Petroleum Science 20 (2023) 295–308

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

A systematic machine learning method for reservoir identification and production prediction

Wei Liu [a, *], Zhangxin Chen [a], Yuan Hu [b], Liuyang Xu [c]

[a] Department of Chemical and Petroleum Engineering, University of Calgary, AB, Canada
[b] Rockeast Energy Ltd., Calgary, AB, Canada
[c] Jilin Oilfield, CNPC, Changchun, Jilin, 130000, China

ARTICLE INFO

Article history:
Received 15 February 2022
Received in revised form
1 September 2022
Accepted 5 September 2022
Available online 9 September 2022

Edited by Yan-Hua Sun

Keywords:
Reservoir identification
Production prediction
Machine learning
Ensemble method

ABSTRACT

Reservoir identification and production prediction are two of the most important tasks in petroleum exploration and development. Machine learning (ML) methods are used for petroleum-related studies, but have not been applied to reservoir identification and production prediction based on reservoir identification. Production forecasting studies are typically based on overall reservoir thickness and lack accuracy when reservoirs contain a water or dry layer without oil production. In this paper, a systematic ML method was developed using classification models for reservoir identification, and regression models for production prediction. The production models are based on the reservoir identification results. To realize the reservoir identification, seven optimized ML methods were used: four typical single ML methods and three ensemble ML methods. These methods classify the reservoir into five types of layers: water, dry and three levels of oil (I oil layer, II oil layer, III oil layer). The validation and test results of these seven optimized ML methods suggest the three ensemble methods perform better than the four single ML methods in reservoir identification. The XGBoost produced the model with the highest accuracy; up to 99%. The effective thickness of I and II oil layers determined during the reservoir identification was fed into the models for predicting production. Effective thickness considers the distribution of the water and the oil resulting in a more reasonable production prediction compared to predictions based on the overall reservoir thickness. To validate the superiority of the ML methods, reference models using overall reservoir thickness were built for comparison. The models based on effective thickness outperformed the reference models in every evaluation metric. The prediction accuracy of the ML models using effective thickness were 10% higher than that of reference model. Without the personal error or data distortion existing in traditional methods, this novel system realizes rapid analysis of data while reducing the time required to resolve reservoir classification and production prediction challenges. The ML models using the effective thickness obtained from reservoir identification were more accurate when predicting oil production compared to previous studies which use overall reservoir thickness.

1. Introduction

Big data of reservoir and oil production is exponentially expanding. The traditional methods employed to identify reservoirs and predict their production cannot efficiently use historical information and new data. Geologists conduct reservoir identification based on large amounts of geophysical data and numerical simulators cannot take full advantage of all the reservoir information due to model scaling (Rodríguez et al., 2014; Siddiqi and Andrew, 2002). Traditional reservoir simulators require a fixed set of parameters, some of which are very difficult to obtain (e.g., skin factor, compression index and capillary force) and other useful parameters cannot be incorporated into these simulators (e.g., dynamic oil level). The combination of inaccurate data and missing information results in miscalculations being used for prediction. In the era of big data, it is increasingly necessary to develop an effective and reliable technique that will maximize the benefits of data explosion while making full use of massive reservoir data, which will help resolve reservoir identification and production prediction validity challenges.

Machine learning (ML) as a subdivision of artificial intelligence

* Corresponding author.
E-mail address: wei.liu2@ucalgary.ca (W. Liu).

https://doi.org/10.1016/j.petsci.2022.09.002

Nomenclature

AC	Acoustic log
AI	Artificial intelligence
ANN	Artificial neural network
CART	Classification and regression tree
DL	Dry layer
DT	Decision tree
FN	False negative
FP	False positive
GR	Gamma ray
GBDT	Gradient boosting decision trees
KNN	*k*-nearest neighbors
LLD	Deep laterolog
LR	Logistic regression
MAE	Mean absolute error
ML	Machine learning
PER	Permeability
POR	Porosity
R^2	Correlation coefficient
RF	Random forest
RMSE	Root mean squared error
SP	Spontaneous potential
S_w	Water saturation
TN	True negative
TP	True positive
WL	Water layer
XGB	XGBoost
IO	I oil layer
IIO	II oil layer
IIIO	III oil layer
Day	The number of production days
$F_k(\boldsymbol{x})$	Predicted value in GBDT
f_k	Tree structure
H_h	Hidden layers
$h_\theta(x)$	Prediction function of LR
k_{avg}	The mean of permeability values
L_{do}	A dynamic oil level
L_p	Distance between samples
l	Loss function
n	Number of classifications
$\boldsymbol{R}^n$	Feature space
S_{rw}	The residual water saturation
T	Thickness of the predicted IO and IIO
T_r	Real thickness of IO and IIO
T_o	Overall reservoir thickness
T'	Number of leaf nodes
V_k	A permeability variation coefficient
Wr	The water content ratio in the first month
w	Score on leaf node
w_{hj}	Weight matrix between a hidden layer and the output layer
w_{ih}	Weight matrix between the input layer and a hidden layer
x	Original sample parameter
x'	New parameter
X_{min}	Minimum value of sample
X_{max}	Maximum value of sample
Y_j	Output layer
$\widehat{y}_i$	Prediction
y_i	Observation
y_k	True value in GBDT
y_i^{obs}	Observed data
y_i^{pred}	Predicted data
ϕ_{avg}	Mean porosity value for each well
$\boldsymbol{\theta}_h$	Threshold matrix associated with a hidden layer
$\boldsymbol{\theta}_j$	Threshold matrix associated with the output layer
Ω	Regularization term
γ	Parameter to control the regularization
λ	Parameter to control the regularization

(AI) has been applied in various fields with a positive impact for many years. Practical applications of ML techniques have been widely investigated in petroleum engineering, including reservoir characterization (Anifowose et al., 2017; Chaki et al., 2018), prediction of reservoir properties (Helmy et al., 2013; Anifowose et al., 2015; Priezzhev and Stanisalav, 2018) and production prediction (Chakra et al., 2013; You et al., 2019). Some studies have applied ML techniques to petroleum geology (Raeesi et al., 2012; Merembayev et al., 2018) where several input parameters are selected related to geological characteristics of a reservoir and its operating conditions. To predict lithofacies or reservoir properties (e.g., porosity and permeability), related well log data has been used to train a predictive model. After learning the underlying relationship between input variables and an output target, this data-driven model is finally applied to forecast specific lithofacies or reservoir properties.

In the previous research, several ML techniques have been introduced to solve classification and regression problems in petroleum engineering and geology. Among them, artificial neural network (ANN) and random forest (RF) are commonly used due to their remarkable performance. The ANN technique has been applied to lithology identification and recognition with well log data (Ren et al., 2019; Kamenski et al., 2020) using the back propagation neural network (BPNN) to find patterns that identify the lithology. This technique has been used to predict oil production of existing wells (Awoleke and Lane, 2011; Van and Chon, 2018). To forecast the oil production, well log data related to reservoir geological characteristics and dynamic operation data have been used to build a representative prediction model. The RF technique has been applied to geological and geochemical data for lithology identification by previous researchers. It outperforms other ML algorithms in this area (Harris and Grunsky, 2015; Cracknell and Reading, 2012). In addition to the classification performance, this technique provides reliable predictions for geological mapping applications (Radford et al., 2018).

Reservoir identification is the fundamental work necessary for production forecasting. Most previous ML studies about reservoirs were focused on lithofacies classification or reservoir properties and very few mention reservoir identifications, let alone the combination of reservoir identification and production prediction. Historically, the prediction of oil production leaned heavily on the overall thickness of reservoir as a key data input (Guo et al., 2021). This technique oversimplified the reservoir which contains not only the good oil layer, but other layers (water, dry and poor oil) that do not or cannot produce oil. There is no proven direct relationship between the overall thickness of reservoir and the volume of oil production. For example, other conditions being equal, a 20-m-thick reservoir A is theoretically predicted to produce more oil than 10-m-thick reservoir B. In fact, if reservoir A is mainly composed of water and dry layers and reservoir B is formed by almost all good oil

layers, the production of reservoir B will be better than that of reservoir A. The heavily used method of predicting reservoir production based on the overall thickness of reservoir is not accurate. To solve this problem, before predicting the production in this study, the reservoir was first classified by the five types of layers found within a reservoir: water, dry and three different levels of oil layers. Among these, I oil layer (IO) and II oil layer (IIO) were defined as effective reservoirs producing industrial oil flows that are perforated during production. In the subsequent production prediction, the thickness of IO and IIO (effective thickness) was used as an important variable instead of the overall reservoir thickness commonly used by predecessors.

In this study, several ML methods were used and compared to identify effective reservoirs in oilfields. To further predict their production, this study implemented the prediction of cumulative production for new wells and existing wells. This is different from most previous studies on oil production which focused on subsequent production and production decline rates of existing wells. The prediction of effective reservoirs and other production variables were used to train a predictive model for production. In this way, an integrated ML system was developed that integrates the whole industrial process from the reservoir identification to prediction of oil production, increasing the accuracy and efficiency of production prediction by making full use of the obtained results from the reservoir identification process. Moreover, two reference models were built to compare and prove that the prediction results from the reservoir identification process were reliable enough to be used in production models. In reference model I (RM I), real data for effective thickness (thickness of IO and IIO) was fed into ML production models to compare with the production models based on reservoir identification models. Reference model II (RM II) used the overall reservoir thickness to compare the prediction result with production models using effective thickness.

A ML model was used to predict and classify potential reservoirs into several known reservoir types according to the selected input features. Then, the predictive results from the reservoir identification were fed into an ML model to predict oil production. The full use of reservoir information increased the prediction accuracy of the production models. This systematic ML method for reservoir identification and prediction of oil production reduces required human resources and thus reduces the volume of human errors.

