Journal of Petroleum Exploration and Production Technology (2021) 11:3509–3519
https://doi.org/10.1007/s13202-021-01253-8

ORIGINAL PAPER-PRODUCTION ENGINEERING

AI/ML assisted shale gas production performance evaluation

Fahad I. Syed[1] · Temoor Muther[1] · Amirmasoud K. Dahaghi[1] · Shahin Negahban[1]

Received: 30 January 2021 / Accepted: 30 July 2021 / Published online: 7 August 2021

Abstract

Shale gas reservoirs are contributing a major role in overall hydrocarbon production, especially in the United States, and due to the intense development of such reservoirs, it is a must thing to learn the productive methods for modeling production and performance evaluation. Consequently, one of the most adopted techniques these days for the sake of production performance analysis is the utilization of artificial intelligence (AI) and machine learning (ML). Hydrocarbon exploration and production is a continuous process that brings a lot of data from sub-surface as well as from the surface facilities. Availability of such a huge data set that keeps on increasing over time enhances the computational capabilities and performance accuracy through AI and ML applications using a data-driven approach. The ML approach can be utilized through supervised and unsupervised methods in addition to artificial neural networks (ANN). Other ML approaches include random forest (RF), support vector machine (SVM), boosting technique, clustering methods, and artificial network-based architecture, etc. In this paper, a systematic literature review is presented focused on the AI and ML applications for the shale gas production performance evaluation and their modeling.

Keywords Shale gas · ANN · Machine learning · SVM · Boosting technique · Clustering

Abbreviations

AI	Artificial intelligence
DCA	Decline curve analysis
ANN	Artificial neural network
EUR	Estimated ultimate recovery
GBT	Gradient boosting tree
IHNN	Integrated hybrid neural networks
LSSVM	Least square support vector machine
LSTM	Long short-term memory
ML	Machine learning
OOB	Out of bag
RF	Random forest
SVM	Support vector machine
TOC	Total organic carbon

Introduction

Shale gas production was estimated to be at 50% of the net total of natural gas production as of 2018 in the USA, which eventually has gone even higher as of today. The reservoir performance modeling depends on the reservoir properties such as absorption flow and desorption flow, complexities of the reservoir, and variation of fracture permeability, etc. It is difficult to acquire property data, and many studies have tried to predict and model the productivity of shale gas reservoirs using decline curve and ML approaches (Han, D et al. 2020). The main challenge and questions that emerge during the production of shale gas include mainly the distance between wells or the well spacing, optimum stage length, rate of exertion of pressure, and the number of clusters for each stage involved. Answering these questions mainly provides a basis to come up with the most suitable AI approach for the set objectives (Mohaghegh, S.D et al. 2017).

Conventional reservoir engineering approaches including the experimental, analytical and numerical modeling do come handy in solving different challenges of shale gas reservoir which involves its characterization, production forecasting, and resource management (Syed, F.I., et al. 2021). However, they have different limitations. The analytical approaches relies on certain assumptions which sometimes

✉ Fahad I. Syed
fsyed@ku.edu

[1] School of Engineering, Chemical and Petroleum Engineering, The University of Kansas, 1450 Jayhawk Blvd, Lawrence, KS 66045, USA

cannot include the complex nature and fluid flow dynamics of the shales. Also, studying the data is tedious and engineers have to rely on trial and error for solutions. For numerical analysis, however, different aspects of the shale performance can be covered, but again, they are computationally expense (Syed, F.I et al. 2019). Also, including the compositional study and high dimensional physics including various complex reservoir, production, and completion characteristics such as natural fractures, desorption, hydraulic fracture cluster, fracture design, operational constraints, flowbacks makes the problem extremely tedious and the process of reaching to an acceptable answer becomes slower (Sprunger, C et al. 2021). Again, running various management, history matching, and optimization study makes the problem costly and worsen the results timings. This is where machine learning plays a huge role. ML can deal with high dimensional problems with accuracy, even the problems on which we do not have any analytical and numerical correlation available (Zargari et al. 2010; Kalantari 2011; Mohaghegh 2011; Mohaghegh 2013; Ertekin et al. 2019; Alatrach et al. 2019).

ML approaches can be implemented in finding the correlation between productivity and other factors like hydraulic fracturing and specific shale gas reservoir properties. They can include the use of supervised learning techniques like regression analysis, gradient boosting, decision trees, and support vector machine (SVM). Unsupervised methods like clustering include k-means clustering can also be used in the analysis and modeling of the production of shale gas (Syed et al. 2020a, b). In addition to neural network-based approaches, like artificial neural network and recurrent neural network for time series data analysis. The productivity properties can be obtained consisting of the prediction of the cumulative production rates and the estimated ultimate recovery (Syed et al. 2021a, b; Han et al. 2020a, b; Vikara et al. 2020; Al-Alwani et al. 2019; Bhattacharya et al. 2019; Shahkarami et al. 2018; Ansari et al. 2018).

Considering such advantage of using ML approaches, this paper specifically aims to discuss ML techniques that are applied in shale gas production and their corresponding results. Also, the pros and cons of implementing these methods are presented to enable the possibilities of future research. The rest of the article includes the systematic literature review including the background information of different ML techniques. In addition to this, the results implementing ML to shale gas reservoirs are summarized. At last, the study is concluded with remarks keeping in view of ML applications to shale gas performances. Specifically, the manuscript tries to discuss on the answers of following major concerns:

i. What are the ML methodologies utilized in performing shale gas production analysis?
ii. What is the data type used in different analyses?
iii. How is the ML employed in this analysis?

Research databases

The study is conducted by looking for scholarly databases including ‘SPE One Petro’ and ‘Google Scholar’, both are reputable sources for shale gas analysis research. Also, the research is been search directly in high-ranked journals including Nature, Sensors, Petroleum Journal, Energies, PLoS ONE, Geophysics & Engineering, and Geoscience Frontiers, etc. For this study, the articles were searched with a set criterion including a couple of exclusion criteria listed below;

- Works not related to both ML and shale gas production.
- Works that do not have any ML application

Similarly, the choice of keywords for building the search plays the most important role in the systematic literature review. For this study, machine learning and shale gas are the most commonly used terminologies, however, artificial neural network, support vector machine, boosting technique, and clustering, etc. are some of the keywords that are also used.

