Classifying basin-scale stratigraphic geometries from subsurface formation tops with machine learning

Jesse R. Pisel1,2 | Michael J. Pyrcz3,4

1Texas Institute for Discovery Education in Science, The University of Texas at Austin, Austin, TX, USA
2College of Natural Sciences, The University of Texas at Austin, Austin, TX, USA
3Cockrell School of Engineering, The University of Texas at Austin, Austin, TX, USA
4Jackson School of Geosciences, The University of Texas at Austin, Austin, TX, USA

Correspondence
Jesse R. Pisel, Texas Institute for Discovery Education in Science, The University of Texas at Austin, Austin, TX 78712, USA. Email: jesse.pisel@austin.utexas.edu

Abstract
Presented here is a transfer-learning model for classifying basin-scale stratigraphic geometries from subsurface formation tops. Support vector, decision trees, random forests, AdaBoost and K-nearest neighbour classification models are evaluated to support this challenge. Each model is trained on labelled synthetic stratigraphic geometry data generated in Python using observable geological principles and concepts. Accuracy is measured using a weighted Jaccard similarity coefficient score, and certainty of each prediction is quantified using margin sampling. The random forest classifier has the highest initial accuracy, and the optimal hyperparameters for the model that yield 88.4% accuracy and 72.8% mean certainty via five-fold cross-validation and active learning are documented on a real-world subsurface dataset. The random forest classifier with optimised hyperparameters is then used to make predictions on the real-world subsurface formation tops dataset. The dataset consists of formation tops for the Upper Cretaceous and Palaeocene strata of the Eastern Greater Green River Basin in south-central Wyoming. Results from model predictions include an area of truncation in the Lance Formation across the basin, and an area of onlap and truncation on the nose of the Rock Springs Uplift that previous studies in the region corroborate. It is believed that this model is most useful for guided interpretation, and identifying regions that warrant further inquiry by domain experts.

KEYWORDS
fluvial, machine learning, stratigraphic, subsurface

1 INTRODUCTION

Basin-scale stratigraphic geometries are generalised shapes of stratigraphic packages ranging from metres to kilometres in thickness (Figure 1). Large-scale stratal geometries (100–1,000 m thick) document a complex interplay of sediment supply, deposition, subsidence and topography. These factors directly affect the shape of sequences of stratigraphic geometries. The shapes of these geometries have been investigated by numerous sequence stratigraphers since the late 1970s (Galloway, 1989; Mitchum et al., 1977; Vail, 1987; Van Wagoner et al., 1988). In subsurface projects, geologists interpret these geometries from thickness and structure maps. At the large scale, these geometries include unconformities,
downlap, onlap and truncation. This study addresses the latter two of these geometries.

Truncation is where an overlying formation truncates the underlying formation with angular discordance between the formations. Onlap is horizontal stratigraphy that laps onto an inclined surface (Figure 1) (Mitchum et al., 1977). Succinctly, onlap is unconformable at the base of the stratigraphy, while truncation is unconformable at the top of the stratigraphy. Despite the difference between these two geometries, it is challenging to differentiate between the two using structure and thickness maps alone. The similarity between the structure and thickness maps for both geometries makes it difficult to understand the depositional and tectonic history of a basin based solely on geological interpretation.

To address this challenge, it is proposed to use a machine learning classifier to differentiate between onlap, truncation and horizontal stratification at the basin scale. The model is trained, optimised and cross validated on synthetically generated samples. Next, the machine learning model is tasked with making predictions on a dataset from the eastern Greater Green River Basin in south-west Wyoming. Last, machine learning predictions are reconciled with field-based geological interpretations to understand the classification model and its predictions. The classification predictions and their implications are then used to interpret the tectonic history of the eastern Greater Green River Basin.

2 | METHODS

A machine learning model was trained to predict stratal geometry from subsurface formation tops. The workflow is as follows:

1. Generate synthetic training data and split into training and validation datasets.
2. Augment the training dataset with non-linear transforms of the features.
3. Normalise the feature set after feature engineering.
4. Test five machine learning model types and select the highest performing model based on the scores from the validation data.
5. Tune the model hyperparameters of the highest performing model based on cross-validation scores.
6. Select the optimal hyperparameters based on highest accuracy and lowest uncertainty with active learning.
7. Apply the trained model to an unlabelled subsurface dataset to interpret the Late Cretaceous and Early Palaeocene tectonic history of the eastern Greater Green River Basin.