2. Methodology

2.1. The systematic ML method

Fig. 1 is the flowchart illustrating the procedures of the study. Firstly, in the reservoir classification process, seven ML models (logistic regression (LR), *k*-nearest neighbors (KNN), decision tree (DT), ANN, RF, gradient boosting decision trees (GBDT) and XGBoost (XGB)) were compared to determine the best method for reservoir classification. Then the predicted thickness of the effective reservoir combined with other production features were fed into the production prediction process. In this process, after comparing all the classification processes, 2 ML models (ANN and XGBoost) were selected to predict the production. After training and testing, the prediction results were compared to two reference models.

2.2. Pre-processing

In the process of the ML technique, various parameters (inputs) fed into the models had different dimensional units that affected the data analysis results. To eliminate the interference between dimensional units, data preprocessing using normalization (Eq. (1)) was required to calculate different parameters (Raschka, 2015):

$$x' = (x - X_{\min})/(X_{\max} - X_{\min}) \tag{1}$$

where x is an original sample parameter; x' is the new parameter; $X_{\min}$ is the minimum value of the sample; and $X_{\max}$ is the maximum value of the sample.

2.3. Approach

The objective of this study was to provide an advanced and alternative approach that accurately identifies a reservoir and predicts oil production reliably. These two prediction tasks are supervised learning problems as the samples have input features and corresponding outputs. According to the types of predicting results, reservoir identification is a supervised classification problem since its reservoir type as the output value is a discrete value. The production prediction is a supervised regression problem because production is a continuous value as output. For a classification problem (reservoir identification), seven classifiers are selected: LR, KNN, DT, ANN, RF, GBDT, and XGB. For a regression problem (production prediction), ANN and XGB were selected to show their predictive results due to their better performance when compared to other ML methods. These classifiers are briefly reviewed below.

2.3.1. Logistic regression

LR (Cox, 1958) is a regression analysis method where a dependent variable is categorical and used for binary classification. It can be generalized to multiclass problems. As shown in Eq. (2), LR uses a nonlinear sigmoid function for classification prediction:

$$g(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

Assume that the eigenvector influencing the prediction result is $\boldsymbol{x} = (1,\ x_1,\ x_2, \cdots, x_n)$ and the regression coefficient is $\boldsymbol{\theta} = (\theta_0,\ \theta_1,\ \theta_2, \cdots, \theta_n)$; then we see that

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n = \sum_{i=1}^{n} \theta_i x_i = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x} \tag{3}$$

We construct the prediction function as:

$$h_\theta(x) = g\left(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}\right) = \frac{1}{1 + e^{-\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}}} \tag{4}$$

If $\boldsymbol{\theta}$ is known, $h_\theta(x)$ can be used to calculate eigenvector $\boldsymbol{x}$. If the result is greater than 0.5, it is classified as 1; otherwise, it is classified as 0.

2.3.2. k-nearest neighbors

The classification rule of KNN is: the label of an unclassified sample point is determined by the label of the maximum class in the nearest k neighboring points (Liu et al., 2022). In this paper, the Minkowski distance L_p was used as the metric for measuring the distance between two sample points. Suppose that the feature space is $\boldsymbol{R}^n$, $\boldsymbol{x}_i,\ \boldsymbol{x}_j \in \boldsymbol{R}^n$, $\boldsymbol{x}_i = (x_i^1,\ x_i^2, \cdots, x_i^n,\ )^{\mathrm{T}}$, $\boldsymbol{x}_j = (x_j^1,\ x_j^2, \cdots, x_j^n,\ )^{\mathrm{T}}$. The distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in L_p is defined as follows:

$$L_p(x_i, x_j) = \left(\sum_{l=1}^{n} \left| x_i^l - x_j^l \right|^p \right)^{1/p} \tag{5}$$

where p is a parameter to determine the distance type; $\boldsymbol{x}_j$ is a training sample point; and $\boldsymbol{x}_i$ is a point needed to predict the output

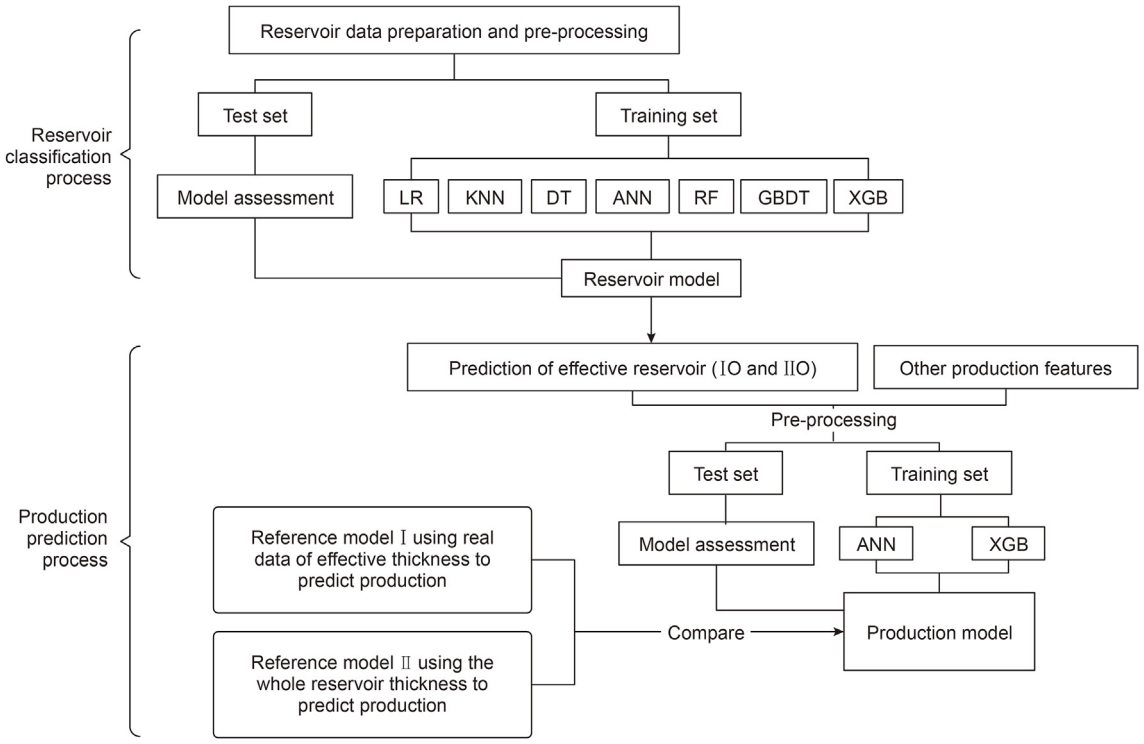

Fig. 1. Flowchart of the research progress.

class. Therefore, p and k are two significant hyper-parameters when building a KNN model.

2.3.3. Decision tree

Classification and regression tree (CART) is the most popular algorithm used to build a DT. CART (Breiman et al., 1984) develops a Gini index to choose a feature for splitting a tree. The Gini index reflects the probability that two samples are randomly selected from a data set and their labels (classes) are different. A lower Gini index indicates greater purity of the data set.

If the sample set $\boldsymbol{D}$ is split into $\boldsymbol{D}1$ and $\boldsymbol{D}2$ using the discrete feature A, the Gini index calculated after splitting is defined as:

$$\text{Gain\_Gini}(\boldsymbol{D},\ A) = \frac{|\boldsymbol{D}1|}{|\boldsymbol{D}|}\text{Gini}(\boldsymbol{D}1) + \frac{|\boldsymbol{D}2|}{|\boldsymbol{D}|}\text{Gini}(\boldsymbol{D}2) \tag{6}$$

Therefore, $\text{Gain\_Gini}(\boldsymbol{D},\ A)$ is the uncertainty after splitting. A smaller $\text{Gain\_Gini}(\boldsymbol{D},\ A)$ value is preferred because this provides greater purity for a data set.

2.3.4. Artificial neural network

ANN is a computing system that imitates the working of the human brain in learning patterns from experience and processing data to solve classification and regression problems. ANN is comprised of an input layer, hidden layers H_h and an output layer Y_j. The hidden layers need to be artificially set according to an actual situation so that a model can achieve the best prediction outcome.

For a classification problem in this study, the activation function used in a hidden layer is a Sigmoid function (Eq. (7)), while the activation function used in the output layer is a Softmax function (Eq. (8)) to solve a multi-class problem:

$$H_h = f\left(\sum_i \boldsymbol{w}_{ih}\cdot\boldsymbol{X}_i - \boldsymbol{\theta}_h\right) = \frac{1}{1+\mathrm{e}^{\boldsymbol{\theta}_h - \sum_i \boldsymbol{w}_{ih}\cdot\boldsymbol{X}_i}} \tag{7}$$

$$Y_j = f\left(\sum_h \boldsymbol{w}_{hj}\cdot\boldsymbol{H}_h - \boldsymbol{\theta}_j\right) = \frac{\mathrm{e}^{\sum_h \boldsymbol{w}_{hj}\cdot\boldsymbol{H}_h - \boldsymbol{\theta}_j}}{\sum_{i=1}^{n}\mathrm{e}^{\sum_h \boldsymbol{w}_{hj}\cdot\boldsymbol{H}_h - \boldsymbol{\theta}_i}} \tag{8}$$

where $\boldsymbol{w}_{ih}$ is the weight matrix of the node connections between the input layer and a hidden layer; $\boldsymbol{w}_{hj}$ is the weight matrix of the node connections between a hidden layer and the output layer; $\boldsymbol{\theta}_h$ is the threshold matrix associated with a hidden layer; $\boldsymbol{\theta}_j$ is the threshold matrix associated with the output layer; and n is the number of classifications.

