

## RESEARCH ARTICLE

# SAMA: Spatially-Aware Model-Agnostic Machine Learning Framework for Geophysical Data

ASMA Z. YAMANI<sup>1</sup>, KLEMENS KATTERBAEUR<sup>2</sup>, (Member, IEEE),  
ABDALLAH A. ALSHEHRI<sup>2</sup>, (Senior Member, IEEE), AND  
RABEAH A. AL-ZAIDY<sup>3</sup>, (Member, IEEE)

<sup>1</sup>Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

<sup>2</sup>EXPEC Advanced Research Center, Saudi Aramco, Dhahran 35713, Saudi Arabia

<sup>3</sup>Center for Integrative Petroleum Research (CIPR), ICS, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

Corresponding author: Rabeah A. Al-Zaidy (rabeah.alzaidy@kfupm.edu.sa)

This work was supported by Saudi Aramco under Grant CIPR2349.

**ABSTRACT** Geophysical data is a form of spatial data that suffers from various limitations when applying conventional machine learning algorithms and evaluation techniques. A key limitation facing models trained on geophysical data is their inability to generalize well when deployed to predict from new unseen data. We address the problem of inaccurate performance assessments of machine learning models, that stems from violating independence assumptions during the feature selection and evaluation phases of the learning process. Our proposed spatially-aware and model-agnostic (SAMA) framework provides a suite of spatially-aware feature generation, feature selection, and model validation algorithms that account for spatial characteristics of geophysical data. The framework is model agnostic, as it tackles data-related challenges that are not affected by the specific machine learning algorithm used to fit the data. To demonstrate the effectiveness of the proposed approach, it is applied to the water saturation mapping problem using a novel geophysical dataset to train a prediction model. The proposed spatially-aware models obtains an  $R^2$  of 0.620, an  $RMSE$  of 0.220 for predicting water saturation for the *Whole Region* of the reservoir model box and an  $R^2$  of 0.161, an  $RMSE$  of 0.263 for the *Interwell Region*. Extensive experiments on 5 additional unseen datasets show that the model maintains stable performance across different datasets, which demonstrates the ability of the SAMA framework to produce robust models that are transferable to new datasets.

**INDEX TERMS** Bias-variance trade-off, cross-fold validation, feature engineering, feature selection, geophysical data, random forest, regression, reservoir characterization, spatial autocorrelation, machine learning, model validation, water saturation mapping.

## I. INTRODUCTION

The geophysical domain involves a wide range of problems that can benefit from machine learning based approaches. Geophysical data is typically represented in the form of spatial data. It provides information on the physical properties of the Earth's surface and subsurface that is essential in analytical approaches for groundwater, mineral, and hydrocarbon discovery [1]. One of the applications that benefit from geophysical data is water saturation distribution mapping, which

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tong.

holds key information for reservoir engineers to maximize oil recovery and reduce costs [2]. Due to its unique properties inherited from spatial data, geophysical data introduces several challenges for Machine Learning (ML) modeling using conventional methods and evaluation techniques [3].

Developing a machine learning based prediction model for a given problem seeks to search the hypothesis space, using the given dataset, to find a function (hypothesis) with highest prediction power compared to other models, i.e. lowest *bias*, while performing equally well when used to predict from new unseen datasets, i.e. lowest *variance*. Bias occurs when the model is *underfitted* to the training dataset due

to inaccurate assumptions regarding the data or the model during the learning process. Variance occurs when a model is *overfitted* to the training data which is commonly caused by an inaccurate assessment of the model performance during the validation process. Failing to address model bias results in a model that produces highly erroneous predictions, which makes a machine learning solution to the problem entirely futile. On the other hand, not accounting for variance leads to a model that cannot generalize, in other words, one that is not transferable to datasets different than the one used for training. Therefore, the objective of the learning process is to find a model that balances the bias-variance trade-off. Although this study aims to address overfitting in spatial data, the proposed approach implicitly addresses underfitting, as we demonstrate in following sections.

Generalizability of a model requires addressing the problem of overfitting through data-based methods, model-based methods, or both. Model-based methods include techniques such as regularization if the model is overly complex, or in the case of neural networks overfitting can be reduced using dropout techniques. However, if the overfitting is caused fully, or even partly, by data-related sources, the data-based approaches are applied to address inconsistencies such as noise or missing data using data augmentation and denoising techniques. Spatial data has certain characteristics that can cause overfitting, such as spatial dependency, spatial heterogeneity, and scale [3].

Spatial dependency causes spatial auto-correlations in the data, restraining the application of ML and statistical models that are designed with the independence and identical distribution assumption for features, i.e. the i.i.d assumption. It also leads to over-promising performance when following conventional evaluation methods. Another property, spatial heterogeneity, introduces obstacles related to the evaluation of the model, which can severely degrade the generalizability of ML models [4], [5]. Spatial evaluation strategies in the literature [5], [6] focus on cross-validation methods, which can be fairly expensive when hyperparameter tuning ML models or performing spatial feature selection, as spatial evaluation is performed instead of a random train-test split [7]. Also, a majority of these studies focus only on interpolation tasks.

In addition to the challenges resulting from spatial data properties described above, the reservoir water saturation mapping problem is challenging in itself. A main challenge is that the data required for the mapping is not always available, especially the area between two well locations, referred to as the *interwell* region, since information about the different characteristics is mostly available only in near-well locations. Crosswell electromagnetic (EM) surveys are introduced with the goal to bridge this gap of information in the interwell area and provide resistivity mapping to regions extending to over 1km between the emitting and receiving boreholes [8].

Although Machine learning techniques have been used as early as 2002 [9] to predict water saturation for near-well locations, recent studies addressing the reservoir water saturation mapping problem mostly follow data assimilation

and modeling approaches. These methods require knowledge of either extra parameters, such as permeability, or the presence of a physical or numerical model. Other studies use history matching techniques that require several snapshots in time to make the predictions [10], [11]. To the best of our knowledge, there are no studies that rely only on machine learning techniques for water saturation distribution mapping from a single snapshot of Crosswell EM surveys and porosity distribution.

This work aims to provide an approach that will provide a more accurate evaluation of geophysical models built using conventional machine learning approaches. The main contributions of this work are summarized as follows:

- 1) **Spatially-aware model-agnostic (SAMA)** learning framework to develop robust ML models for geophysical data that generalize well.
- 2) **Spatial masking algorithm** for the evaluation of ML models built on 3D geophysical data cubes to obtain a realistic performance estimate of the model, and this avoiding over-promising models.
- 3) **Spatial Masking-Random Forest model** for performing water saturation mapping utilizing only the resistivity and porosity of a reservoir data cube.
- 4) Evaluation of the proposed approach based on the water saturation mapping problem requirements.

The rest of this paper is structured as follows: In section II, we introduce background related to the problem and discuss some related work. In section III, we describe the components comprised the proposed *Spatially-aware model-agnostic (SAMA)* framework. In Section IV, we present our experimental setup and results. In section V, significant findings and observations are discussed. Finally, in section VI, we conclude.

## II. BACKGROUND

### A. PROBLEM FORMULATION

Water Saturation ( $S_w$ ) per Schlumberger oilfield's glossary is defined as "The fraction of water in a given pore space", and is measured in percent of saturation units [12]. Water saturation is the most significant parameter to compute hydrocarbon volume in oil-water fields to optimize oil production. It is imperative for the oil production process to assess the availability of hydrocarbon reserves [2]. Ideally, core analysis is performed in order to determine water saturation in the subsurface. However, core data is not always available for most wells in a given field due to the borehole condition or the high cost of obtaining cores. Core data also requires lab analysis, which is typically time-consuming and expensive [13].

To overcome challenges with obtaining core data, Archie's equation [14] for computing water saturation was proposed. To calculate water saturation levels, Archie's equation relies on the resistivity and rock porosity, whose values are obtained from resistivity well-logs and porosity logs, respectively. In ideal situations, specifically clean sandstone formations, this calculation yields accurate results. However, it tends to

make simplifications that can cause erroneous results when dealing with shaly and heterogeneous formations where clay and other minerals can be the source of the conductivity instead of the hydrocarbons [15].

Several machine learning techniques have been used since 2002 [9] to predict water saturation for near-well locations. Methods including Artificial Neural Networks (ANNs) [9], [16], [17], [18], [19], [20], Support Vector Machine (SVM) [18], [19], and Decision Trees (DT) [19] have been used to predict saturation from well log data. As predictions based on well-logs can extend only a few meters surrounding the borehole, seismic data [21], [22], [23], and crosswell electromagnetic surveys [10], [11], [23] are used for the estimates to extend to a few kilometers surrounding the borehole. These approaches allow the engineers to obtain information about the interwell region and provide estimates for the water saturation. At the same time, they are either proposed in the context of history matching or data assimilation, which requires a physical model. Therefore, in this work we aim to develop a water saturation mapping model that does not require any physical model and relies only on the resistivity and porosity at a single snapshot in time.