2.1 | Training Data

The 3D training data for this study are generated from synthetic stratigraphic successions. Three different types of stratal geometries are simulated: horizontal stratification, truncation and onlap (Figure 1). All training data are generated in Python jupyter notebooks using the common packages:

1. NumPy to build and manipulate arrays representing surfaces (Harris et al., 2020; Pérez & Granger, 2007).
2. Pandas to store and manipulate tabular dataset (McKinney, 2010).
3. Matplotlib for plotting stratigraphic cross-sections, structure and thickness maps, and visual quality control (Hunter, 2007).
4. Verde for efficient interpolation of surfaces (Uieda, 2018).

The synthetic horizontally stratified succession is calculated by combining horizontal and subhorizontal stratigraphic surfaces with no angular discordance. The synthetic truncation stratigraphy is calculated by combining surfaces that were parallel to a symmetric sine curve and truncating the surfaces at a specific height (Figure 2). The onlap stratigraphy is calculated by combining horizontal and subhorizontal stratigraphic surfaces that onlap onto a symmetric sine curve that represent a topographic high (Figure 2). For surfaces in all three classes, random noise is added to the depths, and missing values are added by removing values at random to account for areas where a human interpreter might not have picked a contact.

For each class of synthetic stratigraphy, a random number of vertical wells are generated with a set number of formation tops picked at the correct depth in 3D space to mimic the generally sparse dataset available for interpretation. Additionally, the training dataset contains rotations and translations of the data in the xy plane to replicate a
variety of orientations. For example, in one simulation a topographic high might be located on the west side of the region and oriented from north to south with a long wavelength. In a different simulation it might be located on the north side of the region and oriented from west to east with a second topographic high on the south side of the region with a similar orientation. Finally, the datasets are eroded by creating an erosion surface at the top of the model and removing all surfaces above that level. This gives what is believed to be a full representation of the feature space containing a variety of orientations, depths, erosion and random missing values to represent natural conditions and realistic sample data from the subsurface. At the end of stratigraphic surface generation, the dataset consists of sample locations, depth to surfaces at that location and the type of stratal geometry. Ultimately, the machine learning problem needs predictor and response features. In this case, the sample locations and stratigraphic thicknesses will be used as predictor features and the type of stratal geometry as the response feature. To use these predictor features depth must be converted to thicknesses, and the feature set further enhanced.
2.2  Feature Engineering

Feature engineering is the process of either extracting or transforming raw data into inputs for a machine learning model (Zheng & Casari, 2018). This includes the scaling, binning, log transforms and power transforms of features. To enlarge the training dataset, feature engineering was conducted to account for non-independence, due to vertical volume constraints and horizontal spatial continuity, between spatial sample points, and to increase the sensitivity of the model to non-linear changes in the dataset. From the training data, stratigraphic thicknesses were calculated for each sample point by differencing depths of formation tops. Additional features were then added that are natural log and power transformations of the formation thickness at the sample location. Because the sample locations are not independent of other sample locations surrounding them, the formation thicknesses with their log and power transformations for the nearest sample locations were added to each primary sample location. The number of sample locations in the vicinity is a model hyperparameter that is tuned on a withheld testing subset of the subsurface dataset, where categorical accuracy is maximised.

The result is a training dataset with 1,203 features, 28,800 samples including 9,600 samples for each geometry. The predictor feature set is normalised on the interval from 0 to 1, to generalise the model for any scale of interest. The structure of the dataset enables a machine learning classification model to learn the geometries for every sample location and the closest \( N \) sample locations where \( N \) ranges from 0 to 399.

2.3  Machine Learning Workflow

Training data are split into training and validation subsets. Initially, validation data were selected at random and constituted 20% of the training data. Scikit-learn (Pedregosa et al., 2011) was then applied in Python to evaluate five different classification models: support vector classification, decision trees, random forests, AdaBoost and K-nearest neighbours. These classifiers are expected to do reasonably well on this problem because it is assumed that the training data are linearly separable and it is easy to interpret these models results. For this model, the most probable stratal type is selected as the prediction, but additional attention is paid to predictions with slim majorities for any single stratal type.

For this multiclass problem, a weighted Jaccard similarity coefficient score is used to measure the accuracy of each model. The initial weighted accuracy for each model with no tuned hyperparameters is as follows: support vector classifier 52%, decision tree 79%, random forest 82%, AdaBoost 50% and K-nearest neighbour 61%. The random forest classifier had the highest initial accuracy on the withheld validation dataset, and was selected for further hyperparameter tuning. Grid search and fivefold cross-validation was used in scikit-learn to find the optimal parameters for the random forest classification model. Accuracy increases with an increasing number of wells in the vicinity, but the certainty of the predictions also decreases with the number of wells in the vicinity. To avoid overfitting the model on the training data, margin sampling of classification probabilities and active learning on the unlabelled real-world dataset was used to optimise the number of wells in the vicinity.