In Eq. (9), a ReLU activation function was applied in this study to solve a regression problem:

$$\text{ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \le 0 \end{cases} \tag{9}$$

2.3.5. Random forest

RF is an ensemble learning method. It establishes an advanced model based on a bagging technique (Breiman, 1996) and a random feature selection technique (Ho, 1998). Bagging is a method used for sampling randomly with replacement and helps to generate several new single trees to reduce variance. For example, to solve a reservoir identification problem by RF, each tree provides the prediction of a possible reservoir type. The final decision depends on the majority voting from single trees. The random feature selection technique is useful to make all the trees uncorrelated and further reduce the variance of prediction.

2.3.6. Gradient boosting decision trees

Different from RF, GBDT is an ensemble ML tool using a boosting technique, which sequentially generates base models and improves the predictive power of the ensemble through incremental minimization of residual errors in each iteration of construction of a new base model (Brown and Mues, 2012). While building a classification model in GBDT, samples that are misclassified in a previous base model are more likely to be assigned an increased weight in the next step. The new model has improved prediction accuracy compared to a previous model. A loss function is used to measure the difference between the predicted $F_k(\boldsymbol{x})$ and true values y_k to indicate how well a model fits the data.

In a GBDT algorithm for a multi-class problem, the loss function is (Friedman, 2001):

$$L\left(\{y_k, F_k(\boldsymbol{x})\}_1^K\right) = -\sum_{k=1}^{K} y_k \log p_k(\boldsymbol{x}) \tag{10}$$

where $y_k = 1$ (class $= k$) $\in \{0,1\}$, $p_k(\boldsymbol{x}) = P(y_k = 1|\boldsymbol{x})$, and

$$p_k(\boldsymbol{x}) = \exp(F_k(\boldsymbol{x})) \Big/ \sum_{l=1}^{K} \exp(F_l(\boldsymbol{x})) \tag{11}$$

2.3.7. XGBoost

XGB is one of the most popular methods in the ensemble machine learning category today. It performs very well in multiple programming competitions like Kaggle (Chen and Guestrin, 2016). XGB is a machine learning technique for classification and regression problems. It produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Based on the concept of GBDT, XGB uses a regularized model formalization to control over-fitting, which leads to better performance.

Distinct from GBDT, the objective function of XGB consists of a loss function and a regularization term (Chen et al., 2016):

$$L = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{12}$$

$$\Omega(f) = \gamma T' + \frac{1}{2}\lambda \|w\|^2 \tag{13}$$

where l is a loss function as in GBDT, and measures a difference between prediction $\hat{y}_i$ and observation; A regularization term Ω is added in the objective function to control over-fitting and contribute to better performance and flexible complexity; f_k represents a specific tree structure; T' and w denote the number of leaf nodes and the score on each node respectively; γ and λ are parameters to control the regularization.

2.4. Evaluation metrics

Evaluation metrics are essential to measure the quality of a ML model. It is worth noting that different types of evaluation metrics are applicable to different tasks and emphasize different aspects of a model's performance (Andika and Chandima Ratnayake, 2019). For example, accuracy is often selected for its easy-to-use scoring and flexibility for multiclass problems (Hossin and Sulaiman, 2015). The root mean squared error (RMSE), mean absolute error (MAE), and correlation coefficient (R^2) measures are popular for their effectiveness and common use to solve regression problems in many petroleum applications (Abdulraheem et al., 2007; Chakra et al., 2013). In this study, to deal with a reservoir identification–classification problem and a production prediction–regression problem separately, two sets of metrics were selected and introduced below.

2.4.1. Classification problems

Before discussing the metrics for a classification problem, it is necessary to introduce four elements in a binary problem: true positive (TP), false negative (FN), false positive (FP) and true negative (TN). The 'true' and 'false', respectively, mean whether a prediction is correct or not compared to real data. The 'positive' and 'negative' represent whether a prediction class is the same as a specific class or not. To assess the predictive performance of the above classifiers and select the corresponding optimal models, a matrix of accuracy, precision, recall and f1 score are evaluated. Accuracy is based on the ratio of correctly predicted samples to the total samples. Precision is defined as the ratio of correct positive predictions. Recall is the ratio of actual positive results correctly predicted. f1 score is the harmonic mean of the precision and recall. The higher the f1-score value the better precision and recall.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{14}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

$$\text{f1} = 2\frac{1}{{}^1/_{\text{recall}} + {}^1/_{\text{precison}}} = 2\frac{\text{precison} \cdot \text{recall}}{\text{precison} + \text{recall}} \tag{17}$$

2.4.2. Regression problems

In this study, the following measurements were applied to substantiate the statistical accuracy of the performance of ANN and XGB for production prediction: RMSE, MAE, and R^2. The RMSE is a measure of the spread of actual values around the average of the predicted values. It computes the average of the squared differences between each predicted value and its corresponding actual value. It is expressed as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{\text{obs}} - y_i^{\text{pred}}\right)^2} \tag{18}$$

where y_i^{obs} is the observed data; y_i^{pred} is the predicted data; and n is the number of data points. The MAE is a statistical measure of dispersion. It is computed by taking the average of the absolute errors of the predicted values relative to the actual values. It is given by:

$$\text{MAE} = \frac{\sum_{i=1}^{n}\left|y_i^{\text{obs}} - y_i^{\text{pred}}\right|}{n}. \tag{19}$$

R^2 assesses the quality of a model prediction by observing the difference between predicted data and actual data (Nash and Sutcliffe, 1970). It is expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{\text{obs}} - y_i^{\text{pred}}\right)^2}{\sum_{i=1}^{n}\left(y_i^{\text{obs}} - \overline{y_i^{\text{obs}}}\right)^2} \tag{20}$$

where $\overline{y_i^{\text{obs}}}$ is the average value of the observed data.

3. Case studies

3.1. Reservoir identification

The data was acquired from a public domain of China National Petroleum Corp. (CNPC). It is comprised of logging data and layer thickness (eight features) from 124 wells. A total of 2800 samples (single layers) made up the dataset. Seven ML models (LR, KNN, DT, ANN, RF, GBDT and XGB) were constructed involving one output variable reservoir classification for each single layer and a total of eight input features (see description in Table 1 below) including *LLD, GR, AC, SP, POR, PER,* S_W and the thickness of each layer. Table 1 shows details and statistical descriptions of input features used for reservoir identification. The output reservoir classification includes five classes: 1) Dry layer (DL) - no oil, gas or formation water, 2) Water layer (WL) - only containing formation water, and 3) three levels of oil layer - IO, IIO and III oil layer (IIIO). Among these three types of oil layers, IO and IIO were defined as effective reservoir for their high and medium industrial value and IIIO is defined as a worthless reservoir due to its poor industrial value as assessed by CNPC.

After data pre-processing, the samples were randomly split into a training set of 2500 samples and a testing set of 300 samples. The training set is used to develop a model to perform a single layer classification and the testing set is applied to the trained model to estimate how well the model has been trained. A hyper-parameter tuning process and 10-fold validation were used in the seven classifiers to choose the best combination of hyper-parameter values for each model. In this study, *Grid Search* is used to find the optimal hyper-parameters of a model that results in the most accurate predictions. *Grid Search* is a function that comes in *Scikit-learn*'s *model_selection* package. Firstly, the values of hyper-parameters were passed to the *Grid Search* function by defining a dictionary containing a particular hyper-parameter along with the values it can take. Then *Grid Search* tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the 10-fold validation method. After using this function, the optimal combination of hyper-parameters with the highest prediction accuracy can be selected.

3.2. Prediction of production

In this case study, five different ML methods (DT, ANN, RF, GBDT and XGB) were compared for the prediction of production. After comparison, ANN and XGB demonstrated the best performance amongst the methods. These two methods were employed to forecast the oil recovery performance for a series of producing wells. ANN and XGB models were respectively constructed with one output variable which is the single well's first five months cumulative oil production. Eight input variables were used: 1) T = the total thickness of IO and IIO (effective thickness) obtained from the results of reservoir identification, 2) ϕ_{avg} = the mean porosity value for each well, 3) k_{avg} = the mean of permeability values, 4) V_k = a permeability variation coefficient (Eq. (21)), 5) S_{rw} = the residual water saturation, 6) Wr = the water content ratio in the first month, 7) L_{do} = a dynamic oil level, and 8) Day = the number of production days. To be as representative as possible, ϕ_{avg} and k_{avg} are weighted averages in T. Table 2 shows the unit and statistical descriptions of input features in prediction of oil production.

$$V_k = \frac{\sqrt{\sum_{i=1}^{n}(K_i - \overline{K})^2 / n}}{\overline{K}} \tag{21}$$

where K_i is the permeability of a single layer; $\overline{K}$ is the average permeability of all layers; and n is the number of layers in the well.

All 124 well records were subjected to the pre-processing stage mentioned above. They were all used in the training and testing phases of ANN and XGB. The best constructions of ANN and XGB models were determined by a grid search. Two thirds of the original data were used as a training data set. The remaining one third was employed as the testing data set. Finally, to verify whether the prediction result in Section 3.1 was reliable enough to be used for production prediction and to show its superiority compared to previous studies, the final prediction results were compared to two reference models. In RM I, T_r is the thickness of real IO and IIO. In RM II, T_o is the overall reservoir thickness.

4. Results and analysis

4.1. Comparative analysis of classification models for reservoir identification

4.1.1. Results of training and validation

In a hyper-parameter tuning process, accuracy is used as the metric to measure the model performance. Table 3 shows the optimal hyper-parameter combination of each classification model builds the best predictive model.