Background on ML

Machine learning approaches follow various steps including data acquisition, preprocessing and normalization, selection of the models, training, testing and the validation of the performance. The following are some of the commonly applied ML algorithms;

Random forest

Random forest (RF) is an ML approach used in producing predictions by the construction of more than one decision tree where every single tree consisting of different learning data set. This method ensembles all trees including the regression and classification trees based on specific type of problem presented. The data that are not used in the construction of a decision tree is used for the model validation; a technique called out-of-bag (OOB). The approach is considered advantageous such that despite dropping the accuracy of a single tree, the final accuracy of the combining trees will be higher (Temizel et al. 2020). A large RF also leads to a generalization of errors, and it is robust to noise and outliers since randomly constructed data from the entire dataset is used to retrieve individual decision trees as schematically shown in Fig. 1.

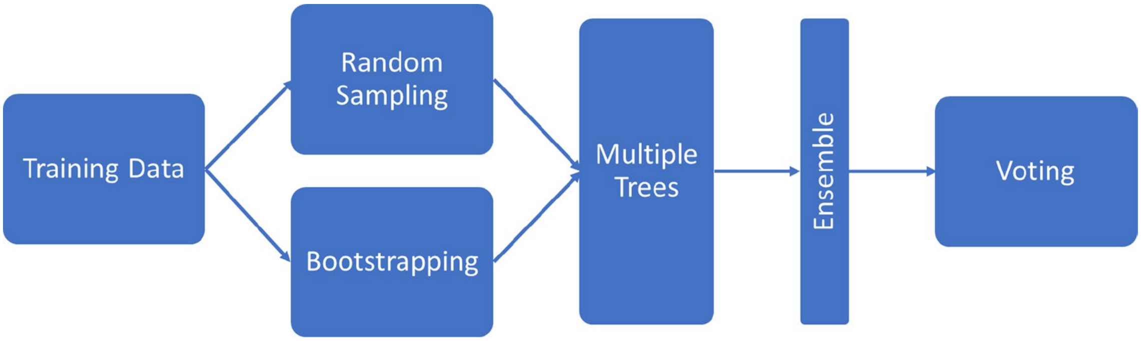


Fig. 1 Process modeling of random forest algorithm

Gradient boosting tree

Gradient Boosting Tree (GBT) ensembles weak predictive decision trees to develop a strong predictive model. In this methodology, the weak link of one tree is covered by the other tress, and hence, the generative capabililty of the final model is much better than the single tree model (Breiman 1997). The support of optimization algorithm in such technique allows the minimization of cost function on gradient descent for optimized values. The optimization stops when the updated residual values reaches to the minimum loss which indicates a perfect fitting of model values to the actual values (see Fig. 2).

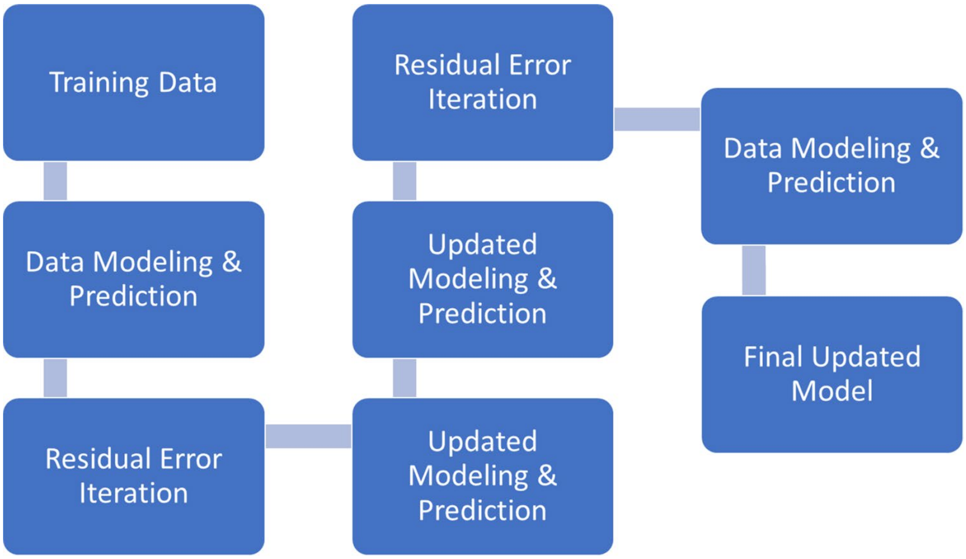
Support vector machine

Support vector machine (SVM) is an ML technique used for the classification and regression analysis that finds a hyperplane in the space of high dimensional and performs a corresponding classification. It utilizes nearly all entities and provisions of a non-overlapping segmentation as a method used in learning based on linear discriminant analysis (Saberioon et al. 2020). Besides, It performs classification on nonlinear data set by linking it to a cross-dimensional opening known as kernel trick to determine the decision of the hyperplane as shown in Fig. 3**.**

Clustering analysis

It can include *k*-means clustering that partitions data in k non-overlapping clusters by decreasing the deviation of the variance with each group. It finds solutions that maximize cluster cohesion in the cluster and out of cluster separation. Furthermore, Hierarchical clusters form dendrograms. The different distances can be used in clustering, and they include Euclidean distance for the same scale attribute and Minkowski distance (Han, D., et al. 2020).

