence of human factors on fracturing pressures because operators usually learn experiences from precedents and perform better in subsequent fracturing stages.

The categorical data encoding and feature scaling are carried out during the data pre-processing. There are three fluid types and one special fluid condition (the pump off, denoted as N/A) in our dataset, which are string variables and can not be used for algorithm training directly. Therefore, the

fluid type is encoded by the OneHotEncoder and transformed into a combination of three binary digits (i.e. 0, 0 and 1) after deleting the dummy variable (Bisong, 2019). The fluid type used to be one column of strings is then transformed into three columns of binary code when they are fed into the algorithm. The proppant type is presented by its averaged diameter. All input variables are standardized by removing the mean and scaling to unit variance. The datasets, inputs and outputs are summarized in Table 2. The pump rate, fluid type, proppant concentration, averaged proppant diameter and BI are used as inputs for the algorithm training, in which the hydraulic parameters are set as control variables. The machine learning algorithm is trained by inputs to predict the fracturing pressure, then are compared with field records to carry out error analyses. Both of the Basic and BI datasets contain all stages in Table 1.

Table 2 Summary of datasets, inputs and outputs for the data processing workflow

	Inputs (Control Variate Method)		Outputs	Error Analyses (Based on field records)	
	Brittleness	Hydraulic Parameters		Output Error	Error Improvement by BI
Basic Dataset	/	Pump rate; fluid type; Proppant concentration and mesh.	Pressure (Baseline)	Baseline Error	/
BI Dataset	Brittleness Index	Pump rate; fluid type; Proppant concentration and mesh.	Pressure (BI-based)	BI-based Error	BI optimization and derivation

2.2 Brittleness index for optimization

Six representative BI calculations are selected from MBI, LBI and EBI categories for evaluations, summarized in Table 3. BI₁ and BI₂ are widely used for shale gas fracturing. BI₃ are empirical equations based on data from the shale gas wells in the Permian Basin (the US). The sum of the calculations is utilized for BI₃. BI₄ is the energy dissipation per unit area during the process of new fracture creation. The fracture toughness (*K_{IC}*) represents the resistance of rock to fracture propagation. BI₅ and BI₆ are deformations and extensions of BI₂. The calculations in Table 3 are integrated into the machine learning workflow for data pre-processing. Each BI is calculated by original data and fed into the algorithm as an input feature. Then, the prediction errors based on each correlation are analyzed and used as the criterion for BI optimization.

Table 3 Summary of the BI correlations for evaluation

No.	Correlation	Category	Original parameters collected for calculations	References
BI ₁	$BI = \frac{Qtz + Carb}{Qtz + Carb + Cly + TOC}$	Mineral-based	<i>Qtz</i> –quartz; <i>Carb</i> –carbonate; <i>Cly</i> –clay; <i>TOC</i> –total organic carbon	(Glorioso & Rattia, 2012)
BI ₂	$BI = \frac{1}{2}(E_n + \nu_n)$	Elastic-based	E_n –normalized Young’s modulus; ν_n –normalized Poisson ratio	(Rickman et al., 2008)
BI ₃	$BI = -0.012DTC + 1.4921$ $BI = -1.5314\phi + 0.8575$	Logging-based	<i>DTC</i> –compressional slowness; ϕ –porosity	(Xiaochun Jin et al., 2014)
BI ₄	$BI = (1 - (\nu)^2) * (\frac{KIC^2}{E}) * 10^3$	Elastic-based	E –Young’s modulus; ν –Poisson ratio; <i>KIC</i> –fracture toughness	(X. Jin, Shah, Roegiers, & Zhang, 2015)
BI ₅	$BI = \frac{E\rho}{\nu}$	Elastic-based	ρ –density; E –Young’s modulus; ν –Poisson	(S. Sun et al., 2013)
BI ₆	$\frac{E}{\lambda} = \frac{\nu_s^2(3\nu_p^2 - 4\nu_s^2)}{(\nu_p^2 - 2\nu_s^2)(\nu_p^2 - \nu_s^2)}$ $BI = \frac{(E/\lambda)_n + \nu_n}{2}$	Logging-based	ν_p –P-wave velocity; ν_s –S-wave velocity; λ –Lame’s first parameter; E –Young’s modulus; ν –Poisson ratio	(Pan, Zhang, & Chen, 2020)

2.3 Machine learning algorithm and workflow

Three different machine learning algorithms are deployed to restrict the potential interferences induced by the algorithm performances. The MLR is used for both pressure prediction and variable analysis. The SVR, with a radial basis function (RBF) kernel, is capable of both linear and non-linear regression (Al-Anazi & Gates, 2010; Zhenyu Guo et al., 2018). A three-layer ANN (input, hidden and output layers) is constructed under the Tensorflow frame with 20 neural units in the first and hidden layers referring to previous research (Mahmoud et al., 2017; Shen, Cao, Ruddy, & Teixeira de Moraes, 2020). The dropout regularization (the rate of 0.2) is set after the first and hidden layers to avoid overfitting (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). The Adam routine is selected as the optimizer to compile the model, where a callback function is applied to return and automatically update the learning rate (Kingma & Ba, 2014). Other hyperparameters are optimized by the Grid search and K-fold cross-validation (Bengio & Grandvalet, 2004; Bergstra & Bengio, 2012), including the training epochs (30) and batch size (50).

The data processing workflow, integrating BI calculations and machine learning algorithms, is illustrated in Fig. 1. The original data is pre-processed by the calculations in Table 3 for BIs. Considering the division design in Tables 1 and 2, the geological and hydraulic effects on fracturing

pressure/fracability are restricted, thus boosting the dominance of BI on predictions. Each BI-based workflow processes the original measurements respectively. The predicted pressures based on the Basic (donated as the Baseline prediction) and BI datasets are compared with field records for error analyses and BI optimization. The remaining errors are improved by the derivation of a new BI based on the backward elimination strategy. The geological features are selected based on the variable analysis, then introduced into the new BI calculation for promotions. Finally, the workflow is executed again to verify the performance of the new BI.