The hyperparameters selected to tune for this model are divided into two categories. The first category relates to sampling and includes: the number of wells in the vicinity, sample bootstrapping, the minimum samples per leaf and the minimum samples per split. The second category are hyperparameters related to the size of the forest and the tree split quality criterion. Ultimately, the first category controls the accuracy of the random forest, while the second category controls the compute time of each iteration. Optimal hyperparameters and the grid search parameters are in Table 1.

### Table 1  Model hyperparameter grid search values

| Hyperparameter name       | Grid search value 1 | Grid search value 2 | Grid search value 3 | Optimal value |
|---------------------------|---------------------|---------------------|---------------------|---------------|
| Bootstrap                 | True                | False               | False              | False         |
| Criterion                 | Gini                | Entropy             | Entropy             | Entropy       |
| Max depth                 | 1                   | 10                  | 100                 | 100           |
| Min samples per leaf      | 10                  | 100                 | 1,000               | 10            |
| Min samples per split     | 10                  | 100                 | 1,000               | 10            |
| Number of estimators      | 10                  | 100                 | 1,000               | 1,000         |
| Wells in vicinity         | 0–25 (step of 1)    | 25–100 (step of 25) | 100–399 (step of 50) | 2             |
certainty, it was then used to predict the stratal geometries for a subsurface dataset from the Eastern Greater Green River Basin.

3 | GEOLOGICAL SETTING

The eastern Greater Green River Basin is a Sevier and Laramide age intermontane basin located in south-central Wyoming and north-central Colorado that is bounded by basement-cored uplifts (Figure 4A) (Dickinson et al., 1988). From the early through much of the Late Cretaceous, thin-skinned deformation of the Sevier orogeny created the Wyoming Fold and Thrust Belt and associated Sevier Basin (Jordan, 1981). During the Late Cretaceous and early Eocene, the structural deformation changed to thick-skinned, basement-cored reverse faulting known as the Laramide orogeny. Laramide-aged positive topographic features partitioned the once continuous Sevier Basin into numerous discontinuous basins, disrupting sediment dispersal and deposition patterns (Dickinson et al., 1988). The uplift and erosion of sediment from the Rock Springs Uplift truncated the underlying Upper Cretaceous Lance Formation. Following this, there was deposition of the overlying Palaeocene Fort Union Formation (Figure 4B; Kirschbaum et al., 1994). While steep westward dips characterise the west side of the Rock Springs Uplift, the eastern limb has 4 degrees of angular discordance between the Lance and Fort Union formations (Kirschbaum et al., 1994).

Hettinger and Kirschbaum (1991) and later Lynds and Lichtner (2016) used well-log correlations to interpret the angular discordance on the eastern limb of the Rock Springs Uplift between the Lance and Fort Union formations as truncation. This implies erosion of the entire basin to varying degrees after deposition of the Lance Formation, and before deposition of the Fort Union Formation.

3.1 | Subsurface Data

Lynds and Lichtner (2016) correlated subsurface formation tops across Eastern Greater Green River Basin, which are available as a supplementary data download from the Wyoming State Geological Survey. The dataset includes correlations of the top of the Palaeocene Fort Union Formation, Late Cretaceous Lance Formation and the Late Cretaceous Fox Hills Formation. This dataset also includes correlations of members within each formation, but are beyond the scope of this study. The correlations for the three formations include 887 wells across the basin (Figure 5). For further information on picking criteria for formation tops see Lynds and Lichtner (2016). From this dataset, the stratigraphic thicknesses are calculated for the Fort Union and Lance formations in all 887 wells. Additionally, the feature engineering above is used on this subsurface dataset. After calculating the thicknesses, completing the feature engineering and normalisation, the classification model is used with optimal parameters to make predictions for all 887 wells and both formations.

4 | RESULTS

The classification results for the Lance Formation are presented first followed by the results for the Fort Union Formation. For the Lance Formation, the classification model predicts that a majority of the wells in the Great Divide Basin axis contain a horizontally stratified Lance Formation (Figure 6, Table 2). The classification model documents onlap in the south-central portion of the Great Divide Basin where it predicts onlap geometries that trend north-west to south-east (Figure 6). Moving further to the south the classification model predicts a broad swath of truncation geometries in the Lance Formation. After which the model predicts onlap geometries in the Washakie Basin (Figure 6). Overall the classification model documents a mix of stratal geometries in the Lance Formation (Figure 6, Table 2).

