Data-Centric Engineering (2023), 4: e12
doi:10.1017/dce.2023.9

RESEARCH ARTICLE

Machine learning approaches for the prediction of serious fluid leakage from hydrocarbon wells

Mehdi Rezvandehy[1] and Bernhard Mayer[2]

[1]Department of Chemical and Petroleum Engineering, University of Calgary, Calgary, AB, Canada
[2]Department of Geoscience, University of Calgary, Calgary, AB, Canada
Corresponding author: Mehdi Rezvandehy; Email: mehdi.rezvandehy@ucalgary.ca

Received: 23 August 2022; **Revised:** 05 April 2023; **Accepted:** 14 April 2023

Keywords: Energy wells; imbalanced class classification; imputation; probability estimation; resampling

Abstract

The exploitation of hydrocarbon reservoirs may potentially lead to contamination of soils, shallow water resources, and greenhouse gas emissions. Fluids such as methane or CO_2 may in some cases migrate toward the groundwater zone and atmosphere through and along imperfectly sealed hydrocarbon wells. Field tests in hydrocarbon-producing regions are routinely conducted for detecting serious leakage to prevent environmental pollution. The challenge is that testing is costly, time-consuming, and sometimes labor-intensive. In this study, machine learning approaches were applied to predict serious leakage with uncertainty quantification for wells that have not been field tested in Alberta, Canada. An improved imputation technique was developed by Cholesky factorization of the covariance matrix between features, where missing data are imputed via conditioning of available values. The uncertainty in imputed values was quantified and incorporated into the final prediction to improve decision-making. Next, a wide range of predictive algorithms and various performance metrics were considered to achieve the most reliable classifier. However, a highly skewed distribution of field tests toward the negative class (nonserious leakage) forces predictive models to unrealistically underestimate the minority class (serious leakage). To address this issue, a combination of oversampling, undersampling, and ensemble learning was applied. By investigating all the models on never-before-seen data, an optimum classifier with minimal false negative prediction was determined. The developed methodology can be applied to identify the wells with the highest likelihood for serious fluid leakage within producing fields. This information is of key importance for optimizing field test operations to achieve economic and environmental benefits.

Impact Statement

Field test operations to detect methane and CO2 leakages from hydrocarbon wells can be costly. Most wells do not have leaks or are categorized as non-serious, which means that no repair is needed until they are abandoned. However, it is crucial to identify and prioritize serious leakages for immediate remediation to prevent environmental pollution. This study developed a reliable predictive model by correlating the results of historical field tests with various well properties, including age, depth, production/injection history, and deviation, among others. The trained model can predict the likelihood of serious leakage for untested wells, allowing for the prioritization of wells with the highest probability of leaks for field testing. This approach leads to cost-effective field testing and environmental benefits.

1. Introduction

Exploitation of oil and gas reservoirs has raised public concerns regarding potential contamination of soils, shallow groundwater, and increases in greenhouse gas emissions (Shindell et al., 2009; Brandt et al., 2014; Cherry et al., 2014). Improperly sealed hydrocarbon wells may lead to the migration of gases such as methane and CO_2 to shallow aquifers, soils, and the atmosphere, or vented through surface casing vent flows (SCVFs). Regulators require monitoring of such gas migration (GM) and vent flows to detect leakage and prioritize the repair of the most severe cases (Watson and Bachu, 2009; Montague et al., 2018; Abboud et al., 2021). The Alberta Energy Regulator (AER) in Alberta, Canada, conducts such field tests for energy wells within the province. The AER applies two field tests for the identification of fluid migration after a well is completed to produce hydrocarbon or to inject any fluid:

1. SCVF is the flow of gas (methane, CO_2, etc.) out of the casing annulus or surface casing. SCVF is often referred to as internal migration. Wells with positive SCVF are considered serious in the province of Alberta under one or several of the following conditions: (a) gas-flow rates higher than 300 m^3/d, (b) stabilized pressure >9.8 kPa/m, (c) liquid-hydrocarbons, and (d) hydrogen sulfide (H_2S) flow (see Alberta Energy Regulator, 2003, for more information).
2. GM is a flow of any gas that is detectable at surface outside of the outermost casing string. GM is often referred to as seepage or external migration (Alberta Energy Regulator, 2003). A GM is serious if there is a high flow rate or public safety hazard or off-lease environmental damage, such as groundwater contamination (see Alberta Energy Regulator (2003) for more information).

Wells with positive SCVF/GM are classified as nonserious if none of the conditions for the serious category are met. In Alberta, repair for serious SCVF/GM leakage is required within 90 days; otherwise, repair is deferred to abandonment (Alberta Energy Regulator, 2003; Watson and Bachu, 2009; Kang et al., 2014; Montague et al., 2018).

Efficient and cost-effective testing of all hydrocarbon-producing wells is a major challenge in areas with large numbers of producing or injection wells. The AER requires testing for all wells only within a small specific area in central and eastern Alberta and only for wells completed since 1995 (Montague et al., 2018; Abboud et al., 2021). There are many wells in other parts of Alberta including abandoned and orphaned wells for which no SCVF/GM test have been conducted. Montague et al. (2018) applied predictive models (machine learning) based on known well properties to generate a binary result for GM: whether the well is positive for a SCVF/GM test or not. Wisen et al. (2020) carried out a descriptive analysis of fugitive gas incidents in British Columbia (BC), Canada, with a particular focus on SCVF. The records from the British Columbia Oil and Gas Commission (BCOGC) were analyzed to uncover frequent leakage pathways, the frequency of such incidents, and the levels of greenhouse gas emissions. Sandl et al. (2021) utilized a basic predictive method to investigate the linkages between various well features and reported instances of GM in BC. There were other database analysis studies to identify wellbore leakage based on a wide range of factors (Fleming et al., 2021; Cahill and Samano, 2022; Iyer et al., 2022).

In this article, we apply several potential improvements to the workflow presented in Montague et al. (2018). The only wells within the small test region (central and eastern Alberta) was included in the study by Montague et al. (2018) and no predictions were made regarding the seriousness of fluid migration. Most leakages are nonserious (Alberta Energy Regulator, 2022) and repair is not required until abandonment; in contrast, serious leakages are critical and should be identified and prioritized for amendment to prevent environmental pollution. All available data within Alberta were used in this article to predict serious fluid leakage.

A wide range of machine learning algorithms was applied using Python's scikit-learn package (Pedregosa et al., 2011). The algorithms are *Stochastic Gradient Descent*, *Logistic Regression*, *Support Vector Machine*, *Random Forest*, *Adaptive Boosting*, and *Deep Neural Network.* See Géron (2019) for more information about these algorithms. The key goals of this study were to address obstacles encountered when implementing predictions, including the following:

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

1. The first challenge faced by this study was the high number of missing data. Replacing missing data with a constant value such as the mean or median of the feature can result in unreliable predictions and artifacts due to a large population of data having similar values. The *K* nearest neighbors (KNN) is an algorithm that applies feature similarity to impute missing values. The KNN algorithm is prone to being affected by outliers. Moreover, it does not take into account the potential uncertainty in imputed values. There are complex techniques such as multivariate imputation by chained equation (MICE) (van Buuren and Groothuis-Oudshoorn, 2010) and deep learning (DataWing, 2022). However, the imputation using these techniques can be quite slow and computationally expensive for large datasets. They may also need special software, distributional assumption, and the uncertainty in imputation of missing data cannot be considered. Therefore, a new approach was developed to impute missing values by conditioning using available data and quantifying the uncertainty of imputed values. This approach is fast and efficient for big data sets and easy to implement (in Python).
2. The second challenge was a highly skewed distribution toward the negative class. This leads to forcing the predictive models to an unrealistically high classification for the negative class (Montague et al., 2018; Brownlee, 2020) and underestimating the positive class. Resampling techniques were used to adjust the class distribution of training data to feed more balanced data into predictive models, thereby creating a new transformed version of the training set with a different class distribution. Two main approaches for random resampling an imbalanced dataset are: (a) *Undersampling*: this approach deletes random instances of a majority class from the training set. A drawback of undersampling is eliminating the instances that may be important, useful, or critical for fitting a robust decision boundary (He and Ma, 2013; Brownlee, 2021). (b) *Oversampling:* this approach adds random instances (duplicates) of the minority class to the training set. A disadvantage of this technique is that it increases the likelihood of overfitting because of including the exact copies of the minority class examples (Fernández et al., 2018; Brownlee, 2021). In this article, integration of oversampling and undersampling approaches was applied to resolve the issue of imbalanced data.
3. Finally, the main aim was to achieve a higher-performance classifier. Ensemble Learning was utilized to integrate multiple models to build a stronger predictor. It works by aggregating the predictions of a group of predictors. Hard Voting is a simple Ensemble Learning that aggregates the predictions of each classifier and predicts the class that gets the most votes. Soft Voting is another Ensemble Learning that works by averaging the probability of each class and predict a class with the highest probability (Géron, 2019). Both Hard Voting and Soft Voting were applied in this article to enhance performance.

The overall objective of this study was to apply the above-described novel solutions to machine learning approaches to generate a final trained model that can subsequently be applied to predict the probability of serious fluid leakage for the wells with known properties for which SCVF/GM tests are not available.

2. Field Test Data

Figure 1 shows a location map of classification for SCVF/GM test results obtained by AER (Alberta Energy Regulator, 2022) between January 1984 and November 2021 within Alberta, Canada. Based on 27,404 tests that were conducted, 83.8% of the wells were classified as nonserious, while 16.2% of the tested wells had serious leakage that required immediate fixing to avoid environmental impacts. We considered known physical properties of the wells as training features to predict the probability of serious fluid leakage in the province of Alberta, Canada. The properties were retrieved from geoSCOUT, a large database of well characteristics in Alberta (geoSCOUT, 2022). Table 1 shows the 22 physical properties that were considered for each well displayed in Figure 1. They include the following: properties 1 and 2 define deviated and horizontal wells (True/False). Properties 3–10 describe surface casing and production casing specifications of each well. Production-casing and surface-casing grades are string