## B. RELATED WORK

### 1) REGRESSION FOR GEOPHYSICAL DATA

Machine learning algorithms such as support vector machines, decision trees, and random forests have gained considerable attention in the field of spatial data applications [3]. While recent studies focus on evaluation strategies for ML models built for spatial data, early studies focused on modifying ML algorithms to be used for spatial data. In 2005, a study proposed an extension to Support Vector Machine (SVM), in an algorithm named Support Vector Random Fields (SVRF) [24]. Unlike SVMs, that make the “*independent and identically distributed*” assumption, SVRFs allow for spatial dependencies to be modeled using Conditional Random Fields (CRF). CRF is a statistical model that leverages the contextual information in the data. To capture the relationship between the features and the class’s label, the SVRF model makes use of an observation-matching function. As for capturing the relationships between the neighboring data points, and a local-consistency function, respectively [24]. Forms of decision trees have also been developed to account for spatial dependency and autocorrelation, namely, spatially aware Predictive Clustering Trees (PCT). This method follows the same approach of decision trees, where the test criteria at the nodes is the main difference, in PCT it selects the split that maximizes the inter-cluster variance reduction [25]. In order to leverage spatial heterogeneity a study [26] proposes Geographical Random Forest (GRF), which converts the global process of building trees into a decomposition of multiple local sub-models.

### 2) WATER SATURATION MAPPING

Predicting water saturation using machine learning has been extensively studied in the literature through various

approaches that are guided by the data available for the location.

Well-log data such as sonic, density, neutron porosity, and resistivity logs are used to predict water saturation in reservoirs. Several models exist that are based on neural networks on their own [9], or in combination with other techniques such as fuzzy logic [16], [27], ensemble structures [28], Mutual Information (MI) [20], least-squares support vector machine (LS-SVM) [29], and subtractive clustering [2]. Traditional machine learning models such as multilayer perceptron (MLP), Support Vector Machine, Decision Tree Forest, and Tree Boost methods were used, in another study, to train models for predicting water saturation in tight gas reservoirs [19].

Core data provides a different set of features that can be used to predict water saturation. A study using core porosity, deep resistivity log, neutron porosity, density log, sonic, and gamma-ray logs as input parameters [17], shows to improve prediction accuracies over the dual water model [30]. In another study, ANN is also used to predict the water saturation using the porosity, permeability of sample extracted from the core of the well, and height above the free water level [31].

In addition to well-log data and core data, seismic data contain useful information that can be used to predict porosity and water saturation. Studies adapting machine learning approaches for learning from seismic data use models such as support vector Regression (SVR) [21], ANNs [22] and adaptive neuro-fuzzy inference system optimized by a genetic algorithm [18].

To overcome some challenges of seismic data such as low resolution, crosswell electro-magnetic surveys are proposed to better understand the fluid types and saturation in the inter-well region of a reservoir [32]. Several studies are dedicated to mapping water saturation from these surveys using machine learning models [10], [11], [23], [33].

## C. MODEL VALIDATION

### 1) K-FOLD CROSS VALIDATION

In *K-fold cross validation*, random  $1/K$  of the points of the data cube are reserved for testing and the rest for training. This process is repeated  $K$  times and can be with or without replacement.

### 2) RELATED WORKS ON SPATIALLY-AWARE VALIDATION

Multiple methods have been proposed to give a more realistic estimation of ML models built for spatial data by having spatially dependent data in train and test sets. One study [34] proposed Spatial leave-one-out cross-validation. It considers a single data point for validation, and leaves out all spatially dependent surrounding data points within a certain threshold and trains the ML model using the remaining points. It repeats this step and then computes the average performance over all repetitions. Although the approach addresses the issue of spatial autocorrelation, it remains computationally costly.

Other works discuss the idea of cross-validation in spatial dependent data or data with spatial autocorrelation. One approach proposes spatial k-fold cross-validation (SKCV) [6] that introduces a dead zone around the testing data points, in which data points to be removed and the final model is trained on the remaining points. Another study investigates cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic dependencies [5]. The study introduces the concept of spatial blocking for the different data structures. It presents a guide on how to perform the blocking, starting with analyzing the type of dependency, then assessing the objective of the final model, for both interpolation and extrapolation. The next step is to block the data according to the objective and structure, perform the cross-validation, and finally make the final prediction. As most spatial analysis studies focus on interpolation, this study points out that blocking induces extrapolating and is key in building ML models that extrapolate [5]. In addition, it mentions that in the case where the approach aims to avoid inducing extrapolation in the model, the size of the blocks should be minimized. Geographical covariant maps and buffer distances can be used to ensure that data points with high spatial dependency are not co-located in the same train or test split [35]. Although the resulting model is unbiased, the process is computationally expensive and storage demanding. Therefore, it is not recommended for data sets with greater than 1000 points. Moreover, with the covariant maps added as features, the resulting model is optimal for interpolating only and not to extrapolate or explain the structural dependencies. Whereas in this study we focus on generalization.

In the remote sensing field, studies emphasize that in addition to spatial validation strategies, spatial dependency should be considered in different aspects of ML model development, including spatial feature selection [7]. Using spatial feature selection ensures that only features that generalize beyond the training data are included. The study also concludes that statistical evaluation for the models is not enough to evaluate ML models built for remote sensing problems and that a visual assessment must be performed [7].

Recently, a study pointed out the issue of over-promising RF models, and RF models that perform well in random cross-validation but have poor transferability [4]. As an example, the study evaluated an RF model that was trained to predict total volatile organic compounds from different borehole geophysical logs. It demonstrates that RF models built with random cross-validation might be suitable for interpolating missing well-logs, however, they may not be suitable for generalization [4].

#### D. RANDOM FOREST REGRESSION

Random Forests (RF), formally introduced by Leo Breiman [36], is an ensemble of tree predictors in which each tree is fit to an independent bootstrap sample drawn from the data. The results of those trees are aggregated by unweighted voting in the case of classification, whereas in regression,

it is performed by calculating the mean of the individual predictions [36], [37].

Assuming an unknown joint distribution  $P_{XY}(X, Y)$ , where  $X$  is a  $p$ -dimensional random vector  $X = (X_1, \dots, X_p)^T$  representing the predictor variables and  $Y$  is the independent variable, the goal of the algorithm is to find the prediction function  $f(X)$  for predicting  $Y$ . The prediction function is set to minimize the expected value of the loss function [37].

$$E_{XY}[L(Y, f(X))] \quad (1)$$

where  $E_{XY}$  is the expectation with respect to the joint distribution of  $X$  and  $Y$ . In regression, usually,  $L$  is the squared error loss.

$$L(Y, f(X)) = (Y - f(x))^2 \quad (2)$$

By that, minimizing  $E_{XY}[L(Y, f(X))]$  for squared error loss gives the conditional expectation.

$$f(x) = E[Y|X = x] \quad (3)$$

The ensemble predictor  $f(x)$  in regression is a result of averaging the results of the base learners  $h_1(x), \dots, h_J(x)$  where  $J$  is the number of base learner trees in the ensemble.

$$f(x) = \frac{1}{J} \sum_{j=1}^J h_j(x) \quad (4)$$

The tree base learners leverage an independent bootstrap sample from the data given at random and make the fitting using binary recursive partitioning. In binary recursive partitioning, all the training data points are put in a single node. Then, until the stopping criteria is met, the node(s) are split into two descendant nodes based on the value of the predictor variables, and this is done recursively. In order to determine the split in regression, each possible split is considered, and the selected split point is the one that minimizes the mean squared residuals at the nodes ( $Q$ ), that are defined as:

$$Q = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (5)$$

where  $n$  is the number of training data points at a node and  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ , which is the predicted value at the node [37].

When training a Random Forests Regressor, some of the most common hyperparameters to tune are (using sklearn library parameter names): *n\_estimators*, which determines the number of decision trees building the forest, *max\_depth*, which sets the maximum number of levels in each decision tree, and *max\_features*, which sets the maximum number of features to consider when splitting a node. There are also other parameters that are less common to tune, such as *bootstrap* to determine whether sampling the data point is with or without replacement, *min\_samples\_split* which sets the minimum number of points to be placed in a node before the node is split, and *min\_samples\_leaf* which sets the minimum number of points to be allowed in a leaf node. There are several advantages of using Random Forests such as its robustness to outliers and noise, simplicity, and that it can be

**Algorithm 1** *Spatial Masking* - Mask Selection Algorithm

```

procedure SelectMask
  trainData  $\leftarrow$  layers with the symbol 0
  testData  $\leftarrow$  layers with the symbol 1
  model  $\leftarrow$  Fit the model using the trainData
  SpatialMaskErr  $\leftarrow$  RMSE of testing the model using the testData
  ConsecutiveErr  $\leftarrow$  0
  K  $\leftarrow$  LCount / MaskLen
  for i in range(K) do
    trainData  $\leftarrow$  layers with symbol 0 and  $!(i < LNumber / LCount \leq i + 1)$ 
    for j in range(1, MaskLen-1) do
      testLayer  $\leftarrow$  LNumber == (i * K) + j
      if LNumber > MaskLen/2 or LNumber < LCount - MaskLen/2 then
        ConsecutiveErr += RMSE of testing the model using the testLayer
  if SpatialMaskErr - ConsecutiveErr > e then
    AcceptMask

```

implemented with parallelization easily. Also, that it provides variable importance and internal estimates of errors [36].