In the Fort Union Formation, there is a similar pattern as in the Lance Formation. The axis of the Great Divide Basin is classified as horizontally stratified (Figure 7; Table 2). The classification model documents onlap in the south-central portion of the Great Divide Basin where it predicts onlap geometries that trend north-west to south-east (Figure 7). Moving further to the south the classification model predicts a broad swath of truncation geometries in the Lance Formation. After which the model predicts onlap geometries in the Washakie Basin (Figure 6). Overall the classification model documents a mix of stratal geometries in the Lance Formation (Figure 6, Table 2).
FIGURE 4  (A) Location map of the Greater Green River Basin and sub-basins in Wyoming, Colorado and Utah (modified from Lynds & Lichtner, 2016). The dashed black rectangle is the study area depicted in Figure 5. (B) Chronostratigraphic chart for the Upper Cretaceous through Eocene in the Eastern Greater Green River Basin (modified from Lynds & Lichtner, 2016)
documents a wider area of horizontal stratification for the Fort Union Formation in the Great Divide and Washakie basins than for the Lance Formation. In the Great Divide and Washakie basins, the Fort Union Formation has horizontal geometries in more wells than the underlying Lance Formation (Figures 6 and 7; Table 2).

**FIGURE 5** Map of the Eastern Greater Green River Basin documenting the location of wells, cross-sections and faults in the Great Divide and Washakie basins (modified from Lynds & Lichtner, 2016)

**FIGURE 6** Map showing the spatial distribution of stratal geometry predictions for the Lance Formation. The thickness of the Lance Formation ranges from 0 to 1,500 m in the basins, and visually aligns with the transitions from horizontal stratification to onlap and then to truncation. The size of the well markers document the certainty of the classification with larger markers corresponding to higher certainty in the classification. The stacked histogram documents the number of wells predicted for each class.

**TABLE 2** Model prediction results for both formations and mean certainty for each
DISCUSSION

5.1 Lance Formation

The classification results for the Lance Formation are geologically reasonable in that the deepest portions of the Great Divide Basin are conformable with both the underlying Fox Hills Formation and overlying Fort Union Formation. The band of wells classified as onlap that trend north-west to south-east across the southern portion of the Great Divide Basin are interpreted as documenting a wide basin margin during deposition of the Lance Formation. Subsidence in the Great Divide Basin could have produced the onlap geometries that the classification model documents in the Lance Formation, similar to the mini-basin subsidence modelled by Sylvester et al. (2015). The lack of onlap classifications along the east-central portion of the Great Divide Basin is believed to document that the Rawlins, and Sierra Madre uplifts were not active, and not influencing the sediment routing system during this time. Hettinger et al. (1991) document age differences between the Lance and Fort Union formations in their measured section (J) using palynological analyses. However, they do not document soil formation, or other diagenetic processes typically associated with exposure during uplift and erosion of sedimentary units. This lends credence to the onlap and horizontally stratified classifications predicted by the classification model. Moving south towards the Wamsutter Arch area, areas where the classification changes from onlap to truncation are interpreted to be areas that were originally onlap, but were subsequently uplifted and eroded to form a truncated geometry.

This area of north-west to south-east oriented band of wells classified as truncation encompasses the following structural features: the Wamsutter Arch, the eastern limb of the Rock Springs Uplift, the Dad Arch and the eastern Washakie Basin. Uplift on these structural features after deposition of the Lance Formation explains the truncation classification. This in turn means that the uplifts would have started to partition the Eastern Greater Green River Basin into the separate sub-basins after the deposition of the Lance Formation.

The prediction of truncation of the Lance Formation across the centre of the Eastern Greater Green River Basin is justified from outcrop observations by Roehler (1983) and Hettinger and Kirschbaum (1991) who document a palaeosol at the top of the Lance Formation along the eastern limb of the Rock Springs Uplift. Additional 2D seismic data presented by Ryder et al. (1989) and Rudolph et al. (2015) confirm that the Lance Formation is truncated in this portion of the basin which lends further validation to the classification results.

Throughout the eastern Greater Green River Basin, there are areas where there are outlier wells classified as onlap surrounded by wells classified as truncation. These wells are likely the result of underlying data issues. The issues could be different human interpreters using different well-log criteria, or it could be due to other depth variations such as incorrect surface elevations or other errors while digitising well data.

Visually, some of the low-certainty classifications are for wells classified as onlap interspersed with wells classified as truncation. These wells are likely the result of underlying data issues. The issues could be different human interpreters using different well-log criteria, or it could be due to other depth variations such as incorrect surface elevations or other errors while digitising well data.

5

FIGURE 7 Map showing the spatial distribution of stratal geometry predictions for the Fort Union Formation. The thickness of the Fort Union Formation ranges from 0 to 1,500 m in the basins. The size of the well markers document the certainty of the classification with larger markers corresponding to higher certainty in the classification. The stacked histogram documents the number of wells predicted for each class.
classification model assists in constraining the timing, location and orientation of uplift and erosion across the Eastern Greater Green River Basin in a manner that is consistent with orientations documented by early workers in the region (Ritzma, 1955; 1968).