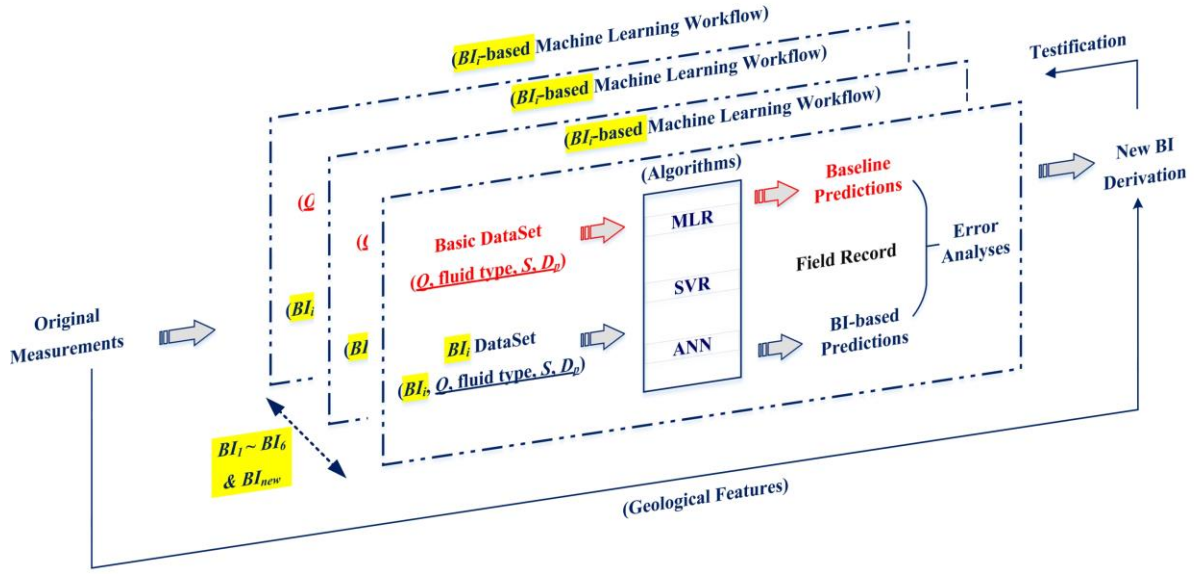


Fig. 1. Schematic of the data-driven workflow for BI optimization and derivation. Q , S , D_p represent the pump rate, proppant concentration and averaged proppant diameter, respectively.

3 Results

The prediction of pressure/fracability based on the Basic dataset is set as the reference. The error between the reference and field-recorded pressure is denoted as the baseline error, and is used to measure the promotion based on the BI dataset. The error improvement is applied as the criterion of BI optimization, and is quantified by the mean absolute error (MAE) and root mean squared error (RMSE). Based on the optimization, a new BI equation is derived *via* variable analysis and backward elimination strategy, which is then verified based on field applications.

3.1 Predictions of fracturing pressure

The predicted fracturing pressure/fracability based on different algorithms are presented in Fig. 2 using Stage 5 from W2 as an example. The predictions based on BI₃ (the blue dashed curve) and only

hydraulic parameters (the red dashed curve) are compared with field measurements (the solid black curve). Generally, predictions based on BI₃ fluctuate around the field curve and exhibit higher precision than baseline predictions. BI₃ introduces the logging data and porosity of the targeting formation, thus improving the interpretation of fracability and reducing the error significantly.