Fig. 2 Process modeling of boosting tree algorithm


Training Data
Data Modeling & Prediction
Residual Error Iteration
Residual Error Iteration
Updated Modeling & Prediction
Updated Modeling & Prediction
Data Modeling & Prediction
Final Updated Model

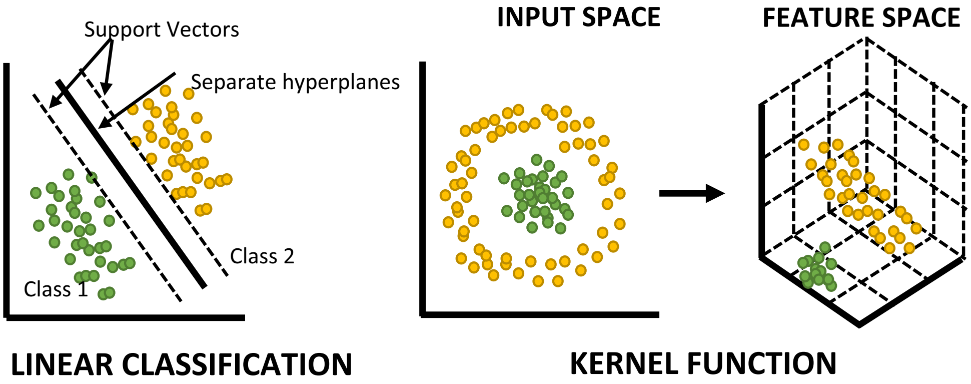


Fig. 3 Boundary decision function of SVM algorithm (Saberioon et al. 2020)

Artificial neural networks

They are biologically inspired techniques consisting of interlinked neurons used in processing with sets of adaptable weights allowing passage of signals. It consists of entrances that welcome data, hidden layers used in the extraction of patterns and output that produces and presents the resultant networking output (Esmaili et al. 2012a, 2012b; Zhou et al. 2014; Shahid, N et al. 2019). The weight and bias are updated regularly through optimization algorithms which works on minimizing the cost or loss functions such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and others to reach at an optimized value. These neural network ranges from shallow networks with smaller number of neurons and hidden layers to large deep complex neural networks consisting of many hidden layers and large number of neurons.

Results of systematic literature review

In this section, the systematic literature review is presented that is mainly distributed based on a different basis including annual publications, databases, citation and research methods, etc.

Annual publication distribution

The articles used in this review were distributed over different years, with most of the articles published recently. The recent publications were chosen due to the advances in the various ML techniques used, and this is informed by the fact that ML is an ever-changing field with more advanced algorithms and methods being researched and implemented every day.

Figure 4 shows the number of articles published between the years 2015 and 2020. The search confirms that more research has not been done on the ML techniques for shale gas production over the year. There has been a recent surge in research papers in this field, and this can be attributed to the research of more robust ML techniques that can now be

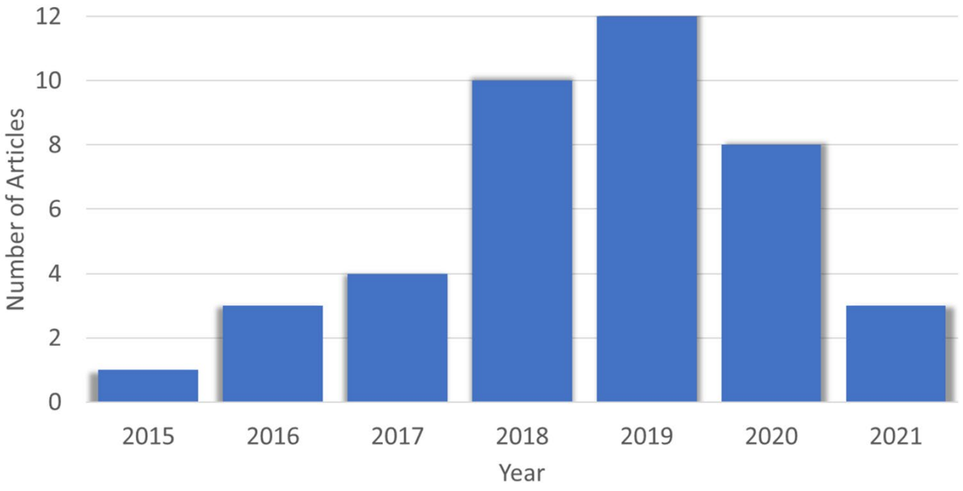


Fig. 4 Distribution of articles published over the years between 2015 and 2020

applied in different fields. There may have been a growing interest in research areas on shale gas production using ML since the year 2015. The fact may be due to the vast amount of data generated from shale gas mining and advances in ML algorithms.

Database publication distribution

Figure 5 is a breakdown of the papers per each of the given paper databases. Most of the articles were published in the Journal of the Society of Petroleum Engineers and Petroleum Journals. Of the nineteen papers shortlisted, 13 were published in journals, 5 in conferences, and 1 is a report as shown in Fig. 6. The research papers were taken from different journals, including the Society of Petroleum Engineers, which is the top journal with three publications, while the remaining journals have a single publication in the subject discipline. The remaining journals include Applied Sciences, Energies, Petroleum, and Geoscience Frontiers. Most of the journals belonged to the petroleum engineering domain.

The research papers that were attributed to conferences were three in total, with each conference having a single article. The conferences included the SPE Hydraulic Fracturing Technology Conference and Exhibition, URTec among others. Also, these papers were mainly associated with the petroleum engineering domain, specifically discussing aspects of shale gas production and the application of ML in this field.

Citation analysis

It is essential to consider the citation of an article since it determines the number of times that other articles have cited a given article. The higher the number of citations, especially of other articles that have been published in top-tier conferences might contribute to the relevance and the quality of the given article. The articles were obtained from google scholar and the number of citations obtained from this site. This is a reputable site, and it could also give access to related articles, and it becomes easy to analyze the impact of an article.

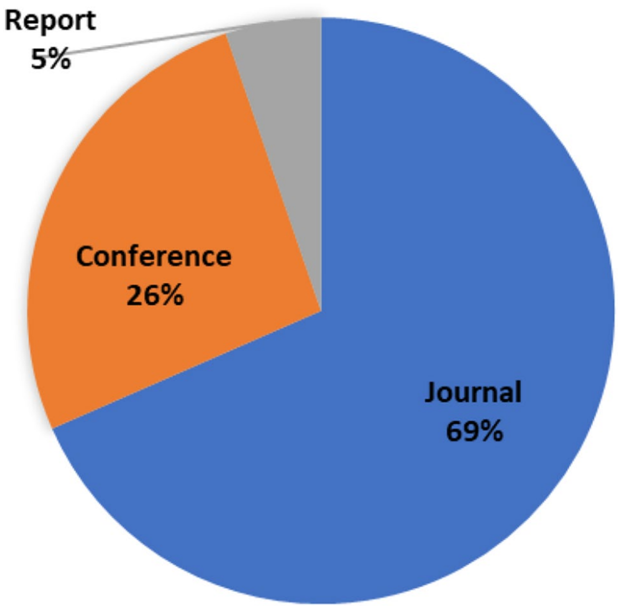


Fig. 6 Breakdown of the types of articles

The most cited article was by Esmaili, S. and Mohaghegh, S.D (2016) that discusses several data analytics techniques for modeling shale assets in more then 45 citations. The average number of citations for the research papers was found around 12.05. The detailed information is presented in Table 1.

