Journal of Petroleum Science and Engineering 219 (2022) 111058

Contents lists available at ScienceDirect

Journal of Petroleum Science and Engineering

journal homepage: www.elsevier.com/locate/petrol

Impact of data quality on supervised machine learning: Case study on drilling vibrations

Saket Srivastava [a,*], Rushit N. Shah [b], Catalin Teodoriu [a], Aditya Sharma [a]

[a] University of Oklahoma, Norman, 63079, OK, USA
[b] University of Illinois at Chicago, Chicago, 60607, IL, USA

ARTICLE INFO

Keywords:
Data quality
Supervised learning
Classification
Drilling vibrations
Sampling frequency
Data imbalance
Feature extraction
Data labeling

ABSTRACT

Training complex machine learning and deep learning models has become straightforward with the advent of highly efficient, open-source machine learning libraries. Supervised classification techniques such as logistic regression, random forests, and neural networks have also gained popularity in the drilling industry on the back of promising results. As a result, these techniques have been increasingly researched, especially in the domain of drilling vibrations. However, much of this research interest has been limited to finding the best classification model for estimating severity of downhole vibrations. While the choice of classification model is important, we argue that the successful implementation and adoption of machine learning technologies is equally dependent on correctly studying, cleaning, pre-processing the vibration drilling data before applying machine learning techniques. We show that, in certain cases, correctly pre-processing the data guarantees competitive classification performance regardless of the choice of classification model. Specifically, we empirically investigate how factors such as data sampling frequency, data labeling technique, feature extraction technique, and class imbalance impact the performance of different popular classifiers, when dealing with drilling data. We make recommendations specific to vibration classification and highlight pitfalls of certain techniques in that context. Finally, we also develop a step-by-step workflow which enables users to select the correct parameters and techniques at every step, from data collection to model training.

1. Introduction

Timely detection and mitigation of drilling vibrations for improving drilling efficiency and performance is a well-understood challenge in the drilling industry. Vibration control strategies fall into the passive or active control category based on the extent of intervention in the drilling process. An example of the passive control strategy involves the driller to optimize input drilling parameters such as torque, rotary speed, and weight on bit to find an optimal drilling zone with reduced drill string resonance (Dong and Chen, 2016). However, this approach limits the rate of penetration, an evaluation metric that often takes the highest precedence during drilling. Active control strategies embrace the dynamic nature of drilling vibrations to introduce sophisticated feedback control systems that help dampen drilling vibrations by exerting an equal and opposite force to the external vibrations in the drill string (Javanmardi and Gaspard, 1992; Jansen et al., 1995; Karkoub et al., 2009; Krama et al., 2021).

In recent years, another area of development has been the implementation of machine learning (ML) and advanced analytical techniques in characterization and prediction of downhole drilling vibration severity. This progress is fueled by the recent advancements in data collection, computing capabilities and readily available ML models. Significant progress has been made in dealing with structured downhole drilling data for machine learning applications such as stuck-pipe detection, implementing drilling optimization models, estimating downhole shocks etc. Noshi and Schubert (2018) Similarly, unstructured data available in text format through drilling reports have also been successfully leveraged to automate drilling anomaly and dysfunction detection (Zhang et al., 2020; Srivastava et al., 2021).

Recent literature on application of ML models in drilling vibrations highlight classification as the core task where the presence of a target label is integral in searching for possible patterns between the remaining data set features and itself. Recent work on vibration severity classification can be divided into neural network models (Zha and Pham, 2018) and non-neural network models (Hegde et al., 2019). Interestingly, some models utilize only surface data or downhole data to build a classifier (Baumgartner and van Oort, 2014; Srivastava and Teodoriu, 2020), while some utilize both surface and downhole data to

* Corresponding author.
E-mail addresses: dr.saketsrivastava@gmail.com (S. Srivastava), rshah231@uic.edu (R.N. Shah), cteodoriu@ou.edu (C. Teodoriu), aditya.sharma@ou.edu (A. Sharma).

https://doi.org/10.1016/j.petrol.2022.111058
Received 4 March 2022; Received in revised form 30 August 2022; Accepted 11 September 2022
Available online 20 September 2022