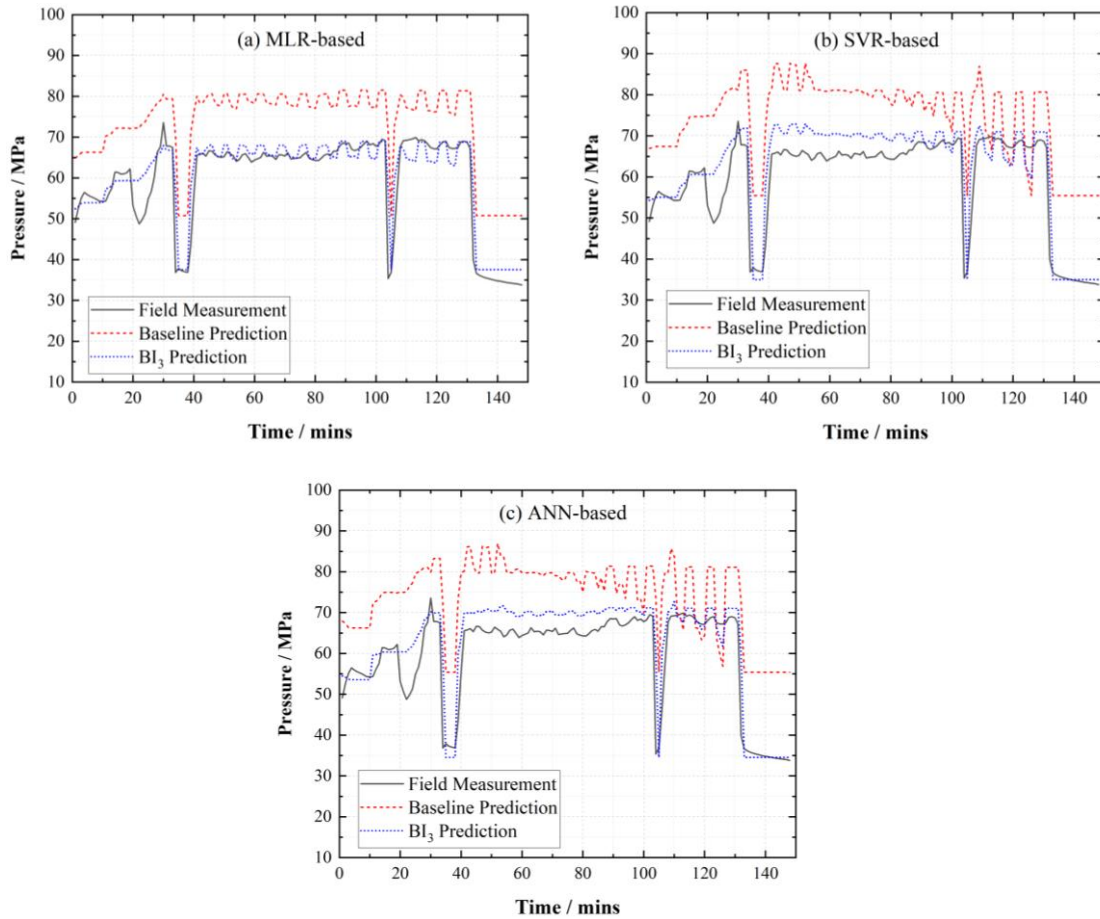


Fig. 2. The predictions of fracturing pressure (fracability) produced by MLR, SVR, and ANN algorithms using Stage 5 from W2 as an example. The red dashed curve represents the prediction based on the Basic dataset and is used as a baseline. The blue dashed curve is the result based on the BI₃-processed dataset. The solid black curve is the original field record.

The BI₃-based workflow (used as an example) is deployed to process the data from all eight testing stages. The errors of the testing stages are presented in Appendix A. The testing errors are averaged and used to characterize the error improvement by introducing the BI₃, as shown in Fig. 3. The averaged MAE and RMSE produced by different algorithms based on the Basic dataset are 0.22 and 15.1, respectively, which are reduced down to 0.18 and 12.8 based on the BI₃. The maximum values of the error are mainly obtained in the last four stages from W5~W7 (Appendix A).

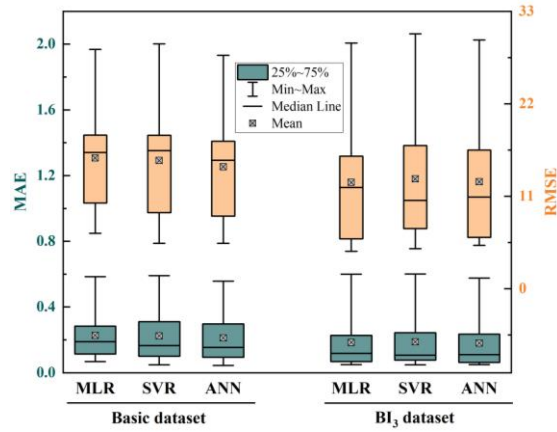


Fig. 3. The comparison of averaged errors (eight testing stages) produced by MLR, SVR and ANN algorithms based on the Basic and BI₃ datasets.

3.2 Error analyses and BI optimization

The error analyses in Table 4 (Appendix A) and Fig. 3 are conducted five more iterations based on BI_i-based workflow (Fig. 1). The algorithms are trained by BI_i datasets ($i=1, 2, 4, 5$ and 6), respectively. The errors are analyzed and presented in Fig. 4, where the error produced by the Basic dataset is used as a reference. Among all the BIs, BI₆ yields the largest MAE and RMSE, followed by BI₄. The performances of BI₂ and BI₃ are comparable, which reduces around 12%~16% of the errors. BI₁ performs slightly better in controlling errors than the performances of BI₂ and BI₃. BI₅ produces the minimum errors compared with others, and diminishes the MAE and RMSE by 20% and 18%, respectively. Besides, the performance of the logging-based BI₃ is unexpected since the equations are fitted using data from a different basin.

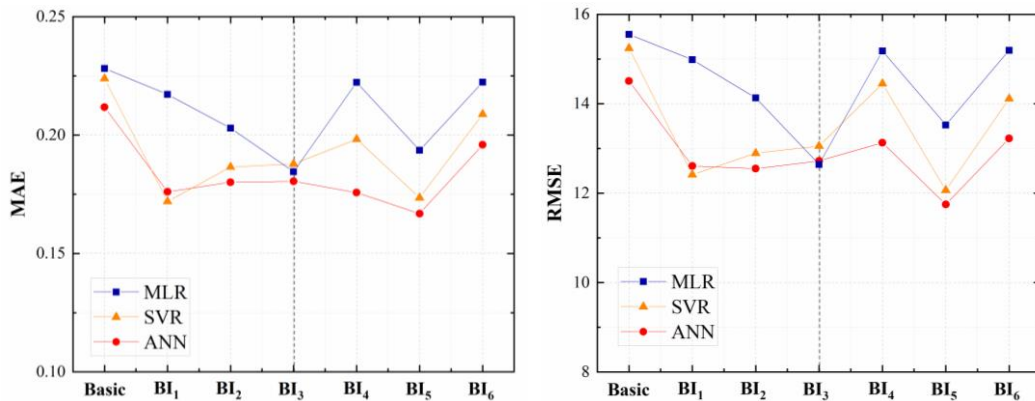


Fig. 4. BI evaluation by comparing averaged errors based on eight testing stages. The blue square, orange triangle and red circle represent results produced by MLR, SVR and ANN algorithms, respectively.

The performances of ANN and SVR are better than that of the MLR algorithm for most of the predictions, as shown in Fig. 4. Therefore, the predictions based on ANN and SVR algorithms are used for the stage-by-stage error analysis. Noteworthy, an appropriate BI may restrain the difference in the algorithm performances as the results based on BI₃ in Fig. 4 (vertical dashed line). The results based on the BI₅-based workflow (with the minimum errors) are expanded as examples for the stage-by-stage analysis among different wells, as shown in Fig. 5.