**III. APPROACH**

In this section we describe the components of the proposed SAMA machine learning framework. The framework is designed to tackle overfitting, and implicitly under-fitting, of regression models that arise from data-related inconsistencies. Therefore, the framework is considered model-agnostic, as any regression algorithm can be used for fitting the data.

**A. SPATIAL FEATURE ENGINEERING****1) SPATIAL FEATURE GENERATION**

In order to learn from nearby points, features are derived based on the grid to which the points belong in order to help the model learn from its region. These features are the mean of the resistivity or porosity of a 2D plane or a 3D area which a point belongs to. A description of the 19 feature IDs considered in this study is shown in table 1. For the 2D features, the featureID naming consists of the number of data points on the two axes that the grid is constructed on. For the 3D features, the Feature\_id naming consists of 3D followed by one or two digits on the size of the 3D block for the  $x$  and  $y$ -axis, followed by one digit, which is the height of the block across the  $z$ -axis. For each feature ID, the resistivity  $res\_featureID$  and porosity  $poro\_featureID$  are calculated, amounting to a total of 24 derived features.

**2) SPATIAL FEATURE SELECTION**

Feature selection is performed to reduce the complexity of the model, to improve the accuracy of the prediction, and to better interpret the model. In our case study, it is performed after adding the derived features that were a result of averaging the feature from the grid in which the data point belongs, Table 1. This experiment aims to compare two approaches

**TABLE 1.** Different grid sizes for generated features.

Feature_id	Feature plane	Block size (x,y,z)
6050	2D, x,y	2 x 2 x 1
3025	2D, x,y	4 x 4 x 1
3004	2D, x,z	4 x 1 x 4
2504	2D, y,z	1 x 4 x 4
6008	2D, x,z	2 x 1 x 2
5008	2D, y,z	1 x 2 x 2
3D42	3D	4 x 4 x 2
3D44	3D	4 x 4 x 4
3D48	3D	4 x 4 x 8
3D22	3D	2 x 2 x 2
3D24	3D	2 x 2 x 4
3D28	3D	2 x 2 x 8

considered for spatial feature selection and feature selection approaches that use evaluation methods that less enforce spatial generalization:

- *Spatial Forward Feature Selection (SFFS-6L)* The features are added one at a time, and after that the prediction's performance is evaluated using the mask testing layer following the selected Mask during the mask selection step. After evaluation, if the performance increases, the feature is added. This process is to be repeated multiple times until the performance of the model stops increasing.
- *Spatial Backward Feature Selection (SBFS-6L)* All features are added at first. Then they are removed one at a time until the performance stops increasing while validating on the mask testing layer following the selected Mask during the mask selection step.
- *Spatial Forward Feature Selection (SFFS-0L)* Features are added one at a time, then the prediction's performance is evaluated using the mask testing layer with the mask 0000010000, in which there are no leave-out layers. After evaluation, if the performance increases, the feature is added. This process is to be run multiple times until the performance of the model stops increasing.

- **Random Forward Feature Selection (RFFS)** The features are added one at a time, validating randomly on 10% of the data points. After evaluation, if the performance increases, the feature is added. This process is to be run multiple times until the performance of the model stops increasing.

The random state is set to 42, and the  $R^2$  is used for the comparison during the feature selection process. For the testing of the feature selection approaches, a Random Forests model is trained and tested by shifting the Mask by one each time (e.g., 00xxx1xxx0, 000xxx1xxx, x000xxx1xx, etc.). For each mask shift, the model is run 5 times without setting the random state. This results in 50 model runs.

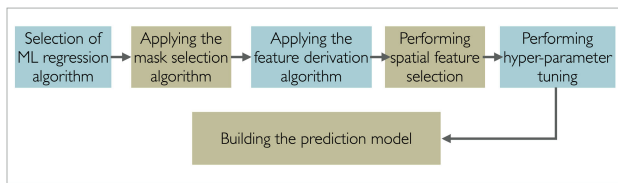


FIGURE 1. SAMA framework pipeline.

## B. SPATIALLY-AWARE VALIDATION STRATEGIES

### 1) SPATIAL BLOCKING

Building on the spatial blocking works in the literature [5] and the analogy of spatial k-fold cross-validation [6] and CV-block [38], a similar idea is applied to spatial data cubes.

#### a: K-FOLD RANDOM CROSS-SECTION CROSS-VALIDATION

In *Random cross-sections cross-validation*,  $1/K$  random cross-sections across the x, y, or z-axis of the data cube are reserved for validation, while the rest for training. This process is repeated  $K$  times and can be with or without replacement. No neighboring points on the cross-section used for validation are included in the training data.

#### b: K-FOLD CONSECUTIVE CROSS-SECTION CROSS-VALIDATION

In *K-fold Consecutive cross-sections cross validation*  $1/K$  layers of consecutive x, y, or z-axis *cross-sections* are used for validation, and the remaining data is used for training. This process is repeated  $K$  times by changing the validation layers until the data cube is exhausted. This approach provides a much stricter evaluation approach as  $1/K$  data cubes of the original spatial data cube is held out during testing. No neighboring points on the cross-section used for validation are included in the training data, nor are the points of the neighboring cross-sections except for two layers.

### 2) SPATIAL MASKING

Although *K-fold Consecutive cross-sections cross validation* seems strict in enforcing spatial generalization, in datasets with spatial heterogeneity or in small datasets, important information may be lost in removing consecutive layers that represent  $1/K$  of the data. Therefore, the need for an approach

that takes into account the spatial heterogeneity of the data arises. In addition, as a previous study concluded, when dealing with spatial data, spatial feature selection should be performed to avoid the selection of features that do not generalize when extrapolating [7], which results in high computation cost when using the existing cross-validation spatial evaluation techniques for tuning the model. Therefore we propose *Spatial Masking* as an approach to train-test split spatial data. In *Spatial Masking*, conforming to a certain pattern, cross-sectionals(layers) are either included in the training subset, left out, or used for testing. The aim of this approach is to provide a more accurate estimate of the model performance and to reduce the spatial auto-correlation in the dataset from the spatial dependencies while accounting for spatial heterogeneity.

In order to select the Mask, *K-fold Consecutive cross-sections cross validation* is performed. The length of the fold is set to the length of the mask pattern, in our case,  $K = 10$ .  $K$  is the length axis from which the cross-sectionals are taken over the length of the Mask, in our case, ten, too. Then the nine training folds are trained and evaluated using *Spatial Masking*, and the tenth fold is used for testing. The average *RMSE* of the layers resulting from the *K-fold Consecutive cross-sections cross validation* after removing the first and last layers (as guards) in the *K-fold Consecutive cross-sections cross validation* are compared against the *RMSE* resulting from the Mask. The Mask that had an *RMSE* equivalent to the *K-fold Consecutive cross-sections cross validation's RMSE* is selected. An illustration of the mask selection approach is shown in Figure 9. The implementation details are in Algorithm 1, where *LCount* is the total number of layers, *LNumber* is the layer's sequential number, and *MaskLen* is the Mask's length.

## C. TRAINING WATER SATURATION MAPPING MODEL

In this section we describe our proposed water saturation mapping model that is developed using the SAMA framework, illustrated in Figure 1. Given the above components, the SAMA framework now combines them using the following pipeline:

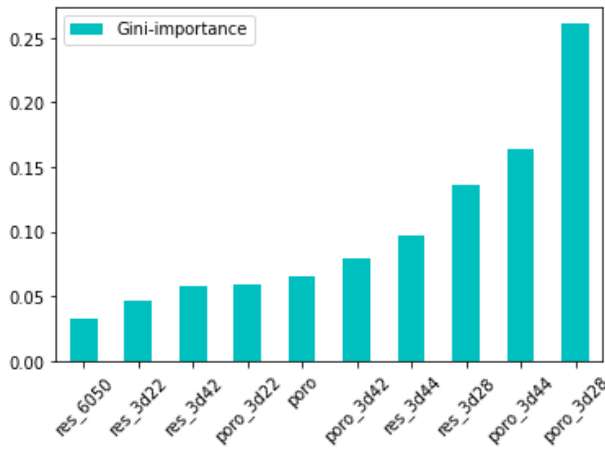
- 1) The machine learning algorithm(s) that are going to be used for regression are selected.
- 2) The mask selection algorithm is used to generate a validation mask using the raw features.
- 3) The feature derivation algorithm is executed. This is performed by applying the algorithm proposed in the feature generation section above.