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

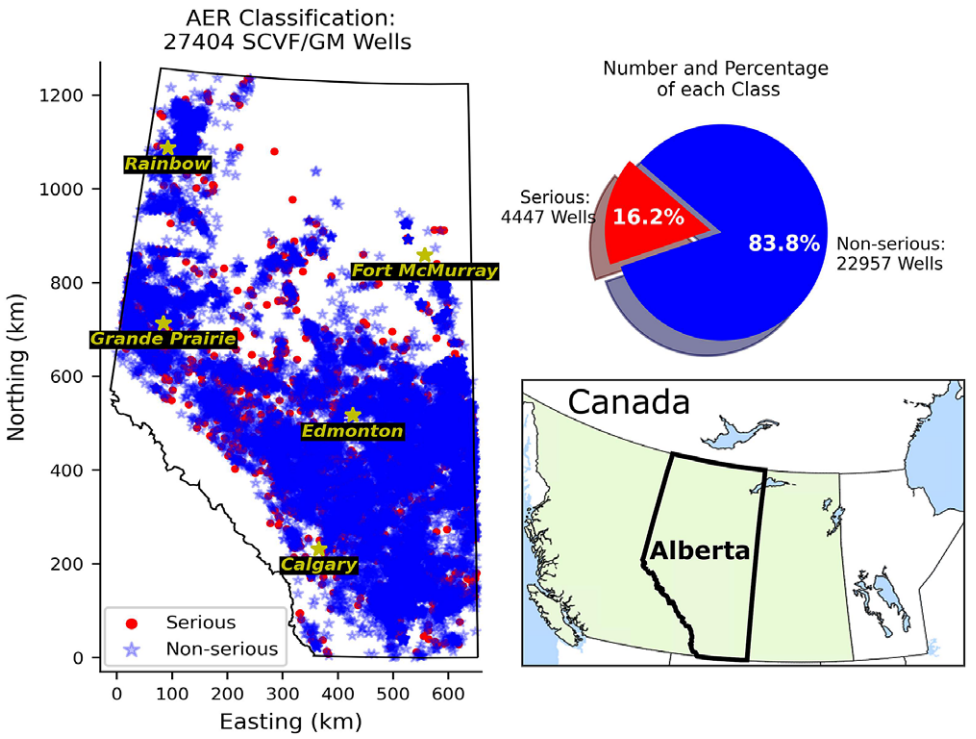

Figure 1. *Location map of Alberta Energy Regulator (AER) classification for test results (serious and nonserious) of surface casing vent flow (SCVF) and gas migration (GM) for energy wells in Alberta, Canada. The majority of the wells are classified as nonserious (83.8%).*

Table 1. *Twenty-two physical properties for each well in Figure 1 retrieved from geoSCOUT (2022).*

Physical properties of wells	
1. Deviated hole (T/F)	12. Borehole temperature (°C)
2. Horizontal hole (T/F)	13. Prod./Inject. formation
3. Surface-casing depth (m)	14. Well status
4. Surface-casing size (mm)	15. Month well spudded
5. Surface-casing weight (kg/m)	16. Well density (n/10 km × 10 km)
6. Production-casing depth (m)	17. Surface abandonment type
7. Production-casing size (mm)	18. Surface abandonment month
8. Production-casing weight (kg/m)	19. Cumulative GAS prod. (e^3m^3)
9. Production-casing grade	20. Cumulative OIL prod. (m^3)
10. Surface-casing grade	21. Cumulative WATER prod. (m^3)
11. Measured depth (m)	22. Total production month

variables (text). Properties 11 and 12 are measured depth and the temperature of the borehole, respectively. Property 13 is the geological formation (text) targeted for production or injection. Property 14 shows the status of the well in text such as suspended, issued, and abandoned. Property 15 is the age of each well in months (counted from January 2022). Property 16 is the regional well density calculated as the total number of hydrocarbon wells with positive SCVF/GM test within 10 km × 10 km area around each well.

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

Property 17 indicates the type of surface abandonment such as plate or cement; 18 is time in month since abandonment (counted from January 2022). Properties 19–22 are cumulative gas, oil, and water production and total months in production.

3. Workflow

Binary classification was applied using the 22 physical properties in Table 1 as training features, while using the SCVF/GM test results (AER classification) as target with serious leakage as positive class (value 1) and nonserious leakage as negative class (value 0). Figure 2 shows the workflow used in this article. First, the dataset is split into a training set, validation set, and test set since a model should be trained first and then reasonably evaluated. The data shown in Figure 1 were split into training (72%), validation (18%), and test sets (10%) as shown in Figure 3. The percentage of nonserious (83.8%) and serious (16.2%) classes for the training, validation, and test sets should be identical to the results for the entire dataset shown in Figure 1. The reason for having a test set as well as a validation set is to avoid overfitting and to evaluate the model based on a never-before-seen dataset. Developing such models always involves tuning hyperparameters; feedback that signals the performance of the model on a validation set is used for tuning. Although the model is never directly trained on the validation set, tuning the configuration of the model based on its performance on the validation set can quickly result in overfitting to the validation set, suggesting that some information about the validation data leaks into the model whenever a hyperparameter is tuned. Therefore, the model should not have any access to any information about the test set; it is only applied at the end of the project to evaluate the performance of the final model.

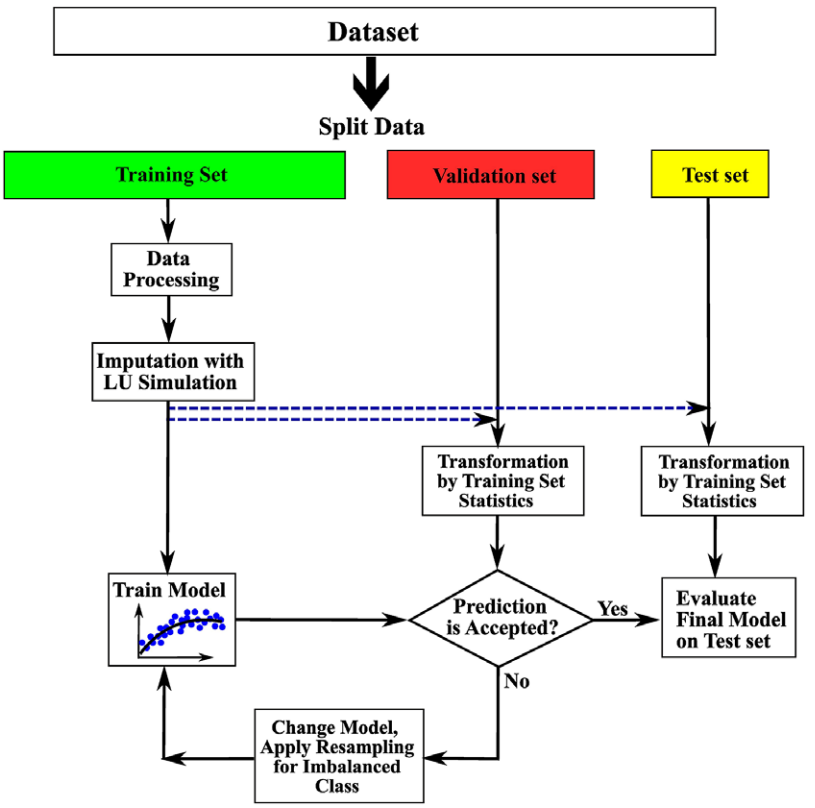

Figure 2. *Schematic illustration of the workflow used in this article.*

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

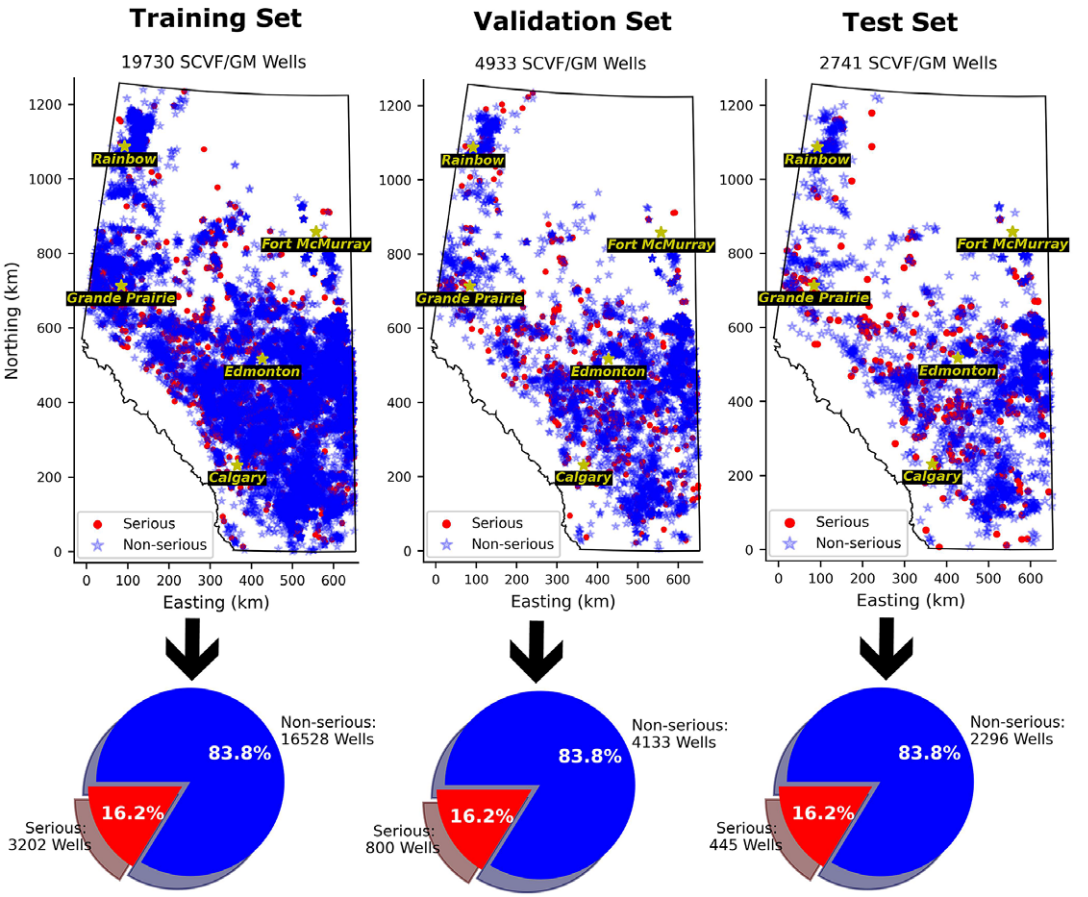

Figure 3. *AER classification shown in Figure 1 is separated into training, validation, and test sets. Location map (top) and pie chart (bottom) for training (left), validation (middle), and test sets (right).*

The next step is data processing for the training set including normalization, text handling, and imputation. Text handling should be efficiently done to convert text to numbers before feeding predictive algorithms. To enhance the performance of algorithms, a target encoding technique (Kaggle, 2022) was utilized in this study for converting text to numbers. Due to high number of missing data for well properties, an efficient technique was developed by pursuing a method of imputing missing data using LU (lower–upper) simulation based on triangular decomposition of the correlation (standardized covariance) matrix. Many theoretical developments of LU simulation have been pursued in geostatistics for geomodeling and spatial resampling (Davis, 1987; Deutsch and Journel, 1998; Journel and Bitanov, 2004; Khan and Deutsch, 2016; Rezvandehy and Deutsch, 2017). A modified LU conditional simulation (Davis, 1987) is suggested here for imputation by conditioning nonmissing values to impute missing data for each well. Section 4 provides an in-depth explanation on how to carry out the imputation technique. The same statistics and models for data processing of the training set should be applied to transform the validation set and test set as shown in dashed lines in Figure 2. The transformation includes normalization, text handling, and imputation.