Fig. 5 Distribution of articles published in different conference proceedings and journals

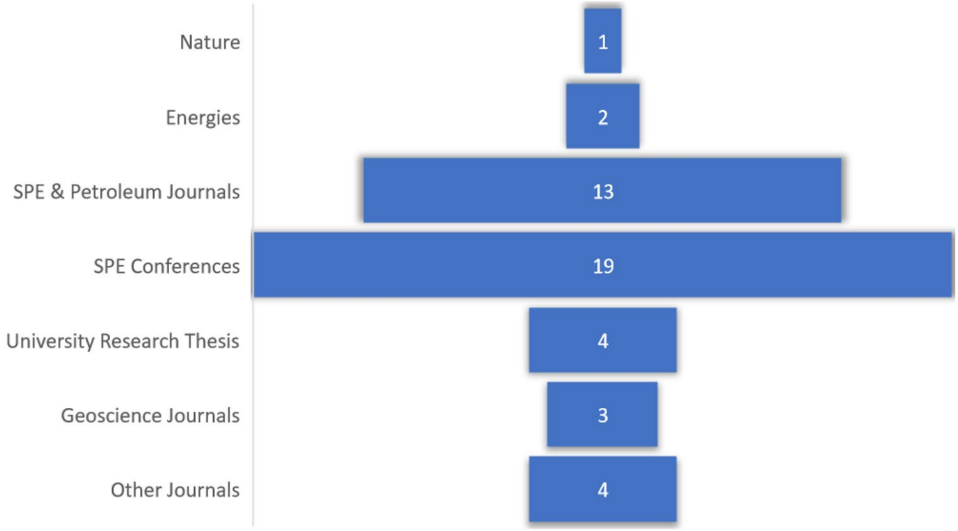


Table 1 Summary of articles with their citations (accessed in May 2020)

References	Publication year	Citations
Lee, K et al. (2019)	2019	5
Mohaghegh, S. D et al. (2017)	2017	25
Tian, Y et al. (2018)	2018	9
Luo, G et al. (2018)	2018	11
Han, D et al. (2020)	2020	3
Alabboodi & Mohaghegh (2016)	2016	3
Han, D et al. (2019)	2019	2
Lee, K et al. (2019)	2019	7
Panjaa, P et al. (2018)	2018	22
Esmailia & D.Mohaghegh (2016)	2016	43
Alqahtani, M (2015)	2015	3
Heaven, D (2019)	2019	19
Matsumori, K et al. (2018)	2018	2
Saberioon, M et al. (2018)	2018	10
Shahid, N et al. (2019)	2019	32
Asala, H. I et al. (2017)	2017	5
Zhu, L et al. (2018)	2018	16
Kamari, A et al. (2017)	2017	23
Qian, K.-R et al. (2018)	2018	1

Research methods analysis

The papers that were analyzed were published between 2015 and 2020. Table 2 shows the summary of the references used in the analysis with the different data types and machine learning algorithms used in the publication and a brief description of what was accomplished. Figure 7 shows the preference of the different ML algorithms in shale gas production with most papers using neural network-based methods at 52%, followed by clustering and SVM at 19%. A pie chart between the different ML algorithms can be drawn to depict the different methods applied in the papers.

Artificial neural network

In the paper by Lee et al. (2019), long short-term memory (LSTM) has been used in the prediction of future production of shale gas since the shale gas data are considered as time-series data. The information is from a commercial database in Canada. The preprocessing of the data follows seven steps that include data cleaning, train and test set splits, feature extractions, normalization, and sorting that will be input in the LTSM model in addition to cutoffs for the well list. This analysis applied LSTM on the data and trained by features of production data for 300 wells during the shut-in period. The testing of the methods was done on unseen data on 15 wells and showed better prediction accuracy as compared to hyperbolic decline curve analysis (DCA). The method has the advantage of providing stable results for all the time series data. The DCA method is just an empirical regression analysis method that is used in the prediction of future good production, although it is limited by high transient flow in shale gas production (Lee et al. 2019). Similar research was done on using LSTMs to forecast local Shale natural gas demand and form as guidance for re-frac candidature to maximize profitability (Asala, H.I et al. 2017). The disadvantage of using recurrent networks like LSTM that have not been addressed in the papers is the inherent vanishing and exploding gradients that LSTMs suffers from and how such a limitation was discussed in the design of the network.

Mohaghegh, Gaskari, and Maysami (2017) use shale analytics, mainly artificial intelligence is applied in more than 3000 wells in Marcellus, Niobrara, and Utica to gain insights into the impacts of district reservoir and the completion parameters on production. It also analyzes the impact on the quality of prediction made by the ANN technologies and the production of blind wells by implementing pattern recognition (Mohaghegh, S.D et al. 2017).

Luo et al. (2018) proposed and also reported by Syed, F.I et al. (2020, 2021) in a couple of comprehensive review papers that petrophysical analysis using deviation in thickness, water saturation, and porosity. The study then applied an ANN to relate the first-year production to essential features (Luo et al. 2018). Alabboodi and Mohaghegh (2016), proposes the use of ANN was used to find the relation of the trend between estimated ultimate recovery (EUR) and corresponding parameters. It utilizes the use of pattern recognition to discover any hidden but potentially useful patterns within the shale dataset using 34 parameters, with the aim of the prediction of EUR. Due to the complexity of the models used, essential performance analysis methods were used in determining the parameters with the highest impact on EUR. The pattern recognition method that was used in this study was fuzzy pattern recognition (FPR) that has the capability of deducing hidden and non-obvious trends and patterns from a large and complex dataset. Some of the parameters selected to be applied on EUR prediction included soak time which is the time between the completion of well and when it is used for production, injected proppant per stage, cluster spacing, the total number of steps, Young's Modulus, and total organic carbon (TOC) (Alabboodi, M.J 2016).