TABLE 2. hyper-parameter tuning search values.

hyperparameter	values	optimum
n_estimators	[25,50,100]	100
max_feature	['sqrt', 0.1,0.2,0.3,0.5,0.7,1]	'sqrt'
max_depth	[10,30,50]	30
min_samples_leaf	[1,2,5,10]	2
min_samples_split	[2,5,10,15]	2

**TABLE 3.** Final feature set and performance values on the validation set for different Feature selection approaches.

Approach	Feature Set	$R^2$ score	RMSE
FFS with Random K-Fold cross validation (RFFS)	'res_3d44', 'res_3d48', 'poro_3d48', 'poro', 'res_3025', 'poro_3d42'	0.987	0.040
FFS with Mask 0000010000 (SFFS-L0)	'res_3d44', 'poro_3d48', 'res_3d48', 'poro', 'poro_3d44', 'res_3025', 'poro_3d42'	0.968	0.064
FFS with Mask 00xxx1xxx0 (SFFS-L6)	'poro_3d44', 'res_3025', 'res_3d22', 'res_2504', 'poro_3025', 'poro_3d42', 'res_6050', 'poro_6050', 'poro_3d22', 'poro_3d24', 'res_3d42', 'res_3d24', 'poro', 'res_3d28', 'poro_3d28'	0.634	0.217
BFS with Mask 00xxx1xxx0 (SBFS-L6)	'poro', 'res', 'poro_6050', 'res_6050', 'poro_3025', 'res_3025', 'poro_5008', 'res_5008', 'poro_2504', 'res_2504', 'poro_6008', 'res_6008', 'res_3004', 'poro_3d22', 'res_3d22', 'res_3d28', 'poro_3d42', 'res_3d42', 'poro_3d44', 'poro_3d48', 'res_3d48',	0.631	0.218



**FIGURE 2.** Feature importance of the SM-RF ( $S_w$ ) mapping model.

**TABLE 4.** Test evaluation results of the feature selection approaches per region.

Feature Set	Whole Region		Focus Region		Interwell Region	
	$R^2$	RMSE	$R^2$	RMSE	$R^2$	RMSE
All features	0.612	0.222	0.225	0.273	0.126	0.268
RFFS	0.542	0.241	0.116	0.292	-0.002	0.288
SFFS-L0	0.578	0.232	0.172	0.282	0.067	0.277
SFFS-L6	0.620	0.220	0.245	0.269	0.161	0.263
SBFS-L6	0.623	0.219	0.250	0.269	0.152	0.264

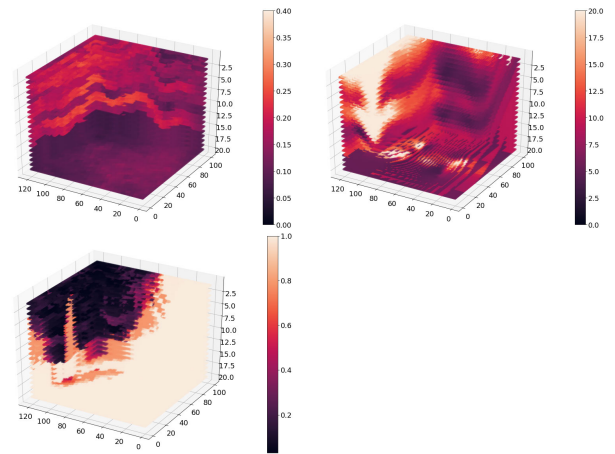
- 4) The selected mask and model are used to perform feature selection using one of the feature selection algorithms.
- 5) Hyper-parameter tuning is performed at this step for model-specific parameters, if any.
- 6) The prediction model is built using the selected features, selected validation mask, and selected hyperparameters.

## IV. RESULTS

### A. DATASET

#### 1) RESERVOIR BOX DATASETS

This work introduces one real-world dataset and five synthetic datasets to evaluate the proposed approach. The real-world dataset is a 3D data cube containing features of a realistic reservoir box model, specifying porosity, resistivity,



**FIGURE 3.** Water saturation (top-left), resistivity (top-right), porosity (bottom) for the RBMD data cubes.

and water saturation values. The reservoir box model dimensions are (122 x 100 x 20) in terms of (x,y,z), representing an area with dimensions 2km x 2km x 70ft (0.21 km) in depth. We refer to this dataset as RBMD. Figure 3 displays the water saturation, resistivity, and porosity for the 3D data cube. Figure 4 displays the water saturation, resistivity, and porosity taken along the z-axis. Figure 5 displays the water saturation, resistivity, and porosity taken along the y-axis.

Five synthetic datasets (RBMD-02, RBMD-05, RBMD-10, RBMD-20, RBMD-50) are generated for evaluation by adding white Gaussian noise at different rates: 2%, 5%, 10%, 20%, and 50% of the standard deviation of the original distribution of the feature, respectively. Visualization of the resistivity and porosity with the added noise for three of the synthetic datasets are presented in Table 5.

The distribution of the features and target in the dataset along the z-axis is shown in Figure 6. We can notice that the water saturation increases along the z-axis. The least water-saturated layers have a mean of 0.4 water saturation; 17 out of 20 layers along the z-axis have a mean of water saturation higher than 0.6. When performing Shapiro-Wilk Test on the target 'swat' water saturation, it fails the test with a p-test score of 0.00, indicating the water saturation data is highly skewed.

Figure 7 shows the correlation between the features 'poro', 'res' and the target 'swat'. It is observed that

TABLE 5. Porosity and resistivity for synthetic datasets at layer 4 across the z-axis.

Dataset name	noise (% of std of feature)	resistivity	Porosity
RBMD-05	5%		
RBMD-20	20%		
RBMD-50	50%		

both features have a negative correlation, with 'res' having  $-0.19$  Pearson correlation and 'poro' having  $-0.42$  Pearson correlation.

2) PRODUCTION AND INJECTION WELLS

The reservoir model box *Whole Region* contains one horizontal injection well, located between 35 and 42 on the x-axis and 23 and 70 on the y-axis. It also contains one horizontal production well, located between 83 and 90 on the x-axis and 27 and 80 on the y-axis. Locations are illustrated in Figure 8. The EM transmitters and receivers are embedded within the wells.

There are two regions that we evaluate separately due to their importance for the problem: we refer to the first one as the *Focus Region* and is located between layers 30 and 95 with respect to the x-axis, layers 23 and 80 with respect to the y-axis, and layers 3 and 13 with respect to the z-axis. The other region is the *Interwell Region*, and is located between layers 43 and 82 with respect to the x-axis, layers 23 and 80 with respect to the y-axis, and layers 3 and 13 with respect to the z-axis.

B. EXPERIMENT SETTINGS

1) EXPERIMENT DESIGN

The experiment procedure first starts by applying the mask selection algorithm. Then using the selected mask evaluate *Spatial Masking* as a train test split against other cross-validation strategies in terms of providing a better estimate for the extrapolation error provided by *K-Consecutive Cross-section Cross-validation*. Then different feature selection spatial and aspatial feature selection approaches are evaluated. Finally, two approaches for fitting the data are evaluated. For all the aforementioned experiments, Random Forests are used as they have been used in multiple studies concerning the evaluation strategies of ML for spatial data [5], [7]. Random Forests do not make the i.d.d. assumption, making them suitable for spatially dependent data. However, they would easily over-fit in the case of spatial data with high auto-correlation since it allows for distributing nearby data points with high auto-correlation within the training and testing subsets, which causes over-estimating the model's performance [5]. Random Forests do not require an extensive amount of hyperparameter tuning to get models with high



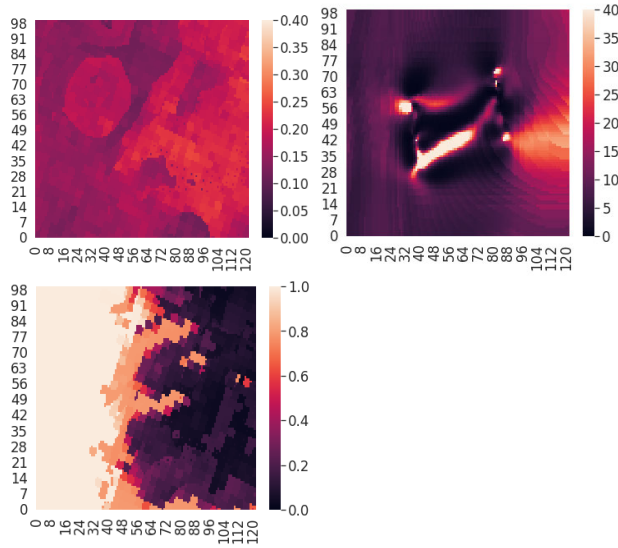


FIGURE 4. Water saturation (top-left), resistivity (top-right), porosity (bottom) for layer 4 taken at the z-axis.