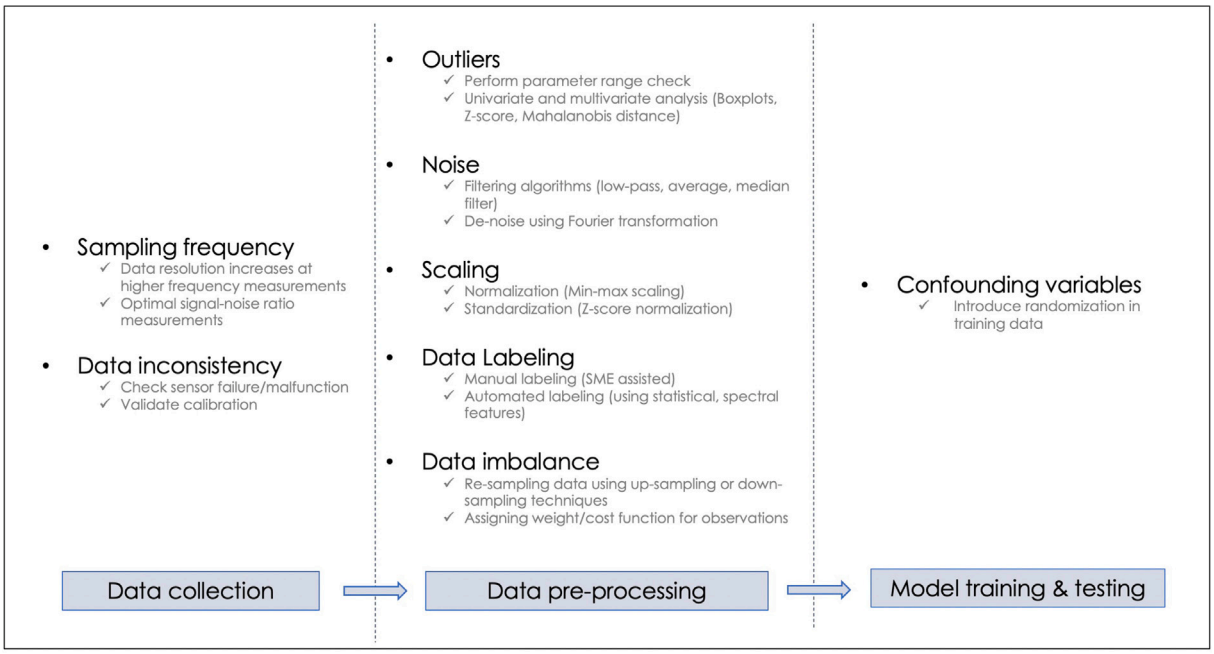

Fig. 1. Workflow to address data quality challenges.

build an exhaustive dataset (Millan et al., 2019). While the precision and classification accuracies of algorithms used in literature are very similar, the data used for training ML models is different for each case. Ultimately, a classifier performs in unison with the quality of data it is trained upon. There is variability in terms of data cleaning and pre-processing tasks before training ML models.

The bottleneck of ML models in the drilling industry is the quality of data available through surface and downhole measurements (Al Gharbi et al., 2018). To interpret and take decisions from real time data, the trustworthiness of drilling data needs to be assessed. Otalvora et al. (2016) developed a data quality index based on uniformity, sensibility, completeness, resolution, format, and structure of drilling data to develop a deeper understanding of its quality. The paper extends the discussion on data quality a step further to evaluate the impact of having high resolution and uniformity within training dataset on classification accuracy and performance. The centerpiece of the paper remains to find an optimal balance between data quality assessment and a high performing classifier for drilling vibrations. Furthermore, there is variability in terms of data cleaning and pre-processing in the industry. The paper explores the impact of different pre-processing techniques on model performance improvement, if any.

2. Data quality challenges

The following section addresses the data quality challenges associated with drilling data at every step of a machine learning task from data collection to model training and testing. Possible solutions to the issues have been provided in Fig. 1. The data quality issues mentioned in Fig. 1 are also discussed in greater detail in the following section:

2.1. Sampling rate

Sampling rate or sampling frequency is the frequency at which the sensor samples the analog signal to obtain the digital signal. It is an important metric of data quality assessment since a higher sampling rate enables more accurate identification of critical high-frequency downhole phenomenon that are not captured at lower sampling rates. Some examples of such phenomena include downhole shocks, pipe sticking, and drastic RPM changes due to torsional vibrations. An important consideration when choosing a sampling frequency for an application is the Nyquist theorem which states that the sampling frequency of a digital signal must be at least two times the highest frequency of interest in the signal (Abu-Mahfouz, 2003; Chandel and Patel, 2013). For example, in order to accurately detect vibration phenomena which manifest at frequencies under 100 Hz, the corresponding vibration signal must be sampled at least 200 Hz.

The earliest utilization of high sampling rates for downhole measurements dates back to 1994 when a wired measurement sub was used to better understand drilling dysfunctions, particularly stick–slip vibrations (Pavone and Desplans, 1994). The only drawback to high sampling downhole measurements is data storage. A popular strategy to work around data storage issues is to optimize data collection with long periods of low frequency measurements with short bursts of high frequency collection (>400 Hz) (Baumgartner and van Oort, 2014). Currently, no literature exists about an optimal data sampling frequency for downhole measurements that is critical for capturing drilling dysfunctions without comprising on data storage abilities.

Fig. 2 compares torsional vibration signature recorded experimentally at 1 Hz vs. 100 Hz to point out the incomplete information observed at lower sampling rates. Incomplete information can lead to misrepresentation of critical events that could potentially affect a classifiers performance. However, in literature, vibration severity classifiers have been successfully trained with both low and high frequency data with high precision. So, is high frequency data measurement justified from an ML perspective? This question is addressed in Section 5.1.

2.2. Outliers

Outliers are a common occurrence in most datasets. Whether they arise from expected/unexpected changes in system behavior, or are artifacts of instrument error, the presence of outliers in a dataset almost inevitably affects the performance of classification algorithms owing to their extreme value (Hodge and Austin, 2004). Indeed, popular classification algorithms such as support vector machine (SVM) and k-nearest neighbors (KNN) are particularly sensitive to datapoints on the convex hulls of each class of data in the dataset (Burges, 1998). In fact, it is for this very reason that KNN has been adopted as an outlier removal technique (Ramaswamy et al., 2000).

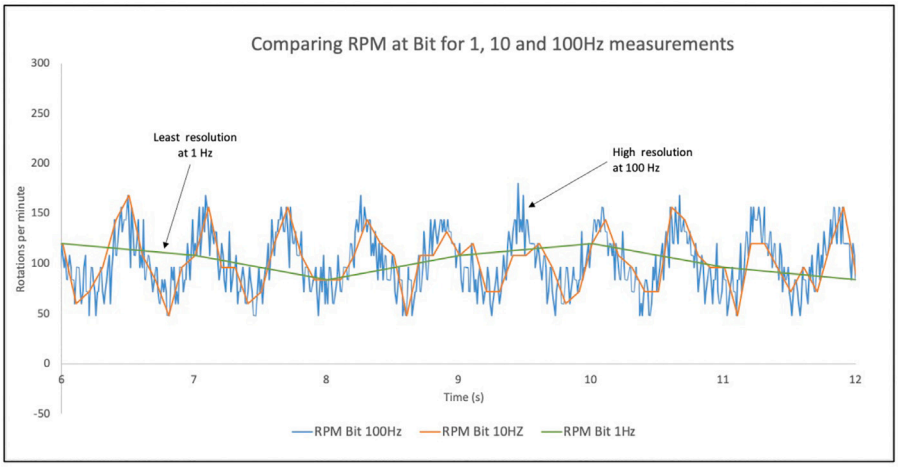

Fig. 2. Torsional vibration patterns at different sampling frequencies.