Alqahtani, M (2015) proposes an inverse ANN model to design the maximum reservoir contact (MRC) wells in a shale gas reservoir. It uses the characteristics of the reservoir and the hydrocarbon production profile to model architecture that achieves the desired gas recoveries. It is aimed at reducing the project risks, especially in projects with high capital and operational costs by guiding and placing the project on the right trajectory. The desired outputs of the project include the mother-well length, lateral spacing, direction, and length (Alqahtani, M 2015).

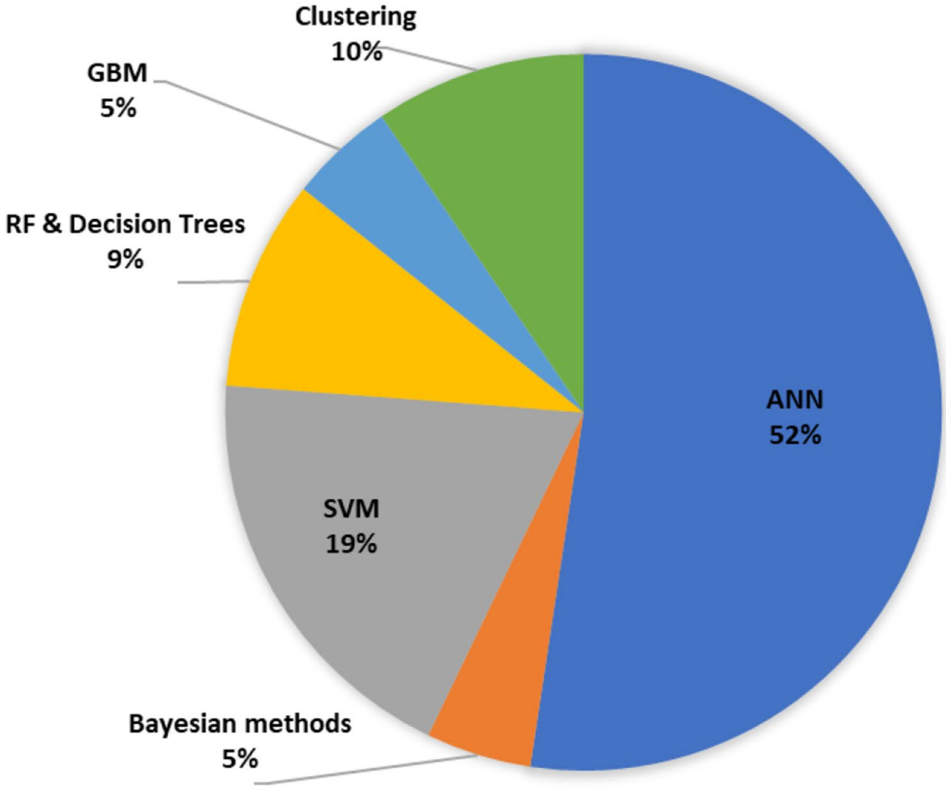
Table 2 Summary of papers that talks about machine learning for analysis and modeling of shale gas performance

Reference	Type of ML	Year	Data used	Description
Lee, K et al. (2019)	LSTM, regression method	2019	Real data	Prediction of the future shale gas production on the time series data obtained
Mohaghegh, S. D et al. (2017)	ANN	2017	Real data	It also analyzes the impact on the quality of prediction made by the ANN technologies and the production of blind wells
Tian, Y et al. (2018)	Bayesian methods, Markov-chain Monte Carlo, Gaussian regression methods	2018	Real data	It relates the production of shale gas to the essential geological controls like depth and total organic carbon in the production of shale gas
Luo, G et al. (2018)	ANN	2018	Real data	It finds the effectiveness of the completion strategy on shale gas production
Han, D et al. (2020)	SVM, RF, GBT, Clustering analysis	2020	Real data	Compares the different supervised learning methods on productivity forecasting on shale gas production
Alabboodi & Mohaghegh (2016)	ANN	2016	Real data	Uses pattern recognition to find the trend relation between estimating ultimate recovery and parameters of a well for the shale gas production
Han, D et al. (2019)	ANN	2019	Real data	The prediction of the production rates of shale gas using ANN
Lee, K et al. (2019)	Clustering-based methods	2019	Review	It reviews reservoir uncertainty based on distance-based clustering
Panjaa, P et al. (2018)	SVM, ANN	2018	Real data	Forecasting of hydrocarbon production from shale gas reservoir
Esmaili, S. and Mohaghegh, S.D (2016)	ANN	2016	Real data	Matching history by considering hydraulic fracture design
Alqahtani, M (2015)	ANN	2015	Real data	To predict desired gas recoveries based on hydrocarbon production profile
Asala, H. I et al. (2017)	LSTM, Feedforward NN	2017	Real data	It optimizes shale gas supply chain
Zhu, L et al. (2018)	ANN(IHNN)	2018	Real data	Performance of shale reservoir based on TOC using IHNN
Kamari, A et al. (2017)	SVM, ANN, decision tree	2016	Real data	Analysis of performance on cumulative gas production
Qian, K.-R et al. (2018)	SVM	2018	Real data	It predicts multiple attributes for sweet spots in reservoirs which can enable an objective characterization of shale gas potential

Panjaa et al. (2018) apply ANN to estimate the general hydrocarbon production forecasting for the shale gas reservoir. The neural network techniques are applied to optimize production and well placement using pattern recognition as well as to help in the reservoir characterization efforts. The main aim was to explore the potential of the application of artificial intelligence in the oil and gas industry. The parameters used to build the reservoir includes the slope of gas, rock permeability, and compressibility, initial dissolved gas-oil ratio, among others (Panja, P et al. 2018). Another paper analyzed the performance of shale reservoirs based on TOC using integrated hybrid neural networks (IHNN) (Zhu, L et al. 2018).

