International Journal of Engineering & Technology, 10 (2) (2021) 124-133

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Data mining based investigation of the impact of imbalanced dataset over fractured zone detection

Haleh Azizi [1] *, Hassan Reza [1]

[1] *School of Electrical Engineering and Computer Sciences, University of North Dakota, Grand Forks, ND, 58201, USA*
**Corresponding author E-mail: haleh.azizi@UND.edu*

Abstract

Several studies have been conducted in recent years to discriminate between fractured (FZs) and non-fractured zones (NFZs) in oil wells. These studies have applied data mining techniques to petrophysical logs (PLs) with generally valuable results; however, identifying fractured and non-fractured zones is difficult because imbalanced data is not treated as balanced data during analysis. We studied the importance of using balanced data to detect fractured zones using PLs. We used Random-Forest and Support Vector Machine classifiers on eight oil wells drilled into a fractured carbonite reservoir to study PLs with imbalanced and balanced datasets, then validated our results with image logs. A significant difference between accuracy and precision indicates imbalanced data with fractured zones categorized as the minor class. The results indicated that the accuracy of imbalanced and balanced datasets is similar, but precision is significantly improved by balancing, regardless of how low or high the calculated indices might be.

Keywords: *Accuracy; Classifier; Fractured Reservoirs; Random Forest; Support Vector Machine.*

1. Introduction

Many classification techniques have been developed based on the assumption that balanced and imbalanced class datasets are equal. Evaluating imbalanced classification is required to determine the positive and negative accuracy of the classes. The need to define classification procedures to determine which are suitable for imbalanced class detection is critical, especially in the case of the smaller class, or minority, recognition. We have focused our work on the fractured zones, or minor classes, since current algorithms ignore the imbalanced datasets.
The issue with imbalanced classification has been studied recently by Kotsiantis et al. [1]. Oil spillage detection using satellite images [2], protein sequence detection [3], face recognition [4], medical diagnosis [5], fault detection [6], anomaly detection [7], and text classification [8] are all examples of imbalance classification research as approached by different disciplines. Sun et al. and Ali et al. thoroughly reviewed the existing literature on imbalanced data classification and described the significant drawbacks associated with the classification of data using this class distribution [9, 10]. These authors examined application domains, the nature of specific problems, learning difficulties, learning objectives, evaluation measures, solutions, and multiple class recognition, enforcing the importance of using imbalanced problems in different disciplines. Ali and associates also reviewed issues related to the application of machine learning techniques with imbalanced class data sets. They discussed recent research trends and discovered many real-world applications for these advancements, especially in medicine and social media.
Ouyang et al. studied 34 classifiers to detect oil spills using imbalanced learning methods. The authors reported that the selected methods returned results with the highest accuracy; however, their approach may not be reliable [11].
Johnson and Khoshgoftar examined existing deep learning techniques to address imbalanced class data [12]. The authors mentioned that real-word classification applications based on imbalanced data had been a challenge when examining fraud and cancer detection. They studied available reports regarding class imbalance and deep learning to show the efficacy of deep learning classifiers when applied to imbalanced data. The authors discussed the strengths and weaknesses of the studied journal articles, resulting in a determination that most current work focuses on computer vision tasks, which are significantly affected by big data. Their survey concludes that various gaps in deep learning using imbalanced data needs to be studied further.
Limited investigations have been conducted to signify the effect of imbalanced data in geoscience and oil and gas. Klyuchnikov et al. proposed a classification approach to identify rock type at the drilling bit [13]. The authors studied samples consisting of approximately 13.5% of shale and 86.5% sand, which created highly imbalanced classes. They utilized three classifiers, Logistic Regression, Neural Networks, and Gradient Boosting on Decision Trees, to overcome the challenges posed by imbalanced data; however, none of the studied methods performed adequately. Pirizadeh et al. discussed various Enhanced Oil Recovery (EOR) methods developed for extracting oil from heavy and extra-heavy reservoirs [14]. The authors also investigated the effectiveness of EOR methods categorized by reservoir type. They proposed an ensemble learning-based approach to overcome class imbalance at the algorithmic level instead of the data level, which effectively improved accuracy by 1.5%. The most valuable research in the oil industry was completed by Brownlee [15], who discussed the development of a classifier for the detection of oil spills or slicks in satellite images. He developed Python scripts and has utilized

Copyright © Haleh Azizi, Hassan Reza. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Logistic Regression to apply the classifier to databases with different preparations. The author determined that the geometric mean (G-mean) calculated using a balanced dataset was slightly higher than in an imbalanced dataset in all classes.
Literature reviews indicate that while valuable information was attained with studies completed on fractured zone detection using PLs and data mining approaches, all ignored the effects of imbalanced datasets. Fracture detection by applying wavelet transformations on PLs [16], [17] and especially porosity logs [18] are examples of fracture detection approaches. Daiguji et al. and Behrens et al. applied data mining techniques on seismic data for fault cognition [19], [20]. The problems associated with these studies is that the seismic attribute resolution is not suitable for fractured zone detection.
Martinez-Torres created a composite fracture log by integrating caliper, gamma-ray, sonic, self-potential, and resistivity logs with fuzzy logic [21]. Tran integrated different logs and classified them to discriminate fractured zones from non-fractured zones [22]. Both studies mentioned that these methods were not reliable without image logs to validate results.
Tokhmechi and his associates have published several papers where they applied wavelet transform and different classifiers and data fusion techniques to discriminate fractured zones from non-fractured zones [23 - 25]. These authors validated their results by comparing them to image logs, confirming that the proposed approaches were repeatable with accuracies higher than 70%. Mazaheri et al. [26], [27] developed the Fracture Measure (FM) technique, a novel fractured zone detection criterion calculated by aperture, fracture type, azimuth, and apparent distance. Artificial Neural Networks were used to estimate FM and cell size, optimized with a fracture zone detection accuracy of 80%. Aghli et al. utilized PLs for fractured zone detection to develop a rapid Velocity Deviation Log (VDL) method that affects porosity and permeability, useful in identifying fracture apertures [28]. The authors found that the differentiation method could easily recognize fractured zones in high fracture density zones. Zarehparvar Ghoochaninejad et al. developed a Sugeno fuzzy inference system that estimates the hydraulic aperture of detected fracture zones utilizing PLs [29].
Ignoring the effects of imbalanced datasets on fractured zone detection accuracy and precision has been the common shortcoming of every cited peer-reviewed journal article. We used a machine learning approach to discriminate fractured zones from non-fractured zones using PLs, and applied them to imbalanced and balanced datasets. This method will allow us to investigate the importance of creating balanced data, especially in fractured zone detection, the minor and the most interesting class.