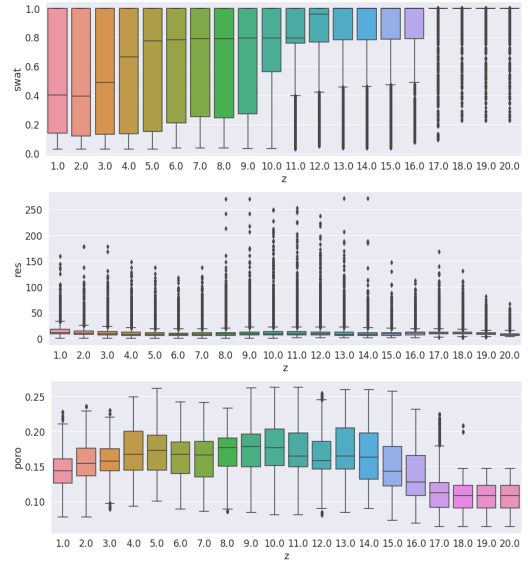


FIGURE 6. Top: water saturation, middle: porosity, bottom: resistivity per layer (layers across the z-axis).

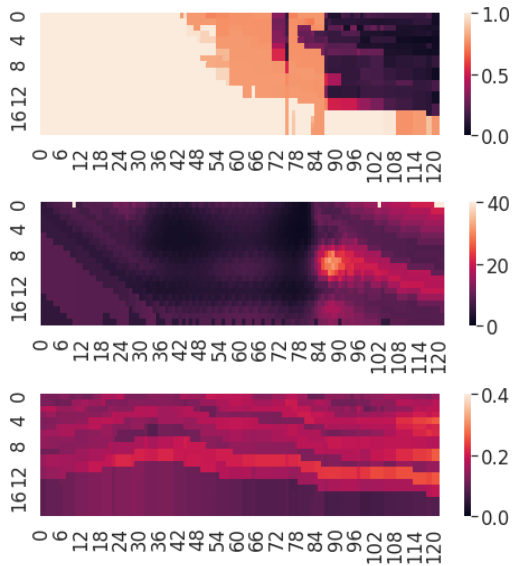


FIGURE 5. Water saturation (top-left), resistivity (middle), porosity (bottom) for layer 76 across the y-axis.

performance, which will direct our focus to the proposed approach for evaluation, as opposed to hyperparameter optimization.

## 2) EVALUATION METRICS

In order to evaluate and report the performance of regression models, several metrics can be used. Here, we will present the evaluation metrics that are used for evaluation; namely, the root mean square error (*RMSE*), and the coefficient of determination  $R^2$ . Those metrics are the top two metrics used for evaluating the  $S_w$  prediction in the surveyed studies are the *RMSE* and  $R^2$ , where 69% of the surveyed studies for water saturation prediction use *RMSE*, and 50% of the studies used  $R^2$  for evaluation.

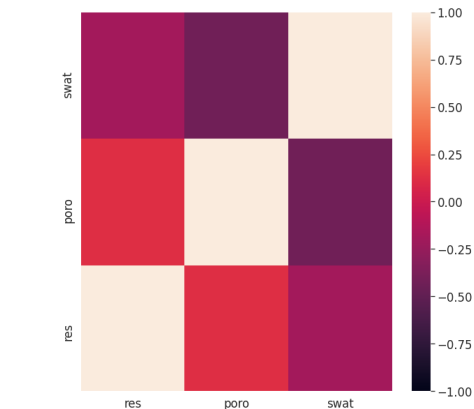


FIGURE 7. Pearson correlation heatmap for the variables of the RBMD dataset.

### a: ROOT MEAN SQUARED ERROR (RMSE)

*RMSE* measures the average magnitude of the errors in a set of predictions. It calculates the squared difference between actual and predicted values.

$$RMSE = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2 \quad (6)$$

where  $n$  is the number of points,  $y_j$  is the actual value, and  $\hat{y}_j$  is the predicted value for a point  $j$ . Squaring the difference penalizes more for errors with a larger magnitude, and it is commonly used in engineering problems.

### b: COEFFICIENT OF DETERMINATION ( $R^2$ )

The coefficient of determination, also known as the goodness of fit, measures the extent to which the output of the regression model is better than a straight line and is given by

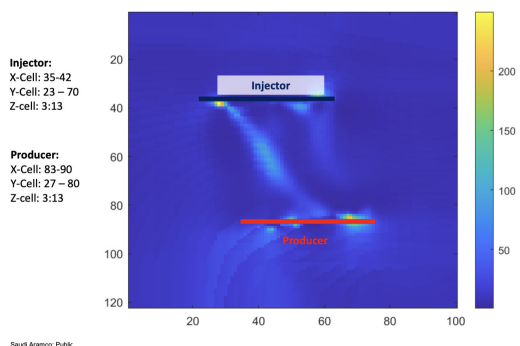


FIGURE 8. Production and injection wells' location.

the formula:

$$R^2 = 1 - \frac{\sum_{j=1}^n (y_j - \bar{y})^2}{\sum_{j=1}^n (y_j - \hat{y}_j)^2} \quad (7)$$

where  $n$  is the number of points,  $y_j$  is the actual value, and  $\hat{y}_j$  is the predicted value for point  $j$ , and  $\bar{y}$  is the mean of the observed values.

### C. EXPERIMENTAL RESULTS

#### 1) SPATIAL FEATURE ENGINEERING

The validation results of applying different feature selection approaches to our case study are presented in Table 6. As shown in the table, the models built on the selected features of the two approaches that did not use spatial feature selection or used a less strict mask as with the Mask 11110111, had a high  $R^2$  score and a fewer number of features, mostly of the larger block sizes. However, when evaluating them using the selected Mask (00xxx1xxx0), which more accurately estimates the extrapolation error and using the evaluation strategy, a drop in  $R^2$  score and the rise in  $RMSE$  error is noticed. This is accompanied by very low performance for the *Focus Region* and the *Interwell Region*, which indicates an overpromising and overfitting models. Evaluation results are shown in Table 7.

On the other hand, when applying spatial feature selection while using the selected Mask (00xxx1xxx0), whether by using SFFS or SBFS, validation scores, in Table 6, and testing scores, in Table 7, are very close. In addition, a higher  $R^2$  score is observed, especially in the *Focus Region* and *Interwell Region*, than when not applying spatial feature selection. Although SFFS and SBFS have almost equal performance, the SFFS-L6 feature set contains 15 features while the SBFS-L6 feature set contains 21 features. Therefore, the SFFS-L6 feature set is used in building the final model. These results corroborate results from previous studies such in [7] that when dealing with spatial data, spatial feature selection should be performed to avoid the selection features that do not generalize when extrapolating. Also, that SFFS is more suitable for avoiding overfitting in spatial data than SBFF.

#### 2) SPATIALLY-AWARE VALIDATION STRATEGIES

##### a: MASK SELECTION

This experiment aims to select the Mask to be used for *Spatial Masking* for this *RBMD* dataset, following the Algorithm 1. The masks used for the experiments are 0000010000,  $0000 \times 1x000$ , 000xx1xx00, x00xx1xx00, 00xxx1xxx0, 0xxxx1xxxx, x0xxx1xxx0, where 1: training layer, x: leave out layer, 0: training layer. The masks are combined with adding a mesh-like structure to include only (50%, 25%, 12.5%, and 6.25%) of the data points in the training layers in training. All cross-sectionals are taken across the y-axis, and the optimal hyperparameters in Table 2 are used. The random state is set to 42, and the  $RMSE$  is used for the comparison.

##### b: SPATIAL MASKING STRATEGY EVALUATION

This experiment aims to compare K-fold cross-validation [39], spatial blocking for 3D data cubes, and *Spatial Masking*. To compare the approaches, Random Forests models are built. The percentage of the training data is 90%, and for the testing, the data percentage is 10% for the cross-validation strategies. As for *Spatial Masking*, since the Mask 00xxx1xxx0 is used, based on the results from the Mask Selection Experiment, 30% of the data is utilized for training, and 10% is used for testing. The random state is set to 42, and the  $R^2$  and  $RMSE$  are used for the comparison.

##### c: SPATIAL MASKING AS A MODEL FITTING APPROACH

This experiment aims to evaluate *Spatial Masking* as a model-fitting approach, in which the model is only fit on the 0 layers of the Mask instead of the whole data for the final build. *K-Consecutive Cross-section Cross-validation* was used for the evaluation, and the model is fit once using the Mask and once all the data points reserved for training are fit.

#### 3) WATER SATURATION MAPPING MODEL

To build SM-RF water saturation ( $S_w$ ) mapping model, the hyperparameters are tuned via a grid search on the different hyperparameter values in Table 2. The derived features selected by the SFFS-L06 approach are utilized, and the model is built and evaluated using the Mask 00xxx1xxx0. The model has a  $R^2$  of 0.620 and a 0.220  $RMSE$  when evaluating the *Whole Region*. As for the *Focus Region* and the *Interwell Region*, an improvement of 33% and 32.6% in terms of  $RMSE$  difference is observed compared to using only the resistivity and porosity without the derived features, which we consider here as the baseline model, shown in Table 8. Visualization of some of the predictions are in Table 9.