Along with these studies, various works have also been presented on hydraulic fracturing design, and optimization in shale gas reservoirs. Esmailia and Mohaghegh (2016) propose the use of an ANN expert system for the history-matching process considering different hydraulic fracture design. The aim was to model hydrocarbon production from the Marcellus shale reservoir. The paper allows the well logs, production history, and hydraulic fracture data to determine the characteristics and the behavior of the model. The model is unique since it uses hard data directly in the reservoir model, which are field measurements, and it can help to optimize the hydraulic fracture process. Bowie et al. (2018) in his study, trained a neural network on shale gas formation consisting of 262 wells and a total of 21 fracture and other well completion variables for completion design optimization for improved well performance. They found fracture tonnage as a prominent

Fig. 7 Different ML techniques talking about ML in shale gas production



variable in well performance. Along with this, fracture pump rate also controls the production performance. They also determined that expensive fracture design proppants including ceramics and resin coated have no significant impact on well performance.

In another literature, Luo et al. (2018) implemented deep neural network to predict Bakken shale performance under large network of hydraulic fracture design parameters. They found proppant mass, and stage count highly impact production performances. Also, Han et al. (2020a, b) propose an ANN-based model to predict production rates during transient flow. The aim was to predict the future production rates using hydraulic fracture factors, reservoir properties, production management, and well completion factor. A total of 150 wells with 8560 variables were used in training an ANN. The prediction of the production rates was made using a machine learning and clustering approach to find the similarity between the dataset.

Gu et al. (2016), also used ANN to predict fracture geometry, and optimize propped length based on different factors including the injected volume, pump rate and perforation position. Shen et al. (2020) also used the power of deep learning to determine the fracture stage start and endpoints, and ball seat placement while considering the injection rate and well pressure dataset. Their model predicted the variables with a accuracy of 99.7%. In another application, the ANN has been trained to determine rock elastic moduli including the young's modulus and brittleness for fracture design operation (Gong et al. 2019).

Deep learning approaches are known to be hard to understand a significant area of focus currently is making the models explainable. They act as a black box, and it is hard to know what is being learned before getting the given prediction that can lead to inaccuracies. The current studies have not tried to state how they overcame the weaknesses of artificial neural network techniques like exploding gradients and vanishing gradients in the LSTM network or the justification on the use of artificial neural network techniques based on the amount of data provided (Heaven, D 2019).

Bayesian methods

A Bayesian approach is used for inference on limited amounts of data. The solutions are based on Bayes' theorem that assumes prior probabilities and likelihood functions (Matsumori, K et al. 2018). Tian et al. (2018) use Bayesian methods with Markov chain Monte Carlo (MCMC) machine learning algorithms have been used in the integration of data set and the prediction of geological properties at the production of well production sites. The paper related the production and geological parameters in shale gas production, and the spatial Gaussian process regression modeling was applied in the investigation of the given primary controls on the production. The results showed that the parameters most significant to the production of shale gas include the depth and the total organic carbon (TOC). The parameters can help in making decisions on shale production and transferring it to other production firms (Tian, Y et al. 2018).

The weakness of applying Bayesian inference is that if the prior information is wrong, the resulting inference will be incorrect. The study did not perform a thorough analysis of the sanity of the data used in the modeling, and thus the resulting conclusion might be wrong based on that.

In one of the recent studies, Ma et al. (2020) developed a multiscale-parameterization method to model complex and irregular fracture network. They presented a novel history matching approach using surrogate-assisted cooperative swarm optimization (SACOSO) data-driven evolutionary algorithm to reduce model parameters' uncertaininties and explore the Bayersian posterior space. Their method found to be efficient with high prediction and modeling capabilities on fracture network. Such method can be applied on understanding the complex fracture network in shale gas resources.

Random forest

Han, Jun, and Kwon (2020) propose the use of RF is the productivity forecasting of shale reservoirs. The aim was to find the optimal input features to the reservoir for the cumulative gas production in 36 months. The method used in the selection of parameters included the application of a penalty function and using a kernel parameter. The results showed on the whole data set and after the application of the variable importance method (VIM), which says how important a variable is in any machine learning data set. The results showed that performing the productivity forecasting on the whole data set yielded a MAPE of 22.94% and 16.80% after the VIM was selected. The method used to provide a robust way of selecting the most useful variables to be used in modeling productivity. The downside is that the complexity and computational costs of RF in performing productivity forecasting are high and thus it is not easily scaled up in this case (Han, D., et al. 2020).

Support vector machine

Han, Jun, and Kwon (2020) propose the use of SVM in the productivity forecasting of shale reservoirs. The aim was to find the optimal input features to the reservoir for the cumulative gas production in 36 months. The results showed that performing the productivity forecasting on the whole data set yielded a MAPE of 20.03% which was a good performance compared to other methods (Han, D., et al. 2020).

Panjaa et al. (2018) apply the least square support vector machine (LSSVM) to estimate the hydrocarbon production forecasting for the shale gas reservoir. It aims to help in learning the physical and thermodynamic signatures in the data in the shale gas production. A better approach could be to try other methods that account for the temporal aspects of forecasting since SVM does not do that (Panja, P et al. 2018).

Kamari, Leed, and Bahadori (2016) applies three machine learning strategies to analyze the cumulative and initial gas production and total gas flow rate in a shale gas reservoir. The study compares ANN, LSSVM, and decision trees in the analysis. It found out that LSSVM was the best performing model in predicting the performance of gas production. The paper however does not try to discuss the probable reasons for the better performance of LSSVM compared to the other models and how much data set was used (Kamari, A et al. 2017).

Qian et al. (2018) apply SVM to predict and analyze shale gas sweet spots and comprehensive characterization of shale gas reservoirs. It predicts multiple attributes for sweet spots in reservoirs which can enable an objective characterization of shale gas potential. The paper does not analyze the prediction with other ML techniques that rely on temporal information like LSTMs to analyze performance in the identification of the sweet spots (Qian, K.R et al. 2018).

Gradient boosting tree

Han, Jun, and Kwon (2020) also propose the use of GBM in the productivity forecasting of shale reservoirs aimed at finding the optimal input features to have an optimal cumulative gas production in 36 months. The results showed that performing the productivity forecasting on the whole data set yielded a MAPE of 22.87%. Boosting techniques always perform better than other algorithms by performing a combination of optimization techniques to obtain the best performance. The disadvantage of using GBM in this paper is that the effect of outliers on the algorithm has not been covered or methods of dealing with outliers. GBM is sensitive to outliers and this can make it perform poorly in productivity forecasting (Han, D., et al. 2020).