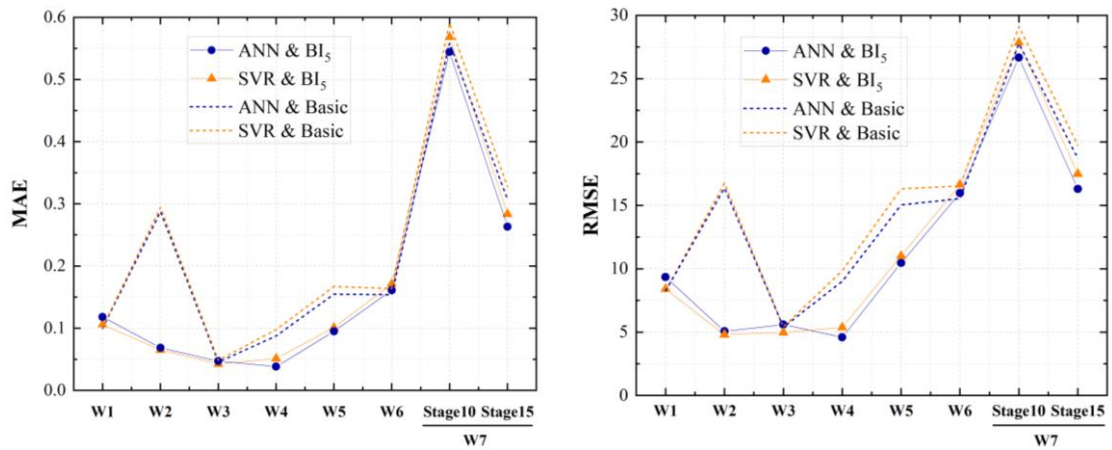


Fig. 5. Stage-by-stage analysis of MAE and RMSE among the testing wells based on the BI₅-based workflow. The dashed curves are the errors based on the Basic dataset. The solid triangle and circle represent the errors based on the BI₅ dataset and different algorithms (SVR and ANN).

The MAEs of the first four stages are controlled under 10% because the algorithms are trained by data from the same wells (W1~W4). The errors of W5 mildly increase, which may be explained by the adjacent relationship between W5 and W3, sharing similar geological properties. The algorithms produce the largest errors for W6 and W7, wells from a far distance and drilled through a different subdivided formation, respectively. The error improvements based on classic BIs mainly occur in the first four wells. Therefore, a new BI is necessary to promote its interpretation capacity for broader deploying scenarios, then improve the errors for W6 and W7. The BI₁, BI₂, BI₃, and BI₅ perform better according to the error analyses (Fig. 4) and are optimized for the new BI derivation.

3.3 New BI derivation

The new BI is designed to involve as many characteristics as possible to deliver more features into the algorithm training, thus improving the prediction. The backward elimination strategy (based on SVR) and independent variable analysis (based on MLR) are employed to optimize the expression

and new parameters. The derivation process is explained in Appendix B. The expression of the new BI is given by

$$BI_{New} = \frac{(BI_{1n} + BI_{3n} + BI_{5n})}{3} \times \frac{Stress_{min}}{Stress_{max}} \quad (9)$$

where BI_{in} is the normalized BI_i ($i=1, 3$, and 5), in which the mineral component (BI_{1n}) exerts essential contribution according to the backward elimination (Appendix B, Table 5); $Stress_{min}$ and $Stress_{max}$ are minimum and maximum horizontal principal stress, and are optimized by variable analyses (Appendix B, Table 6).

The application range of the new BI is constrained by the data source used for the derivation and verification. It is proposed based on field measurements from the shale gas wells in the Sichuan Basin, China. Broader assessments based on different basins are currently infeasible due to the data limitation. However, the BI_3 fitted by the field data from the US exerts relatively high performance in processing the data from China (Fig. 4). We believe that the universality of the new BI (Eq. 9, involving BI_3) can also be extensive.

3.4 Performance of New BI

The performance of the new BI (Eq. 9) is demonstrated by averaging the errors based on eight testing stages and three algorithms (MLR, SVR, and ANN), as shown in Fig. 6. The single $BI_1 \sim BI_6$ decreases MAE and RMSE by around 20% compared with the baseline errors. However, the new BI cuts down 50% of MAE and 40% of RMSE, respectively. The new BI also restrains the differences in algorithm performances, demonstrating the promotion of predictions based on Eq. (9).

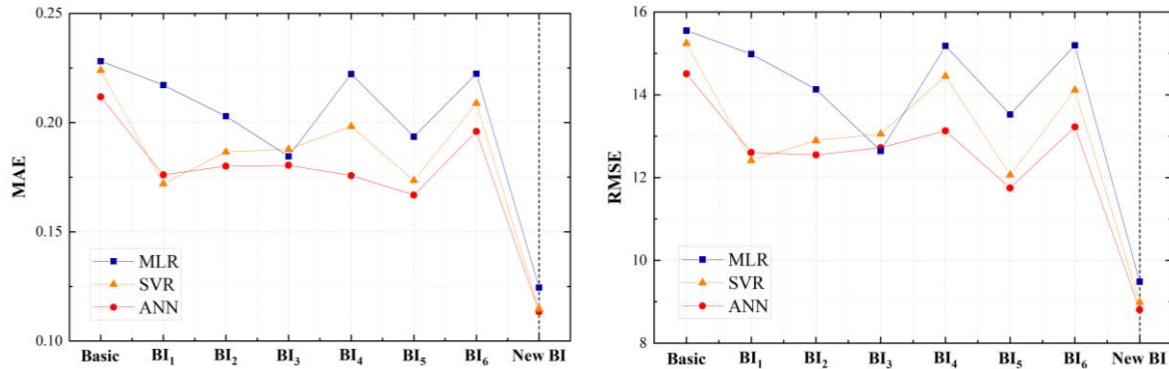


Fig. 6. The averaged MAE and RMSE produced by the new BI based on eight testing stages and three algorithms.