Formally, outliers can be categorized into two types—point (or global) outliers whose values are significantly outside the range of the entire dataset it belongs to; and contextual (or conditional) outliers whose values are significantly outside the range of other data that appears in a similar context. Generally, global outliers are a result of instrument error and can be safely removed via popular data pre-processing techniques which have been extensively studied and surveyed (Chandola et al., 2007; Domingues et al., 2018; Wang et al., 2019) and have robust, open-source implementations available (Zhao et al., 2019). Contextual outliers on the other hand may sometimes be indicative of change in system behavior and removing them could result in the failure to detect and act upon potentially critical system information. For example, sudden extreme fluctuations in drillstring downhole RPM may be indicative of stick–slip. The "context" in such cases is almost always temporal i.e., outliers are classified as such based on the values of other datapoints that occur temporally close together. Thus, as important as it is to remove undesirable outliers from a dataset, it is equally important to consider the nuances of the underlying physical process before doing so.

2.3. Noise

Noisy data is characterized by random deviations from the true underlying process data. It is one of the most commonly occurring data quality issue. Noise can be broadly classified into two categories—class and attribute noise. Class noise occurs when the data points are incorrectly labeled as belonging to a particular class. The source of such noise is typically human error and is introduced during the data labeling process. Attribute noise on the other hand originates in the data collection process and can be attributed to factors such sensor error/variability, stochastic nature of physical processes, and high sampling rates. In this study we focus on attribute noise, which is the more common of the two types. Given that noise results in the inclusion of data points that deviate from, and are not representative of the underlying physical process, it is naturally detrimental to the performance of machine learning classifiers when ignored. It has also been shown that noisy training data results in longer training times (Zhu and Wu, 2004). Noise in data has been studied extensively over the decades, and a multitude of techniques exist to characterize the extent of noisiness in training data, and to further eliminate such noise via signal processing techniques such as signal smoothing and filtering (Mohan et al., 2014). In fact, several modeling techniques even systematically account for some noise in the training data in order to build models and classifiers that are robust to such noise (Middleton, 1999; Vaseghi, 2008).

2.4. Data labeling

Data labeling is the characterization of data into identifiable categories that accurately describe a phenomenon under occurrence. In the case of drilling vibrations, it includes creating labels of vibration severity based on drilling parameters and measured responses. The process of data labeling is critical to the success of every supervised machine learning task. Often done manually, this process requires subject matter expertise supervision and is exhaustive in nature. Data labeling for drilling vibrations can be divided into labeling methodology and criteria.

The methodology of data labeling can be either point-based or time window-based. Point-based labeling process assigns a data label for every instance of drilling data whereas, window-based labeling process assigns a data label for the entire duration of the fixed time window. Point-based labeling is simpler to implement but fails to capture the temporal patterns in the data. Window-based data labeling is cumbersome to implement and optimize but incorporates temporal characteristics of data. In the case of drilling vibrations, windows-based labeling process is more widely utilized. Fig. 3 provides a visual explanation of time window sampling where the Nth time window w_N successfully captures both bit sticking (RPM = 0 at 101 s) and bit release (RPM = 1600 at 106 s) to confirm the presence of a stick–slip cycle. This phenomenon would have been extremely difficult to capture when considering point-based sampling.

A correctly-labeled dataset is crucial for both, training and refining, the classification model. Moreover, while the determined labels should be descriptive of the data features and the occurring phenomenon they correspond to, they should not be derived directly from any one such data feature since this would artificially induce improvements in classifier performance. Under such restrictions, the choice of labeling criteria is crucial. The following are criteria which are used to label vibration data

2.4.1. Stick–slip index (SSI)

Stick–slip index quantifies the intensity of torsional vibrations in the drillstring through downhole RPM measurements. It is defined as

$$\mathrm{SSI}(w_N) = \frac{\text{Bit RPM}_{max}(w_N) - \text{Bit RPM}_{min}(w_N)}{2 \times \text{Bit RPM}_{avg}(w_N)} \quad (1)$$

where min, max, and avg subscripts denote the maximum, minimum and average RPM respectively (as measured in the Nth window w_N. By measuring the extent of bit RPM deviation from the average RPM, the stick–slip index can help estimate the severity of torsional vibrations and can be used as a data labeling criterion (Hegde et al., 2019). SSI values ≥ 1 are considered to represent pure sticking (Patil, 2013). In