Clustering analysis

Kang et al. (2019) discuss the use of distance-based clustering techniques to assess uncertainties in the shale reservoir. Distance-based clustering is defined using distance, dimensional reduction, distance matrix construction, and clustering. The distance metric that was used in this paper is the Minkowski distance and the k-means clustering used in the clustering method. The advantage of using this technique is that it can be easily integrated with deep learning algorithms for feature extraction. The disadvantage is that most of the modeling makes lots of assumptions since there is no ground truth thus it makes it hard to evaluate model performance in measuring uncertainties in the shale reservoirs (Lee, K et al. 2019).

Conclusion

This paper presents a systematic review covering machine learning approaches in shale gas production performance evaluation and modeling. The subject topic is of great importance, especially in assessing the production efficiency of the different approaches used in shale gas production. There are few papers and research that has been performed on the use of machine learning approaches for shale gas production. Different ML approaches have been successfully applied in shale gas reservoirs for different applications including production forecasting, fracture design oprimization, uncertaininty determination, history matching, sweet spot identification, shale gas characterization, TOC impacts and well analysis. These all predictions relied on large dataset and different ML approaches with the most prominent ANN is able to identify the predictive values more accurately. However, there is still concern with the smaller number of data points, since, all ML algorithms work efficiently on larger datasets. Hence, the ML algorithms with the physics driven approach can become useful in obtaining more efficient predictive results with smaller data samples available for shale gas development and production studies. Under current limitations, following recommendations for future studies are made including;

- Performing ensemble learning using multiple ML techniques to see whether the performance of shale gas production can be further improved.
- Carry out more analysis of the best ML approach that can be used as a baseline for shale gas productivity modeling.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations

Conflicts of interest The authors declare that there is no conflict of interest.

References

Alabboodi, M.J. and Mohaghegh, S.D., 2016, September. Conditioning the Estimating Ultimate Recovery of Shale Wells to Reservoir and Completion Parameters. In *SPE Eastern Regional Meeting*. OnePetro.

Al-Alwani, M.A., Britt, L., Dunn-Norman, S., Alkinani, H.H., Al-Hameedi, A.T. and Al-Attar, A., 2019, June. Production performance estimation from stimulation and completion parameters using machine learning approach in the marcellus shale. In 53rd US Rock Mechanics/Geomechanics Symposium. OnePetro.

Alatrach, Y., Saputelli, L., Narayanan, R., Mohan, R., Alklih, M.Y. and Rubio, E., 2019, November. Data-driven vs. traditional reservoir numerical models: A case study comparison of applicability, practicality and performance. In Abu Dhabi International Petroleum Exhibition & Conference. OnePetro.

Alqahtani, M., 2015. *Shale gas reservoirs development strategies via advanced well architectures*. The Pennsylvania State University.

Ansari, A., Fathi, E., Belyadi, F., Takbiri-Borujeni, A. and Belyadi, H., 2018 March. Data-based smart model for real time liquid loading diagnostics in Marcellus Shale via machine learning. In SPE Canada Unconventional Resources Conference. Society of Petroleum Engineers. Doi: https://doi.org/10.2118/189808-MS.

Asala, H.I., Chebeir, J., Zhu, W., Gupta, I., Taleghani, A.D. and Romagnoli, J., 2017, October. A machine learning approach to optimize shale gas supply chain networks. In SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Bhattacharya S, Ghahfarokhi PK, Carr TR, Pantaleone S (2019) Application of predictive data analytics to model daily hydrocarbon production using petrophysical, geomechanical, fiber-optic, completions, and surface data: a case study from the Marcellus Shale, North America. J Petrol Sci Eng 176:702–715

Bowie, B., 2018, March. Machine learning applied to optimize Duvernay well performance. In SPE Canada Unconventional Resources Conference. OnePetro.

Breiman, L., 1997. Arcing the edge. Technical Report 486, Statistics Department, University of California at Berkeley.

Ertekin T, Sun Q (2019) Artificial intelligence applications in reservoir engineering: a status check. Energies 12(15):2897

Esmaili S, Mohaghegh SD (2016) Full field reservoir modeling of shale assets using advanced data-driven analytics. Geosci Front 7(1):11–20

Esmaili, S., Kalantari-Dahaghi, A. and Mohaghegh, S.D., 2012a, October. Modeling and history matching of hydrocarbon production from Marcellus shale using data mining and pattern recognition technologies. In SPE Eastern Regional Meeting. OnePetro.

Esmaili, S., Kalantari-Dahaghi, A. and Mohaghegh, S.D., 2012b, October. Forecasting, sensitivity and economic analysis of hydrocarbon production from shale plays using artificial intelligence & data mining. In SPE Canadian Unconventional Resources Conference. OnePetro.

Gong, Y., Mehana, M., Xiong, F., Xu, F. and El-Monier, I., 2019, September. Towards better estimations of rock mechanical properties integrating machine learning techniques for application to hydraulic fracturing. In SPE Annual Technical Conference and Exhibition. OnePetro.

Gu M, Gokaraju D, Chen D, Quirein J (2016) Shale fracturing characterization and optimization by using anisotropic acoustic interpretation, 3D fracture modeling, and supervised machine learning. Petrophys SPWLA J Form Eval Reserv Descr 57(06):573–587

Han D, Jung J, Kwon S (2020a) Comparative study on supervised learning models for productivity forecasting of shale reservoirs based on a data-driven approach. Appl Sci 10(4):1267

Han, D., Kwon, S., Son, H. and Lee, J., 2020, February. Production Forecasting for Shale Gas Well in Transient Flow Using Machine Learning and Decline Curve Analysis. In Asia Pacific Unconventional Resources Technology Conference, Brisbane, Australia, 18–19 November 2019 (pp. 1510–1527). Unconventional Resources Technology Conference.

Heaven D (2019) Why deep-learning AIs are so easy to fool. Nature 574(7777):163–166

Kalantari, M., 2011. Modeling, History Matching, Forecasting and Analysis of Shale Reservoirs Performance Using Artificial Intelligence. SPE Digital Energy Conference and Exhibition.