### 5.2 Fort Union Formation

Throughout the Eastern Greater Green River Basin, the Fort Union Formation is classified as mostly horizontally stratified. On the western margin of the basin along the Wamsutter Arch, the classification model predicts truncation of the Fort Union Formation along the Wamsutter Arch. However, to the west of the predicted truncation geometry there are wells that are classified as onlap. While this might seem strange to see onlap to the west of truncation, the prediction is consistent with outcrop and subsurface studies by Hettinger and Kirschbaum (1991) who document the Fort Union onlapping the underlying Lance Formation in these areas. Structurally, these wells classified as truncation are east of a series of right-lateral strike-slip faults. The predicted truncation is either related to Laramide age compression, or Neogene extension interpreted by Bader (2008). Ritzma (1968) proposed movement on the Wamsutter Arch anywhere from Palaeocene through Pliocene, which would be coeval with movement on this strike-slip fault system. Additionally, Barlow (1961) documents a palaeohigh area during the deposition of the Almond Formation and Lewis Shale (Upper Campanian through lower Maastrichtian) in this area of Fort Union truncation. If this area was a fault-bounded palaeohigh, it is probable that the faults were active through the Palaeocene, producing the truncation in these wells and leaving the wells further west with onlap geometry. Given that the subsurface dataset used does not contain formation picks any younger than Palaeocene, the truncation of the Fort Union in this area can only be constrained to either late or middle Palaeocene at the earliest. Hettinger and Kirschbaum (1991) and Roehler (1979) document an intraformational unconformity in the Fort Union based on a lack of middle Palaeocene palynomorphs and a well-developed palaeosol in outcrop. This means that the classification model is potentially picking up the thickness changes due to intraformational erosion.

Another confounding factor on the nose of the Rock Springs Uplift is the Deadman coal zone. From a process-based perspective, the Deadman coal zone is equivalent to a histosol. The Deadman coal zone histosol in the Fort Union Formation overlies the well-developed palaeosol at the top of the Lance Formation. This would appear to be onlap with the Fort Union onlapping onto the basin margin composed of truncated Lance Formation. In this case, the classification model has the choice of one of three classes, and cannot predict the relative combination of both truncation and onlap. It simply classifies the wells as one of the three classes and provides a certainty of that well belonging to the predicted class. All of the wells on the nose of the Rock Springs Uplift were classified with varying degrees of certainty by the random forest classification model.

The other area classified as a mix of onlap and truncation in the Fort Union Formation is in the Washakie Basin. It is in a triangular zone west of the Sierra Madre Uplift, and north of the Cherokee Ridge Arch. If we assume that the orientation of the principal stress direction at the end of the Palaeocene is what Bader (2008) proposes (ca 250°), then the entire area of the basin would have undergone some form of vertical motion to accommodate uplift of the Sierra Madre. This in turn would result in truncation of the Fort Union Formation if the motion was after deposition, or truncation if the motion was before deposition. Because of the similarity of the two classes, the interspersed truncation and onlap predictions are seen. However, note that the truncation predictions follow a generally north-east to south-west trend which could be fault controlled. A subsurface study by Colson (1969) supports the results of the classification model in this region. Colson (1969) interprets truncation of the Fort Union along the northern side of the Cherokee Arch that becomes conformable moving west and north-west across the Washakie Basin, the same as the predictions of the classification model. Additionally, Lynds and Lichtner (2016) interpret truncation of the Fort Union Formation in this region, and discuss previous work by others, but remain uncertain about the spatial extent of the unconformity between the Fort Union and overlying Wasatch Formation. This model lends not only classification, but also for the first time, certainty to the spatial extent of the unconformity.

### 5.3 Machine Learning Model

Overall, the classification model is useful from both a machine learning and a geological perspective. The model has high classification accuracy on the training dataset, and qualitatively it does a reasonable job of classifying the different stratal geometries in the Eastern Greater Green River Basin. The subsurface classifications are consistent with spot checks and visually align with previous interpretations of the Eastern Greater Green River Basin. However, there are still areas with misclassifications where wells with one class surround a well with a different class. In these misclassified wells, the certainty measure is useful to interpret the classification models predictions. The misclassifications are easier to interpret in the Lance Formation than in the Fort Union Formation. In the Lance Formation, the predictions with low probabilities are spatially located at the boundary between the three classes (Figure 7). This means that in these areas the geometries look very similar to one another. This
is interpreted to be a consequence of the increased structural deformation that occurred during and after the deposition of the Fort Union Formation.