2. Database

The database used in this study was built from the data from eight wells, which were drilled in one of the largest carbonate fractured reservoirs: the Asmari formation in South-West Iran. Four hundred fifty wells were drilled in the studied reservoir, eight of which have both PLs and ILs. Interpreted ILs are the source of discrimination between F/NFZs since full sets of PLs were not completed in the studied wells; therefore, 10 out of the 29 PLs were selected for our studies. Selected logs and their availability are listed in Table 1.

Table 1: The Most Available Pls in Studied Wells

PLs		Wells							
		1	2	3	4	5	6	7	8
Caliper	Cali	*	*	*	*	*	*	*	*
	CGR	*	*	*	*	*	*	*	*
Gamma Ray	Thorium	*	*	*		*			*
	Uranium	*	*	*		*			*
	SGR	*	*	*		*	*	*	*
Sonic	DT	*	*	*	*	*	*	*	*
Density	RHOB	*	*	*	*	*	*	*	*
Photoelectric Factor	PEF	*	*	*	*	*	*	*	*
Neutron Porosity	NPHI	*	*	*		*	*	*	*
Water Saturation	SW	*			*		*	*	*
* means available									

The correlation coefficient between PLs and FZs, the fracture label, was utilized for log selection. Table 2 displays the correlation coefficient from well 3. The correlation coefficients between FZs and all logs are less than 0.4, which confirms the complexity of FZ detection by using PLs (Table 2).

Table 2: Cross Correlations between All Petrophysical Logs, Fracture Label, and Depth in Well 3

	DEPTH	Fracture	CALI	CGR	RHOB	SGR	DT	PEF	NPHI	POTA	URAN	THOR
DEPTH	1	-0.07336	-0.57776	-0.09963	-0.05912	-0.32121	0.227034	-0.0844	0.161339	-0.11378	-0.32386	-0.03263
Fracture	-0.07336	1	-0.07464	0.013317	-0.16876	0.149943	0.198929	-0.28905	0.154407	0.022499	0.161401	-0.00121
CALI	-0.57776	-0.07464	1	0.170612	0.041254	0.123247	-0.1629	0.068334	-0.18031	0.163828	0.081759	0.091154
CGR	-0.09963	0.013317	0.170612	1	-0.01025	0.436472	-0.04695	-0.10667	-0.07539	0.859329	0.142732	0.766563
RHOB	-0.05912	-0.16876	0.041254	-0.01025	1	-0.03041	-0.78033	0.6722	-0.9072	-0.03469	-0.02825	0.013941
SGR	-0.32121	0.149943	0.123247	0.436472	-0.03041	1	-0.04126	-0.22342	-0.02338	0.398561	0.948783	0.367711
DT	0.227034	0.198929	-0.1629	-0.04695	-0.78033	-0.04126	1	-0.46904	0.824402	-0.03039	-0.03018	-0.04583
PEF	-0.0844	-0.28905	0.068334	-0.10667	0.6722	-0.22342	-0.46904	1	-0.63973	-0.12778	-0.20908	-0.04885
NPHI	0.161339	0.154407	-0.18031	-0.07539	-0.9072	-0.02338	0.824402	-0.63973	1	-0.04937	-0.00321	-0.06076
POTA	-0.11378	0.022499	0.163828	0.859329	-0.03469	0.398561	-0.03039	-0.12778	-0.04937	1	0.128948	0.435153
URAN	-0.32386	0.161401	0.081759	0.142732	-0.02825	0.948783	-0.03018	-0.20908	-0.00321	0.128948	1	0.122347
THOR	-0.03263	-0.00121	0.091154	0.766563	0.013941	0.367711	-0.04583	-0.04885	-0.06076	0.435153	0.122347	1

The availability and usability, as well as the physical effect of fractures in the PLs were the critical factors for PL selection. The physical effects of FZs over PLs are briefly discussed as follows:

- Fractures cause lower density in rocks; therefore, density log (RHOB) might be lower in FZs.
- The P-wave velocity is lower in porous rock and fluid; therefore, sonic log (DT) might be increased in FZs.
- Gamma Rays, natural radioactivity, are produced by Uranium, Thorium, and Potassium.
- Potassium, Thorium, and Cumulative Gamma Ray (CGR), which is the summation of Potassium and Thorium, are all found in clayey formations. Clay behaves plastically, which could indicate NFZ.

- Uranium can dissolve in water and be deposited in fractures; therefore, high Uranium and Summation of Gamma Ray (SGR) levels could indicate FZs.
- Higher water saturation (Sw) could be an indicator of FZs.
- If brine is trapped in the fractures, resistivity decreases, meaning the NFZ has a high resistivity (RT).
- Water traps indicate high Photoelectric Factor (PEF) levels since connate water elements, such as Barite, have a higher atomic number than the surrounding rock.
- Neutron Porosity (NPHI) is increased if an FZ is filled with fluid or Hydrogen.
- Dolomite and limestone are brittle, with elastic behavior, capable of fracturing.
- Shale and Anhydrite are ductile, capable of plastic behavior, and could be an indicator of NFZs.
- Caliper tools are used to measure the diameter of boreholes instead of bit size; however, rock edges are chipped away during drilling in FZs. As a result, mud accumulation in open fractures may result in lower borehole measurements, to the point where the measurement is less than the bit size.