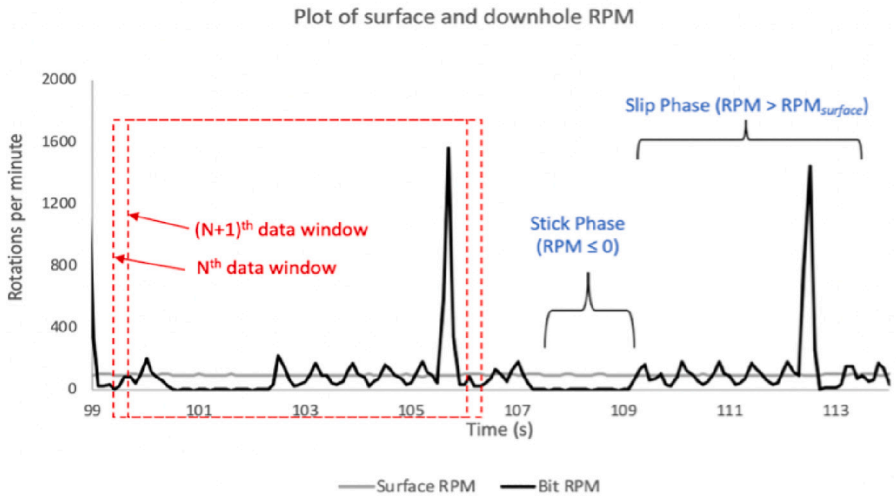

Fig. 3. Visualizing time window sampling as a data labeling procedure.

the case of pure sticking, the bit RPM fluctuates between 0 (or even negative) and at least twice the average bit RPM. SSI values greater than 1 represents a higher bit RPM swing due to longer bit sticking period. SSI values less than 1 indicates torsional oscillations without bit RPM reaching 0 indicating no bit sticking.

2.4.2. Radial acceleration at bit

Downhole sensors (accelerometers) help provide accurate radial acceleration values at the bit. Radial acceleration at bit can potentially fluctuate between 0 and 10 g for stick–slip vibrations (Baumgartner and van Oort, 2014). Vibration severity (in terms of standard gravity, g) can also be calculated using RMS amplitude of axial and lateral acceleration components using Eq. (2) below (Okoli et al., 2019).

$$a_{(\text{in g})} = \sqrt{a_{\text{ax}} - a_{\text{lat}}} \tag{2}$$

Where a_{ax} represents axial and a_{lat} represents lateral bit acceleration. While bit acceleration is a true representation of abrupt changes in bit rotation and resultant torsional vibrations, the challenge remains to associate severity of bit sticking to resultant amplitude of bit acceleration. This impacts the generality of labeling torsional vibrations, making the procedure case-specific.

2.4.3. Rate of penetration (ROP)

One of the most concerning impacts of severe drillstring vibrations is the reduction in rate of penetration, usually defined in ft/hour. As a result, the instantaneous loss in rate of penetration can be related to the severity of drilling vibrations in the system (Srivastava and Teodoriu, 2020). Another holistic indicator of drilling efficiency that includes dependency on ROP amongst other parameters is the mechanical specific energy (MSE). MSE is defined as the mechanical work required to drill a unit volume of rock and depends on various parameters as seen in Eq. (3) below. With greater loss of energy through vibrations, the energy required to drill rock becomes higher. Hence, the labeling strategy correlates vibration severity to drilling efficiency.

$$\text{MSE} = f(\text{Torque}_{\text{bit}}, \text{WOB}, \text{RPM}, \text{Diameter}_{\text{bit}}) \tag{3}$$

2.5. Feature extraction

Feature extraction is data pre-processing undertaken before classification to extract useful information from the data (and often reduce its dimensionality) before training. It is similar to dimensionality reduction to the extent that both approaches seek a mapping $x \rightarrow x'$ where $x \in \mathbb{R}^d$ is a d-dimensional vector in the original feature space, $x' \in \mathbb{R}^p$ is a p-dimensional vector in the reduced feature space ($p \leq d$).

However, in feature selection the reduced features are a subset of the original features i.e., $\{x'_i\}_{i=1}^{p} \subseteq \{x_j\}_{j=1}^{d}$. This is not the case with feature extraction where the new feature set, although derived from the original feature set, is completely different from it (Ghojogh et al., 2019). Extraction of features that represent the data well for a particular task is important to the final classification performance. In fact, one reason why deep neural network architectures work so well in practice is because their deep architecture allow their hidden layers to extract and save high- and low-level features from the input data before finally performing classification (Navamani, 2019; Bello et al., 2020). Since this is not the case with traditional models such as SVM, KNN, etc., feature extraction must be performed as a separate pre-processing step before training. Feature extraction is not generally considered a data quality issue since it is not a part of the data collection or cleaning process, however, from the perspective of training a machine learning model, we still view it as a vital pre-processing step and study its effectiveness through dedicated experiments later in the paper.

While there are a multitude of techniques available to extract features (Zebari et al., 2020), the types of features that are generally extracted for applications involving vibration data can be broadly split into two categories—statistical and spectral.

2.5.1. Statistical features

Statistical features aim to compress the information contained in a time-domain signal into a set of statistics describing the distribution of data contained in the signal. They are obtained by computing aggregate statistics for a sliding time window over the time-domain signal. Typically, the first k-statistical moments of the signal are used as features. Often these are enough to adequately characterize the location and shape of the data distribution. Here, the first moment is the mean μ; the second moment is variance σ^2, quantifying the spread around the mean; the third moment is skewness γ describing whether the data distribution is lopsided to one side or symmetric; and the fourth moment is kurtosis κ, that captures the heaviness of the tail of a distribution. There also exists higher-order moments which can be estimated and included as features. However, a larger sample size is required to estimate higher-order moments with the same precision as lower-order moments. Additionally, higher-order moments are also harder to interpret.