Kamari A, Mohammadi AH, Lee M, Bahadori A (2017) Decline curve based models for predicting natural gas well performance. Petroleum 3(2):242–248

Kang B, Kim S, Jung H, Choe J, Lee K (2019) Efficient assessment of reservoir uncertainty using distance-based clustering: a review. Energies 12(10):1859

Lee K, Lim J, Yoon D, Jung H (2019) Prediction of shale-gas production at Duvernay Formation using deep-learning algorithm. SPE Journal. https://doi.org/10.2118/195698-PA

Luo, G., Tian, Y., Bychina, M. and Ehlig-Economides, C., 2018, September. Production optimization using machine learning in Bakken shale. In Unconventional Resources Technology Conference, Houston, Texas, 23–25 July 2018 (pp. 2174–2197). Society of Exploration Geophysicists, American Association of Petroleum Geologists, Society of Petroleum Engineers.

Ma X, Zhang K, Yao C, Zhang L, Wang J, Yang Y, Yao J (2020) Multiscale-network structure inversion of fractured media based on a hierarchical-parameterization and data-driven evolutionary-optimization method. SPE J 25(05):2729–2748

Matsumori K, Koike Y, Matsumoto K (2018) A biased Bayesian inference for decision-making and cognitive control. Front Neurosci 12:734

Mohaghegh SD (2011) Reservoir simulation and modeling based on artificial intelligence and data mining (AI&DM). J Nat Gas Sci Eng 3(6):697–705

Mohaghegh, S.D., 2013, August. A critical view of current state of reservoir modeling of shale assets. In SPE Eastern Regional Meeting. OnePetro.

Mohaghegh, S.D., Gaskari, R. and Maysami, M., 2017, January. Shale analytics: Making production and operational decisions based on facts: A case study in marcellus shale. In SPE Hydraulic Fracturing Technology Conference and Exhibition. Society of Petroleum Engineers.

Panja P, Velasco R, Pathak M, Deo M (2018) Application of artificial intelligence to forecast hydrocarbon production from shales. Petroleum 4(1):75–89

Qian KR, He ZL, Liu XW, Chen YQ (2018) Intelligent prediction and integral analysis of shale oil and gas sweet spots. Pet Sci 15(4):744–755

Saberioon M, Císař P, Labbé L, Souček P, Pelissier P, Kerneis T (2018) Comparative performance analysis of support vector machine, random forest, logistic regression and k-nearest neighbours in rainbow trout (oncorhynchus mykiss) classification using image-based features. Sensors 18(4):1027

Shahid N, Rappon T, Berta W (2019) Applications of artificial neural networks in health care organizational decision-making: A scoping review. PloS one. 14(2):e0212356

Shahkarami, A., Ayers, K., Wang, G. and Ayers, A., 2018, October. Application of Machine Learning Algorithms for Optimizing Future Production in Marcellus Shale, Case Study of Southwestern Pennsylvania. In SPE/AAPG Eastern Regional Meeting. OnePetro.

Shen, Y., Cao, D., Ruddy, K. and Teixeira De Moraes, L.F., 2020, January. Deep learning based hydraulic fracture event recognition enables real-time automated stage-wise analysis. In SPE Hydraulic Fracturing Technology Conference and Exhibition. OnePetro.

Sprunger C, Muther T, Syed FI, Dahaghi AK, Neghabhan S (2021) State of the art progress in hydraulic fracture modeling using AI/ML techniques. Model Earth Syst Environ. https://doi.org/10.1007/s40808-021-01111-w

Syed FI, Boukhatem M, Al Kiyoumi AA (2019) Lean HC gas injection pilots analysis and IPR back calculation to examine the impact of asphaltene deposition on flow performance. Petroleum Research 4(1):84–95

Syed FI, AlShamsi A, Dahaghi AK, Neghabhan S (2020a) Application of ML & AI to model petrophysical and geo-mechanical properties of shale reservoirs–A systematic literature review. Petroleum. https://doi.org/10.1016/j.petlm.2020.12.001

Syed FI, Alshamsi M, Dahaghi AK, Neghabhan S (2020b) Artificial lift system optimization using machine learning applications. Petroleum. https://doi.org/10.1016/j.petlm.2020.08.003

Syed FI, Alnaqbi S, Muther T, Dahaghi AK, Negahban S (2021) Smart shale gas production performance analysis using machine learning applications. Petroleum Res. https://doi.org/10.1016/j.ptlrs.2021.06.003

Syed FI, Negahban S, Dahaghi AK (2021b) Infill drilling and well placement assessment for a multi-layered heterogeneous reservoir. J Petroleum Explor Prod 11(2):901–910

Temizel, C., Canbaz, C.H., Saracoglu, O., Putra, D., Baser, A., Erfando, T., Krishna, S. and Saputelli, L., 2020, July. Production Forecasting in Shale Reservoirs Using LSTM Method in Deep Learning. In SPE/AAPG/SEG Unconventional Resources Technology Conference. Unconventional Resources Technology Conference.

Tian Y, Ayers WB, Sang H, McCain WD Jr, Ehlig-Economides C (2018) Quantitative evaluation of key geological controls on regional Eagle Ford shale production using spatial statistics. SPE Reservoir Eval Eng 21(02):238–256

Vikara D, Remson D, Khanna V (2020) Gaining perspective on unconventional well design choices through play-level application of machine learning modeling. Upstream Oil Gas Technol 4:100007

Zargari, S. and Mohaghegh, S.D., 2010, October. Field development strategies for bakken shale formation. In SPE Eastern Regional Meeting. OnePetro.

Zhou, Q., Kleit, A., Wang, J. and Dilmore, R., 2014, August. Evaluating gas production performances in marcellus using data mining technologies. In Unconventional Resources Technology Conference, Denver, Colorado, 25–27 August 2014 (pp. 20–36). Society of Exploration Geophysicists, American Association of Petroleum Geologists, Society of Petroleum Engineers.

Zhu L, Zhang C, Zhang C, Wei Y, Zhou X, Cheng Y, Huang Y, Zhang L (2018) Prediction of total organic carbon content in shale reservoir based on a new integrated hybrid neural network and conventional well logging curves. J Geophys Eng 15(3):1050–1061