Overall, caliper, CGR, SGR, RHOB, DT, PEF, NPHI, and Sw measurements are useful; therefore, total porosity was selected for FZ detection. Classifiers will be defined to discriminate between F/NFZs in 10D feature space.

3. Methodologies

3.1. Data pre-processing and imbalance checking

Depth shifting and matching, tolls pickup correction, tolls malfunction correction, cycle skipping correction, and wash out error removing methods were preprocessing procedures completed on PLs. The imbalance index was introduced as a relative measure of imbalance and is defined as the difference between the non-fractured and fractured data, divided by the number of fractured zones (Equation 1). This index approaches zero if the two classes are balanced, and increases in imbalanced situations:

$$Imbalance\ Index = \frac{No.of\ NonFractured - No.of\ Fractured}{No.of\ Fractured} \quad (1)$$

Table 3 compares the fractured and non-fractured zones - indicating that the number of FZs are less than NFZs for all wells.
The imbalance indices are highest for wells 3 and 4, at 2.18 and 5.73, respectively. The imbalance index for wells 2 and 6 were approximately 0.1; therefore, a comparison between classification in imbalance and balance scenarios for those four wells could indicate the importance of balancing. The Near Miss Under balance sampling algorithm was applied instead of NFZ to create a balanced dataset.
Random Forest (RF) and Support Vector Machine (SVM) were scripted in Jupyter iPython Notebook. The database was decomposed to train (70 %), test (30 %), and classify imbalanced and balanced data. The confusion matrix (CM), accuracy or correct classification rate (CCR), precision, and recall were calculated as the classifier's performance index. These evaluation parameters are introduced as follows:

Table 3: Comparison between the Numbers of Fractured and Non-Fractured Zones in Studied Wells and Their Imbalance Indices

Well	Non-Fractured Fractured	Imbalance Index
1	1215 945	0.28
2	1608 1450	0.11
3	1103 347	2.18
4	1778 264	5.73
5	1056 690	0.53
6	1493 1353	0.10
7	965 625	0.54
8	1408 864	0.63

CM is a square matrix in which entry diagonals represent the number of accurate classified data, with the rest representing misclassified data (Fig 1). The current study is the two-class problem, in which C00 represents the count of true positive, which means true classified fractured zones. C01 represents true negatives, or non-fractured zones misclassified as fractured. C10 represents false positives, or non-fracture zones misclassified as fractured. C11 represents true positives, or the zones are correctly classified [30].

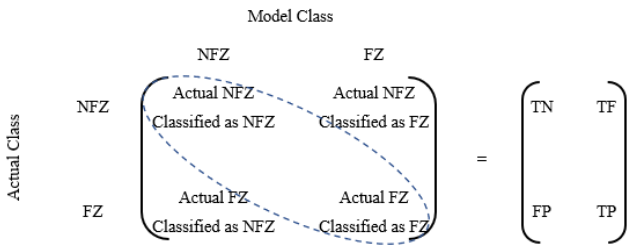


Fig 1: Trace of Confusion Matrix Used to Indicate Classification Accuracy.

Accuracy, or CCR, represents the number of correctly classified data divided by the total data [30] (Equation 2):

$$Accuracy\ or\ CCR\ = \frac{TN+TP}{TN+FP+TP+FN} \tag{2}$$

In Equation 2, TN/TP and FN/FP are true negative or positive results, or false negative or positive results, respectively. Accuracy may not be a good measure if the dataset is not balanced, where fractured and non-fractured classes have different amounts of data. Precision, or positive predictive values, might be better validation tools [30] (Equation 3):

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

In this Equation, one represents the highest precision, which happens when FP equals zero.
Recall is known as sensitivity or true positive rate. The highest recall is also equal to one, which means FN is zero [30] (Equation 4):

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

3.2. Random forest

RF is a tree-based classifier that can classify with high accuracy, stability, and ease of interpretation. This approach employs features, or modes, of categorical features. RF is a popular classification and regression algorithm in sci-kit-learn; it combines several decision trees trained by different sets of observations. The final prediction, illustrated in Fig 2, is created by averaging the prediction of each tree.

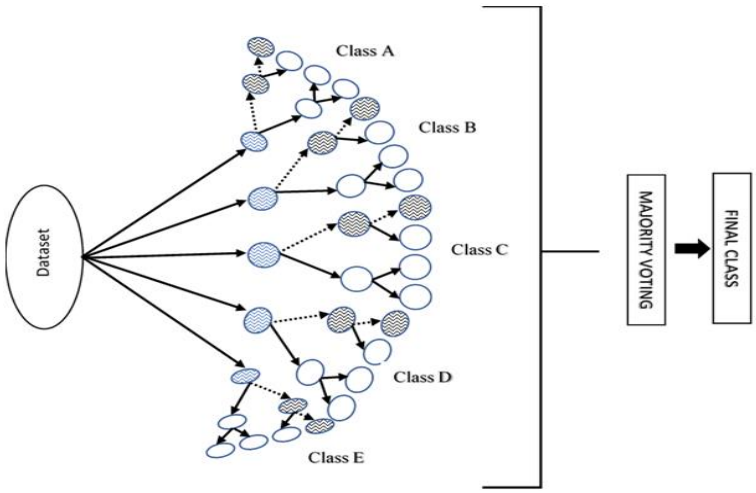


Fig. 2: A schematic view of RF.

The best splitting node in Decision Tree classification refers to obtaining nearly-homogenous sub-nodes or child nodes upon splitting a parent node. The Gini Index (GI), used in this paper, is one of the best methods to complete this work. Gini is an index of the number of random data points misclassified, varying between 0 and 0.5. A lower Gini indicates a lower chance of obtaining misclassified random data points. This index allows us to make better decisions with lower ambiguity [31], [32] (Equation 5).

$$G = \sum_{i=1}^{k} p_i(1-p_i) = 1 - \sum_{i=1}^{k} p_i^2 \tag{5}$$

3.3. Support vector machine

SVM is one of the most popular methods of machine learning. This method classifies datasets by developing hyperplanes in multidimensional space. It iteratively generates the best-general hyperplane that contains a minimum error and a maximum margin, which, in turn, results in discrimination between different classes (Fig 3).