Accuracy, precision, recall and f1-score were used to evaluate the performance of the seven ML methods using the 10-fold-cross-validation. In Fig. 2, a box-and-whisker plot was utilized to assess the statistical dispersion of each classifier's accuracy on all folds based on their optimal hyper-parameters, respectively. Single classifiers showed reasonable performance in terms of average value of accuracy – ANN (82.61%), KNN (82.58%), DT (80.87%) and LR (73.89%). XGB, RF and GBDT had the top three average accuracies

Table 1
Description of input features in reservoir identification.

Feature	Nomenclature	Unit	Min	Mean	Max
LLD	Deep laterolog	Ω m	8.2	58.5	24,430
GR	Gamma ray	API	23.2	68.1	127.6
AC	Acoustic log	μs/m	169.1	236.3	411.5
SP	Spontaneous potential	mV	−157	−15.9	28.2
POR	Porosity	%	0.1	10.3	23.3
PER	Permeability	mD	0.01	14.7	1976.9
S_w	Water saturation	%	0.1	76.2	291.6
Thickness	Thickness of each layer	m	0.7	3.2	40.8

Table 2
Description of input features in production prediction.

Feature	Unit	Min	Mean	Max
T	m	3.4	39.9	107.5
ϕ_{avg}	%	5.4	12.1	17.5
k_{avg}	mD	0.7	30.2	533.4
V_k	/	58.2	0.1	936.6
S_{rw}	%	30.3	48.7	66.3
Wr	%	0.5	23.9	100
L_{do}	m	70.1	1518.3	2243.5
Day	day	6	109	145

Table 3
Tuned optimal hyper-parameter values of seven classification methods.

Classification method	Tuned hyper-parameter	Optimal hyper-parameter setting
LR	Penalty parameter determining the strength of regularization (*penalty*)	L2
KNN	The number of neighbors (*k*)	7
	The number used to calculate distance (*p*)	6
DT	The maximum number for tree depth (max *depth*)	10
	The minimum number of samples required to split an internal node (min *samples split*)	5
ANN	Learning rate	0.01
	Maximum number of learning iterations (max *iter*)	500
	Solver for weight optimization (*solver*)	Adam
RF	The number of trees in the forest (*n estimators*)	200
	The maximum depth of the individual estimators (max *depth*)	7
GBDT	Learning rate	0.3
	The number of estimators in a model (*n estimators*)	500
	The maximum depth of individual estimators (max *depth*)	15
XGB	Learning rate	0.1
	The number of estimators in a model (*n estimators*)	500
	The maximum depth of individual estimators (max *depth*)	10

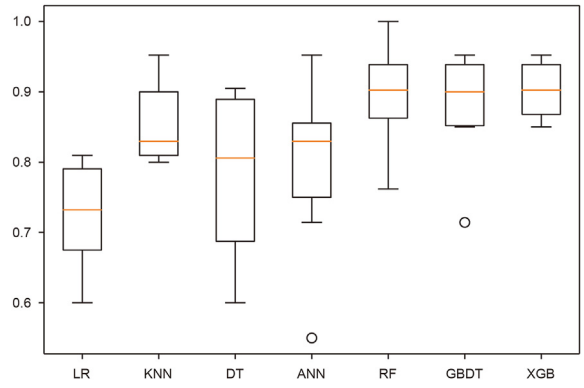

Fig. 2. Box plots for accuracy of seven ML methods from 10-fold cross validation.

(91.74%, 89.36%, and 89.25%, respectively). This is interpreted as validation of the ensemble classification techniques (RF, GBDT and XGB) to generally produce better results than single classifiers (LR, KNN, DT and ANN). Despite some outlier points, the accuracy value of GBDT and XGB had the least variance and the highest stability. XGB had the highest average accuracy.

Table 4 shows the precision, recall and f1-score of each reservoir class for seven ML methods. In the identification of the five reservoir classes, DL and IO had the highest prediction accuracy. The IIIO class is likely to be misclassified as another kind of reservoir because there were very few original samples labeled as class IIIO. Table 5 shows the number of samples used for each type of reservoir during the training and testing process. The IIIO class reservoir makes up about 10% of the samples compared to the other types of reservoirs. The lack of samples leads to poor learning performance for every ML method. The overall performance of three ensemble ML methods was better than the other four single methods for identification of the five reservoir classes. After ensemble methods, ANN was the best classifier as a single method, but its precision of IIIO was only 0.58 and prediction performance of other reservoir types were all much weaker than the ensemble classifiers. As the effective reservoir, the identification of IO and IIO is more important than that of other reservoir classes. XGB had the best performance with all metric values $\geq$ 0.92 in the IO class identification and up to 0.86 in the IIO class identification, followed by GBDT. The results of Fig. 2 and Table 4 indicate that the ensemble methods

Table 5
Number of samples in each type of layer.

	DL	WL	IO	IIO	IIIO
Training set	739	506	628	554	73
Testing set	79	74	72	66	9

Table 4
Precision, recall and f1-scores for 10-fold cross validation for seven classifiers.

Method	Output reservoir	Precision	Recall	f1-score	Method	Output reservoir	Precision	Recall	f1-score
LR	DL	0.82	0.87	0.84	KNN	DL	0.87	0.90	0.89
	WL	0.76	0.57	0.65		WL	0.78	0.68	0.73
	IIIO	0.40	0.21	0.28		IIIO	0.45	0.60	0.51
	IIO	0.69	0.72	0.70		IIO	0.77	0.78	0.77
	IO	0.82	0.83	0.82		IO	0.88	0.85	0.85
DT	DL	0.85	0.81	0.83	DT	DL	0.88	0.90	0.89
	WL	0.61	0.68	0.64		WL	0.82	0.74	0.78
	IIIO	0.38	0.25	0.30		IIIO	0.58	0.64	0.61
	IIO	0.74	0.76	0.75		IIO	0.77	0.81	0.79
	IO	0.86	0.82	0.84		IO	0.89	0.88	0.88
RF	DL	0.88	0.94	0.91	GBDT	DL	0.89	0.93	0.91
	WL	0.83	0.70	0.80		WL	0.81	0.77	0.79
	IIIO	0.73	0.68	0.70		IIIO	0.80	0.76	0.78
	IIO	0.80	0.84	0.82		IIO	0.82	0.85	0.83
	IO	0.92	0.88	0.90		IO	0.92	0.91	0.91
XGB	DL	0.90	0.94	0.91					
	WL	0.83	0.84	0.83					
	IIIO	0.82	0.81	0.82					
	IIO	0.84	0.86	0.85					
	IO	0.93	0.92	0.92					

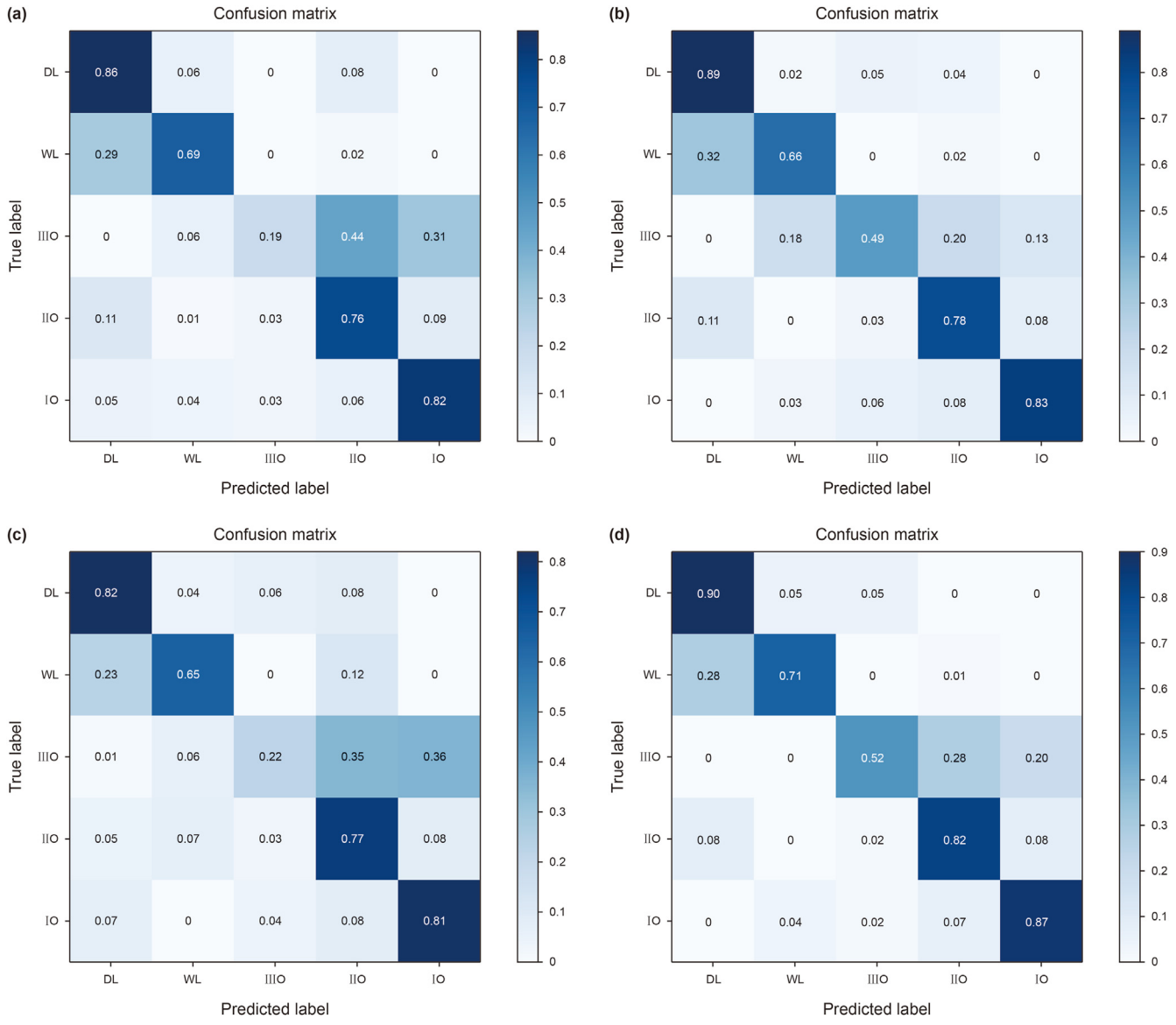

Fig. 3. Confusion matrix plots of single methods on the test dataset: (**a**) LR model; (**b**) KNN model; (**c**) DT model; (**d**) ANN model.

achieved better classification results compared to the other four single methods. Among the ensemble methods, XGB and GBDT are preferred for their higher accuracy, precision, recall and f1-scores.