To evaluate the feature engineering and classification model two strategies were used. First, on the training and test datasets, feature groups (location, vicinity well stratigraphic thickness, natural log transform of thickness, power transform of thickness) were iteratively removed and the model retrained. The model’s predictions were then tested against the ground-truth values in the test dataset and used to calculate the accuracy of the model. The model scores 88% accuracy without location data and 85.3% and 88% accuracy without stratigraphic thickness and power transforms of thickness, respectively. When training the model without the natural log transform of stratigraphic thickness group, the accuracy on the test dataset is at best 62%. This means that the natural log transform are the most important features for predicting stratigraphic geometry at this scale.

Next, to interpret the potential misclassifications, t-distributed stochastic neighbour embedding (t-SNE) is used to reduce the dimensions of the feature set to visualise the classifications in a low-dimensional space. The idea behind t-SNE is that it uses local relationships between data points to map from high to low-dimensional space. It does this by using a Gaussian distribution to map the similarities between points in high-dimension space before using a Student t-distribution to project the probability distribution to low-dimensional space (Maaten & Hinton, 2008). This method of mapping the probability distribution to low-dimensional space means that t-SNE is able to deal with the curse of dimensionality and crowded data points after dimension reduction. After running t-SNE on the data from the Eastern Greater Green River Basin, a scatterplot was created of the t-SNE embedding dimensions coloured by the probability of each predicted class and symbolised by the two different formations (Figure 8). Figure 8 documents two primary groups of predictions, and additional smaller clusters. One large group of predictions contains wells classified with low certainty as onlapping and truncated. These are the wells where the onlap and truncation geometries appear similar. The other large cluster contains exclusively horizontally stratified predictions. The random forest classification model has low confidence in the wells where truncation and onlap appear similar (upper right cluster) (Figure 8).

To interpret how a single sample is classified the distribution of the natural log thicknesses was compared for the sample and the training data. The more wells in the vicinity that are included in the feature set, the closer the sample distribution gets to the distribution of the idealised training data classes. The Kullback–Leibler (K–L) divergence is used to measure the similarity between a single sample’s natural log thickness distribution and the training data distribution. Visually, a histogram is plotted of a training sample and its similarity to each one of the three classes to have a general idea of how a sample will be classified (Figure 9). The lower the divergence value the more similar the sample’s distribution is to the class. As the number of wells in the local vicinity increases the K–L divergence decreases. This explains why the accuracy steadily increased with the number of wells in the vicinity during model training. With the optimal number of two wells in the local vicinity, the K–L divergence is a computationally inexpensive way to estimate classes at a glance. This gives another layer of interpretability to understand how the random forest is making its decisions. After interpreting the performance of the model for classifying each well, its practical applications in subsurface studies is investigated.

5.4 Guided Interpretation

The greatest use of the classification model in the Eastern Greater Green River Basin is that it aids in geological interpretation of stratal geometries. It documents the different spatial distributions of horizontal stratification, onlap and truncation. The classification model aids the geologist in searching for unique patterns that are not easily interpreted in non-unique structure and thickness maps. Of course, traditional geological interpretation of wireline logs and core
descriptions such as that by Anderson and Longman (2018) tend to be much more accurate and thorough for individual wells. However, the use of the random forest classification model produces a broad scale interpretation of stratal geometries across a basin in a matter of seconds. This method is ultimately most useful for identifying large-scale geometries. Other cases of large-scale geometries such as thinning due to condensing of the strata across a low-accommodation zone remain to be tested with a new training dataset. In the condensed section case, the classification model would most probably classify the wells as truncation, yet the stratigraphic surfaces would be conformable. In the case of the Eastern Green River Basin, the focus is on the field-documented geometries of the Upper Cretaceous and Palaeocene, and condensed sections can be ruled out, as the outcrop documented palaeosols at the truncation surfaces document erosion and exposure rather than conformable thinning of the formations. Additionally, the uncertainty in classification from this model is useful for geologists. The model can be run on an entire basin in a matter of minutes, and then geologists can further investigate areas of low probability classifications. With this iterative approach, a single geologist can map stratigraphic geometries and assign probabilities in a much shorter time than traditional methods. Future work on classifying stratal geometries should focus on higher resolution stratigraphic surfaces. This includes multiscale models with intraformational members, sequence stratigraphic surfaces, and bed and bedset boundaries.