Significant improvements are observed in cases of W6 and W7 based on the new BI, as shown in Fig. 7. Those errors are rarely reduced by BI₅ as shown in Fig. 5. The new BI suppresses the MAE of the neighbour well (W6) under 10%, which will benefit the evaluation of fracability for the well-factory mode fracturing (several neighbour wells are drilled and fractured in one platform). The experience of previous fracturing operations can be extracted by the new BI for pressure prediction and schedule optimization for new neighbour wells. Meanwhile, approximately 55% of MAE and 52% of RMSE, on average, are diminished for the W7 case. The new BI provides better interpretations of the new subdivided formation and helps the machine learning algorithms to yield more accurate predictions. Therefore, the higher performance of the new BI is demonstrated according to the error comparisons with classic BIs.

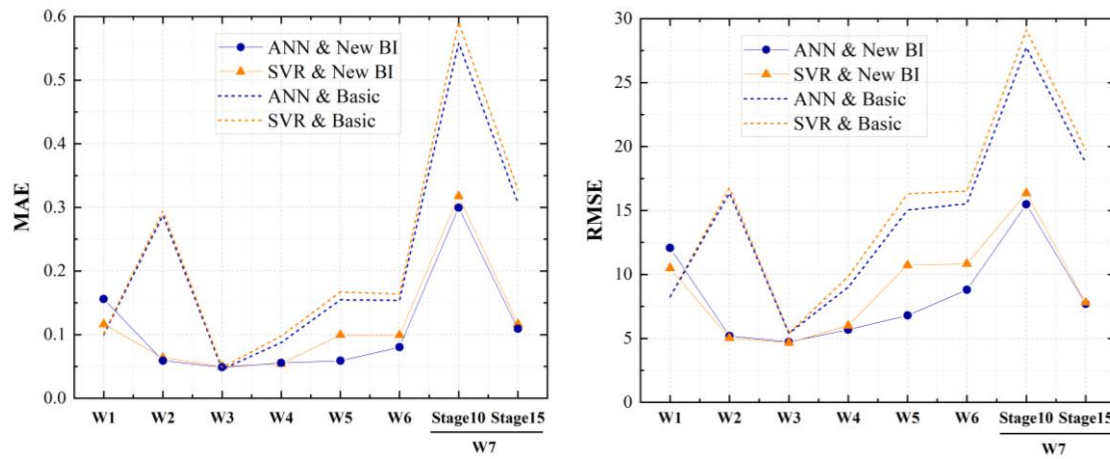


Fig. 7. Stage-wise improvements of MAE and RMSE based on the new BI.

4 Discussion

4.1 Analysis of remaining errors

We use error reductions (relative to the baseline reference) as the criterion of BI optimization. Therefore, the final errors in Fig. 7 may remain high for the W7 case, especially for Stage 10. The precise prediction of fracturing pressure is beyond the scope of this work since the performance of the new BI is already demonstrated by the significant improvement in MAE and RMSE. Future studies focusing on the pressure/fracability prediction may concern the remaining errors in Fig. 7 and restrict the errors by introducing more dominating features or advanced machine learning algorithms. In addition, slight error increases are observed for W1. The hydraulic parameters may determine the

prediction of fracturing pressure for W1, referring to the relative low MAE and RMSE yield by the Basic dataset where only pumping parameters are involved. The introduction of BI may provide limited contribution or even interference to the algorithm training and the predictions. Similar results are inspected in the case of W3, where low errors and limited promotion of the BIs are obtained.

4.2 Limitation and implication

The new BI is tested by eight fracturing stages from seven different wells, resulting in satisfactory performances. The limitations of this work may still exist in the expression and usage of the new BI. The expression of the new BI (3-BIs-combination) involves mineral, logging and elastic parameters, thus increasing the assignment of data collection and calculations. We, therefore, mention that a 2-BIs-combination (Eq. 6 in Appendix B, Table 5) may be a convenient substitute for on-site manual computations. Noteworthy, the capacity of the 2-BIs-combination is likely to be restrained, especially when the missing features (compared with the 3-BIs-combination) are determinants of the brittleness. Regarding the usage, the new BI is more complicated (based on a data-driven approach) than the traditional value comparison. Fitting the correlation between the new BI and fracability (i.e. the maximum fracturing pressure) based on a mass of field cases may produce a similar value-based criterion as the traditional usage. However, new data is unavailable for this work due to the data limitation. Moreover, we believe that it is more important to make better and broader usages of the brittleness index based on data science (rather than simple value comparison), in order to boost the contribution of BI to the “sweet point” recognition and hydraulic fracturing design.

5 Conclusions

Six classical mineral-, logging- and elastic-based BI correlations are evaluated by processing field measurements of shale gas fracturing based on machine learning workflows. The MLR, SVR, and ANN algorithms are employed for the evaluations to mitigate the interferences of algorithm performances. A new BI is derived based on the optimization and is verified by field applications of eight fracturing stages from seven different wells. The new BI provides a more reliable option and may be essential for recognizing formations, optimizing fracturing designs and predicting operation pressures. The major conclusions may be generalized as follows:

(1) The fracability is redefined as the fracturing pressure dominated by rock mechanical (mainly brittleness), geological, and hydraulic (pumping parameters) conditions, which clarifies the confusion between brittleness and fracability. The control variate method is applied to restrain the effects of geological and injecting aspects and enhance the dominance of brittleness on fracability. Classic BIs are evaluated by predicting the fracturing pressure based on a machine learning workflow. The data-driven approach provides a promising application scenario, which makes better and broader usages of the brittleness index than the conventional way of value comparison.

(2) The elastic-based BI produces the minimum errors for pressure predictions, followed by the logging-based BI, which is unexpected since the logging-based BI is empirical equations based on data from a different basin. The mineral-based BI also exerts a significant contribution to predictions and is crucial for the new BI derivation according to the backward elimination analysis (Appendix B). Moreover, an appropriate BI may restrain the different performances of various machine learning algorithms.