4.1.2. Results of testing

Figs. 3 and 4 present the reservoir classes that were correctly classified or misclassified in the test dataset for single models and ensemble models. The testing result showed that LR, KNN and DT predict DL and IO with an accuracy greater than 0.8. ANN provided the best identification performance for a single method, but the prediction accuracy for each reservoir was still lower than the ensemble methods. This was consistent with the results from a 10-fold cross validation. In the confusion matrices below, XGB is the optimal ML method for overall performance. It identified IO, IIO and DL with $\geq$0.9 accuracy. Even training with very few samples in IIIO class, XGB still predicted this reservoir with an accuracy up to 0.76. The prediction of XGB was selected as the final reservoir identification result from all the ML methods.

The confusion matrix result of XGB (Fig. 4c) shows the probability of a real IO reservoir being predicted as the IO reservoir was 92% and the probability of being predicted as the IO or IIO reservoir was up to 99%. This was interpreted to mean the probability of being predicted as another type of reservoir (not an IO or IIO reservoir) was only 1%. The probability of the real IIO reservoir being predicted as the IIO reservoir was 90% and the probability of being predicted as the IO or IIO reservoir was 95%, which means the probability of being predicted as another type of reservoir (any reservoir except for IO or IIO) was only 5%. Although the prediction accuracy of the XGB model for each type of reservoir was $\leq$92%, the prediction success rate for effective reservoirs (IO and IIO) was very high - up to 99%. Using the feature importance method, Fig. 5 shows importance score ranking of different input features calculated by XGB. S_w is the most valuable feature for reservoir identification. *POR* and *PER* were the second and third most relevant features contributing to accurate prediction. Thickness of reservoir is the least important feature in this case.

To train and test the predictive model for reservoir identification, the ANN model needs several minutes to finish the work, while the other six ML models require several seconds. By contrast, traditional methods used to identify a reservoir could be much time consuming and require additional manpower. An experienced geologist requires days or even weeks to complete the

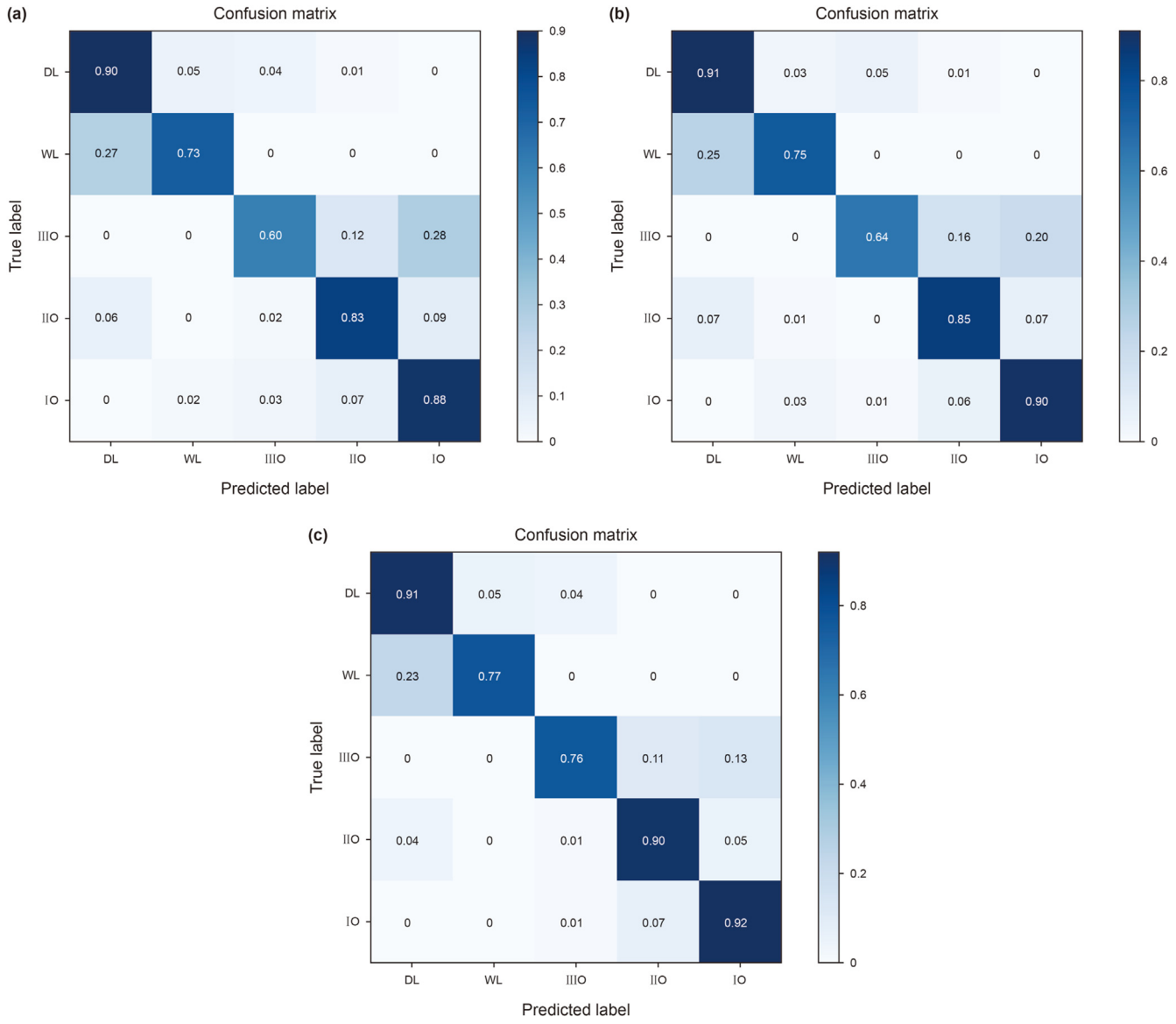

Fig. 4. Confusion matrix plots of ensemble methods on the test dataset: (**a**) RF model; (**b**) GBDT model; (**c**) XGB model.

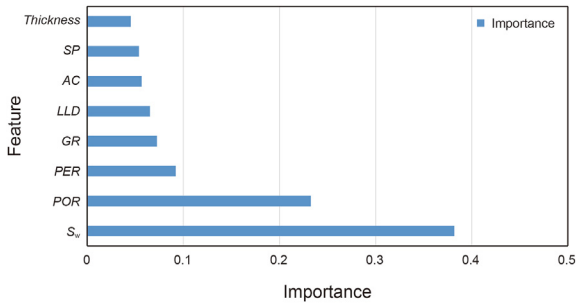

Fig. 5. Importance score rank of input features in reservoir identification.

identification of thousands of reservoirs and the accuracy may not reach 99% due to the human error. Other traditional methods based on logging data need a lot of calculations, which are time consuming and often as not as accurate as the identification determined by a geologist. The ML method's speed of seconds or minutes for a highly accurate reservoir identification completely supersedes traditional methods.

4.2. Analysis of regression models for production forecasting

4.2.1. ANN model

The ANN model configuration was set to have two hidden layers or three hidden layers, each with six nodes per layer. In Fig. 6, the loss shown is the value of MAE plotted per number of epochs for each ANN model configuration in the training and validation phases. The loss stabilized after 800 epochs and there was little further decrease in the mismatch between ANN prediction and real target values. To increase the accuracy of the ANN prediction, increasing the number of hidden layers was necessary to learn more about the relationship between the input variables and the output target.

A new configuration for ANN was built with four hidden layers, each with six nodes, running 1000 epochs to train the model and estimate the target. In Fig. 7, the loss curve in the training data declines continuously and the loss of validation data begins to increase after 100 epochs. This challenge is called overfitting. The

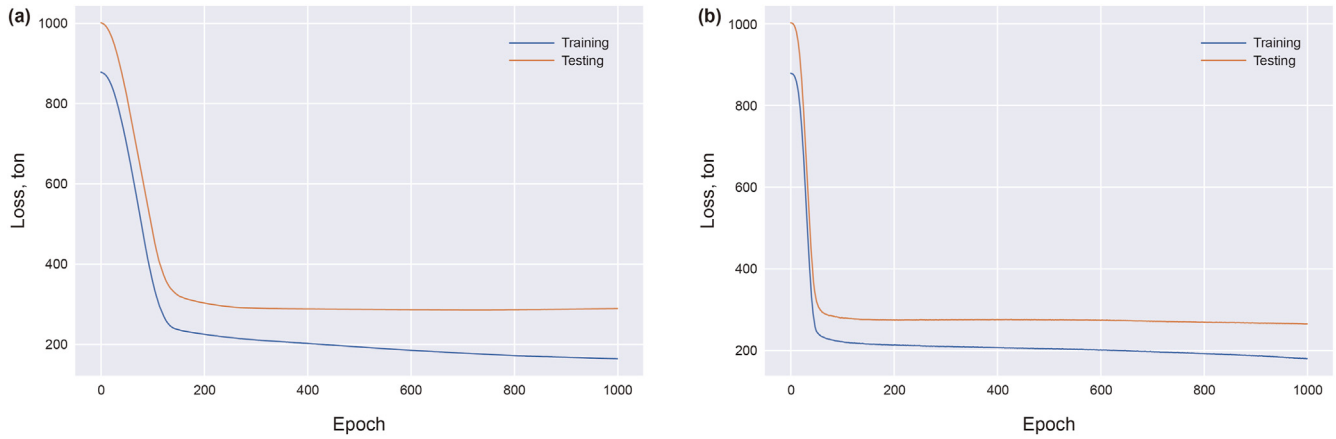

Fig. 6. Loss vs. epoch curve in the training data and the validation data: (**a**) Two hidden layered ANN model; (**b**) Three hidden layered ANN model.