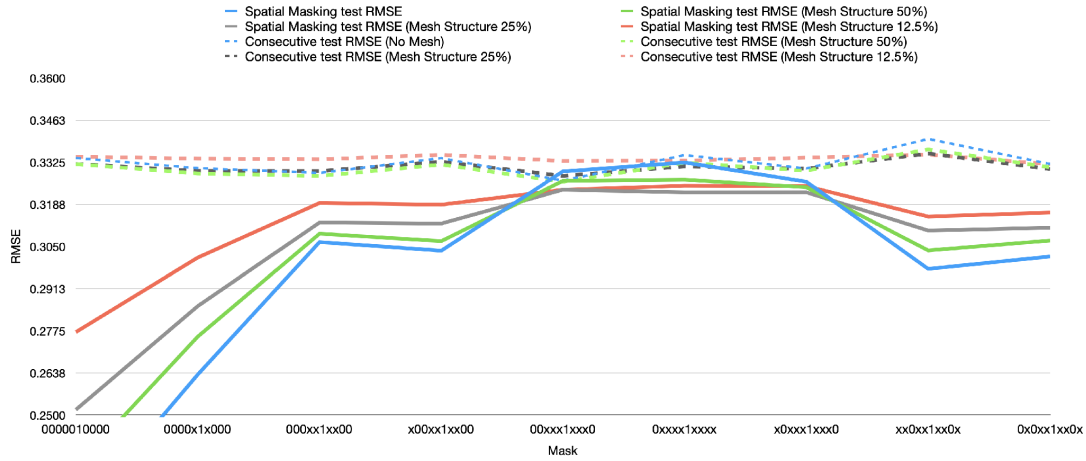
## V. DISCUSSION AND ANALYSIS

### A. MASK SELECTION

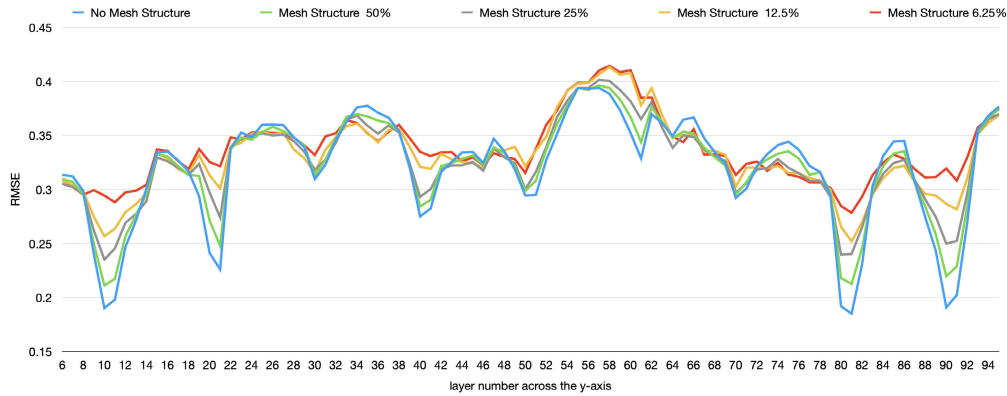
Figure 9 illustrates how as leave-out layers surrounding the testing layers are added, the difference between the  $RMSE$  of using K-Consecutive cross-sections cross-validation and *Spatial Masking* diminishes. At masks 00xxx1xxx0, 0xxxx1xxxx, and x0xxx1xxx0 the  $RMSE$  is

**TABLE 6.** Final feature set and performance values on the validation set for different Feature selection approaches.

Approach	Feature Set	$R^2$ score	RMSE
FFS with Random K-Fold cross validation (RFFS)	'res_3d44', 'res_3d48', 'poro_3d48', 'poro', 'res_3025', 'poro_3d42'	0.987	0.040
FFS with Mask 0000010000 (SFFS-L0)	'res_3d44', 'poro_3d48', 'res_3d48', 'poro', 'poro_3d44', 'res_3025', 'poro_3d42'	0.968	0.064
FFS with Mask 00xxx1xxx0 (SFFS-L6)	'poro_3d44', 'res_3025', 'res_3d22', 'res_2504', 'poro_3025', 'poro_3d42', 'res_6050', 'poro_6050', 'poro_3d22', 'poro_3d24', 'res_3d42', 'res_3d24', 'poro', 'res_3d28', 'poro_3d28'	0.634	0.217
BFS with Mask 00xxx1xxx0 (SBFS-L6)	'poro', 'res', 'poro_6050', 'res_6050', 'poro_3025', 'res_3025', 'poro_5008', 'res_5008', 'poro_2504', 'res_2504', 'poro_6008', 'res_6008', 'res_3004', 'poro_3d22', 'res_3d22', 'res_3d28', 'poro_3d42', 'res_3d42', 'poro_3d44', 'poro_3d48', 'res_3d48',	0.631	0.218



**FIGURE 9.** Training and testing scores based on different masks.



**FIGURE 10.** RMSE scores across layers when using 10-Consecutive cross-sections cross validation.

**TABLE 7.** Test evaluation results of the feature selection approaches per region.

Feature Set	Whole Region		Focus Region		Interwell Region	
	$R^2$	RMSE	$R^2$	RMSE	$R^2$	RMSE
All features	0.612	0.222	0.225	0.273	0.126	0.268
RFFS	0.542	0.241	0.116	0.292	-0.002	0.288
SFFS-L0	0.578	0.232	0.172	0.282	0.067	0.277
SFFS-L6	0.620	0.220	0.245	0.269	0.161	0.263
SBFS-L6	0.623	0.219	0.250	0.269	0.152	0.264

very close. However, at Mask 00xxx1xxx0, the RMSE of the K-Consecutive cross-sections cross-validation is at its

**TABLE 8.** Results of evaluating the models performance on different regions in the reservoir.

Model	Whole Region		Focus Region		Interwell	
	$R^2$	RMSE	$R^2$	RMSE	$R^2$	RMSE
Baseline Model	0.162	0.327	-0.333	0.358	-0.477	0.349
SM-RF Model	<b>0.620</b>	<b>0.220</b>	<b>0.245</b>	<b>0.269</b>	<b>0.161</b>	<b>0.263</b>

lowest value. The difference in RMSE between using the Mask 00xxx1xxx0 while utilizing 50% of the training layers data and the K-Consecutive cross-sections cross-validation RMSE is at 0.006. Therefore, the Mask 00xxx1xxx0 and the

TABLE 9. True values, prediction values, and residual errors of layers 6, 36, and 76 across the y-axis.

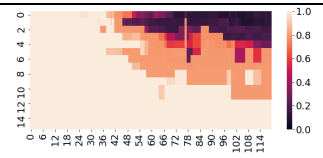
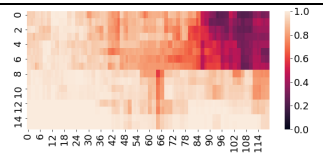
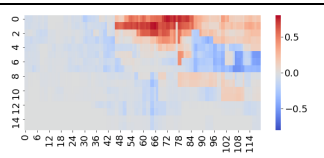
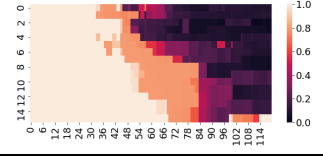
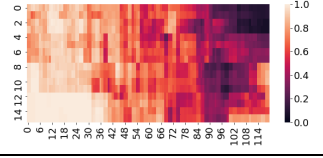
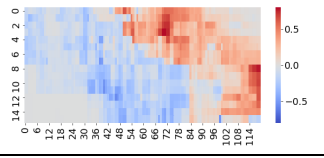
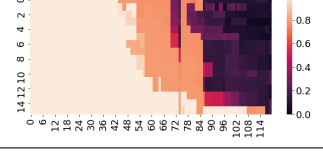
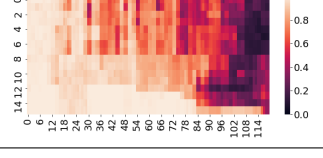
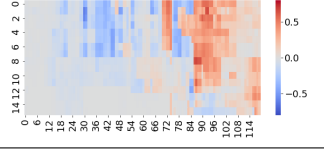
y-cross sectional layer number	RMSE	True Value ( $S_w$ )	Predicted Value ( $S_w$ )	Residual Error
6	0.187			
36	0.237			
76	0.186			

TABLE 10. Results of testing the Baseline model and the SM-RF model on the synthetic datasets, with *random\_state* = 42.

Dataset	SM-RF Model		
	$R^2$	RMSE	RMSE increase
RBMD	0.633	0.218	-
RBMD-02	0.632	0.218	0%
RBMD-05	0.629	0.219	0.04%
RBMD-10	0.617	0.222	1.8%
RBMD-20	0.610	0.224	2.6%
RBMD-50	0.533	0.245	11.0%

utilization ratio of 50% are selected to build the Spatially Masked Random Forests (SM-RF) models for predicting water saturation.