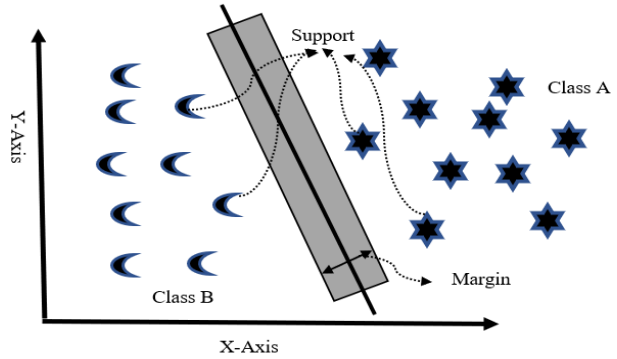


Fig. 3: A Schematic View of SVM and Support Vectors.

The kernel in SVM has a vital role because it is implemented to transform input data into a higher-dimensional space with optimal class discrimination. The wrong kernel choice may lead to higher error. The most popular kernels are 1) Linear, 2) Polynomial, and 3) Radial Basis Function (RBF). We used the RBF kernel [31], [32] (Equation 6).

$$K(x, x_i) = exp\ (-gamma * sum(x - x_i)^2) \quad (6)$$

In (Equation 6), gamma is in the range of 0 to1.
Data needs to be normalized before classification with the SVM method. Equation (7) depicts the relation used to normalize an SVM [33]:

$$\widehat{x_i} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \widehat{x_i} \rightarrow [0\ ,1] \quad (7)$$

In Equation 7, $\widehat{x_i}$ is the normalization, and xi is the primary amount of one log at the same depth, and $\boldsymbol{x_{min}}$ and $\boldsymbol{x_{max}}$ are the minimum and maximum ranges of data, respectively.

4. The experimental results

We applied RF and SVM to balanced and imbalanced PLs to discriminate F/NFZs.

4.1. Classification using imbalanced data

4.1.1. RF

Table 4 lists the results of the discrimination of F/NFZs using RF. The table indicates that CCRs or accuracy = 91 – 98%, precision = 83 – 95%, recall = 89 – 97%, and support = 435 - 918.

Table 4: The Results of Discrimination of Fzs from Nfzs Using RF. Imbalanced Versions of Selected Pls are used.

Well	Confusion Matrix	Accuracy or CCR	Precision	Recall	Support
1	$\begin{bmatrix}342 & 20\\ 39 & 247\end{bmatrix}$	0.91	0.86	0.92	648
2	$\begin{bmatrix}466 & 28\\ 30 & 394\end{bmatrix}$	0.94	0.93	0.93	918
3	$\begin{bmatrix}338 & 9\\ 15 & 73\end{bmatrix}$	0.94	0.83	0.89	435
4	$\begin{bmatrix}539 & 3\\ 12 & 59\end{bmatrix}$	0.98	0.83	0.95	613
5	$\begin{bmatrix}309 & 9\\ 15 & 191\end{bmatrix}$	0.95	0.93	0.96	524
6	$\begin{bmatrix}442 & 17\\ 20 & 375\end{bmatrix}$	0.96	0.95	0.96	854
7	$\begin{bmatrix}279 & 16\\ 9 & 173\end{bmatrix}$	0.95	0.95	0.92	477
8	$\begin{bmatrix}407 & 7\\ 18 & 250\end{bmatrix}$	0.96	0.93	0.97	682

The difference between accuracy and precision is significant. Wells 3 and 4 have the highest Imbalance Index.

4.1.2 SVM

Table 5 displays the results of applying SVM on imbalanced data. CCRs are higher than 91%, confirming that FZ detection using petro-physical logs is possible. CCR and precision for well 4 are 95% and 78%, respectively, indicating that for well 4, SVM is biased toward NFZs. CCR and precision are above 98% for well 8. Therefore, we can report that for well 8, a synthetic image log has been created.

Table 5: The Results of Discrimination of Fzs from Nfzs Using SVM. Imbalanced Versions of Selected Pls are Used.

Well	Confusion Matrix	Accuracy or CCR	Precision	Recall	Support
1	$\begin{bmatrix}311 & 33\\ 24 & 280\end{bmatrix}$	0.91	0.92	0.89	648
2	$\begin{bmatrix}452 & 24\\ 24 & 418\end{bmatrix}$	0.95	0.94	0.94	918
3	$\begin{bmatrix}325 & 17\\ 10 & 83\end{bmatrix}$	0.94	0.89	0.83	435
4	$\begin{bmatrix}518 & 16\\ 17 & 62\end{bmatrix}$	0.95	0.78	0.79	613
5	$\begin{bmatrix}303 & 9\\ 20 & 192\end{bmatrix}$	0.95	0.90	0.96	524
6	$\begin{bmatrix}437 & 13\\ 16 & 388\end{bmatrix}$	0.97	0.96	0.97	854
7	$\begin{bmatrix}264 & 6\\ 10 & 197\end{bmatrix}$	0.97	0.95	0.97	477
8	$\begin{bmatrix}400 & 2\\ 6 & 274\end{bmatrix}$	0.99	0.98	0.99	682

4.1.3. Comparison of classification using imbalanced data

In Fig 4, CCR as well their average for two classifiers in all studied wells are presented. Surprisingly, the best results have been for well 4, which has the highest imbalance index. This clearly shows that in an imbalanced situation, CCR might be perfect compared to the precision of the previous study with low precision. Therefore, making a judgment based on the result using CCR could be misleading. In average, SVM performance is a little bit better than RF.