Another important consideration is that purely statistical features such as moments of a distribution can be estimated efficiently only when the data distribution is unimodal. Multimodally-distributed data is first modeled as a mixture of unimodal distributions e.g., as a mixture of Gaussians, before its moments can be estimated. Estimating the

parameters of such mixture model requires techniques such as Maximum Likelihood Estimation (MLE) (Le Cam, 1990; Myung, 2003) which makes computing statistical features for such data prohibitively expensive. Alternatively, in cases where specific characteristics of frequently-occurring vibration patterns are known beforehand, other statistical quantities besides moments may also be used as features. These include the mode, median, minimum, maximum, and quantities obtained by combining them. For instance, in torsional vibration applications, it is known that a value of Stick–Slip Index (SSI) greater than 1 is indicative of the occurrence of the stick–slip phenomenon (Hegde et al., 2019). The stick–slip index is just a combination of more fundamental statistics of the data, also defined in the previous section.

2.5.2. Spectral features

In general, when rotating machinery is operated under variable speed and fluctuating load, the resulting time-domain vibration signal typically contains multiple frequency components. As a result, the spectral make-up of the instantaneous vibration signal is a good indicator of the state under which the machinery is operating. This is very relevant to drilling with long drillstrings equipped with heavy bottom hole assembly components. The spectral signature obtained from downhole RPM is also a good indicator of drilling conditions where abrupt changes in spectral signal can be associated with severe downhole vibrations. Thus, even when the data belonging to different classes may not be linearly separable in the time domain, it may be so after transformation to the frequency domain. In such cases, simply including the best estimates of the frequency components in the signal at any instant, as a feature in the training data can yield significant gains in classifier performance. A variety of techniques to extract spectral features are available. Simple methods like Fast Fourier Transform (FFT) can efficiently extract all the frequency components and their magnitudes, which can then be used as features in the training data (Millan et al., 2019). More sophisticated time–frequency analysis methods such as Wavelet Packet Transform (WPT) (Yen and Lin, 2000) and Continuous Wavelet Transform (CWT) captures the change in the frequency components in the original signal (Yang et al., 2003; Zebari et al., 2020).

Principal Component Analysis (PCA) (Pearson, 1901; Wold et al., 1987) is unique in that although it is a feature extraction category, the features extracted by PCA are a linear combination of the original features, and cannot be classified into either of the two categories discussed above. PCA is also commonly classified under dimensionality reduction techniques since the number of principal components (PC) used is often less than the number of original features.

2.6. Class imbalance

Even though the importance of having high sampling measurement has been discussed, an integral data quality issue remains unaddressed. Even if access to a sufficient volume of data is assumed, good performance may not be guaranteed if each phenomenon under investigation is not sufficiently represented in the dataset.

This is also referred to the class imbalance issue in training machine learning models. This occurs when there exists an imbalance in class representation in the data and is a commonly-occurring phenomenon in real world datasets. For example, when dealing with drilling vibrations, the occurrence of simple torsional vibrations outweighs severe sticking conditions. Training a classifier using an imbalanced training dataset biases it towards the majority class; this inevitably results in inferior performance on the test dataset. Srivastava and Teodoriu (2020) implemented a multi-level classifier for drilling vibrations using surface measurements on an imbalanced dataset and observed a drastic reduction in model performance.

Data imbalance can be dealt with upsampling minority class, downsampling majority class or by assigning a weight component based on class imbalance severity. Hegde et al. (2019) observed an increase in model precision when class imbalance ratio (ratio of minority to majority labels) was closer to 1. This paper discussed the impact of class imbalance on performance of different classifiers and the interdependence of sampling frequency and class imbalance on model performance.

3. Stick–slip vibrations

Stick–slip vibrations is the most concerning vibration issue faced during drilling due to the drastic downhole RPM changes caused. The paper focuses on stick–slip vibration classification amongst other vibration modes in the drill string. These vibrations are torsional in nature where the drillstring lacks the torsional strength to overcome the frictional force between the formation and the bit cutters. In some cases, this phenomenon can also happen between stabilizers and the borehole wall (Baumgartner and van Oort, 2014).

Fig. 4 displays stick–slip vibrations, a severe form of torsional oscillations. Unlike torsional oscillations, stick–slip vibrations comprises of the “stick” and “slip” phase in which the bit speed varies from zero (bit sticking) to several times the surface RPM. Bit sticking can lead to potential bit and tool damage, over-torquing of connections, drill string twist-off and lower rate of penetration. An increment in RPM and a decrement in WOB is an industry practice to deal with stick–slip vibrations.

A common practice for identification of stick–slip vibrations is to analyze surface measurements for peaks in surface torque because of sudden release of built-up energy in the string. However, due to delayed and dampened surface responses, oftentimes, severe stick–slip situations can be interpreted as simple torsional vibrations. While torsional vibrations are inevitable while drilling, stick–slip vibrations could have long periods of sticking causing high spikes in downhole RPM. Sharma et al. (2020), through experimental tests showed the susceptibility of lower rotational speeds to severe sticking. Additionally, during severe sticking situations, while surface rotations are constant, downhole RPM was observed up to 15 times the surface RPM.

Fig. 5 highlights another technique to detect downhole vibrations in which a frequency-domain model is used instead of a time-domain model to isolate vibrations in surface torque due to changing parameters and drilling conditions. The frequency plots are fine-tuned to pick up frequencies below 5 Hz. When plot for the two drilling sections highlighted in green, the frequency plots show a clear difference in normal drilling vs. drilling during stick–slip period. The presence of low frequency peaks highlights the presence of stick–slip vibrations. The amplitude peaks from the frequency plots can be further utilized as an additional feature for training a classifier (Millan et al., 2019). Amongst other vibration modes, stick–slip vibrations are the lowest frequency vibrations in the range of 0.001 to 5 Hz. However, surface data collection rate in the industry is generally 1 Hz for continuous measurement. The sampling frequency for downhole data acquisition ranges from 0.5 Hz to 400 Hz as seen in literature. With lack of standardized data sampling frequency, this paper addresses the impact, if any, of sampling frequency on the data collection and ultimately, the accuracy of ML classification models for torsional vibration severity.