To investigate that further, we look at the errors per layer when running a K-Consecutive cross-sections cross-validation. Errors are illustrated in Figure 10. We can notice the errors stabilize at their highest around layers 4-7 with a sharp decline in RMSE error in the layers before and after for every ten layers. This indicates that layers 4-7 are more representative of the actual performance of the model built, and the high performance of the outer three layers from each side is due to overfitting from having nearby points in training and validation data. This validates the mask choice by the algorithm as it selected a mask with three leave-out layers from each side of the testing layer to avoid this sort of contamination to the testing data. Applying the mesh-like structure further treats the overfitting problem, as we can notice from the illustration and by the reduction in the standard deviation of the RMSE error across the layers from 0.05 when not using the mesh structure to 0.043 and 0.3 when using 50% and 12.5% of the data respectively. However, when combining the mesh structure with the *Spatial Masking* method and a Mask that has a total of 6 leave-out layers reduces the training

TABLE 11. Results of testing different the RF model on different spatial blocking method.

Spatial Evaluation Strategy	$R^2$ score	RMSE
K-Consecutive cross-sections cross-validation <i>Baseline</i>	<b>0.120</b>	<b>0.323</b>
Random 10-fold cross validation	0.736	0.183
Random cross-sections cross-validation	0.638	0.214
Spatial masking with Mask 11xxx0xxx1 and Mesh structure	<b>0.166</b>	<b>0.327</b>

data significantly and throws out too much of the data, which reduces the overall performance. Therefore, the elimination of the data to create the mesh-like structure was capped at 50%.

**B. SPATIAL MASKING AS AN EVALUATION STRATEGIES**

Table 11 shows that using *Spatial Masking* as an evaluation approach gives the closest error estimate in terms of RMSE to the baseline of 10-consecutive cross-section cross-validation layer. Thus providing a more accurate estimate for the extrapolation error.

**C. WATER SATURATION MAPPING MODEL**

The SM-RF model robustness when testing on the 5 synthetic datasets, the model’s performance is steady to a great extent. From Table 10, there is a 0.008 increase in RMSE for the RBMD-20 dataset, where 20% of noise is added. This equates to a decrease of 3.6% in  $R^2$  for the SM-RF model as opposed to a 59% decrease in  $R^2$  for the Baseline Model. As the error difference percentage is bound by the percentage of noise added to generate the synthetic datasets presented in Table 10, mathematically, the model is considered stable [40].

## VI. CONCLUSION

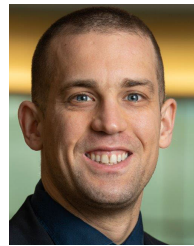
In this work, we tackled the issue of over-promising evaluation and over-fitting Forests models when built on spatially auto-correlated data. We proposed *Spatial Masking* as a train-test split approach to build and evaluate Random Forests models while minimizing overfitting and delivering a more realistic evaluation of the models. Through a series of experiments on a case study of water saturation ( $S_w$ ) mapping in oil/water reservoirs, *Spatial Masking* did provide a more accurate error estimate of the extrapolation error than random *K-Fold cross validation* and Random *K-Fold cross-section cross validation*. Combining *Spatial Masking* with a Mesh-like structure did reduce the extrapolation error further when fitting the model. An SM-RF ( $S_w$ ) mapping model was developed by fitting the data on the Mask 00xxx1xx00 with a 50% utilization rate of the training data and the features selected by the spatial forward selection approach using the same mask. In the future, we will investigate the different sources of the errors found in the mapping and work on further improvement in the mapping related to the *Interwell* region.

## REFERENCES

- [1] N. Anderson and A. Ismail, "A generalized protocol for selecting appropriate geophysical techniques," in *Geophysical Technologies for Detecting Underground Coal Mine Voids Forum*. Princeton, NJ, USA: Citeseer, 2003, pp. 28–30.
- [2] S. Asante-Okyere, C. Shen, Y. Y. Ziggah, M. M. Rulegeya, and X. Zhu, "Principal component analysis (PCA) based hybrid models for the accurate estimation of reservoir water saturation," *Comput. Geosci.*, vol. 145, Dec. 2020, Art. no. 104555, doi: 10.1016/j.cageo.2020.104555.
- [3] B. Nikparvar and J.-C. Thill, "Machine learning of spatial data," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 9, p. 600, 2021. [Online]. Available: <https://www.mdpi.com/2220-9964/10/9/600>
- [4] N. Terry, C. D. Johnson, F. D. Day-Lewis, B. Parker, and L. Slater, "Beware of spatial autocorrelation when applying machine learning algorithms to borehole geophysical logs," *Groundwater*, vol. 59, no. 3, pp. 315–319, Feb. 2021, doi: 10.1111/gwat.13081.
- [5] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Aroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, no. 8, pp. 913–929, Mar. 2017, doi: 10.1111/ecog.02881.
- [6] J. Pohjankukka, T. Pahikkala, P. Nevalainen, and J. Heikkonen, "Estimating the prediction performance of spatial models via spatial k-fold cross validation," *Int. J. Geographical Inf. Sci.*, vol. 31, no. 10, pp. 2001–2019, Jul. 2017, doi: 10.1080/13658816.2017.1346255.
- [7] H. Meyer, C. Reudenbach, S. Wöllauer, and T. Nauss, "Importance of spatial predictor variable selection in machine learning applications—Moving from data reproduction to spatial prediction," *Ecological Model.*, vol. 411, Nov. 2019, Art. no. 108815, doi: 10.1016/j.ecolmodel.2019.108815.
- [8] S. Alsaif, A. Alkhatib, and A. Marsala. (2017). *Advanced Uncertainty Quantification Methods Deployed on Electromagnetic Dataset for Reservoir Saturation Mapping*. [Online]. Available: <https://www.earthdoc.org/content/papers/10.3997/2214-4609.201701751>
- [9] H. B. Helle and A. Bhatt, "Fluid saturation from well logs using committee neural networks," *Petroleum Geosci.*, vol. 8, no. 2, pp. 109–118, Jun. 2002, doi: 10.1144/petgeo.8.2.109.
- [10] K. Katterbauer, S. Arango, S. Sun, and I. Hoteit, "Enhanced characterization of reservoir hydrocarbon components using electromagnetic data attributes," *J. Petroleum Sci. Eng.*, vol. 140, pp. 1–15, Apr. 2016, doi: 10.1016/j.petrol.2015.12.015.
- [11] K. Katterbauer and A. Marsala, "A novel sparsity deploying reinforcement deep learning algorithm for saturation mapping of oil and gas reservoirs," *Arabian J. Sci. Eng.*, vol. 46, no. 7, pp. 6859–6865, Oct. 2020, doi: 10.1007/s13369-020-05023-2.
- [12] *Explore the New Oilfield Glossary*. Accessed: Jul. 2022. [Online]. Available: [https://glossary.oilfield.slb.com/en/terms/w/water\\_saturation](https://glossary.oilfield.slb.com/en/terms/w/water_saturation)
- [13] A. Adeniran, M. Elshafei, and G. Hamada, "Functional network soft-sensor for formation porosity and water saturation in oil wells," in *Proc. IEEE Instrumentation Meas. Technol. Conf.*, May 2009, doi: 10.1109/imtc.2009.5168625.
- [14] G. Archie, "The electrical resistivity log as an aid in determining some reservoir characteristics," *Trans. AIME*, vol. 146, no. 1, pp. 54–62, Dec. 1942, doi: 10.2118/942054-g.
- [15] A. Alimoradi, A. Moradzadeh, and M. R. Bakhtiari, "Methods of water saturation estimation: Historical perspective," *J. Petroleum Gas Eng.*, vol. 2, no. 3, pp. 45–53, Mar. 2011.
- [16] N. Al-Bulushi, P. R. King, M. J. Blunt, and M. Kraaijveld, "Development of artificial neural network models for predicting water saturation and fluid distribution," *J. Petroleum Sci. Eng.*, vol. 68, nos. 3–4, pp. 197–208, Oct. 2009, doi: 10.1016/j.petrol.2009.06.017.
- [17] M. Mardi, H. Nurozi, and S. Edalatkhah, "A water saturation prediction using artificial neural networks and an investigation on cementation factors and saturation exponent variations in an Iranian oil well," *Petroleum Sci. Technol.*, vol. 30, no. 4, pp. 425–434, Feb. 2012, doi: 10.1080/10916460903452033.
- [18] C. H. Sambo, M. Hermana, A. Babasari, H. T. Janjuhah, and D. P. Ghosh, "Application of artificial intelligence methods for predicting water saturation from new seismic attributes," in *Proc. Offshore Technol. Conf. Asia*, Mar. 2018, pp. 1–8, doi: 10.4043/28221-MS.
- [19] S. Baziar, H. B. Shahripour, M. Tadayoni, and M. Nabi-Bidhendi, "Prediction of water saturation in a tight gas sandstone reservoir by using four intelligent methods: A comparative study," *Neural Comput. Appl.*, vol. 30, no. 4, pp. 1171–1185, Dec. 2016, doi: 10.1007/s00521-016-2729-2.
- [20] M. I. Miah, S. Ahmed, and S. Zendejboudi, "Connectionist and mutual information tools to determine water saturation and rank input log variables," *J. Petroleum Sci. Eng.*, vol. 190, Jul. 2020, Art. no. 106741, doi: 10.1016/j.petrol.2019.106741.
- [21] S. R. Na'imi, S. R. Shadzadeh, M. A. Riahi, and M. Mirzakhani, "Estimation of reservoir porosity and water saturation based on seismic attributes using support vector regression approach," *J. Appl. Geophys.*, vol. 107, pp. 93–101, Aug. 2014, doi: 10.1016/j.jappgeo.2014.05.011.
- [22] A. U. Osarogiabon, C. C. Udeze, and I. J. Imorame, "Modeling petrophysical property variations in reservoir sand bodies using artificial neural network and object based techniques," presented at the SPE Nigeria Annu. Int. Conf. Exhib., Aug. 2015, doi: 10.2118/178278-ms.
- [23] Y. Zhang and I. Hoteit, "Feature-oriented joint time-lapse seismic and electromagnetic history matching using ensemble methods," *SPE J.*, vol. 26, no. 03, pp. 1341–1365, Oct. 2020, doi: 10.2118/203847-pa.
- [24] C.-H. Lee, R. Greiner, and M. W. Schmidt, "Support vector random fields for spatial classification," in *Proc. PKDD*, 2005, pp. 121–132.
- [25] D. Stojanova, M. Ceci, A. Appice, D. Malerba, and S. Džeroski, "Global and local spatial autocorrelation in predictive clustering trees," in *Discovery Science*, T. Elomaa, J. Hollmén, and H. Mannila, Eds. Berlin, Germany: Springer, 2011, pp. 307–322.
- [26] S. Georganos, T. Grippa, A. Niang Gadiaga, C. Linard, M. Lennert, S. Vanhuyse, N. Mboga, E. Wolff, and S. Kalogirou, "Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling," *Geocarto Int.*, vol. 36, no. 2, pp. 121–136, Jan. 2021, doi: 10.1080/10106049.2019.1595177.
- [27] A. Mollajan, "Application of local linear neuro-fuzzy model in estimating reservoir water saturation from well logs," *Arabian J. Geosci.*, vol. 8, no. 7, pp. 4863–4872, Jul. 2014, doi: 10.1007/s12517-014-1526-4.
- [28] S. A. Jafari Kenari and S. Mashohor, "Robust committee machine for water saturation prediction," *J. Petroleum Sci. Eng.*, vol. 104, pp. 1–10, Apr. 2013, doi: 10.1016/j.petrol.2013.03.009.
- [29] A. N. Okon, S. E. Adewole, and E. M. Uguma, "Artificial neural network model for reservoir petrophysical properties: Porosity, permeability and water saturation prediction," *Model. Earth Syst. Environ.*, vol. 7, pp. 2373–2390, Oct. 2020, doi: 10.1007/s40808-020-01012-4.