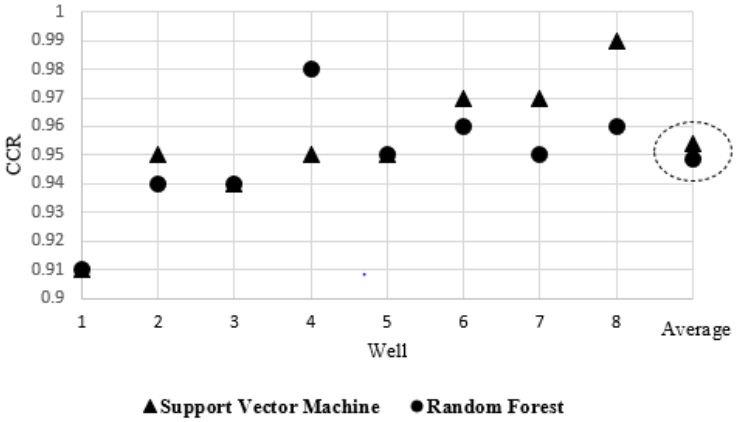


Fig. 4: Comparison between the CCR of F/Nfzs of Imbalanced Data Differentiation by Applying RF and SVM to Studied Wells.

4.2. Classification using balanced data

4.2.1. RF

Table 6 displays the results of applying RF to balanced data. CCR and precision for all wells are higher than 88%, representing the method's effectiveness for the proper differentiation of F/NFZs. The result of well 4 is interesting, as it indicates that the precision is higher than the accuracy, while in the case of the imbalanced data (Table 4), the precision was 15% lower than the accuracy. This approach reveals the advantage of the F/NFZs methods.

Table 6: The Results of Discrimination of Fzs from Nfzs Using RF. Balanced Versions of Selected Pls are used

Well	Confusion Matrix	Accuracy or CCR	Precision	Recall	Support
1	$\begin{bmatrix}237 & 36\\ 33 & 261\end{bmatrix}$	0.88	0.89	0.88	567
2	$\begin{bmatrix}405 & 28\\ 42 & 395\end{bmatrix}$	0.92	0.90	0.93	870
3	$\begin{bmatrix}104 & 6\\ 7 & 92\end{bmatrix}$	0.94	0.93	0.94	209
4	$\begin{bmatrix}73 & 6\\ 1 & 79\end{bmatrix}$	0.96	0.99	0.93	159
5	$\begin{bmatrix}198 & 19\\ 9 & 188\end{bmatrix}$	0.93	0.95	0.91	414
6	$\begin{bmatrix}383 & 16\\ 19 & 394\end{bmatrix}$	0.96	0.95	0.96	812
7	$\begin{bmatrix}178 & 11\\ 4 & 182\end{bmatrix}$	0.96	0.98	0.94	375
8	$\begin{bmatrix}238 & 14\\ 6 & 261\end{bmatrix}$	0.96	0.98	0.95	519

4.2.2. SVM

The results of applying SVM to balanced data are displayed in Table 7. The average accuracies and precisions are above 90%.

Table 7: The Results of Discrimination of Fzs from Nfzs Using SVM. Balanced Versions of Selected Pls are used.

Well	Confusion Matrix	Accuracy or CCR	Precision	Recall	Support
1	$\begin{bmatrix}253 & 24\\ 23 & 267\end{bmatrix}$	0.92	0.92	0.92	567
2	$\begin{bmatrix}414 & 35\\ 29 & 392\end{bmatrix}$	0.93	0.93	0.92	870
3	$\begin{bmatrix}97 & 15\\ 10 & 87\end{bmatrix}$	0.88	0.90	0.85	209
4	$\begin{bmatrix}83 & 2\\ 14 & 60\end{bmatrix}$	0.91	0.81	0.97	159
5	$\begin{bmatrix}182 & 32\\ 7 & 193\end{bmatrix}$	0.91	0.96	0.86	414
6	$\begin{bmatrix}414 & 18\\ 12 & 368\end{bmatrix}$	0.96	0.97	0.95	812
7	$\begin{bmatrix}183 & 11\\ 6 & 175\end{bmatrix}$	0.96	0.97	0.94	375
8	$\begin{bmatrix}261 & 7\\ 1 & 250\end{bmatrix}$	0.98	1	0.97	519

4.2.3. Comparison of classification using balanced data

In Fig 5, CCR and their average for two classifiers in all studied wells are presented. Based on the results, in average, performance of RF has been better than SVM. In well 4, with the highest imbalance index, RF has performed significantly better than SVM. Overall, RF was better discriminated F/NFZs for balanced dataset.

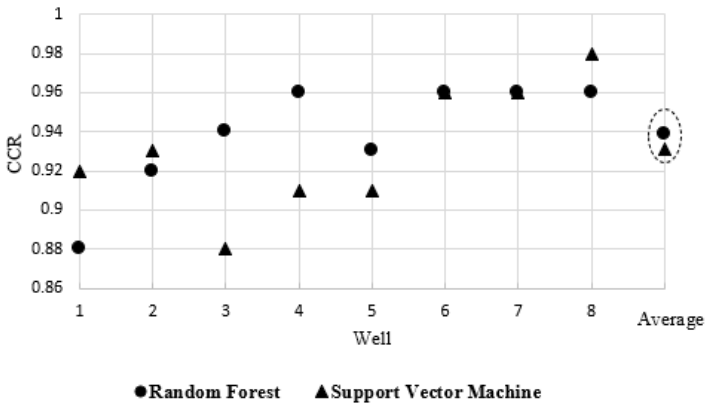


Fig. 5: Comparison between the CCR of RF and SVM to Differentiate F/Nfzs Using Balanced Data.