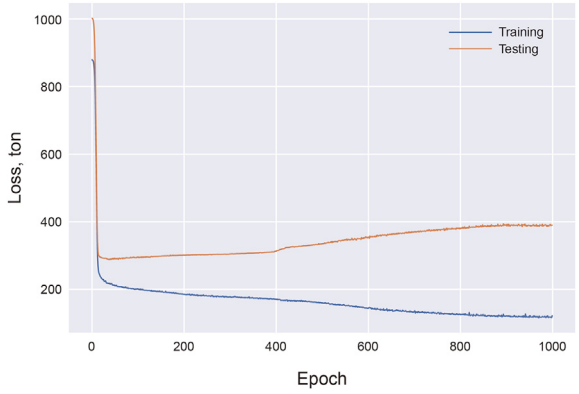

Fig. 7. Loss vs. epoch curve of four hidden layered ANN model in the training data and the validation data.

Table 6
Model results of ANN and two reference models.

Metrics	Training			Testing		
	ANN	RM I	RM II	ANN	RM I	RM II
RMSE, ton	246.347	231.916	298.038	321.711	315.228	399.703
MAE, ton	174.183	166.374	235.917	258.414	250.341	356.315
R^2	0.879	0.895	0.822	0.795	0.801	0.704

model learns a great degree of error or random noise within training data and then its predictive power is reduced. Finally, the construction with four hidden layers, each with six nodes and running 100 epochs were selected for the ANN model to forecast the oil production.

Table 6 illustrates the model performance of ANN and two reference models (RM I and RM II) in the prediction of production.

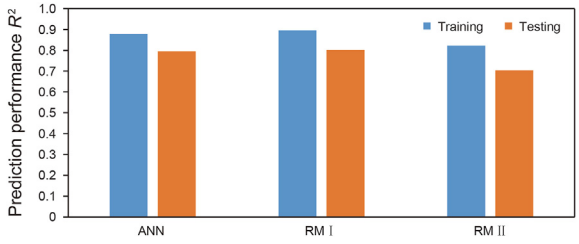

Fig. 8. Comparison of prediction performance (R^2) of ANN and two reference models.

The R^2 of three models in training and testing process is compared in Fig. 8. For the ANN model, the configuration with 100 epochs provided a very reliable performance for estimating oil production using predicted reservoir information in the training process, where R^2 was 0.879, MAE was 174.183 ton and RMSE was 246.347 ton. The testing result was satisfied with R^2 being 0.795, MAE being 258.414 and RMSE being 321.711. To verify that the predicted effective reservoir thickness could replace the true data, the training and testing performance of RM I (using the real effective reservoir thickness) was compared to the ANN model. In the training and testing sets, the performance of RM I was only slightly better than the ANN model in three metrics. The negligible difference between these two models proved the practicability of the predicted effective reservoir. By contrast, RM II using the overall reservoir thickness instead of effective reservoir thickness had a weaker prediction performance than the ANN model in the training and testing processes. In the RM II testing set, R^2 was at 0.704, MAE at 356.315 ton and RMSE at 399.703 ton, which meant RM II accounted for 70.4% of the production variance in the research area and on average there was more than 350-ton uncertainty in the prediction of first 5 months cumulative oil production for each well. Therefore, the ANN model using predicted effective reservoir thickness was applicable for its similar performance compared to RM I and was much better than RM II with higher prediction accuracy. Using R^2 as the metric of accuracy for the testing process, the prediction accuracy of the ANN model with effective reservoir thickness was 13% higher than that of RM II.

4.2.2. XGB model

To determine the best combination of all hyper-parameter values in an XGB model, a grid search was used in this study. A limited number of values for each hyper-parameter were selected because it is not feasible to try the entire range of possible values. In the XGB model, three hyper-parameters were examined: (1) a learning rate, (2) the number of estimators in the model, and (3) the maximum depth of the individual regression estimators. In the learning rate, a value range of [0.1, 1], with 0.1 as the distance between two adjacent values, was assessed. For the number of estimators, a value range of [10, 50, 100, 250, 500] was evaluated, while a range of [1, 10], with 1 as the spacing between values was examined for the maximum depth. After going through all the possible combinations of the three hyper-parameters, the XGB model was built with a learning rate of 0.1, 50 estimators and a maximum depth of 4.

Table 7 and Fig. 9 show the prediction performance of the XGB

Table 7
Model results of ANN and two reference models.

Metrics	Training			Testing		
	XGB	RM I	RM II	XGB	RM I	RM II
RMSE, ton	26.04	25.881	69.475	237.753	234.792	298.924
MAE, ton	18.807	18.094	57.153	189.962	187.028	257.936
R^2	0.999	0.999	0.913	0.857	0.861	0.782

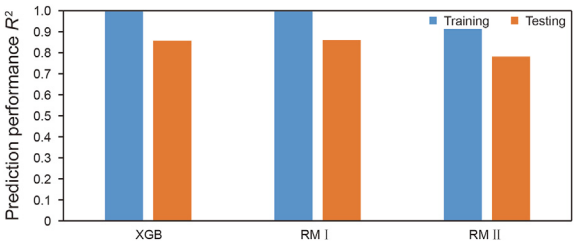

Fig. 9. Comparison of prediction performance (R^2) of XGB and two reference models.

model and its two reference models in three different metrics. XGB had a very reliable performance estimating oil production with the predicted reservoir data in the training and testing processes, with R^2 being 0.999 for training and R^2 being 0.857 for testing. Like the ANN model, the prediction results from training and testing of the XGB model using the predicted effective reservoir thickness was very similar to the prediction result of RM I with the real effective reservoir thickness. RM II with overall reservoir thickness had a lower prediction accuracy compared to the XGB model in the training and testing sets. The XGB model using predicted effective reservoir thickness was considered reliable because of its similar performance compared to RM I and had a more accurate prediction than RM II. In test datasets, the prediction accuracy (R^2) of XGB model with effective reservoir thickness was about 10% higher than that of RM II.

4.2.3. Comparative analysis of ANN and XGB

Table 8 expresses the comparative performance of ANN and XGB. XGB is preferred because it outperforms ANN in every evaluation metric for the training data and validation data sets. The two methods perform better in the training dataset when compared to the validation dataset. Different from ANN, the performance of the training data in XGB was significantly better than that of the validation data. This is because in ANN, a larger epoch can increase the learning and training accuracy, but at the same time it leads to an overfitting problem where the model learns a great degree of error or random noise within the training data and then its predictive power is reduced. To avoid overfitting and to find the best validation accuracy, the learning accuracy must be limited. Fig. 9 shows the cross plots of real oil production against predictions using the ANN model and XGB model. In Fig. 10a and b, the prediction of the ANN model performs well with training data and testing data, with R^2 being 0.8790 and 0.7950, separately. Fig. 9c and d provides the prediction performance of the XGB model. Higher values of R^2 in the training set (0.9986) and testing set (0.8575) prove the superiority of XGB in the prediction of production compared to ANN.

Table 8
Comparative performance of ANN and XGB.

Metrics	Training		Testing	
	ANN	XGB	ANN	XGB
RMSE, ton	246.347	26.04	321.711	237.753
MAE, ton	174.183	18.807	258.414	189.962
R^2	0.879	0.999	0.795	0.857

Since XGB performed better than ANN for prediction of oil production, the importance score of features in this case was calculated based on the prediction of XGB. Fig. 11 shows the importance score of different features using feature importance method for the XGB, RM I and RM II models. All three models had the same top 3 important features. L_{do} was the most valuable feature for the prediction of cumulative oil production; this cannot be used in traditional simulations. *Day* and *Wr* were the second and third most important features contributing to the prediction. In the XGB model and RM I, T and T_r were the fourth most important features and had very little difference. T_o in RM II was the least important feature. T_o contributed less to the prediction of cumulative oil production compared to T and T_r. The result was consistent with the prediction performance of XGB model and reference models.

From the algorithm perspective, it makes sense that XGB performed better than ANN in these case studies. The two methods are important and widely used in data science research and by industry. These different machine learning methods perform differently for different types of tasks. ANN captures image, voice, text and other high-dimensional data by modeling a spatiotemporal location. The tree-based XGB handles tabular data well and has some features that ANN does not have, such as interpretability of a model, easier hyper-parameter tuning and a faster calculating speed. In this study, it became obvious that compared to XGB, it is tough work to find the best construction of ANN without overfitting. The calculating speed of XGB was nearly 100 times faster than ANN in the process of production forecasting. XGB needed only 2 or 3 s, while ANN required several minutes. The difference in computational speed is attributed to: (1) using a backpropagation process, the convergence rate of ANN is particularly slow and easily falls into the local minimum (Ren et al., 2020); (2) compared to ANN, XGB has a lower number of hyperparameters to be tuned; (3) sparsity-aware split finding of XGB makes it find the optimal direction and only non-missing observations are visited; (4) cache-aware access and blocks for out-of-core computation make XGB fast. Although the computing speed of ANN is slower than XGB, XGB and ANN models have much faster prediction speed when compared to traditional methods.