Predictive algorithms were trained by the training set to achieve the most reliable classifier. *K*-fold cross-validation was utilized to get a clean prediction for the training set (to prevent overfitting): it splits the training set into *K*-folds and then predicts each fold using a model trained on the remaining folds. Evaluating a classifier is often more challenging than a regressor. The most common approach for

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

assessment is *Accuracy*, which is calculated by the number of true predicted over the total number of data. However, accuracy alone may not be practical for performance measurement of classifiers, especially in the case of skewed datasets. Accuracy should be considered along with other metrics. A confusion matrix is a much better way to evaluate the performance of a classifier (Géron, 2019). Although the confusion matrix represents a lot of information, sometimes more concise metrics (including accuracy) are preferred as: *Sensitivity*: the proportion of correct positive predictions to the total positive classes, *Precision*: the proportion of correct positive prediction to the total positive predicted values, and *Specificity*: true negative rate or the proportion of negatives that are correctly identified. For imbalanced class dataset, one metric is always significantly higher than another. Depending on the specific needs of the problem, the preference for higher precision or sensitivity may vary. In this study, it was deemed acceptable if the precision is relatively low, which means having noticeable false alerts of serious leakages while they are nonserious. However, we expect the classifier to have a high sensitivity that can detect serious leakages with a minimal number of false negatives. The hyperparameters for each algorithm are fine-tuned to enhance sensitivity. The trained model is evaluated with the validation set. This process is repeated to achieve reasonable sensitivity on the validation set (see Figure 2). If no classifier can reach a reliable sensitivity, class distribution can be adjusted by combination of oversampling and undersampling to reasonably increase sensitivity. Adjustment of class distribution was applied for different ratios: minority class divided by majority class. This process was repeated until a reliable class ratio is achieved based on validation set performance (Figure 2). Resampling should be applied with Ensemble Learning to build a stronger predictor. Section 5 provides an explanation of how to use resampling with ensemble learning to improve prediction performance. The final trained model was tested on never-before-seen data set to make sure the model can generalize well to new unseen data.

4. Imputation

LU conditional simulation (Davis, 1987) was modified to impute missing data by conditioning nonmissing values. This method not only preserves the correlation between features, but also accurately estimates the uncertainty of the imputed missing data. Additionally, it is more efficient and requires less computational resources for handling large datasets, in comparison to other methods. The procedure can be summarized in the following steps:

1. *Normal Score Transformation* (Deutsch and Journel, 1998). Quantile–quantile transformation is applied to convert distribution of each feature $\mathbf{z}$ to a Gaussian distribution with mean = 0 and standard deviation = 1, which is required for LU simulation.
2. *Correlation Matrix of Features.* A correlation matrix (standardized covariance matrix) $\boldsymbol{\rho}$ for n features is shown in Figure 4a. The diagonal elements of this correlation matrix $\boldsymbol{\rho_{11}}, \boldsymbol{\rho_{22}}, \ldots, \boldsymbol{\rho_{nn}}$ are 1 representing the correlation of each feature to itself.
3. *Cholesky Decomposition.* The correlation matrix is then discomposed by Cholesky decomposition as $\boldsymbol{\rho} = \mathbf{LU}$, where $\mathbf{L}$ is the lower triangular matrix with all elements above diagonal elements is 0, and $\mathbf{U}$ is upper triangular matrix with 0 values below diagonal elements. Only $\mathbf{L}$ is required for the LU simulation. Figure 4a shows the lower triangular matrix $\mathbf{L}$ achieved from Cholesky decomposition.
4. *Modified LU Conditional Simulation.* A vector of uncorrelated standard normal deviate $\mathbf{w}$ with mean = 0, standard deviation = 1 is simulated for each feature. The length of $\mathbf{w}$ for each feature is the number of data (here is the total number of wells for the training set). LU unconditional simulation can be simply calculated by $\mathbf{y} = \mathbf{Lw}$. Figure 4b shows how to generate a LU unconditional simulation achieving correlated Gaussian realization $\mathbf{y}$. The unconditional simulation can be used for oversampling to improve the imbalance number of classes for classification (see Section 5). However, conditional simulation is needed to simulate missing data conditioned based on nonmissing values. For conditioning nonmissing features, each array of $\mathbf{w}$ vector with known values needs to be converted to $\mathbf{w^c}$ that is a function of the nonmissing features. For example, if feature 1 z_1

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

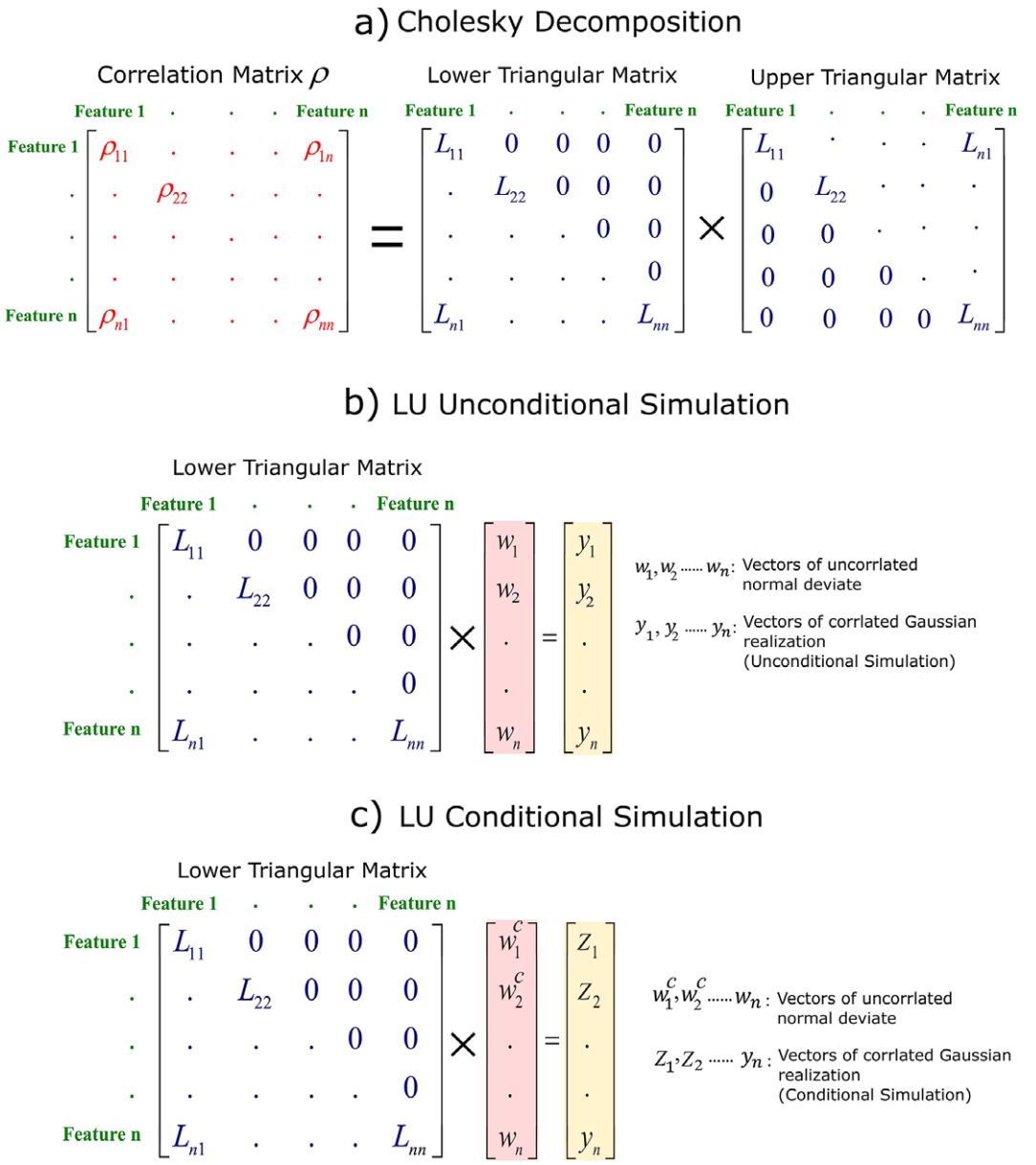

Figure 4. *(a) Cholesky decomposition of correlation matrix for n features (well properties). (b) LU unconditional simulation. (c) LU conditional simulation.*

and feature 2 z_2 are available, LU conditional simulation can be applied to keep z_1 and z_2 unchanged and simulate y_3 to y_n (missing data) conditioned based on z_1 and z_2 as shown in Figure 4c. This conditioning requires to convert w_1 and w_2 to w_1^c and w_2^c as follows:

$$w_1^c = \frac{z_1}{L_{11}}\ ,\ \ w_2^c = \frac{z_2 - L_{21} w_1^c}{L_{22}}, \tag{1}$$

where L_{11}, L_{21}, and L_{22} are elements of the lower triangular matrix **L** from Cholesky decomposition (Figure 4c). This process needs to be repeated to calculate all w^c for conditioning nonmissing features. The nth w is calculated as:

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

$$w_n^c = \frac{z_n - L_{nn-1} w_{n-1}^c}{L_{nn}}, \tag{2}$$

w for missing data should change for each feature and each instance (random sampling from Gaussian distribution) leading to quantification of the uncertainty in missing data after simulation. The main challenge for applying LU conditional simulation for imputation is the ordering of missing and nonmissing features for each row of data. If missing data are placed first followed by nonmissing values, the conditioning cannot be applied since w_1 for missing data are randomly sampled and then w_2^c for nonmissing values are calculated based on equation (1): **Lw** cannot enforce the correlation between simulated and nonmissing values. However, if nonmissing values are placed first followed by missing data, the conditioning will be properly applied because of calculating w_1^c for nonmissing values before w_2. Therefore, nonmissing values must be placed first followed by missing data for each instance (row of data). This requires reconstructing the correlation matrix for each instance to be consistent with the order of features. The ordering is not important within missing and nonmissing features. This is the contribution of this article to modify LU Conditional Simulation for applying correct imputation. Figure 5 shows how to change the order of features and correlation matrix based on nonmissing and missing data. There are four features and four rows. Figure 5b shows how to change the order of raw data in Figure 5a and the related correlation matrix for each row is shown in Figure 5c. All four features of row 1 have nonmissing values and therefore changing the order of features is not required. However, for rows 2–4, the order of features should be changed to start with nonmissing values. The correlation matrices should be consistent for each row of data. For the LU conditional simulation (Figure 4c), Cholesky decomposition must be calculated for each covariance matrix separately.