4. Experimental setup for vibration data set

Downscaled laboratory tests are instrumental in characterizing drilling vibrations, finetuning complex drillstring models and incorporating real-time downhole measurements into vibration analysis. The vibration dataset used in the classification model is experimentally generated at the drilling vibrations laboratory at the University of Oklahoma. The experimental test rig used is the longest vertical setup in the world to study downhole vibrations (Srivastava and Teodoriu, 2019). Some of the unique features of this setup include:

- An embedded drilling simulator for parametric analysis of drilling vibrations.

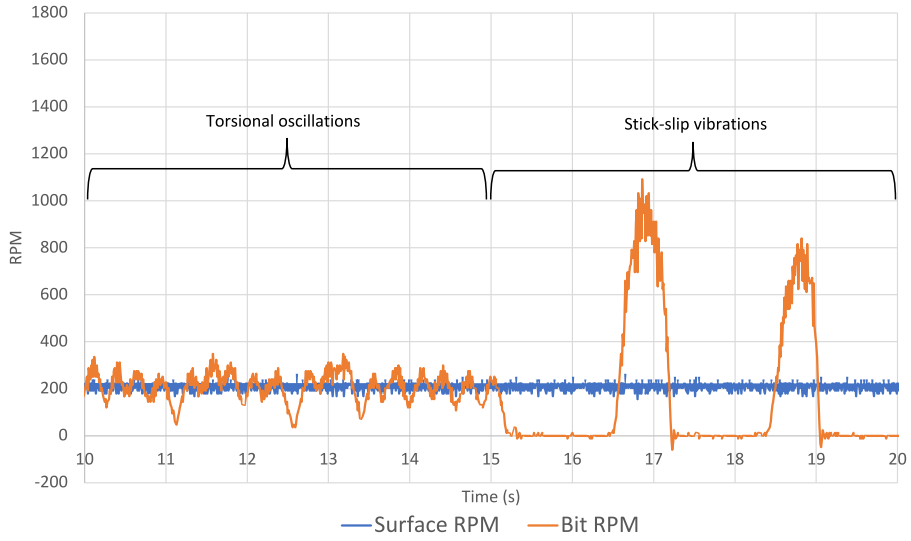

Fig. 4. Torsional oscillations vs. Stick–slip vibrations with stick and slip phase identified.

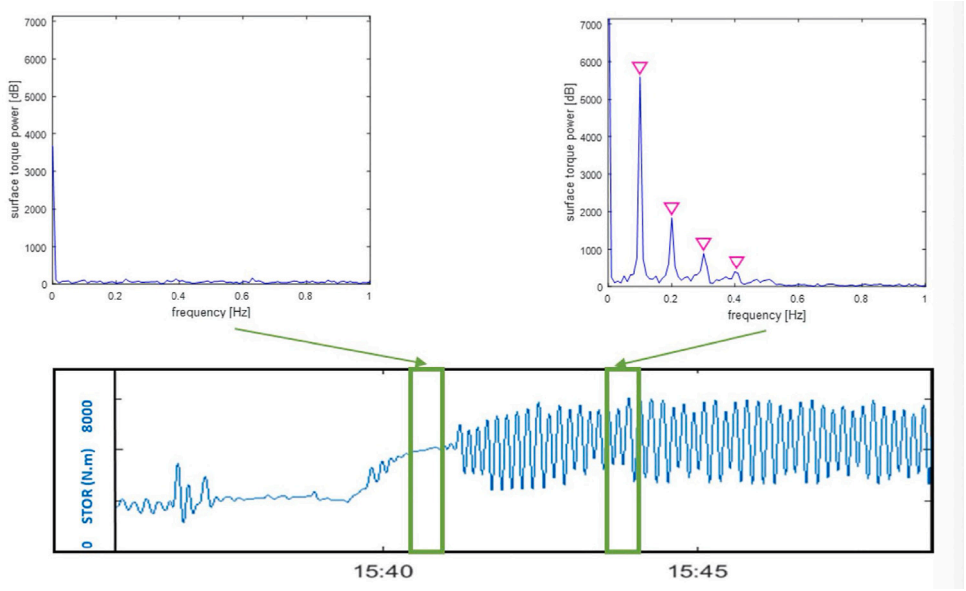

Fig. 5. Frequency plots of surface torque signal during drilling under normal circumstances and during stick–slip vibrations (Millan et al., 2019). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Number of samples collected from each of the three experiments at different sampling frequencies.

Sampling frequency	# Samples/test
1 Hz	133
10 Hz	1310
100 Hz	12,940

- An electromagnetic brake to introduce repeatable and known downhole torque fluctuations to induce torsional vibrations. The brake sits on a hexapod system which is capable of precise, programmable bit movement.
- Advanced sensors and data acquisition capabilities to vary sampling frequencies in the range of 1 to 100 Hz.

The experimental setup (see Fig. 6) is used for varying drilling conditions to study characteristics of torsional vibrations. Using the electromagnetic brake, the sticking period of the bit has been varied and its response has been recorded. The electromagnetic brake also provides an additional feature of repeatability in testing conditions, which is crucial in testing of high frequency and low frequency stick–slip conditions (Sharma et al., 2020). The experimental setup is instrumental in generating training datasets with varying resolution for prediction of stick–slip vibration severity. Details of the experimental design include:

- Testing conditions: RPM = 200, WOB = 5N, Surface Torque = 300–500 mNm
- Testing scenario: Torsional vibrations, Mild stick–slip (bit sticking under 1 s) and Severe stick–slip (bit sticking over 1 s).
- The tests have been performed at sampling frequencies of 1, 10 and 100 Hz, while keeping the total test duration of the constant. This results in a three datasets, each generated using identical testing conditions, but with different number of samples. The sizes of the three resulting datasets are given in Table 1.