In the case of production forecasting, ANN and XGB models show similar performance compared to their RM I. The results of reservoir identification from classification models are reliable and can be used in regression models for prediction of production. In RM II, established ANN and XGB models provide higher accuracy for predicting production. The combination of reservoir identification and production forecasting in this study was meaningful and valuable because the production was correlated with the thickness of effective reservoirs rather than with the overall reservoir thickness.

5. Conclusions

This paper developed an integrated ML system, formed by two interconnected predictive models. It makes full use of historical data and solves reservoir identification and production forecasting problems, making the models faster and less labor-intensive than traditional methods.

The results of reservoir identification revealed that ensemble techniques (RF, GBDT and XGB) perform better than single classifiers (LR, KNN, DT and ANN). The reservoir identification results of XGB were selected because of the outperformance of XGB in all evaluation metrics in the 10-fold-cross-validation and test process when compared to the other methods. The prediction accuracy for

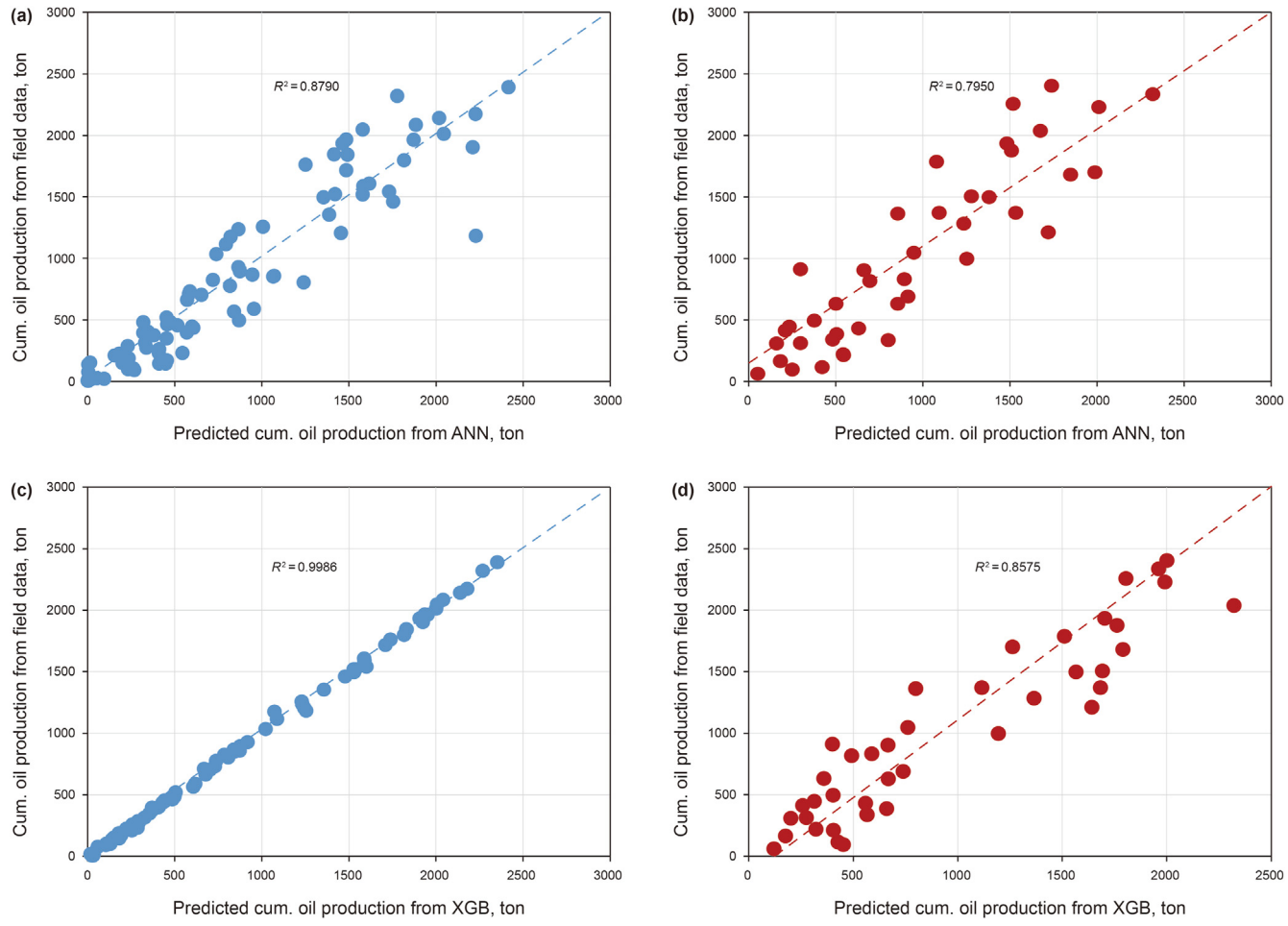

Fig. 10. Cross plots of real field cumulative oil production results vs. forecasts of established ANN model and XGB model: (**a**) Training set results of ANN model; (**b**) Testing set results of ANN model; (**c**) Training set results of XGB model; (**d**) Testing set results of XGB model.

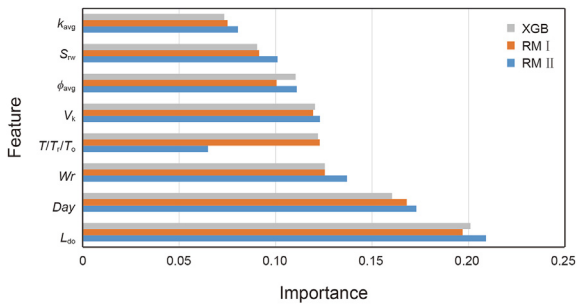

Fig. 11. Importance score rank of input features in prediction of cumulative oil production.

effective reservoirs (IO and IIO) was up to 99%.

Based on the prediction results of IO and IIO obtained from the reservoir identification, the effective thickness (thickness of IO and IIO) was an important input used in the production prediction process to predict the cumulative oil production of single wells. The very little difference (0.01 R^2) between the prediction results of established ANN/XGB model (based on predicted effective thickness) and corresponding RM I (based on real effective thickness) proved that the prediction of reservoir identification was sufficiently accurate and could reliably be used in production forecasting. ANN and XGB models with effective thickness provided higher prediction accuracy than corresponding RM II (based on overall reservoir thickness) in training and testing data sets. In testing process, the R^2 of XGB and ANN model using effective thickness was 10% and 13% higher than that of RM II. The MAE and RMSE of effective thickness-based models were much lower than that of RM II, demonstrating the superiority of effective thickness-based ML models for all metrics. XGB was better than ANN with a higher prediction accuracy and faster computing speed.

In this study, the research data was mainly based on an oil production data set from CNPC. The integrated ML system has proven successful in the predictive test of CNPC's subordinate blocks. In the future, introducing diverse data from different regions may improve the ML models and perhaps make them applicable on a global scale.

Acknowledgments

The research is partly supported by the NSERC/Energi Simulation and Alberta Innovates Chair at the University of Calgary.

Appendix A

Real data and prediction of XGB/ANN model from training process in the case of production prediction.

No.	Real production, ton	Prediction of ANN, ton	Prediction of XGB, ton
1	777	818.033	737.999
2	2173	2227.41	2178.62
3	2046	1579.29	2007.16
4	150	200.014	164.171
5	1963	1871.6	1952.42
6	2012	2044.98	2000.79
7	1845	1415.15	1829.8
8	852	1064.91	847.855
9	286	229.866	289.256
10	1933	1464.19	1906.82
11	104	262.904	107.727
12	704	653.465	703.937
13	1236	865.083	1233.67
14	712	581.236	668.307
15	19	3.78477	23.0181
16	482	319.131	497.908
17	1205	1454.17	1243.27
18	313	329.998	326.537
19	1607	1616.4	1587.17
20	480	475.354	483.264
21	5	1.02	23.8237
22	169	455.467	191.984
23	1716	1484.36	1709.92
24	1761	1250.75	1738.47
25	1115	794.688	1087.56
26	2085	1885.16	2040.46
27	26	54.2055	31.5974
28	376	380.011	372.154
29	188	235.632	178.401
30	519	453.608	504.225
31	146	411.423	180.163
32	2319	1777.83	2268.21
33	1497	1354.75	1529.09
34	1903	2210.43	1924.94
35	2141	2017.69	2138.54
36	566	841.086	606.211
37	18	19.7361	16.5283
38	496	868.637	499.08
39	1541	1731.59	1600.39
40	857	1070.55	871.86
41	396	319.434	369.877
42	260	411.351	257.906
43	1520	1579.48	1528.51
44	1033	738.646	1022.66
45	894	873.995	876.966
46	589	955.536	617.514
47	91	267.938	104.344
48	78	3.78477	58.3632
49	144	447.756	161.496
50	402	346.137	407.909
51	803	1241.18	807.688
52	1588	1582.66	1590.29
53	6	5.45389	32
54	664	571.344	677.822
55	1965	1484.95	1934.8
56	349	453.542	355.716
57	225	182.117	222.646
58	1461	1755.09	1478.51
59	868	946.422	839.416
60	2390	2416.86	2351.98
61	275	338.072	294.482
62	464	456.224	486.921
63	1183	2228.44	1254.79
64	20	93.9628	35.771
65	137	3.78477	141.637
66	1798	1817.35	1816.96
67	100	231.809	129.682
68	397	569	402.12
69	232	408.25	250.134
70	1355	1386.1	1355.86
71	1843	1491.49	1831.53
72	445	600.457	441.883
73	731	587.383	730.651
74	436	607.271	430.751

(continued on next page)

(continued)

No.	Real production, ton	Prediction of ANN, ton	Prediction of XGB, ton
75	455	516.642	443.858
76	1521	1421.59	1539.18
77	1175	820.689	1072.94
78	232	543.801	286.53
79	153	13.6937	151.762
80	211	156	254
81	1257	1007	1229
82	928	865	918
83	825	720	785.178

Appendix B

Real data and prediction of XGB/ANN model from testing process in the case of production prediction.