5. *Back-transform from Gaussian to Original Space.* The simulated values in Gaussian space must be back-transformed to original space. This requires to lookup through the standard Gaussian distribution to find the CDF (cumulative distribution function) probability (P) of each simulated

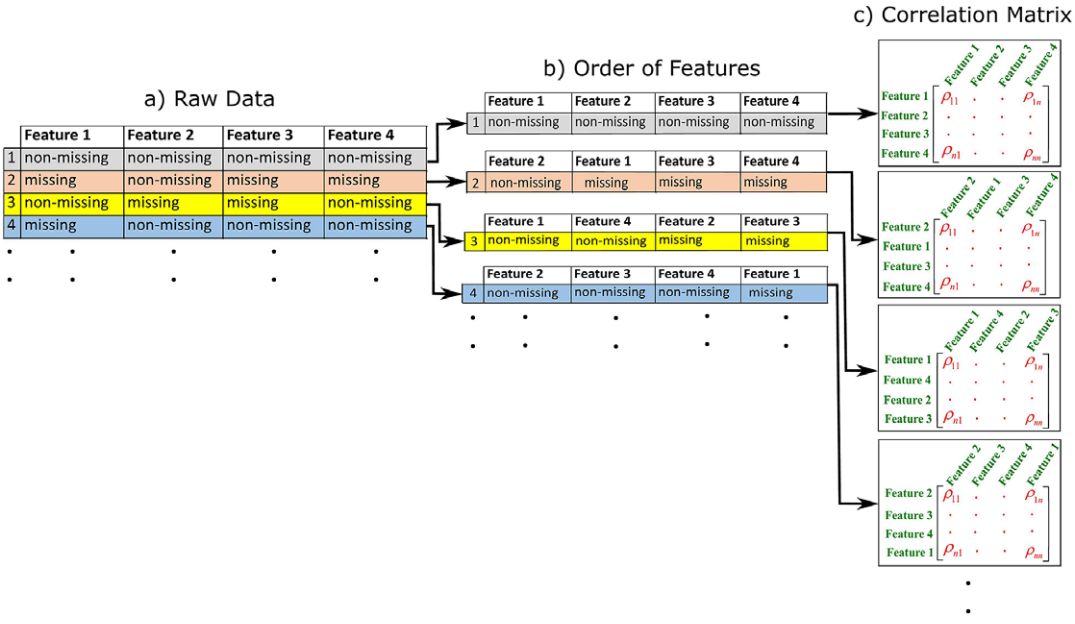

Figure 5. *(a) Schematic illustration of four features with four rows of data with missing and nonmissing values. (b) Change the order of features for raw data to have nonmissing values first followed by missing data. (c) Correlation matrix for each row in (b).*

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

a) Synthetic Data

	Feature 1	Feature 2	Feature 3	Feature 4
1	NaN	1.97	1.60	NaN
2	6.51	NaN	NaN	4.13
3	4.19	0.89	NaN	0.64
4	NaN	1.35	1.93	0.50
•	•	•	•	•
•	•	•	•	•
9998	5.72	NaN	NaN	0.77
9999	NaN	NaN	1.56	2.57
10000	3.32	0.62	3.68	0.13

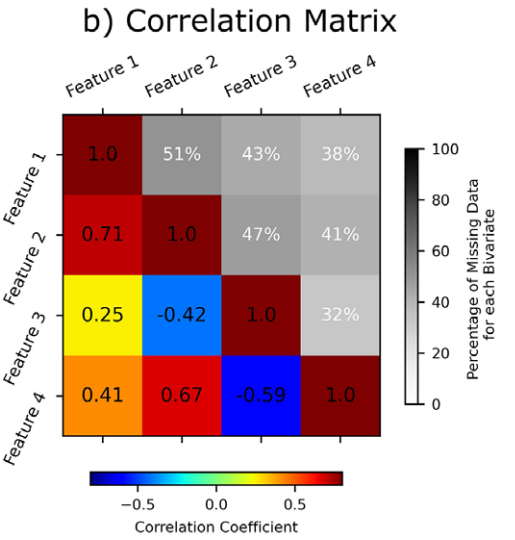

Figure 6. *(a) Synthetic example of four features with 10,000 data. NaNs are missing data. (b) Correlation matrix between features (below diagonal elements) and percentage of missing data for each bivariate feature (above diagonal elements). Maximum percentage of missing data is 51% for bivariate distribution of features 1 and 2.*

value. Then, lookup through the original distribution of related feature to find the P-quantile of the simulated value in the original space. This ensures that nonmissing values remain unchanged.

Steps 4 and 5 are repeated for all data in the training set to impute all missing data. Due to random sampling from the distribution of each feature, this approach quantifies the uncertainty in imputation of missing data by running the process described above many times (e.g., 100 times). Standard normal deviate **w** should be different for each run to simulate different values for missing data while keeping the nonmissing values unchanged and respecting the correlation between features. Moreover, implementation of the above outlined steps is straightforward especially with Python programming language. Since Cholesky decomposition of the correlation matrix should be applied only once per each unique order of features, the process is fast and efficient for big datasets.

To evaluate the efficiency of the proposed imputation technique a synthetic example is considered with four correlated features with 10,000 data as shown in Figure 6a. Features 1 and 2 are Gaussian and lognormal distributions, respectively while features 3 and 4 are triangular distributions with different statistics (mean and mode). Figure 6b shows the correlation matrix between features (below diagonal elements) and percentage of missing data for each bivariate feature (above diagonal elements). The highest percentage of missing data for bivariate distributions is between feature 1 and feature 2 (51%), and the lowest is between feature 3 and feature 4 (32%). Figure 7 shows a scatter plot matrix including histograms of each feature on diagonal elements before imputation (a) and after imputation (b). The correlation between features ($\rho_{x,y}$), the shape of univariate (histograms), and the bivariate distributions are reproduced after imputation. Therefore, the technique is highly suitable for imputation of realistic data.

5. Resampling and Ensemble Learning

Combining both oversampling and undersampling can lead to improved overall performance in comparison with performing one approach in isolation (Brownlee, 2021). Undersampling was considered for the instances with missing data. Although imputation was already applied, it is better to remove random instances that have imputed values instead of real nonmissing instances. We used LU unconditional

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

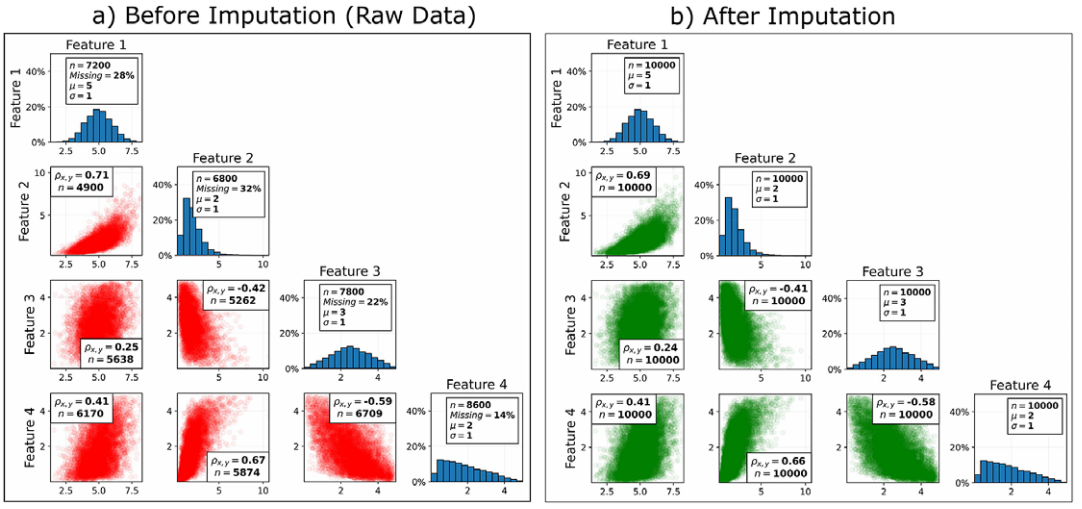

Figure 7. *Scatter plot matrix for synthetic example of four features with 10,000 data before imputation (a) and after imputation (b). Histograms of each feature are shown on diagonal elements.* n *is the number of nonmissing values for each univariate and bivariate distribution and* $\rho_{x,y}$ *is the correlation coefficient for each bivariate distribution.* μ *is the mean and* σ *is the standard deviation.*

simulation for oversampling to avoid the inclusion of exact duplicates as discussed in Section 4 and shown in Figure 4b. This approach is fast and includes associated uncertainty for resampling. Both resampling approaches were combined with equal percentage: undersampling with a selected percentage is applied to the majority class to reduce the bias on that class, while also applying the same percentage for oversampling of the minority class to improve the bias toward these instances. This was applied for different ratios of $\frac{\text{Class 1}}{\text{Class 0}}$, where Class 1 and Class 0 are the proportions of serious and nonserious leakage, respectively. For each ratio, the metrics were calculated. The most reliable ratio is the one that performs similarly across all metrics.

K-fold cross-validation was applied along with resampling to achieve clean prediction. However, using *K*-fold cross-validation on the transformed version of the training set may not be correct since the class distribution of the original data is different from the sampled training set. The correct approach is to resample within *K*-fold cross-validation (Santos et al., 2018). Figure 8 shows a schematic illustration of resampling within 5-fold cross-validation. The training set is divided into 5 stratified splits: the folds of each split have the same class distribution (percentage) of the original data. Resampling was applied on the training folds of each split. A model was trained on the resampled training folds. The test fold, which preserves the percentage of samples for each class in the original dataset, was predicted with the trained model. This process was repeated for all 5-splits that leads to five models. The trained models 1–5 obtained from 5-fold cross-validation (Figure 8) were applied separately to predict the entire validation set. Then, Ensemble Learning was applied by aggregating the predictions from each model to predict the class that gets the most votes (hard voting); since each model is trained on random subsets of training sets with random sampling, it is reasonable to aggregate the predictions. This process was repeated for all classifiers. This leads to achieve a number of good and promising predictors. To build a stronger predictor, soft voting was applied by integrating the promising classifiers to achieve a final trained model at the end.