(3) A new BI calculation is derived by normalizing and averaging three optimized BIs based on the backward elimination strategy. The ratio of minimum and maximum horizontal principal stress, representing the geological factor, is also involved as the correction coefficient in the new equation according to variable analyses. The new BI diminishes around 50% and 40% of the MAE and RMSE, while classical BIs reduce about 20% of the errors. Moreover, the major error improvements are obtained in the distant well and different subdivided formations, indicating the higher performance of the new BI. The performances of different machine learning algorithms based on the new BI are also more uniform with limited variations. The new BI provides a more reliable option for evaluating the brittleness and fracability of the fracturing formation.

315 **Acknowledgement**



316

317 This research has received funding from the European Union's Horizon 2020 research and
318 innovation programme under the Marie Skłodowska-Curie grant agreement No. 846775.

319

References

- Al-Anazi, A., & Gates, I. (2010). A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs. *Engineering Geology*, 114(3-4), 267-277.
- Arora, S., & Mishra, B. (2015). Investigation of the failure mode of shale rocks in biaxial and triaxial compression tests. *International Journal of Rock Mechanics and Mining Sciences*, 79, 109-123.
- Bai, M. (2016). Why are brittleness and fracability not equivalent in designing hydraulic fracturing in tight shale gas reservoirs. *Petroleum*, 2(1), 1-19. doi:10.1016/j.petlm.2016.01.001
- Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep), 1089-1105.
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).
- Bisong, E. (2019). *Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners*: Apress.
- Buller, D., Hughes, S. N., Market, J., Petre, J. E., Spain, D. R., & Odumosu, T. (2010). *Petrophysical Evaluation for Enhancing Hydraulic Stimulation in Horizontal Shale Gas Wells*. Paper presented at the SPE Annual Technical Conference and Exhibition, Florence, Italy. <https://doi.org/10.2118/132990-MS>
- Feng, R., Zhang, Y., Rezagholilou, A., Roshan, H., & Sarmadivaleh, M. (2019). Brittleness Index: From Conventional to Hydraulic Fracturing Energy Model. *Rock Mechanics and Rock Engineering*, 53(2), 739-753. doi:10.1007/s00603-019-01942-1
- Glorioso, J. C., & Rattia, A. J. (2012). *Unconventional Reservoirs: Basic Petrophysical Concepts for Shale Gas*. Paper presented at the SPE/EAGE European Unconventional Resources Conference and Exhibition, Vienna, Austria. <https://doi.org/10.2118/153004-MS>
- Guo, Z., Chen, C., Gao, G., Cao, R., Li, R., & Liu, H. (2018). Integration of support vector regression with distributed Gauss-Newton optimization method and its applications to the uncertainty assessment of unconventional assets. *SPE Reservoir Evaluation & Engineering*, 21(04), 1,007-001,026.
- Guo, Z., Li, X.-Y., Liu, C., Feng, X., & Shen, Y. (2013). A shale rock physics model for analysis of brittleness index, mineralogy and porosity in the Barnett Shale. *Journal of Geophysics and Engineering*, 10(2), 025006.
- Heidari, M., Khanlari, G. R., Torabi-Kaveh, M., Kargarian, S., & Saneie, S. (2013). Effect of Porosity on Rock Brittleness. *Rock Mechanics and Rock Engineering*, 47(2), 785-790. doi:10.1007/s00603-013-0400-0
- Herwanger, J. V., Bottrill, A. D., & Mildren, S. D. (2015). *Uses and Abuses of the Brittleness Index With Applications to Hydraulic Stimulation*. Paper presented at the Unconventional Resources Technology Conference, San Antonio, Texas, USA. <https://doi.org/10.15530/URTEC-2015-2172545>
- Iyare, U., Blake, O., & Ramscook, R. (2020). Brittleness evaluation of Naparima Hill mudstones. *Journal of Petroleum Science and Engineering*, 196, 107737.
- Jarvie, D. M., Hill, R. J., Ruble, T. E., & Pollastro, R. M. (2007). Unconventional shale-gas systems: The Mississippian Barnett Shale of north-central Texas as one model for thermogenic shale-gas assessment. *AAPG Bulletin*, 91(4), 475-499.
- Jin, X., Shah, S. N., Roegiers, J. C., & Zhang, B. (2015). An integrated petrophysics and geomechanics approach for fracability evaluation in shale reservoirs. *SPE Journal*, 20(3), 518-526. doi:10.2118/168589-PA
- Jin, X., Shah, S. N., Truax, J. A., & Roegiers, J.-C. (2014). *A Practical Petrophysical Approach for Brittleness Prediction from Porosity and Sonic Logging in Shale Reservoirs*. Paper presented at the SPE Annual Technical Conference and Exhibition, Amsterdam, The Netherlands. <https://doi.org/10.2118/170972-MS>
- Kim, B.-H., Walton, G., Larson, M. K., & Berry, S. (2021). Investigation of the anisotropic confinement-dependent brittleness of a Utah coal. *International Journal of Coal Science & Technology*, 8(2), 274-290.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Li, C., Xie, H., & Wang, J. (2020). Anisotropic characteristics of crack initiation and crack damage thresholds for shale. *International Journal of Rock Mechanics and Mining Sciences*, 126, 104178.
- Lin, H., Kang, W.-H., Oh, J., & Canbulat, I. (2020). Estimation of in-situ maximum horizontal principal stress magnitudes from borehole breakout data using machine learning. *International Journal of Rock Mechanics and Mining Sciences*, 126, 104199.
- Mahmoud, A. A. A., Elkatatny, S., Mahmoud, M., Abouelresh, M., Abdulraheem, A., & Ali, A. (2017). Determination of the total organic carbon (TOC) based on conventional well logs using artificial neural network. *International Journal of Coal Geology*, 179, 72-80.