No.	Real production, ton	Prediction of ANN, ton	Prediction of XGB, ton
1	1786	1079.75	1511.6
2	688	915.475	739.036
3	446	234.412	313.871
4	816	696.202	492.459
5	2403	1739.3	2000.69
6	1497	1382.06	1565.9
7	2037	1675.9	2322.53
8	996	1252.54	1195.31
9	1045	948.279	762.268
10	340	483.319	565.643
11	630	856.035	668.591
12	164	182.789	175.969
13	1369	1534.89	1682.47
14	631	501.439	359.441
15	2228	2010.97	1992.25
16	115	421.96	428.314
17	495	376.964	404.13
18	430	631.916	559.978
19	309	161.169	201.158
20	312	298.446	275.513
21	911	297.94	400.452
22	1933	1483.01	1703.87
23	336	802.416	567.334
24	831	893.789	588.766
25	2256	1519.4	1804.72
26	2334	2320.35	1961.03
27	212	545.962	405.253
28	1876	1508.64	1761.49
29	61	52.1522	121.8566
30	1363	857.524	800.274
31	1211	1720.96	1642.75
32	95	250.913	452.584
33	1680	1849.23	1790.27
34	385	505.18	660.58
35	413	212.257	259.373
36	1283	1236.68	1365.72
37	1369	1094.31	1115.13
38	1700	1988.77	1263.2
39	1505	1279	1693
40	220	543.427	322.05
41	903	662.845	667.077

References

Abdulraheem, A., Sabakhi, E., Ahmed, M., Vantala, A., Raharja, I., Korvin, G., 2007. Estimation of permeability from wireline logs in a Middle Eastern carbonate reservoir using fuzzy logic. In: 15th SPE Middle East Oil and Gas Show and Conference. https://doi.org/10.2118/105350-MS.

Andika, R., Chandima Ratnayake, R.M., 2019. Machine learning approach for risk-based inspection screening assessment. Reliab. Eng. Syst. Saf. 185, 518–532. https://doi.org/10.1016/j.ress.2019.02.008.

Anifowose, F.A., Labadin, J., Abdulraheem, A., 2015. Ensemble model of non-linear feature selection-based Extreme Learning Machine for improved natural gas

reservoir characterization. Spec. Issue J. Nat. Gas. Sci. Eng. 25, 1561–1572. https://doi.org/10.1016/j.jngse.2015.02.012.
Anifowose, F.A., Labadin, J., Abdulraheem, A., 2017. Ensemble machine learning: an untapped modeling paradigm for petroleum reservoir characterization. J. Petrol. Sci. Eng. 151, 480–487. https://doi.org/10.1016/j.petrol.2017.01.024.
Awoleke, O., Lane, R., 2011. Analysis of data from the Barnett Shale using conventional statistical and virtual intelligence techniques. SPE Reservoir Eval. Eng. 14 (5), 544–556. https://doi.org/10.2118/127919-PA.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.G., 1984. Classification and regression trees. In: Hoecker, A. (Ed.), TMVA–Toolkit for Multivariate Data Analysis. Wadsworth International Group, Belmont, California, USA arXiv preprint physics/0703039.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 26 (2), 123–140. https://doi.org/10.1007/BF00058655.
Brown, I., Mues, C., 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Appl. 39, 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033.
Chaki, S., Routray, A., Mohanty, W.K., 2018. Well-log and seismic data integration for reservoir characterization: a signal processing and machine-learning perspective. IEEE Signal Process. Mag. 35 (2), 72–81. https://doi.org/10.1109/MSP.2017.2776602.
Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min 785–794. https://doi.org/10.1145/2939672.299785.
Chakra, N.C., Song, K.-Y., Gupta, M.M., Saraf, D.N., 2013. An innovative neural forecast of cumulative oil production from a petroleum reservoir employing higher-order neural networks (HONNs). J. Petrol. Sci. Eng. 106, 18–33. https://doi.org/10.1016/j.petrol.2013.03.004.
Cox, D.R., 1958. The regression analysis of binary sequences (with discussion). J. Roy. Stat. Soc. B 20, 215–242. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x.
Cracknell, M., Reading, A., 2012. Machine Learning for Lithology Classification and Uncertainty Mapping. AGU Fall Meeting Abstracts, p. 1511.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232. http://www.jstor.org/stable/2699986.
Guo, Z., Wang, H., Kong, X., Shen, L., Jia, Y., 2021. Machine learning-based production prediction model and its application in Duvernay formation. Energies 14 (17), 5509. https://doi.org/10.3390/en14175509.
Harris, J.R., Grunsky, E.C., 2015. Predictive lithological mapping of Canada's North using random forest classification applied to geophysical and geochemical data. Comput. Geosci. 80, 9–25. https://doi.org/10.1016/j.cageo.2015.03.013.
Helmy, T., Rahman, S.M., Hossain, M.I., Abdelraheem, A., 2013. Non-linear heterogeneous ensemble model for permeability prediction of oil reservoirs. Arabian J. Sci. Eng. 38, 1379–1395. https://doi.org/10.1007/s13369-013-0588-z.
Hossin, M., Sulaiman, M.N., 2015. A review on evaluation Metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process (IJDKP) 5 (2).
Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20 (8), 832–844. https://doi.org/10.1109/34.709601.
Kamenski, A., Cvetković, M., Kolenković Močilac, I., et al., 2020. Lithology prediction in the subsurface by artificial neural networks on well and 3D seismic data in clastic sediments: a stochastic approach to a deterministic method. Int. J. Geom. 11, 8. https://doi.org/10.1007/s13137-020-0145-3.
Liu, W., Chen, Z., Hu, Y., 2022. XGBoost algorithm-based prediction of safety assessment for pipelines. Int. J. Pres. Ves. Pip. 197, 104655. https://doi.org/10.1016/j.ijpvp.2022.104655.
Merembayev, T., Yunussov, R., Yedilkhan, A., 2018. Machine learning algorithms for classification geology data from well logging. 2018 14th International Conference on Electronics Computer and Computation (ICECCO), pp. 206–212. https://doi.org/10.1109/ICECCO.2018.8634775.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, part I, A discussion of principles. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.
Priezzhev, I., Stanisalav, E., 2018. Application of Machine Learning Algorithms Using Seismic Data and Well Logs to Predict Reservoir Properties, vol. 1. European Association of Geoscientists & Engineers, pp. 1–5. https://doi.org/10.3997/2214-4609.201800920.
Radford, D.D.G., Cracknell, M.J., Roach, M.J., Cumming, G.V., 2018. Geological mapping in western Tasmania using radar and random forests. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 11 (9), 3075–3087, September. https://doi.org/10.1109/JSTARS.2018.2855207.
Raeesi, M., Moradzadeh, A., Ardejani, F.D., Rahimi, M., 2012. Classification and identification of hydrocarbon reservoir lithofacies and their heterogeneity using seismic attributes, logs data and artificial neural networks. J. Petrol. Sci. Eng. 82, 151–165. https://doi.org/10.1016/j.petrol.2012.01.012.
Raschka, S., 2015. Python Machine Learning. Packt Publishing Ltd.
Ren, X., Hou, J., Song, S., Liu, Y., Chen, D., Wang, X., Dou, L., 2019. Lithology identification using well logs: a method by integrating artificial neural networks and sedimentary patterns. J. Petrol. Sci. Eng. 182, 106336. https://doi.org/10.1016/j.petrol.2019.106336.
Ren, Y., Mao, J., Zhao, H., Zhou, C., Gong, X., Rao, Z., Wang, Q., Zhang, Y., 2020. Prediction of aerosol particle size distribution based on neural network. Adv. Meteorol. https://doi.org/10.1155/2020/5074192.
Rodríguez, H.M., Escobar, E., Embid, S., Morillas, N.R., Hegazy, M., Larry, W.L., 2014. New approach to identify analogous reservoirs. SPE Econ & Mgmt 6, 173–184. https://doi.org/10.2118/166449-PA.
Siddiqi, S.S., Andrew, K.W., 2002. A study of water coning control in oil wells by injected or natural flow barriers using scaled physical model and numerical simulator. In: SPE Annual Technical Conference and Exhibition. https://doi.org/10.2118/77415-MS.
Van, S.L., Chon, B.H., 2018. Effective prediction and management of a CO_2 flooding process for enhancing oil recovery using artificial neural networks. J. Energy Resour. Technol. 140 (3), 032906. https://doi.org/10.1115/1.4038054.
You, J., Ampomah, W., Kutsienyo, E.J., Sun, Q., Balch, R.S., Aggrey, W.N., Cather, M., 2019. Assessment of enhanced oil recovery and CO_2 storage capacity using machine learning and optimization framework. SPE Europec featured at 81st EAGE Conference and Exhibition. https://doi.org/10.2118/195490-MS.