6. Results and Discussion

Figure 9 shows a correlation matrix for 22 well properties of the AER classification (SCVF/GM test results), before imputation (Figure 9a) and after imputation (Figure 9b) for the training set. The correlations between features after imputation have been reproduced. An example is provided at the

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

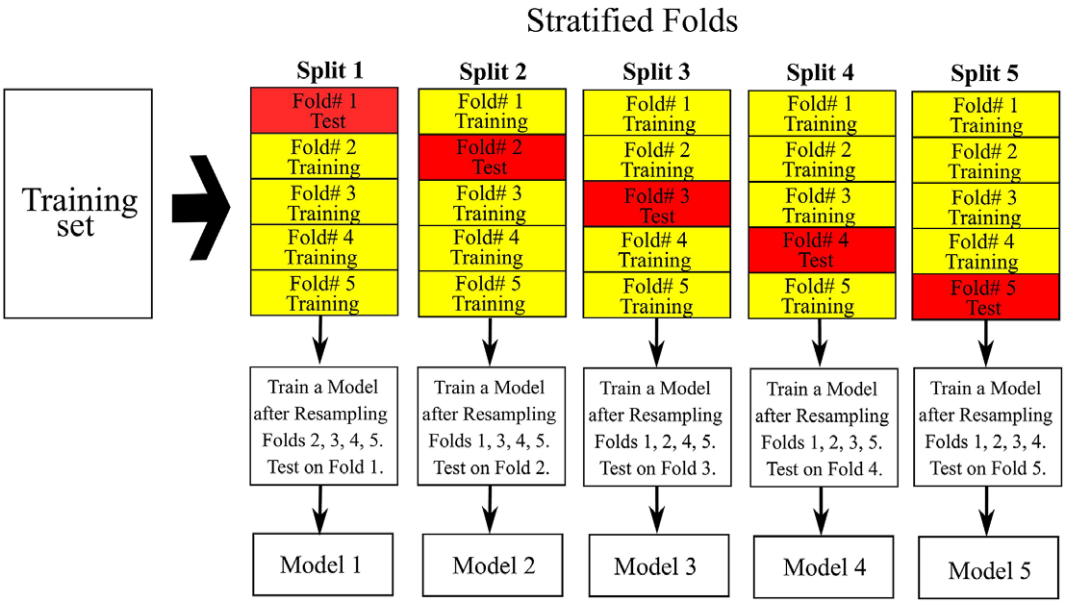

Figure 8. *Schematic illustration of resampling within 5-fold cross-validation that leads to five models (models 1–5). Resampling is applied only on the training folds. A model is trained for the resampled training folds of each split. The trained model is used to predict the test fold which preserves the percentage of samples for each class in the original data set.*

bottom of Figure 9 that shows the crossplot between surface-casing depth (m) and production casing depth before and after imputation: imputed data (stars) have the same correlation as nonmissing values (circles). The imputation can be repeated multiple times to quantify the associated uncertainty of imputed values. The same approach must be applied for imputation of missing data in validation and test sets; however, the correlation matrix and feature distribution of the training set must be used to prevent information leak into these datasets (Figure 2). The training data is subsequently ready to feed a machine learning algorithm for binary classification.

Figure 10 shows the performance of the predictive algorithms based on the four metrics achieved from the confusion matrix for the training set (a), validation set (b), and test set (c). For a sanity test, predictions from simple rule of thumb called Dummy Classifier were compared with the results from these algorithms (Pedregosa et al., 2011). The hyperparameters for each algorithm are fine-tuned to enhance performance. The trained models derived using the training set are applied for prediction of the validation set and test set to confirm that overfitting is not occurring. The comparison between Figure 10a–c shows that the performances are very similar except for Deep Neural Network having a lower precision for the test set indicating overfitting for this algorithm. Specificity is the highest and sensitivity is lowest for almost all classifiers. The Dummy classifier has the lowest values for all metrics except for sensitivity that is close to the predictive algorithms. A highly skewed distribution toward the negative class leads to have sensitivity close to that of the Dummy classifier even for powerful algorithms. In order to achieve better comparison of the classifiers, a tool called receiver operating characteristic (ROC) curve was used to measure performance. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity). Every point on the ROC curve represents a chosen cut-off even though it cannot be seen. The most common way to compare classifiers is to measure the area under the curve (AUC). A perfect classifier has an AUC equal to 1, whereas a purely random classifier has an AUC ≈ 0.5. Figure 11 shows ROC curves of classifiers for the training set (a), validation set (b), and test set (c). The AUC is shown for each algorithm. The Dummy classifier has the lowest AUC (≈ 0.5). Random Forest and Deep Neural Network have the highest AUC; however, due to overfitting, AUC for Deep Neural Network decreased

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

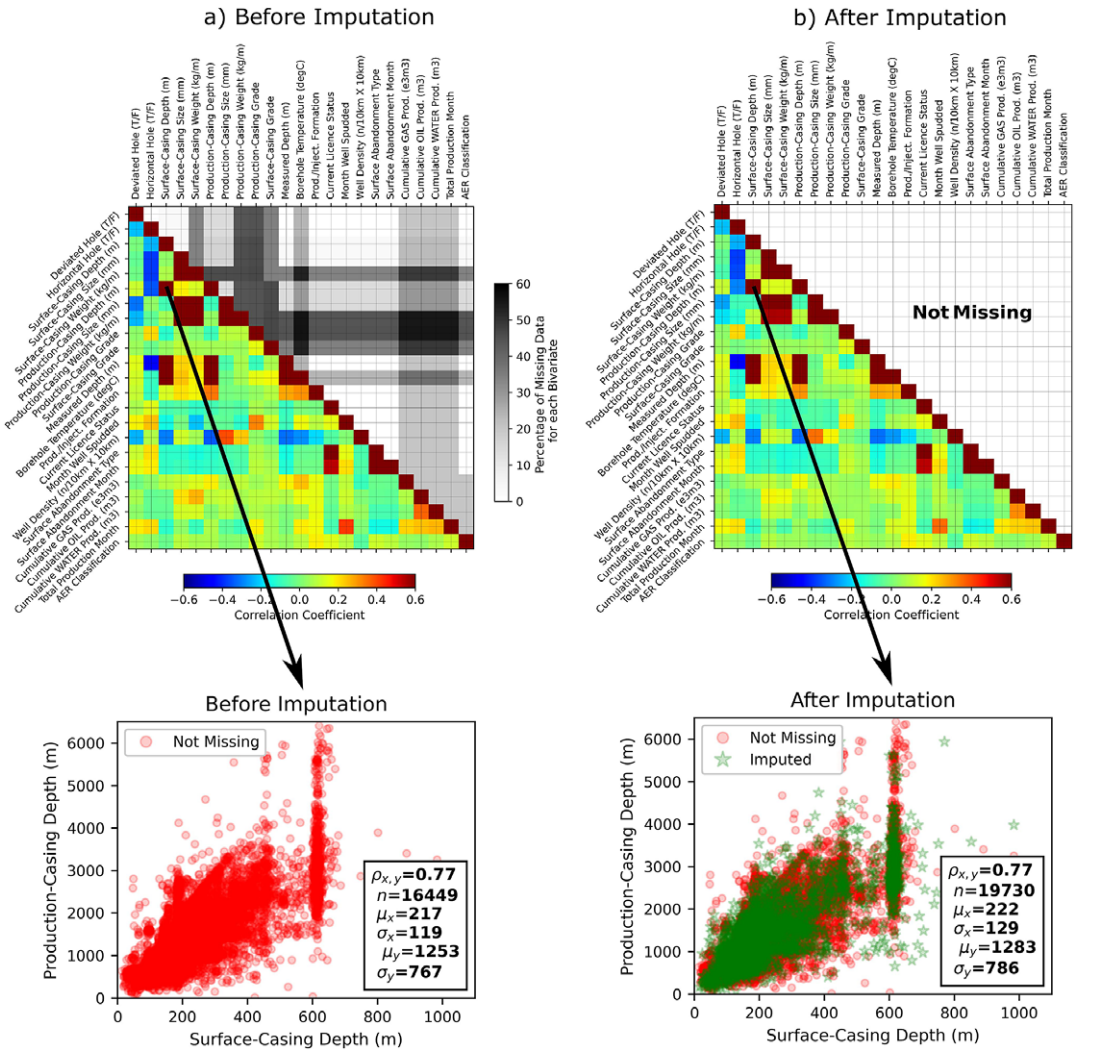

Figure 9. *Correlation matrix (below diagonal elements) for 22 well properties of an AER classification (SCVF/GM test results), before imputation (a) and after imputation (b). The percentage of missing data for bivariate distribution is shown above diagonal elements. Cross plots between surface-casing depth (m) and production-casing depth (m) before and after imputation are shown at the bottom. n is number of nonmissing values for each univariate and bivariate distribution and $\rho_{x,y}$ is the correlation coefficient for each bivariate distribution. μ is the mean and σ is the standard deviation.*

for the test set. Therefore, it is concluded that the Random Forest is the most reliable classifier with AUC ≈ 0.75. Using the above-described approach, the importance of each feature (Table 1) to predict serious leakage was determined by 100 realizations of Random Forest with 100 realizations of imputation, one realization at a time to quantify the importance of each feature with full picture of uncertainty. The results are summarized in Figure 12. Each bar shows the mean percentage of importance with uncertainty interval (variance). Month Well Spudded (e.g., the age of the wells in months) has the highest importance to predict fluid leakage followed by six features that have similar performance: Total Production Month, Surface-Casing Depth, Production Casing Depth, Measured Depth, Cumulative WATER Prod and Cumulative GAS Prod. The features do not have strong linear correlations with the target (see Figure 9 for linear correlation of the well properties with AER classification). However, increasing Month Well

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

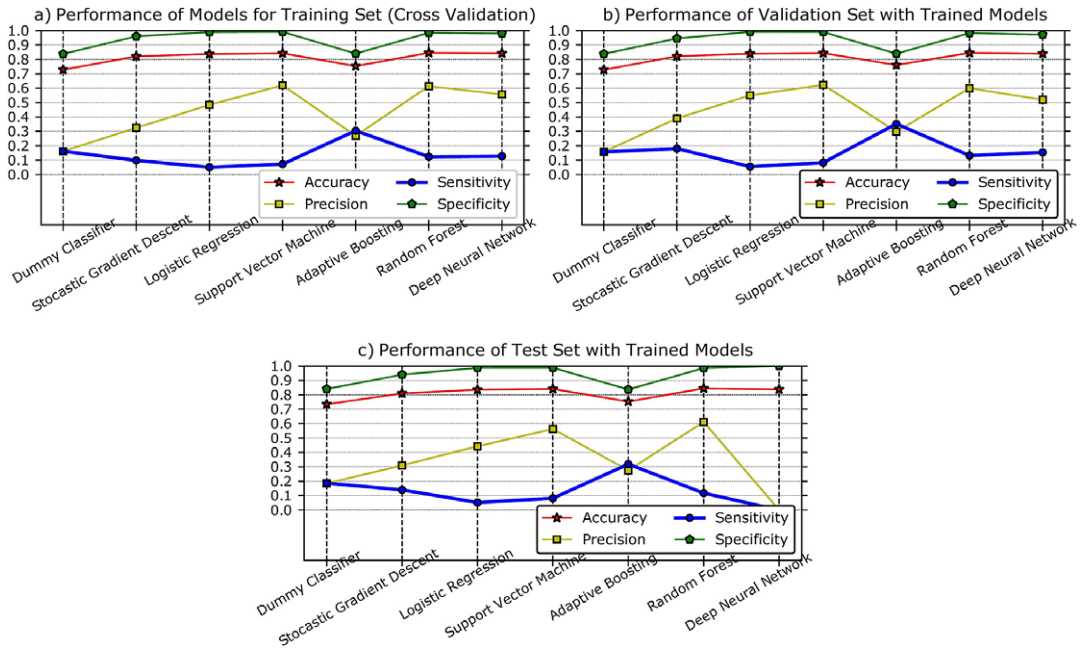

Figure 10. *Performance of predictive algorithms for training set (a), validation set (b), and test set (c) based on the metrics accuracy, sensitivity, precision, and specificity achieved from confusion matrix. Specificity is the highest and sensitivity is lowest for almost all classifiers.*

Spudded, Total Production Month, Measured Depth, Cumulative GAS Prod, and Cumulative WATER Prod probably lead to more serious fluid leakages according to the positive correlation with the target in Figure 9. It is interesting to note that the features Deviated Hole and Horizontal Hole have very low importance on impacting serious leakage. Furthermore, the size of production and surface casing, and the surface abandonment type appear to have negligible influence on the likelihood of serious fluid leakage.