6 | CONCLUSION

A training dataset was developed employing a machine learning model which resulted in a new hypothesis for Late Cretaceous and Palaeocene stratal geometries in the Eastern Greater Green River Basin. The training dataset is synthesised by using observable geological principles of basin-scale stratigraphic geometries, namely horizontal stratification, onlap and truncation. The dataset is believed to sample a majority of the possible feature space. After creating the training dataset, feature engineering is applied to integrate non-linearity, spatial dependence between sample locations and normalisation for generalisation of the prediction model to a variety of system scales. The dataset was then split into different train and validation datasets, and a random forest model was trained on the training dataset.

The model is tested on the validation dataset with 88.4% classification accuracy and 72.8% mean certainty. The trained model classified the stratigraphic geometries for two formations in 887 wells in the Eastern Greater Green River Basin of central Wyoming. The model reveals areas that were horizontally stratified bounded to the south by areas of onlap, which are attributed to basin filling during the Late Cretaceous and Palaeocene. Areas around the Rock Springs and Sierra Madre uplifts further document uplift and truncation of both the Lance and Fort Union formations. This is interpreted to indicate that the Wamsutter Arch underwent multiple episodes of uplift, beginning after the deposition of the Lance Formation, and
at minimum a second period after deposition of the Fort Union Formation. The classification results are corroborated by numerous subsurface, outcrop and core-based geological studies, lending credence to machine learning based subsurface interpretation.

The random forest classification model provides an expedient method to classify basinwide geometries and make interpretations of basin filling patterns and their interaction with structural related uplift and subsidence. Additionally, the model assigns certainty to each well that can be used to constrain areas of further scrutiny. More work is needed for prediction of multiscale surfaces, but this article documents that geometric models developed from observable stratigraphic principles can be used to build and test training datasets that can be used on real-world subsurface data.

ACKNOWLEDGMENTS
This article, associated code and online data benefited from a thorough review by Matt Hall. We thank Mark Kirschbaum and Charles Samra for reviewing and discussing early drafts of this article and providing valuable insight.

DATA AVAILABILITY STATEMENT
The subsurface dataset used in this study is available from the Wyoming State Geological Survey at http://www.wsgs.wyo.gov/products/wsgs-2016-ri-73.zip (Lynds & Lichtner, 2016), and the generated training data are available at https://osf.io/a6cwh/ (Pisel, 2020). Code is available at http://github.com/jessepisel/stratal-geometries and https://zenodo.org/record/4121965.