- Mathia, E., Ratcliffe, K., & Wright, M. (2016). *Brittleness Index-A Parameter to Embrace or Avoid?* Paper presented at the Unconventional Resources Technology Conference, San Antonio, Texas, 1-3 August 2016.
- Meng, F., Wong, L. N. Y., & Zhou, H. (2020). Rock brittleness indices and their applications to different fields of rock engineering: A review. *Journal of Rock Mechanics and Geotechnical Engineering*. doi:10.1016/j.jrmge.2020.06.008
- Mews, Alhubail, & Barati. (2019). A Review of Brittleness Index Correlations for Unconventional Tight and Ultra-Tight Reservoirs. *Geosciences*, 9(7). doi:10.3390/geosciences9070319
- Mullen, M. J., & Enderlin, M. B. (2012). *Fracability Index - More Than Rock Properties*. Paper presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, USA. <https://doi.org/10.2118/159755-MS>
- Pan, X.-P., Zhang, G.-Z., & Chen, J.-J. (2020). The construction of shale rock physics model and brittleness prediction for high-porosity shale gas-bearing reservoir. *Petroleum Science*, 17(3), 658-670. doi:10.1007/s12182-020-00432-2
- Rickman, R., Mullen, M. J., Petre, J. E., Grieser, W. V., & Kundert, D. (2008). *A Practical Use of Shale Petrophysics for Stimulation Design Optimization: All Shale Plays Are Not Clones of the Barnett Shale*. Paper presented at the SPE Annual Technical Conference and Exhibition, Denver, Colorado, USA. <https://doi.org/10.2118/115258-MS>
- Sharma, R. K., & Chopra, S. (2015). Determination of lithology and brittleness of rocks with a new attribute. *The Leading Edge*, 34(5), 554-564.
- Shen, Y., Cao, D., Ruddy, K., & Teixeira de Moraes, L. F. (2020). Near Real-Time Hydraulic Fracturing Event Recognition Using Deep Learning Methods. *SPE Drilling & Completion*, 35(03), 478-489. doi:10.2118/199738-PA
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929-1958.
- Sun, D., Lonbani, M., Askarian, B., Jahed Armaghani, D., Tarinejad, R., Thai Pham, B., & Huynh, V. V. (2020). Investigating the applications of machine learning techniques to predict the rock brittleness index. *Applied Sciences*, 10(5), 1691.
- Sun, S., Wang, K., Yang, P., Li, X., Sun, J., Liu, B., & Jin, K. (2013). *Integrated prediction of shale oil reservoir using pre-stack algorithms for brittleness and fracture detection*. Paper presented at the International Petroleum Technology Conference.
- Tarasov, B., & Potvin, Y. (2013). Universal criteria for rock brittleness estimation under triaxial compression. *International Journal of Rock Mechanics and Mining Sciences*, 59, 57-69.
- Xia, Y., Zhou, H., Zhang, C., He, S., Gao, Y., & Wang, P. (2019). The evaluation of rock brittleness and its application: a review study. *European Journal of Environmental and Civil Engineering*, 1-41. doi:10.1080/19648189.2019.1655485
- Ye, Y., Tang, S., & Xi, Z. (2020). Brittleness Evaluation in Shale Gas Reservoirs and Its Influence on Fracability. *Energies*, 13(2), 388.
- Zhang, D., Ranjith, P. G., & Perera, M. S. A. (2016). The brittleness indices used in rock mechanics and their application in shale hydraulic fracturing: A review. *Journal of Petroleum Science and Engineering*, 143, 158-170. doi:10.1016/j.petrol.2016.02.011

Appendix A – Summary of errors based on BI₃-based workflow

The BI₃-based workflow (used as an example) is performed to process the data from all eight testing stages. The errors are summarized in Table 4. Each value represents the averaged error of the entire fracturing curve. The MAE and RMSE characterize the accuracy and similarity between actual and predicted fracturing curves, respectively. Generally, the errors increase apparently in the last four stages from W5~W7 compared with errors based on W1~W4. Stages 3, 5, 7 and 9 are from the same wells as the training stages, which may constrain the geological uncertainties because of the close location and result in better predictions.

Table 4 Summary of prediction errors based on different datasets and algorithms.

Algorithm		Error	Stage 3	Stage 5	Stage 7	Stage 9	Stage 11	Stage 18	Stage 10	Stage 15
Basic dataset	MLR	MAE	0.08	0.26	0.07	0.14	0.19	0.19	0.59	0.31
		RMSE	7.11	14.57	6.60	13.26	18.13	17.84	28.48	18.41
	SVR	MAE	0.10	0.29	0.05	0.10	0.17	0.16	0.59	0.33
		RMSE	8.25	16.75	5.40	9.84	16.31	16.53	29.12	19.74
	ANN	MAE	0.10	0.29	0.05	0.09	0.16	0.15	0.56	0.31
		RMSE	8.25	16.37	5.40	9.01	15.03	15.53	27.75	18.71
BI ₃ dataset	MLR	MAE	0.08	0.05	0.05	0.13	0.12	0.11	0.60	0.32
		RMSE	7.13	4.45	4.73	12.41	12.16	11.93	29.21	19.14
	SVR	MAE	0.09	0.07	0.05	0.08	0.13	0.12	0.60	0.36
		RMSE	8.68	5.62	4.77	8.75	12.31	12.26	30.31	21.73
	ANN	MAE	0.11	0.07	0.05	0.06	0.12	0.11	0.58	0.35
		RMSE	10.30	5.39	5.15	6.83	12.02	11.51	29.58	20.99

Appendix B – Derivation of the new BI

The BI_1 , BI_2 , BI_3 , and BI_5 produce smaller errors according to the error analyses (Fig. 4) and are optimized for the new BI derivation. The new BI is designed to involve as many characteristics as possible to deliver more features into the algorithm training, thus improving the prediction. Therefore, the initial expression of new BI is proposed by averaging the normalized BIs referring to previous efforts (Pan et al., 2020; Rickman et al., 2008)

$$BI_{New} = (BI_{1n} + BI_2 + BI_{3n} + BI_{5n}) / 4$$

$$BI_{in} = \frac{BI_i - BI_{i\min}}{BI_{i\max} - BI_{i\min}} \quad (1)$$

where BI_{in} is the normalized BI_i ($i=1, 3$, and 5). BI_2 is an average of the normalized Young's modulus and Poisson ratio (Table 3), thus remaining the original value.