An example of torsional vibrations is seen in Fig. 7 where bit RPM oscillates around 200RPM without actually sticking. Mild stick–slip has

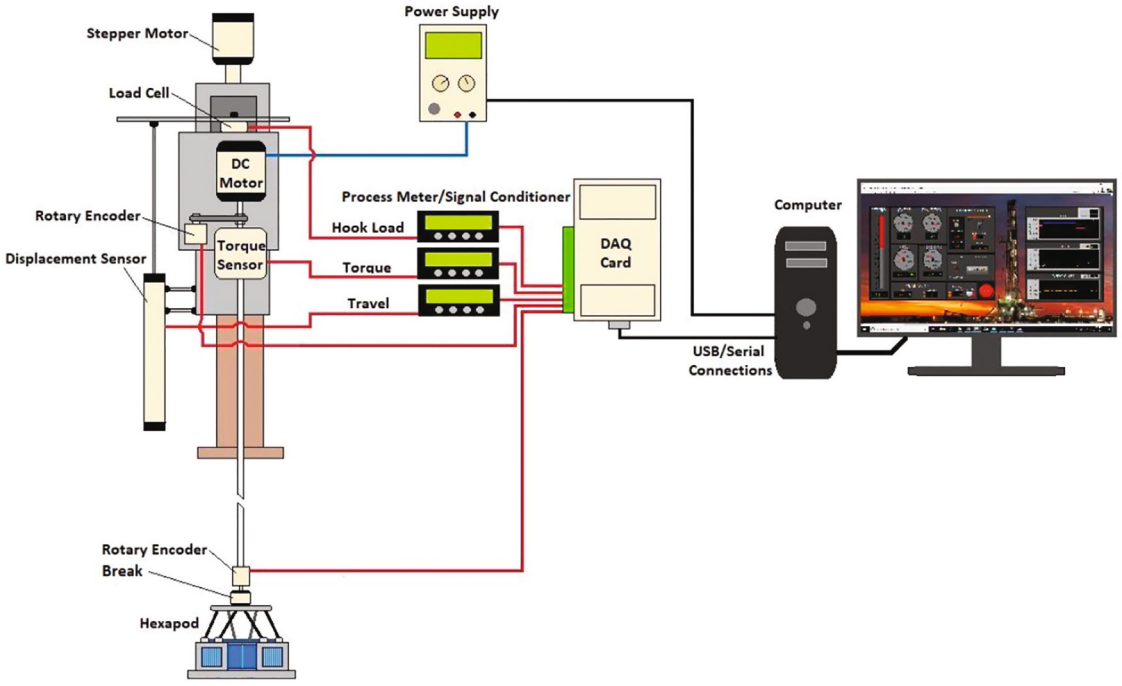

Fig. 6. Schematic of the drilling vibration experimental test rig at the University of Oklahoma.

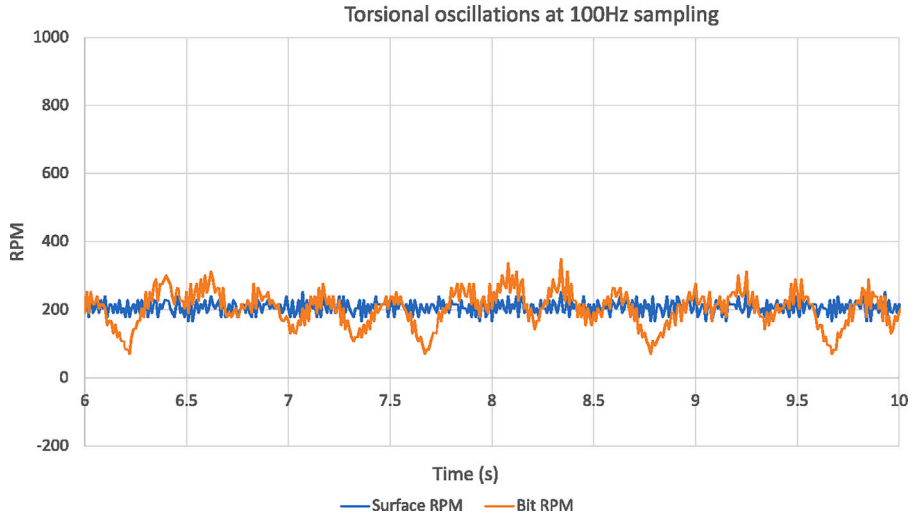

Fig. 7. Example of torsional oscillation test at 100 Hz sampling rate.

been shown in Fig. 8 where bit sticking is observed when bit RPM is momentarily at 0. Bit sticking of 1 s and lower is observed in the mild stick–slip test. Similarly, a case of severe bit sticking is seen in Fig. 9 where up to 3 s of bit sticking is observed with a consequent but drastic bit slip phase.

5. Experiments and results

In this section we describe our classification experiments and their results. We discuss how choices made at different stages in the data collection and pre-processing steps affect the performance of several popular classifier models. We also investigate the reasons for difference in performance of classifiers, wherever applicable. Apart from the experimental variable being examined in each subsection, the rest of the machine learning pipeline remains unaltered across different experiments. Specifically, we use the following six classifiers—Support Vector Machine (SVM) (Boser et al., 1992), Linear Discriminant Analysis (LDA) (Fisher, 1936), Logistic Regression (LR) (Hosmer Jr. et al., 2013), Naïve Bayes (NB) (Rish et al., 2001), Classification and Regression Trees (CART) (Breiman et al., 1984), k-Nearest Neighbors (KNN) (Fix and Hodges, 1989). Further,

- For each classifier we run 100 simulations. Each simulation involves shuffling and resampling the training and testing data, followed by retraining each classifier.
- We utilize a 80:20 split for the training:test ratio.
- We utilize two metrics to evaluate our trained classifiers—Test Accuracy and F1-Score.