4.3. Comparison between classifications using imbalanced and balanced data

The average accuracy and precision achieved using two classifiers on imbalanced and balanced data are represented in Table 8. The difference between average accuracy and precision for imbalanced and balanced data is significant. Table 8 indicates that the average accuracy for balanced data decreased from 92.8% to 91.4%, while the average precision was increased from 88.4% to 91.9%. Using imbalanced data and the inability of FZ recognition are the main shortcomings of classifiers, as illustrated with the CM values found in Tables 4-7. NFZs are dominant in imbalanced data, which is the primary contribution to failed recognition; therefore, the classifiers are biased in their attempts to improve CCR. Creating a balanced dataset solves the problem of minor class (FZ) recognition, as shown in Table 8.

Table 8: Comparison between Average Accuracy and Precision of Different Classifiers for Eight Studied Wells While Using Imbalanced and Balanced Data

Well	Imbalance Data		Balanced Data	
	Average Accuracy or CCR	Average Precision	Average Accuracy or CCR	Average Precision
1	0.895	0.875	0.880	0.880
2	0.912	0.902	0.905	0.900
3	0.927	0.850	0.912	0.922
4	0.962	0.810	0.925	0.900
5	0.935	0.897	0.917	0.945
6	0.915	0.905	0.902	0.902
7	0.942	0.932	0.945	0.960
8	0.937	0.900	0.927	0.940
Average	0.928	0.884	0.914	0.919

The difference between accuracy and precision, which might be an index of the classification's reliability, is illustrated in Fig 6. The low difference between accuracy and precision shows that the classifier has distinguished both F/NFZs. On average, the difference between the accuracy and the precision of balanced data is approximately zero, whereas the accuracy of the imbalanced data is approximately 4% (Fig 6). The classifiers for imbalanced data were biased toward the dominant class, NFZ. The importance of balancing data is highlighted in wells 3 and 4, with imbalance indices of 2.18 and 5.73, respectively (Table 3). The most significant result lies in the difference between average accuracy and precision for well 4. Using original data, the difference was approximately 15% but declined to approximately 2% after balancing. These results indicate that balancing is mandatory if the imbalance index is higher than 1.

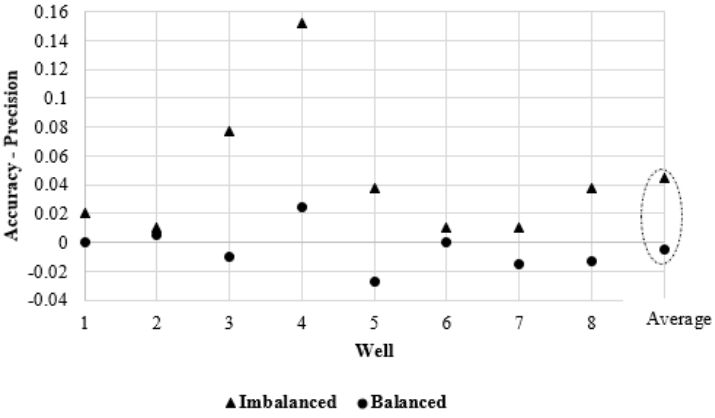


Fig. 6: Average Accuracy minus Average Precision of Utilized Classifiers for Balanced and Imbalanced Data.

One of fundamental graphical evaluation tool for test result is Receiver Operating Characteristic (ROC) curve. This curve plots the sensitivity (true positive rate) of a test versus its specificity (false positive rate) for different cut-off points of a parameter. The accuracy of a test is measured by the area under the ROC curve (AUC).
Sensitivity is the probability of a depth will be positive given as a fracture zone. Specificity is the probability of a depth will be negative given as a nonfracture zone.
In Figs 7-8, the ROC curve for both balance and imbalance data (one well as train and one well as a test), while RF classifiers are utilized, are shown respectively. As these figures show, the closer the apex of the curve toward the upper left corner, the greater the dis-criminatory ability of the RF classifier. This is measured quantitatively by the AUC such that a value of >0.97 indicates excellent discriminatory ability.

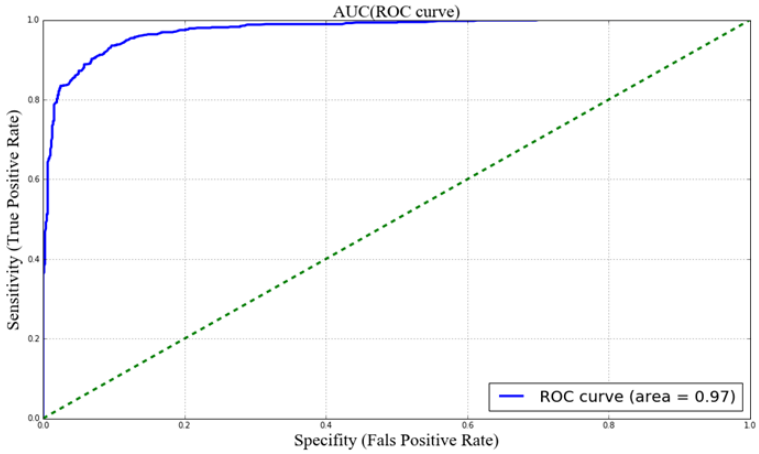


Fig. 7: Receiver Operating Characteristic Curve for RF Classifier on Balanced Data and While One Well Is as A Train an Another Is as A Test.

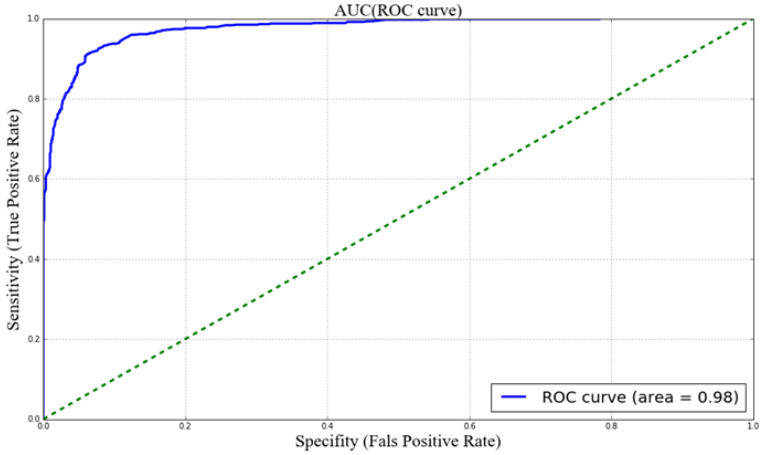


Fig. 8: Receiver Operating Characteristic Curve for RF Classifier on Imbalanced Data and While One Well Is as A Train an Another Is as A Test.