- [30] C. Clavier, G. Coates, and J. Dumanoir, "Theoretical and experimental bases for the dual-water model for interpretation of shaly sands," *Soc. Petroleum Engineers J.*, vol. 24, no. 2, pp. 153–168, Apr. 1984, doi: [10.2118/6859-pa](https://doi.org/10.2118/6859-pa).
- [31] K. Kamalyar, Y. Sheikhi, and M. Jamialahmadi, "Using an artificial neural network for predicting water saturation in an Iranian oil reservoir," *Petroleum Sci. Technol.*, vol. 30, no. 1, pp. 35–45, Jan. 2012, doi: [10.1080/10916461003752561](https://doi.org/10.1080/10916461003752561).
- [32] Z. A. Al-Ali, M. H. Al-Buali, S. AlRuwaiti, S. M. Ma, A. F. Marsala, D. Alumbaugh, L. DePavia, C. Levesque, A. Nalonnil, P. Zhang, and C. Hulme, "Looking deep into the reservoir," *Oilfield Rev.*, vol. 21, no. 2, pp. 38–47, 2009.
- [33] K. Katterbauer, I. Hoteit, and S. Sun, "Synergizing crosswell seismic and electromagnetic techniques for enhancing reservoir characterization," *SPE J.*, vol. 21, no. 3, pp. 909–927, Jun. 2016, doi: [10.2118/174559-pa](https://doi.org/10.2118/174559-pa).
- [34] K. Le Rest, D. Pinaud, P. Monestiez, J. Chadoeuf, and V. Bretagnolle, "Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation," *Global Ecol. Biogeography*, vol. 23, no. 7, pp. 811–820, Apr. 2014, doi: [10.1111/geb.12161](https://doi.org/10.1111/geb.12161).
- [35] T. Hengl, M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler, "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables," *PeerJ*, vol. 6, p. e5518, Aug. 2018, doi: [10.7717/peerj.5518](https://doi.org/10.7717/peerj.5518).
- [36] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324).
- [37] C. Zhang, *Ensemble Machine Learning: Methods and Applications*. New York, NY, USA: Springer, 2012.
- [38] R. Valavi, J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita, "Block CV: An R package for generating spatially or environmentally separated folds for K-fold cross-validation of species distribution models," *Methods Ecology Evol.*, vol. 10, no. 2, pp. 225–232, Nov. 2018, doi: [10.1111/2041-210x.13107](https://doi.org/10.1111/2041-210x.13107).
- [39] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," Tech. Rep., 2001.
- [40] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. IJCAI*, Aug. 1995, vol. 14, no. 2, pp. 1137–1145.

**ASMA Z. YAMANI** received the B.S. degree in computer science from Imam Abdulrahman bin Faisal University, Saudi Arabia, in 2019, and the M.S. degree in computer science from KFUPM, Saudi Arabia, in 2020, where she is currently pursuing the Ph.D. degree in computer science. Her research interests include applied machine learning, related to petroleum engineering, bioinformatics, biomedical engineering, and renewable energy technologies. She has won the first place in the SPE DUPTS Student Competition for Computer Vision to Identify Drill Bits Damage Causes and the second place in the SPE DUPTS Student Competition for Correcting Drilling Data Uncertainties.



**KLEMENS KATTERBAEUR** (Member, IEEE) received the master's degree in petroleum engineering from Heriot Watt University and the Ph.D. degree from the King Abdullah University of Science and Technology. He is an experienced petroleum engineer and a software developer focusing on the development of the latest 4IR technologies for reservoir engineering applications. He has a proven track record having developed data driven uncertainty frameworks for enhancing oil recovery and strengthening sustainability of existing oil and gas reservoirs. He has developed in recent years some major technologies, such as enhanced artificial intelligence technologies for tracking waterfronts in subsurface reservoirs and forecasting their movements. Furthermore, he has developed robotics systems for enabling real-time logging while drilling and subsurface sensing and logging operations. He was an experienced young professional member in several energy related societies. He has been an active member and focused a lot on mentoring of young students that may dream to go into the oil and gas industry. In doing so, he has advised several students and assisted them in broaden their expertise to focus on learning about new digital technologies, code development, and robotics.



**ABDALLAH A. ALSHEHRI** (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, GA, USA. He is currently the Champion of the Deep Diagnostic Team, Reservoir Engineering Technology Division, EXPEC Advanced Research Center, Saudi Aramco. He is also undertaking various industry-leading research projects to develop novel technologies to increase hydrocarbon discovery and reservoirs recovery capitalizing on fourth industrial revolution (4IR) technologies and artificial intelligent (AI) technologies. He received several national and international awards. He deployed many technologies, has more than 30 patents (granted and disclosed), and more than 60 publications, including journals, book chapters, and conferences.



**RABEAH A. AL-ZAIDY** (Member, IEEE) received the master's degree in information systems security from Concordia University, Canada, and the Ph.D. degree in computer science and engineering from The Pennsylvania State University, USA. She was a Postdoctoral Fellow at PennState and then at KAUST. She is currently an Assistant Professor of computer science at KFUPM, Saudi Arabia. At KFUPM, she leads an AI Research Group, where she is a Manager of the Aramco-KFUPM Joint Project "Intelligent Reservoir Mapping." She worked as an AI Research Consultant at the Saudi Data and AI Authority (SDAIA). She is also an Affiliate Faculty with the Center of Integrative Petroleum Research (CIPR), KFUPM's College of Petroleum and Geosciences (CPG). She has authored several papers at top-tier reviewed conferences and journals, such as DIIN, AAAI, TheWebConf, and EMNLP. Her research interests include multi-modal information retrieval, machine learning, deep learning, natural language processing, and semantic structuring of big data and their applications in energy, climate, and environmental problems, and Arabic language AI services. She is a member of AAAI, ACM, and ACL. She is a PC and an editor of leading international AI conferences and journals.

• • •