ORCID
Jesse R. Pisel https://orcid.org/0000-0002-7358-0590

REFERENCES
Anderson, D.S. & Longman, M.W. (2018) Subsurface reinterpretation of Ordovician and Devonian strata in Southwest Wyoming with implications for upwarping across the transcontinental arch. The Mountain Geologist, 55, 91–118.
Bader, J.W. (2008) Structural and tectonic evolution of the Cherokee Ridge arch, south-central Wyoming: implications for recurring strike-slip along the Cheyenne Belt suture zone. Rocky Mountain Geology, 43, 23–40.
Barlow, J.A. Jr. (1961) Almond formation and lower Lewis Shale, East Flank of Rock Springs Uplift, Sweetwater County, Wyoming. In Symposium on Late Cretaceous Rocks, Wyoming and Adjacent Areas, 16th Annual Field Conference Guidebook, pp. 113–115.
Colson, C. (1969) Stratigraphy and production of the Tertiary formations in the Sand Wash and Washakie Basins. In Symposium on Tertiary Rocks of Wyoming, 21st Annual Field Conference Guidebook, pp. 121–128.
Dickinson, W.R., Klute, M.A., Hayes, M.J., Janecke, S.U., Lundin, E.R., McKittrick, M.A. et al. (1988) Paleogeographic and paleotectonic setting of Laramide sedimentary basins in the central Rocky Mountain region. Geological Society of America Bulletin, 100, 1023–1039.
Galloway, W.E. (1989) Genetic stratigraphic sequences in basin analysis I: architecture and genesis of flooding-surface bounded depositional units. AAPG Bulletin, 73, 125–142.
Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020) Array programming with NumPy. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2
Hettinger, R.D., Honey, J.G. & Nichols, D.J. (1991) Chart showing correlations of upper Cretaceous Fox Hills Sandstone and Lance Formation, and lower Tertiary Fort Union, Wasatch, and Green River formations, from the eastern flank of the Washakie Basin to the southeastern part of the Great Divide Basin, Wyoming. US Geological Survey IMAP Report 2151, https://doi.org/10.3133/i2151
Hettinger, R.D. & Kirschbaum, M.A. (1991) Chart showing correlations of some upper Cretaceous and lower Tertiary rocks, from the east flank of the Washakie Basin to the east flank of the Rock Springs Uplift, Wyoming. US Geological Survey IMAP Report, 2152, https://doi.org/10.3133/i2152
Hunter, J.D. (2007) Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9, 90–95.
Jordan, T.E. (1981) Thrust loads and foreland basin evolution, Cretaceous, western United States. AAPG Bulletin, 65, 2506–2520.
Kirschbaum, M.A., Andersen, D.W., Helm, R.L. & Baldwin, R.J. (1994) Paleocene drainage systems, Rock Springs Uplift, Wyoming. The Mountain Geologist, 31, 19–28.
Lynds, R.M. & Lichtner, D.T. (2016) Stratigraphy and hydrocarbon potential of the Fort Union and Lance formations in the Great Divide and Washakie basins, south-central Wyoming. Wyoming State Geological Survey Report of Investigations 73, 70, 2 pls. p.
McKinney, W. (2010) Data structures for statistical computing in python, vol. 445, pp. 51–56. Austin, TX.
Mitchum, R.M. Jr, Vail, P.R. & Sangree, J.B. (1977) Seismic stratigraphy and global changes of sea level: Part 6. Stratigraphic interpretation of seismic reflection patterns in depositional sequences: Section 2. Application of seismic reflection configuration to stratigraphic interpretation. In Payton, C.E. (Ed.), Seismic Stratigraphy – Applications to Hydrocarbon Exploration. American Association of Petroleum Geologists, Memoir, 27, 53–62.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
Pérez, F. & Granger, B.E. (2007) IPython: a system for interactive scientific computing. Computing in Science & Engineering, 9, 21–29.
Pisel, J.R. (2020) Classifying basin-scale stratigraphic geometries from subsurface formation tops with machine learning. OSF, osf.io/a6cwh.
Ritzma, H.R. (1955) Late Cretaceous and early Cenozoic structural pattern, southern Rock Springs uplift, Wyoming. In Wyoming Geological Association Guidebook, Wyoming Geological Association, 10th Annual Field Conference, pp. 135–137.
Ritzma, H.R. (1968) Geology and occurrence of gas, Wamsutter arch, Sweetwater County, Wyoming. In Beebe, B.W. & Curtis, B.F. (Eds.), Natural gasses of North America, AAPG Memoir 9, 1, 817–827.
Roehler, H.W. (1979) Geology and energy resources of the Sand Butte Rim NW Quadrangle, Sweetwater County, Wyoming. Professional Paper Report 1065A. doi:10.3133/pp1065A
Roehler, H.W. (1983) Stratigraphy of Upper Cretaceous and Lower Tertiary outcrops in the Rock Springs Uplift, Wyoming. US Geological Survey IMAP 1500.

Rudolph, K.W., Devlin, W.J. & Crabaugh, J.P. (2015) Upper Cretaceous sequence stratigraphy of the Rock Springs Uplift, Wyoming.

Ryder, R.T., Lee, M.W., Agena, W.F. & Anderson, R.C. (1989) Seismic profile through Patrick draw-table rock area, east flank Rock Springs Uplift, Wyoming. In Gas Resources of Wyoming, v. 40th Annual Field Conference Guidebook, pp. 209–229.

Sylvester, Z., Cantelli, A. & Pirmez, C. (2015) Stratigraphic evolution of intraslope minibasins: insights from surface-based model. AAPG Bulletin, 99, 1099–1129.

Uieda, L. (2018) Verde: processing and gridding spatial data using Green’s functions. Journal of Open Source Software, 3, 957. https://doi.org/10.21105/joss.00957

Vail, P.R. (1987) Seismic stratigraphy interpretation using sequence stratigraphy: Part 1: Seismic stratigraphy interpretation procedure. In Bally, A.W. (Ed.), Atlas of seismic stratigraphy. American Association of Petroleum Geologists, Studies in Geology 27, 1–10.

van der Maaten, L. & Hinton, G. (2008) Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605.

Van Wagoner, J., Posamentier, H., Mitchum, R., Vail, P., Sarg, J., Loutit, T. et al. (1988) An overview of the fundamentals of sequence stratigraphy and key definitions. In Wilgus, C.K., Hastings, B.S., Kendall, C.G.S.G., Posamentier, H.W., Ross, C.A. & Van Wagoner, J.C. (Eds), Sea-level changes: an Integrated Approach. Society of Economic Paleontologists and Mineralogists, Special Publications 42, 39–45.

Zheng, A. & Casari, A. (2018) Feature engineering for machine learning: principles and techniques for data scientists. Sebastopol, CA: O’Reilly Media Inc.

How to cite this article: Pisel JR, Pyrcz MJ. Classifying basin-scale stratigraphic geometries from subsurface formation tops with machine learning. Depositional Rec. 2021;7:64–76. https://doi.org/10.1002/dep2.129