5. Conclusions

We applied two classification methods to PLs in a carbonate-fractured reservoir to differentiate between F/NFZs. We used 10 of 29 PLs for FZ detection using a caliper, CGR, SGR, RHOB, DT, PEF, NPHI, SW, and effective and total porosity measurements. The key contributions of this study can be summarized as follows:

- Statistical studies indicate that the correlation coefficient between fractured zones and various logs differ in the range of -0.25 to 0.25; therefore, fractured zone detection using individual PLs is impossible because of the low correlation. Statistical studies also indicated that the correlation coefficient between some of the PLs is high, such as between CGR, Potassium, and Thorium. In these cases, one of the high correlation PLs could be selected for FZ detection instead of using them all.
- The number of FZs is less than NFZs in studied wells. The imbalance index was defined to represent the imbalance situation in different wells. The results indicated that well 6 is the most balanced well with an index of 0.1, and well 4 is the most imbalanced well with an index of 5.73.
- RF and SVM were applied to discriminate FZs from NFZs by using selected PLs for each well. Results indicated that the average CCR for the imbalanced and balanced datasets was approximately 92%.
- Precision and accuracy for balanced datasets were similar in all cases; however, the precision was inaccurate when using imbalanced data. These results reinforce the importance of creating balanced data when the imbalance index is high.
- In the wells with a high imbalance index, RF performed significantly better than SVM. The difference between accuracy and precision was high, indicating that classifier bias to NFZs, the major class, and FZ, the minor and most important class, might be better recognized. These results reinforce the importance of creating a balanced dataset before applying classifiers.

6. Future work

PLs are usually captured inside oil well boreholes, while ILs rarely are. In the studied oil field, more than 450 wells were drilled. Of these wells, ILs were captured in only eight boreholes. Modelers need fracture data for network modeling and fluid flow simulation, and currently developed models suffer from data limitations. FZ detection by running classifiers on PLs is possible; therefore, classifiers must be trained in the wells with both IL and PL data.

Current studies have reported that both accuracy and precision are high in wells containing low imbalance factors, while precision is low in high imbalance index wells. These results indicate that the data is challenging to balance; therefore, wells must be differentiated between high and low imbalance indices. Training well databases can train the classifiers used to discriminate F/NFZs in the wells with low imbalance indices. The dataset must be balanced before the classifiers are trained to discriminate between the F/NFZs when the imbalance index is high.

Differentiation of high and low imbalance indices in oil wells using PLs, rock type, tectonics, and seismicity, among other methods, could be the subject of future research.

Acknowledgements

I thank the editor, Anna Crowell, for her patience and detailed comments that allowed us to significantly improve the manuscript.