The trained Random Forest model could be applied to predict the probability of serious fluid leakage for the wells without SCVF/GM field tests. The predictions will have high accuracy (>0.8) and specificity (>0.95), relatively low precision (0.6), but sensitivity will not be reliable: similar to prediction of Dummy Classifier (see Figure 10) which is due to imbalanced class. Therefore, we combined under-sampling and oversampling to increase sensitivity by adjusting class distribution. The resampling was applied for different ratios of $\frac{\text{Class 1}}{\text{Class 0}}$, where Class 1 is proportions of serious leakage and Class 0 is the proportion of nonserious leakage. Increasing this ratio may lead to an increase in sensitivity but a decrease in other metrics. The ratio that has similar performance for the metrics must be the most reliable ratio. Random Forest was applied for prediction since it has the highest performance (Figures 10 and 11). *K*-fold cross-validation can be applied on the transformed version of the training set. Figure 13a shows a 5-fold cross-validation after resampling the training set for 22 ratios of $\frac{\text{Class 1}}{\text{Class 0}}$ from 0.25 to 1.95. By increasing the ratio, AUC increases; sensitivity increases significantly; specificity decreases but accuracy almost remains unchanged. The metrics accuracy, specificity, and sensitivity have equal performance (0.8) for the ratio of 1.02.

However, *K*-fold cross-validation after resampling incorrectly leads to overoptimism and overfitting since the class distribution of the original data is different from the sampled training set (Santos et al., 2018). The correct approach is to resample within *K*-fold cross-validation as discussed in Section 5. Figure 13b shows resampling within 5-folds cross-validation for the ratios of $\frac{\text{Class 1}}{\text{Class 0}}$. Compared with Figure 13a, AUCs have decreased significantly which confirms overoptimism and overfitting for applying *K*-fold cross-validation after resampling. The ratio 1.27 in Figure 13b is the point where the

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

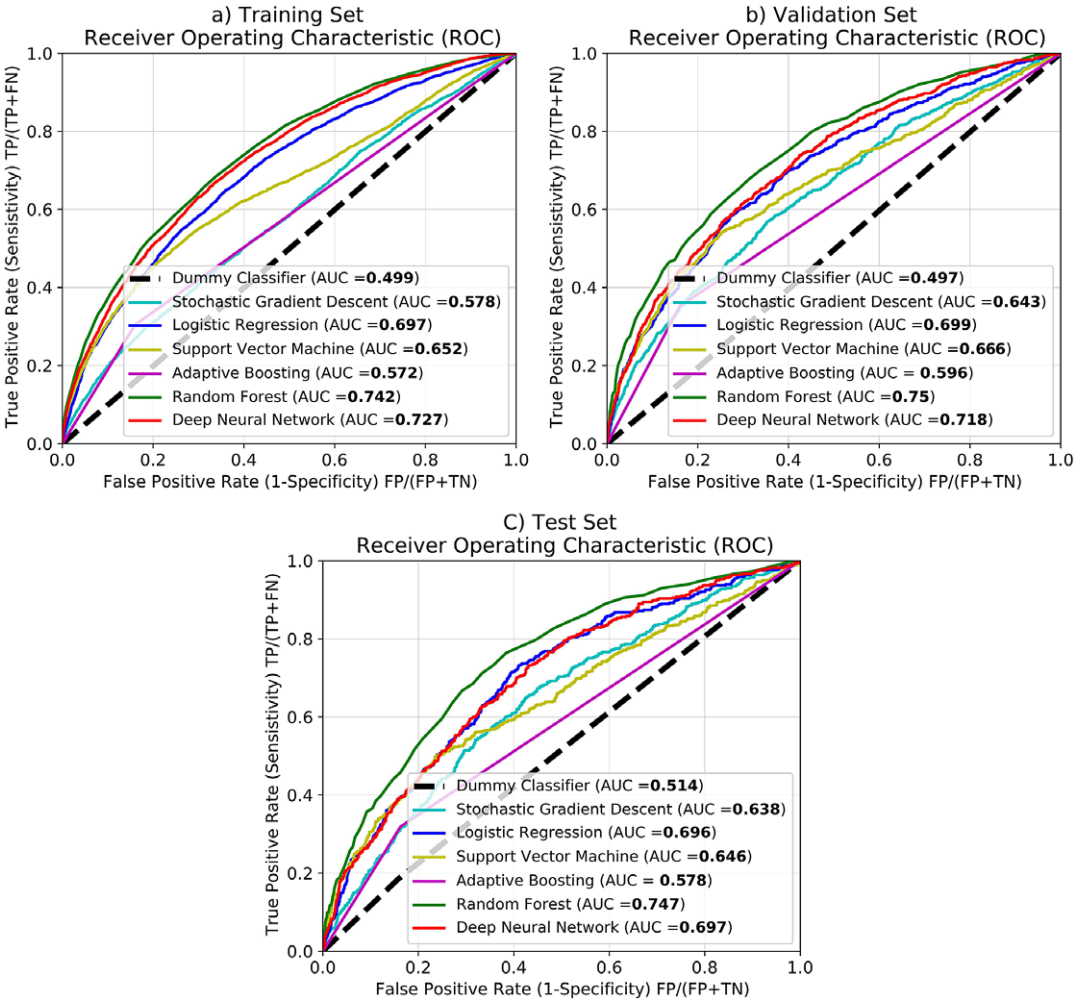

Figure 11. *ROC curves with calculated AUC for training set (a), validation set (b), and test set (c). TP, true positive; TN true negative; FP, false positive; FN, false negative. Random Forest has the highest AUC without overfitting.*

line of all metrics cross: the model performs almost equally for two classes. Therefore, the ratio 1.27, which signifies there are more wells with serious leakage in the training data, is selected as the ratio of two classes to achieve a more reliable sensitivity. The models 1–5, which were obtained through 5-fold cross-validation (Figure 8), were individually used to make predictions on the validation set and test data. The predictions made by each model were then combined and the class with the most votes was chosen as the final prediction. This process was repeated for other classifiers to compare the performances with Random Forest. Figure 14 shows the performance of the classifiers for the validation set (a) and test set (b) for the resampling ratio of $\frac{\text{Class 1}}{\text{Class 0}} = 1.27$ within 5-fold cross-validation. All classifiers for validation and test set have a higher performance than the Dummy Classifier. Due to resampling, sensitivity has significantly increased (compared with Figure 10). This leads to remarkable reduction for false negative predictions of wells with serious leakage. The performance for the test set is a little lower than for the validation set. This may be related to minor overfitting in the validation set while fine-tuning hyperparameters of classifiers. Predicting the probability of serious fluid leakage with known well properties (Table 1) for the wells without field test in Alberta will likely have similar performances as Figure 14b. Random Forest has the

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

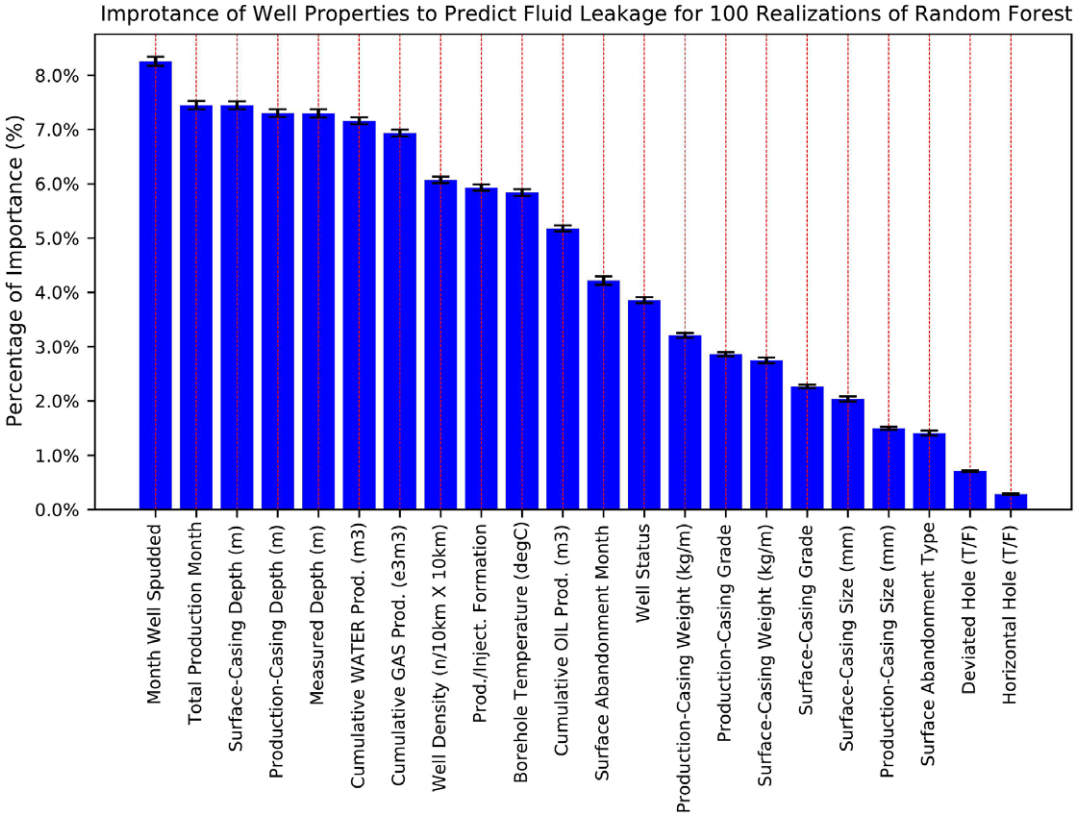

Figure 12. *Hundred realizations of Random Forest with 100 realizations of imputation, one realization at a time to quantify the importance of each feature with uncertainty for predicting target (AER classification). The feature Month Well Spudded (age of the wells in months) has the highest importance; Deviated Hole and Horizontal Hole have the lowest importance.*