The backward elimination strategy is deployed for the new BI derivation. Eq. (1), containing all four selected BIs, is firstly applied for algorithm training, which is denoted as the All-in step. The prediction error in this step is used as a reference. Then, three out of the four BIs are freely assembled for four 3-BIs-combinations, in which the one with the minimum error is chosen for further derivation (Iteration 1 in Table 5). If the minimum errors are not larger than the All-in errors, the chosen 3-BIs-combination will be the new reference for the next round comparison – an elimination process. The elimination process repeats for smaller-scale combinations until the new error is larger than the reference (Iteration 2 in Table 5).

Table 5 Selection of the BI-combinations based on the backward elimination strategy.

Iteration 1			Iteration 2		
Eq.	MAE	RMSE	Eq.	MAE	RMSE
Single BI_5 (reference)	0.174	12.1			
(1) All-in (reference)	0.121	9.25	(3) $BI_{new} = (BI_{1n} + BI_{3n} + BI_{5n}) / 3$ (reference)	0.120	9.20
(2) $BI_{new} = (BI_2 + BI_{3n} + BI_{5n}) / 3$	0.167	11.7	(6) $BI_{new} = (BI_{1n} + BI_{3n}) / 2$	0.130	9.86
(3) $BI_{new} = (BI_{1n} + BI_{3n} + BI_{5n}) / 3$	0.120	9.20	(7) $BI_{new} = (BI_{1n} + BI_{5n}) / 2$	0.136	10.2
(4) $BI_{new} = (BI_{1n} + BI_2 + BI_{5n}) / 3$	0.130	9.85	(8) $BI_{new} = (BI_{3n} + BI_{5n}) / 2$	0.163	11.5
(5) $BI_{new} = (BI_{1n} + BI_2 + BI_{3n}) / 3$	0.123	9.38			

The SVR-based workflow guides the elimination process. The averaged errors based on all eight testing stages are summarized in Table 5. All new BI combinations promote the model performance

compared with the errors based on BI₅. Eqs. (2) and (8), without BI₁, produce the most prominent errors in each iteration, indicating the significance of mineral components for the interpretation of fracability. Eq. (3) results in the minimum errors and is selected for further derivation. Eq. (6) requires fewer parameters (2-BIs-combination) and produces relatively small errors, thus may be a convenient substitute for on-site estimations.

New parameters are tested and introduced into the expression to promote BI performance. A series of geological and engineering candidates are applied for variable analyses (Table 6). The parameters involved in Eq. (3) and hydraulic parameters are set as the control variables. The well depth, pore pressure, stage length, number of clusters, and stress ratio (ratio of minimum and maximum horizontal principal stress) are appended to the control variables and fed into the algorithm independently and successively (Tests 1~5 and All-in test), as shown in Table 6. Correspondingly, a series of MLR workflow is constructed to process the training dataset. The new parameter is optimized referring to the variable coefficient and P values – the significance of a variable to the pressure prediction. Since the feature scaling is performed, the larger coefficient (absolute value) indicates a more significant impact of the corresponding variable on the fracturing pressure.

Table 6 Summary of the independent variable analysis based on MLR algorithm and training dataset.

		Original test	Test 1	Test 2	Test 3	Test 4	Test 5	All-in test
Dependent variable	Fracturing pressure							
Independent variables	Depth		3.14***					2.78***
	P-pore			2.10***				-8.27***
	Stage length				-1.38***			1.81***
	Clusters					-2.43***		-3.02***
	Stress ratio						6.19***	17.44***
Control variables	Fluid	4.77***	4.70***	4.78***	4.74***	4.62***	4.79***	4.59***
	Fluid	3.59***	3.66***	3.61***	3.61***	3.61***	3.61***	3.63***
	Fluid	7.54***	7.46***	7.57***	7.49***	7.27***	7.60***	7.26***
	Quartz	-4.42***	-7.54***	-6.40***	-4.54***	-2.78***	-7.10***	-4.70***
	Clay	-8.28***	-10.66***	-9.32***	-8.75***	-7.73***	-9.56***	-8.60***
	Carb	7.82***	4.12***	6.80***	6.61***	6.01***	0.47	-12.81***
	TOC	1.67***	1.26***	1.13***	1.87***	0.95***	1.19***	0.93**
	Porosity	-1.30***	-2.09***	-0.59*	-1.30***	-0.84***	-0.58**	-2.21***
	Density	4.31***	2.38***	3.16***	5.44***	4.95***	2.36***	0.95**
	Young's M	-12.86***	-10.46***	-13.55***	-11.51***	-12.80***	-12.82***	-9.59***
	Poisson	5.62***	3.73***	4.22***	5.74***	4.21***	3.54***	1.71***
	DTC	-18.94***	-15.57***	-18.96***	-17.01***	-18.16***	-17.56***	-13.55***
	Pump rate	5.08***	5.20***	5.06***	5.16***	5.37***	5.04***	5.37***
	Sand-ratio	-0.67***	-0.47**	-0.66***	-0.65***	-0.58***	-0.67***	-0.45**
	Proppant	-0.14	-0.22	-0.13	-0.14	-0.20	-0.11	-0.24

Notes: The p-value is the significant level (* p<0.05, ** p<0.01, *** p<0.001).

468 The stress ratio, with the largest coefficient in Tests 1~5 and All-in test (Table 6), is chosen as the
469 new parameter and used as the correction coefficient for the new equation. The expression of the new
470 BI is given by

471
$$BI_{New} = \frac{(BI_{1n} + BI_{3n} + BI_{5n})}{3} \times \frac{Stress_{min}}{Stress_{max}} \quad (9)$$

472 where $Stress_{min}$ and $Stress_{max}$ are minimum and maximum horizontal principal stress, respectively.

473