References

[1] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, Handling Imbalanced Datasets: A Review, GESTS International Transactions on Computer Science and Engineering, 30(1) (2006) 25-36.
[2] M. Kubat, S. Matwin, Addressing the Curse of Imbalanced Training Sets: One-sided Selection, ICML. (1997).
[3] A. Al-Shahib, R. Breitling, D. Gilbert, Feature Selection and the Class Imbalance Problem in Predicting Protein Function from Sequence, Applied Bioinformatics, 4(3) (2005) 195-203. https://doi.org/10.2165/00822942-200504030-00004.
[4] L. Yi-Hung C. Yen-Ting, Total Margin Based Adaptive Fuzzy Support Vector Machines for Multiview Face Recognition. in Systems, Man and Cybernetics, 2005 IEEE International Conference. (2005) https://doi.org/10.1109/ICSMC.2005.1571394.
[5] M.A. Mazurowski, P.A. Habas, JM. Zurada, *et al.* Training Neural Network Classifiers for Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance, Neural networks: The Official Journal of the International Neural Network Society, 21(2-3) (2008) 427-436. https://doi.org/10.1016/j.neunet.2007.12.031.
[6] Z.B. Zhu Z.H. Song, Fault Diagnosis Based on Imbalance Modified Kernel Fisher Discriminant Analysis, Chemical Engineering Research and Design, 88(8) (2010) 936-951. https://doi.org/10.1016/j.cherd.2010.01.005.
[7] M. Tavallaee, N. Stakhanova, A.A. Ghorbani, Toward Credible Evaluation of Anomaly-Based Intrusion-Detection Methods, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions, 40(5) (2010) 516-524. https://doi.org/10.1109/TSMCC.2010.2048428.
[8] Y. Li, G. Sun, Y. Zhu, Data Imbalance Problem in Text Classification, Information Processing (ISIP), Third International Symposium, IEEE. (2010) https://doi.org/10.1109/ISIP.2010.47.
[9] Y. Sun, A.K.C. Wong, M.S. Kamel, Classification of Imbalanced Data: A Review, International Journal of Pattern Recognition and Artificial Intelligence, 23(4) (2009) https://doi.org/10.1142/S0218001409007326.
[10] A. Ali, S.M. Shamsuddin, A.L. Ralescu, Classification with Class Imbalance Problem: A Review, International Journal of Advance Soft Computation Application, 5(3) (2013) 1-30.
[11] X.Q. Ouyang, Y.P. Chen, B.H. Wei, Experimental Study on Class Imbalance Problem Using an Oil Spill Training Data Set, British Journal of Mathematics and Computer Science, 21(5) (2017) 1-9. https://doi.org/10.9734/BJMCS/2017/32860.
[12] J.M. Johnson, T.M. Khoshgoftaar, Survey on Deep Learning with Class Imbalance, Journal of Big Data, 6(27) (2019) 1-54. https://doi.org/10.1186/s40537-019-0192-5.
[13] N. Klyuchnikov, A. Zaytsev, A. Gruzdev, *et al.*, Data-driven Model for the Identification of the Rock Type at a Drilling Bit, Journal of Petroleum Science and Engineering, 178 (2019) 506-516. https://doi.org/10.1016/j.petrol.2019.03.041.
[14] M. Pirizadeh, N. Alemohammad, M. Manthouri, *et al.*, A New Machine Learning Ensemble Model for Class Imbalance Problem of Screening Enhanced Oil Recovery Methods, Journal of Petroleum Science and Engineering, Available online, 108214. (2020) https://doi.org/10.1016/j.petrol.2020.108214.
[15] J. Brownlee, How to Develop an Imbalanced Classification Model to Detect Oil Spills, https://machinelearningmastery.com/imbalanced-classification-model-to-detect-oil-spills/ 2020.
[16] M. Nemati, H. Pezeshk, Spatial Distribution of Fractures in the Asmari Formation of Iran in Subsurface Environment: Effect of Lithology and Petrophysical Properties, Natural Resources Research, 14 (2005) 305-316. https://doi.org/10.1007/s11053-006-9000-y.
[17] A.R. Mohebbi, M. Haghighi, M. Sahimi, Using Conventional Logs for Fracture Detection and Characterization in One of Iranian Field, International Petroleum Technology Conference held in Dubai, U.A.E., 4-6 December 2007, Paper IPTC 11186. https://doi.org/10.3997/2214-4609-pdb.147.iptc11186.
[18] M. Sahimi, M. Hashemi Wavelet Identification of the Spatial Distribution of Fractures, Geophysical Reservoir Letters, 28(4) (2001) 611-614. https://doi.org/10.1029/2000GL011961.
[19] M. Daiguji, O. Kudo, T. Wada, Application of Wavelet Analysis to Fault Detection in Oil Refinery, Computers & Chemical Engineering, 21 (1997) S 1117-S 1122. https://doi.org/10.1016/S0098-1354(97)00199-3.
[20] R.A. Behrens, M.K. Macleod, T.T. Tran, *et al.*, Incorporating Seismic Attribute Maps in 3D Reservoir Models, SPE Reservoir Evaluation, 1 (1998) 122-126. https://doi.org/10.2118/36499-PA.
[21] L.P. Martinez-Torres, Characterization of Naturally Fractured Reservoirs from Conventional Well Logs, M.Sc. Thesis, University of Oklahoma, USA. (2002).
[22] N. H. Tran, Characterization and Modeling of Naturally Fractured Reservoirs, Ph.D. Thesis, University of New South Wales, Australia. (2004) https://doi.org/10.1109/ICSMC.2005.1571394.
[23] B. Tokhmechi, H. Memarian, V. Rasouli, *et al.*, Fracture Zones Detection Using Wavelet Decomposition of Water Saturation Log, Journal of Petroleum Science and Engineering, 69 (2009a) 129-138. https://doi.org/10.1016/j.petrol.2009.08.005.
[24] B. Tokhmechi, H. Memarian, H. Ahmadi Noubari, *et al.*, A Novel Approach for Fracture Zone Detection Using Petrophysical Logs, Journal of Geophysics and Engineering, 6 (2009b) 365-373. https://doi.org/10.1088/1742-2132/6/4/004.
[25] S.M. Mazhari, H. Memarian, B. Tokhmechi, A Hybrid Learning Automata and Case-based Reasoning for Fractured Zone Detection, Arabian Journal of Geosciences. (2018) https://doi.org/10.1007/s12517-018-3934-3.
[26] A. Mazaheri, H. Memarian, B. Tokhmechi, *et al.*, Developing Fracture Measure as an Index of Fracture Impact on Well-logs, Energy Exploration and Exploitation, 33(4) (2015) 555-574. https://doi.org/10.1260/0144-5987.33.4.555.

[27] A. Mazaheri, H. Memarian, B. Tokhmechi, *et al.*, Cell Size Optimization for Fracture Measure Estimation in Multi-Scale Studies Within Oil Wells, Carbonates and Evaporites, (2019) 261-272. https://doi.org/10.1007/s13146-017-0378-x.
[28] G. Aghli, B. Soleimani, R. Moussavi-Harami, *et al.*, Fractured Zones Detection Using Conventional Petrophysical Logs by Differentiation Method and its Correlation With Image Logs, Journal of Petroleum Science and Engineering, 142 (2016) 152-162. https://doi.org/10.1016/j.petrol.2016.02.002.
[29] H. Zarehparvar Ghoochaninejad, M.R. Asef, S.A. Moallemi, Estimation of Fracture Aperture from Petrophysical Logs Using Teaching–learning-based Optimization Algorithm into a Fuzzy Inference System, Journal of Exploration and Production Technology, 8 (2018) 143-154. https://doi.org/10.1007/s13202-017-0396-1.
[30] L. Wei-Meng, Python- Machine Learning, Getting Started with Scikit-learn for Machine Learning, Chapter 5, John Wiley & Sons, Inc. (2019) 93-117. https://doi.org/10.1002/9781119557500.ch5.
[31] J. Heaton, I. Goodfellow, Y. Bengio, *et al.*, Deep Learning, Genetic Programming and Evolvable Machines, 19(1-2) (2017) 305–307. https://doi.org/10.1007/s10710-017-9314-z.
[32] A. Zhang, Z.C. Lipton, M. Li, *et al.*, Dive into Deep Learning. https://d2l.ai/index.html, (2021) https://doi.org/10.1021/acs.jcim.0c00073.s001.
[33] S. Theodoridis, K. Koutroumbos, Pattern Classification, 2nd Edition, San Diego: Elsevier/Academic, 2002.