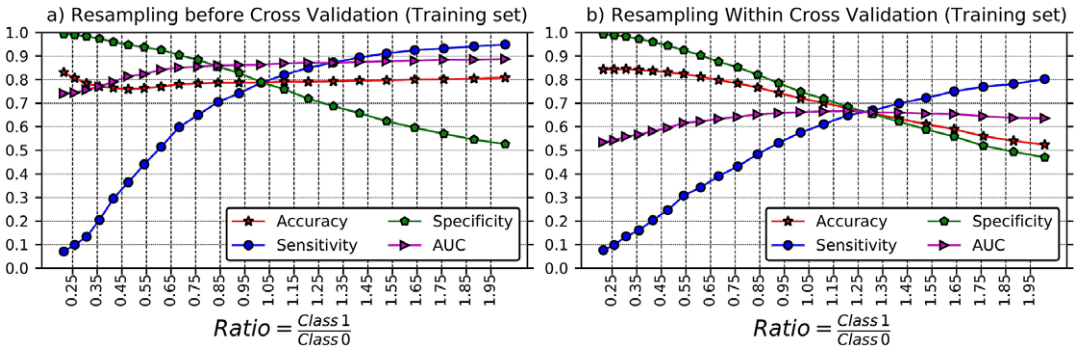

Figure 13. *Resampling before (a) and within (b) 5-fold cross-validation by Random Forest algorithm for training set for the ratios of* $\frac{\text{Class 1}}{\text{Class 0}}$*. Resampling before cross-validation (a) is incorrect due to overoptimism and overfitting. The metrics are accuracy, specificity, sensitivity, and AUC (area under the curve).*

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

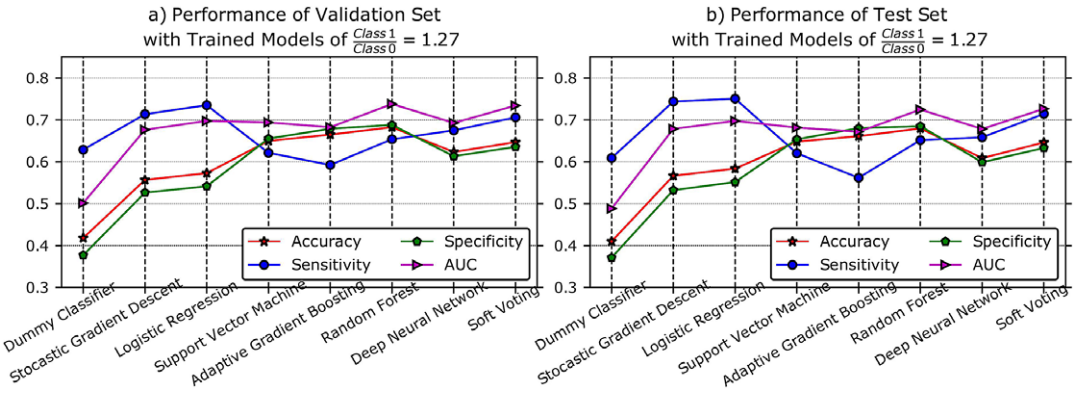

Figure 14. *(a) Performance of the classifiers for validation set (a) and test set (b) for the sampling ratio of* $\frac{\text{Class 1}}{\text{Class 0}} = 1.27$ *within 5-fold cross-validation. All classifiers for validation and test set have reasonably higher performance than Dummy Classifier. There is small overfitting in the validation set because of fine-tuning of hyperparameters. Soft voting, integration of Logistic Regression and Random Forest, is the most reliable classifier with high sensitivity and AUC (area under the curve), and reasonable accuracy and specificity.*

highest AUC but sensitivity is low. Logistic Regression has the highest sensitivity, but accuracy and specificity are relatively low compared with other algorithms. Soft Voting was applied at the end to integrate Random Forest and Logistic Regression by averaging the probability of each class and predict a class with the highest probability. Soft Voting has high sensitivity and AUC, and reasonable accuracy and specificity. Therefore, it is the most reliable classifier to predict the probability of serious fluid leakage for hydrocarbon wells without SCVF/GM field tests.

The specific value of the sampling ratio used in this manuscript, 1.27, was determined through experimentation on the given data set. Therefore, if the data set were to change, it would be necessary to reevaluate and potentially adjust the sampling ratio accordingly. Similarly, the combination of classifiers chosen for ensemble learning, specifically the use of Random Forest and Logistic Regression for soft voting, was based on the characteristics of the current data set. It is important to note that different data sets may require different combinations of sampling ratio and classifiers to achieve optimal performance. However, the developed framework described in this article can be applied for any producing oil and gas field to train an optimum classifier for leakage detection.

To apply the trained classifier for leakage detection, 1,000 hydrocarbon wells without field tests were randomly selected in Alberta, Canada. Figure 15a shows the location map of the wells. The 22 physical properties of the wells (Table 1) used for training the classifier were utilized to predict the probability of serious fluid leakage for each well. Figure 15b shows the percentage of missing data for the well properties. The distribution of the training set for each well property should be applied for normalization, text handling, and imputation of these 1,000 wells. Since some properties have more than 40% missing data, the uncertainty of imputed values should be incorporated in the final prediction. This requires running developed LU conditional simulation for multiple realizations: each realization gives different imputed values. Hundred realizations of imputation were generated. The trained classifier was applied 100 times using one realization of imputation at a time. This results in quantification of the uncertainty for the predicted probability of serious fluid leakage. The 100 predicted probabilities for each well can be aggregated by mean, median or 25th, 75th percentiles for decision-making. Figure 16a shows the calculated mean of probability for serious fluid leakage of the 100 realizations for each well. The histogram of predicted probabilities for two wells (with means of probability for serious fluid leakage of 0.42 and 0.70) in southern Alberta is shown to represent that each well has 100 predicted probabilities achieved by running the classifier with 100 realizations of imputed values. Figure 16b,c shows the

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

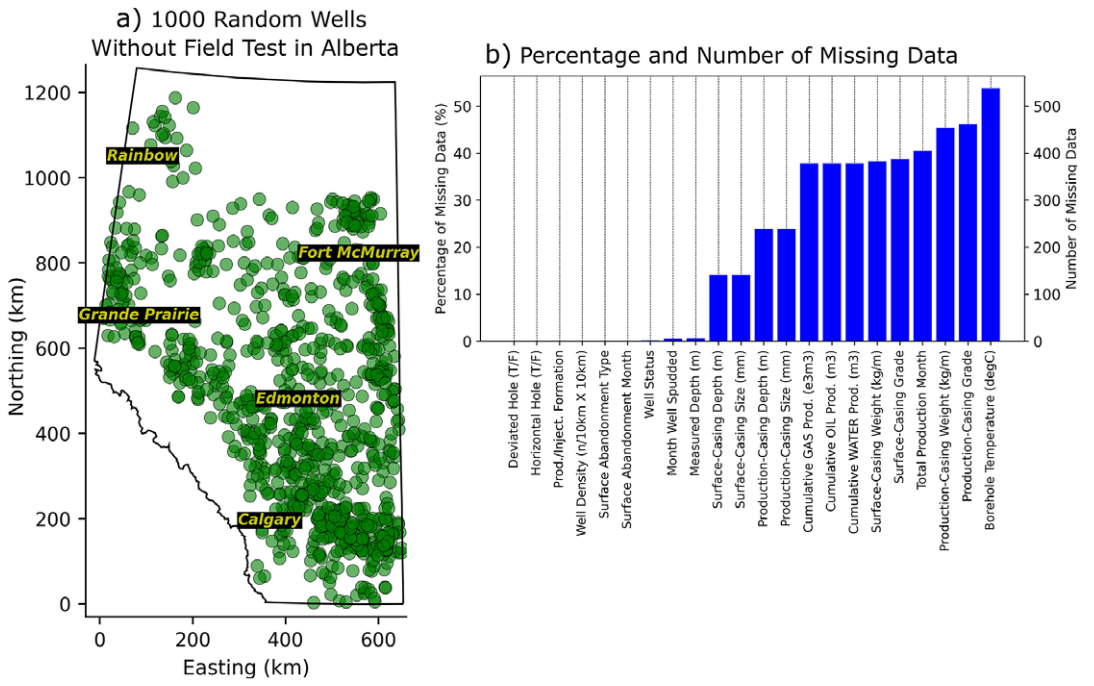

Figure 15. *(a) Location map of 1,000 random wells without field test in Alberta, Canada. (b) Bar chart of missing values of the wells in (a) for 22 well properties retrieved from the geoSCOUT database.*

location map of the wells that have the mean of probability of serious fluid leakage higher than 0.6 and 0.8 in Figure 16a, respectively. There are 136 wells with the probability higher than 0.8 probably representing the leakiest wells for this random selection of 1,000 wells. Therefore, field test operations can be optimized by targeting these 136 wells in Figure 16c first, followed by test for the rest of the wells shown in Figure 16b. To optimize field test procedures, the wells with probabilities of less than 0.4 should not be prioritized for field testing since they may not have serious leakage. There are over 300,000 hydrocarbon wells (Watson and Bachu, 2009) in Alberta. The described approach can be utilized to most efficiently detect serious fluid leakages for previously nontested wells in Alberta.

7. Conclusion

This article describes the development of a framework enabling the prediction of serious fluid leakage from hydrocarbon wells. A wide range of machine learning algorithms was trained based on 22 physical properties from 19,730 wells, evaluated by 4,933 wells (validation set), and finally tested by 2,741 wells (test set). A new imputation technique with low computational cost was developed using a modified LU conditional simulation to overcome the challenge of frequently missing data for selected well properties. Using this approach, the correlations of features were preserved after imputation. Due to random sampling from the distribution of well properties, the associated uncertainty in the imputation of missing data was quantified. Random Forest was found to have the highest performance among all classifiers. The importance of well proprieties to predict serious fluid leakage was quantified based on 100 realizations of Random Forest: each run used different imputed values to incorporate the uncertainty of imputation for feature importance measurement. Month Well Spudded (e.g., the age of the wells) was found to have the highest importance for potential fluid leakage followed by Total Production Month, Surface-Casing Depth, Production Casing Depth, Measured Depth, Cumulative WATER Prod and Cumulative GAS Prod. The well properties do not have a strong positive linear correlation with the target (AER

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

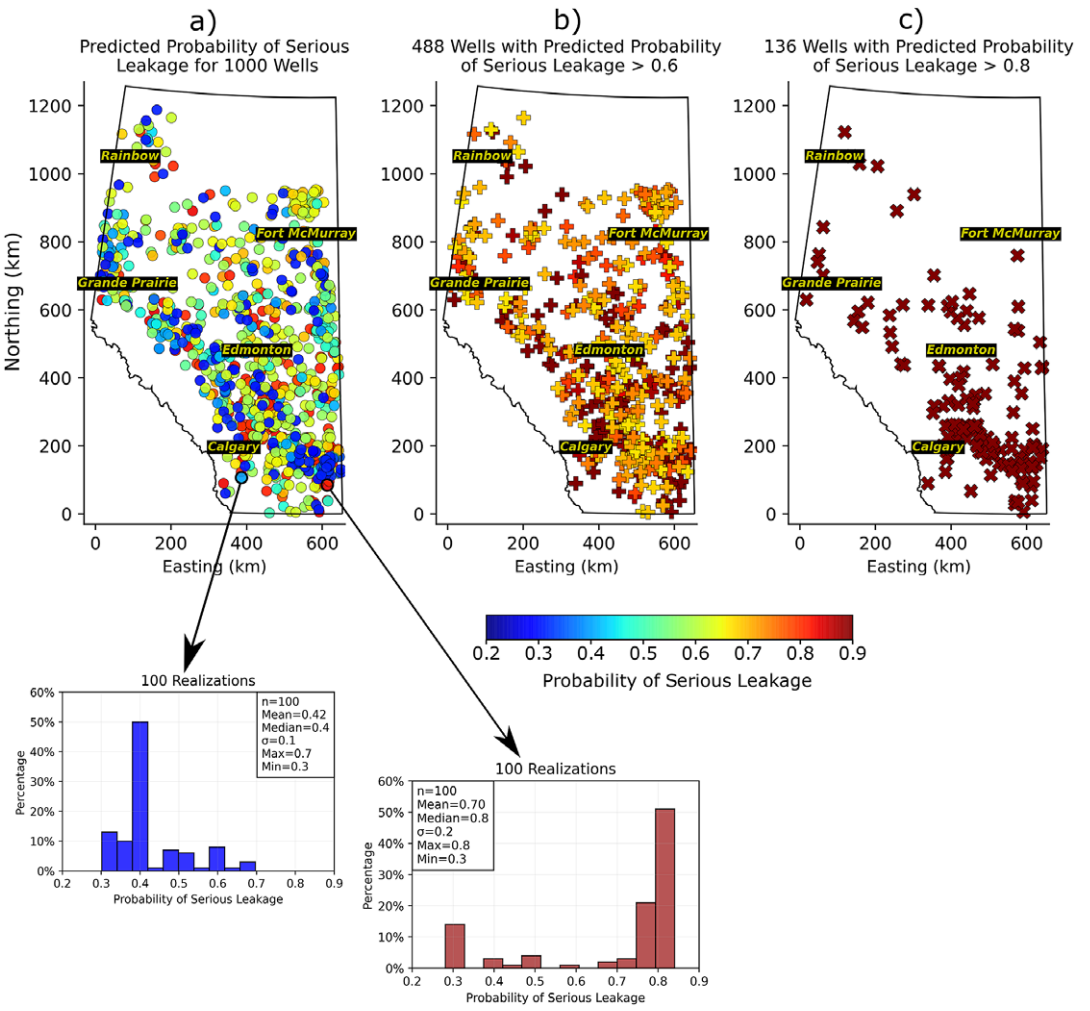

Figure 16. *(a) Mean of 100 predicted probabilities of serious fluid leakage for 1,000 wells. (b) Location map for 488 wells with the probability higher than 0.6 in (a). (c) Location map for 136 wells with the probability higher than 0.8 in (a).*

classification). However, the presented results suggest that the oldest and deepest wells with high production are most likely to have serious fluid leakage issues. It is important to note that Deviated Hole and Horizontal Hole have almost no impact on serious fluid leakage. A challenge is the imbalanced number of classes that leads to low sensitivity forcing predictive models to classify more majority class (nonserious leakage). A combination of undersampling and oversampling techniques was considered to adjust the class distribution of training data and feed more balanced data to predictive model. Under-sampling was randomly applied for the instances with missing well properties. LU unconditional simulation was utilized for oversampling because of reduced computational cost and quantifying uncertainty. *K*-fold cross-validation after resampling resulted in severe overoptimism and overfitting. The correct approach was to resample the training folds of each split for *K*-fold cross-validation to train separate models. The most reliable class ratio was chosen at the point where the metrics have similar performance. The predictions from each trained model on validation and test sets were aggregated by predicting the class that gets the most votes. Model training with more balanced class data reduces false negative estimation and increases the sensitivity. The integration of Random Forest and Logistic

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

Regression as Soft Voting leads to a more reliable prediction. This classifier can be used to detect serious fluid leakage for oil and gas wells in Alberta. However, there is no one-size-fits-all solution and the optimal combination may vary for different data sets. It is crucial to keep in mind that optimal performance can only be achieved by experimenting with various combinations of sampling ratio and classifiers, as different data sets have different characteristics. The method outlined in this article can be utilized to train a robust classifier for any oil and gas production field. This should be helpful for the development of cost-effective field testing approaches that result in environmental advantages by identifying and prioritizing amendment of the leakiest wells.

Acknowledgments. We would like to thank the executive editor and two anonymous reviewers whose valuable suggestions and comments lead to improve the quality of this manuscript.

Author contribution. Conceptualization: M.R.; Data curation: M.R.; Data visualization: M.R.; Formal analysis: M.R.; Funding acquisition: B.M.; Methodology: M.R.; Programming: M.R.; Review and editing: B.M.; Supervision: B.M.; Writing—original draft: M.R. All authors approved the final submitted draft.

Competing interest. The authors declare no competing interests exist.

Data availability statement. The test results of surface casing vent flow (SCVF) and gas migration (GM) for energy wells in Alberta, Canada was retrieved from https://www2.aer.ca/t/Production/views/COM-VentFlowGasMigrationReport/VentFlowGasMigrationReport.csv. Well properties for each well was retrieved from geoSCOUT (2022) software.

Funding statement. This research was supported by Canada First Research Excellence Fund (CFREF) and Global Research Initiative in Sustainable Low Carbon Unconventional Resources (GRI).

References

Abboud J, **Watson T and Ryan M** (2021) Fugitive methane gas migration around Alberta's petroleum wells. *Greenhouse Gases: Science and Technology 11*(1), 37–51.

Alberta Energy Regulator (2003) Interim directive: ID 2003–01. Available at www.aer.ca/documents/ids/id2003-01.pdf (accessed 14 April 2023).

Alberta Energy Regulator (2022) Vent flow/gas migration report. Available at https://www.aer.ca/providing-information/data-and-reports/activity-and-data/general-well-data (accessed 14 April 2023).

Brandt AR, **Heath G**, **Kort E**, **O'sullivan F**, **Pétron G**, **Jordaan S**, **Tans P**, **Wilcox J**, **Gopstein A**, **Arent D**, **Wofsy S**, **Brown NJ**, **Bradley R**, **Stucky GD**, **Eardley D and Harriss R** (2014) Methane leaks from North American natural gas systems. *Science 343* (6172), 733–735.

Brownlee J (2020) Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning. *Machine Learning Mastery*.

Brownlee J (2021) Retrieved from machine learning mastery. Available at https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/ (accessed 14 April 2023).

Cahill AG and Samano PSG (2022) Prioritizing stewardship of decommissioned onshore oil and gas wells in the United Kingdom based on risk factors associated with potential long-term integrity. *International Journal of Greenhouse Gas Control 114*, 103560.

DataWing (2022) "Welcome to DataWig's documentation!" Available at https://datawig.readthedocs.io/en/latest/ (accessed 14 April 2023)

Davis MW (1987) Production of conditional simulations via the LU triangular decomposition of the covariance matrix. *Mathematical Geology 19*(2), 91–98.

Deutsch CV and Journel AG (1998) *GSLIB: Geostatistical Software Library and User's Guide*, 2nd Edn. Oxford: Oxford University Press.

Fernández A, **Garca S**, **Galar M**, **Prati RC**, **Krawczyk B and Herrera F** (2018) *Learning from Imbalanced Data Sets*, vol. 10. Cham: Springer.

Fleming N, **Morais T**, **Mayer K and Ryan MC** (2021) Spatiotemporal variability of fugitive gas migration emissions around a petroleum well. *Atmospheric Pollution Research 12*(6), 101094.

geoSCOUT (2022) "Visualize, analyze, and forecast using an extensive library of premium data" Available at https://www.geologic.com/products/geoscout/ (accessed 14 April 2023)

Géron A (2019) *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, CA: O'Reilly Media.

He H and Ma Y (2013) *Imbalanced Learning: Foundations, Algorithms, and Applications*, 1st edn. Wiley-IEEE Press, New York.

https://doi.org/10.1017/dce.2023.9 Published online by Cambridge University Press

Iyer J, **Lackey G**, **Edvardsen L**, **Bean A**, **Carroll SA**, **Huerta N**, **Smith MM**, **Torsæter M**, **Dilmore RM and Cerasi P** (2022) A review of well integrity based on field experience at carbon utilization and storage sites. *International Journal of Greenhouse Gas Control 113*, 103533.

Journel AG and Bitanov A (2004) Uncertainty in N/G ratio in early reservoir development. *Journal of Petroleum Science and Engineering 44*(1), 115–130.

Kaggle (2022) "Target Encoding" Available at https://www.kaggle.com/ryanholbrook/target-encoding (accessed 14 April 2023).

Kang M, **Kanno CM**, **Reid MC**, **Zhang X**, **Mauzerall DL**, **Celia MA**, **Chen Y and Onstott TC** (2014) Direct measurements of methane emissions from abandoned oil and gas wells in Pennsylvania. *Proceedings of the National Academy of Sciences 111*(51), 18173–18177.

Khan KD and Deutsch CV (2016) Practical incorporation of multivariate parameter uncertainty in Geostatistical resource modeling. *Natural Resources Research 25*(1), 51–70.

Montague JA, **Pinder GF and Watson TL** (2018) Predicting gas migration through existing oil and gas wells. *Environmental Geosciences 25*(4), 121–132.

Pedregosa F, **Varoquaux G**, **Gramfort A**, **Michel V**, **Thirion B**, **Grisel O**, **Blondel M**, **Prettenhofer P**, **Weiss R**, **Dubourg V**, **Vanderplas J**, **Passos A**, **Cournapeau D**, **Brucher M**, **Perrot M and Duchesnay É** (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12*, 2825–2830.

Rezvandehy M and Deutsch CV (2017) Horizontal variogram inference in the presence of widely spaced well data. *Petroleum Geoscience 24*(2), 219–235.

Sandl E, **Cahill A**, **Welch L and Beckie R** (2021) Characterizing oil and gas wells with fugitive gas migration through Bayesian multilevel logistic regression. *Science of the Total Environment 769*, 144678.

Santos MS, **Soares JP**, **Abreu PH**, **Araujo H and Santos J** (2018) Cross-validation for imbalanced datasets: Avoiding over-optimistic and overfitting approaches [research frontier]. *IEEE Computational Intelligence Magazine 13*(4), 59–76.

Shindell DT, **Faluvegi G**, **Koch DM**, **Schmidt GA**, **Unger N and Bauer SE** (2009) Improved attribution of climate forcing to emissions. *Science 326*(5953), 716–718.

van Buuren Sv and Groothuis-Oudshoorn K (2010) MICE: Multivariate imputation by chained equations in *r*. *Journal of Statistical Software 45*, 1–68.

Watson TL and Bachu S (2009) Evaluation of the potential for gas and CO_2 leakage along wellbores. *SPE Drilling & Completion 24*(01), 115–126.

Wisen J, **Chesnaux R**, **Werring J**, **Wendling G**, **Baudron P and Barbecot F** (2020) A portrait of wellbore leakage in northeastern British Columbia, Canada. *Proceedings of the National Academy of Sciences 117*(2), 913–922.

Cite this article: Rezvandehy M and Mayer B (2023). Machine learning approaches for the prediction of serious fluid leakage from hydrocarbon wells. *Data-Centric Engineering*, 4, e12. doi:10.1017/dce.2023.9