Upon evaluating a classifier using the test data, we can obtain the number of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) examples. Given these, the test accuracy and F1-score can be computed as

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

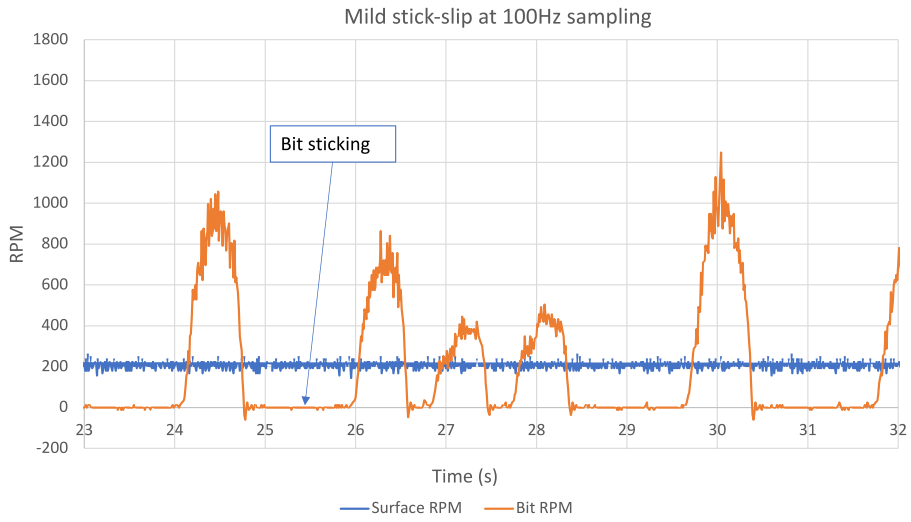

Fig. 8. Example of mild stick–slip test at 100 Hz sampling rate.

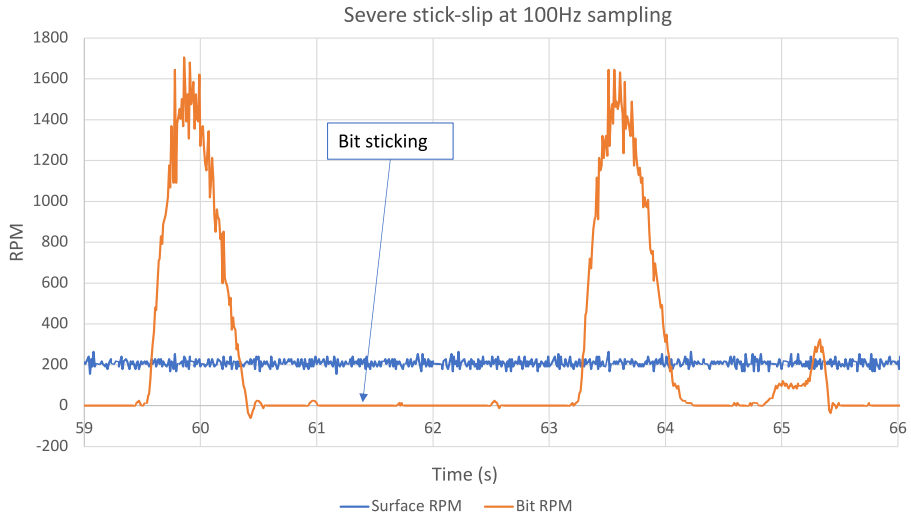

Fig. 9. Example of severe stick–slip test at 100 Hz sampling rate.

$$\text{F1-Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{5}$$

5.1. Sampling frequency

The goal of this experiment is to study the effect the sampling rate of the training data has on final classification accuracy of a variety of classifiers, and test if certain machine learning classifiers are more robust to variations in sampling rate than others. As described in the previous section, our experimental data was collected at three different sampling frequencies – 1 Hz, 10 Hz, and 100 Hz. The results for this experiment are shown in Fig. 10. In general, the mean and median test accuracies of all classifiers are seen to consistently improve when the training data is collected at a higher sampling frequency. But more significantly, there is an even more pronounced trend in the variance of test accuracies at higher sampling rates, which enable higher confidence levels in the final test accuracies. The increase in median test accuracy, and the decrease in its variance is expected since higher sampling frequencies capture more data samples per time interval than lower sampling frequencies.

Further, the difference in the test accuracies between different classifiers with 100 Hz data is insignificant. This can be attributed to the fact that collection of more data enables capturing more patterns that uniquely identify all three types of vibrations – torsional, mild stick–slip, and severe stick–slip. The way this manifests in our data is by way of increased linear separability between the three classes of data at 100 Hz. This is validated by the fact that for the 100 Hz data the performance of linear classifiers such as LDA is comparable to that of non-linear classifiers such as SVM, CART, and KNN . Finally, the conclusions we draw from this experiment are

- All other parameters being equal, better detection and classification of stick–slip can be achieved at higher sampling frequencies.
- Given data sampled at an adequately high sampling frequency, the choice of classifier does not affect the classification results significantly.
- However, a higher sampling frequency also entails higher data storage costs, and the right trade-off between performance and storage costs must be determined on a per-case basis for best results.

5.2. Data labeling

The goal of this experiment is to investigate the impact of different labeling techniques on the performance of various classifiers. Note, the goal is not to investigate if one labeling technique is better than another. In fact, it is desired that the choice of labeling technique

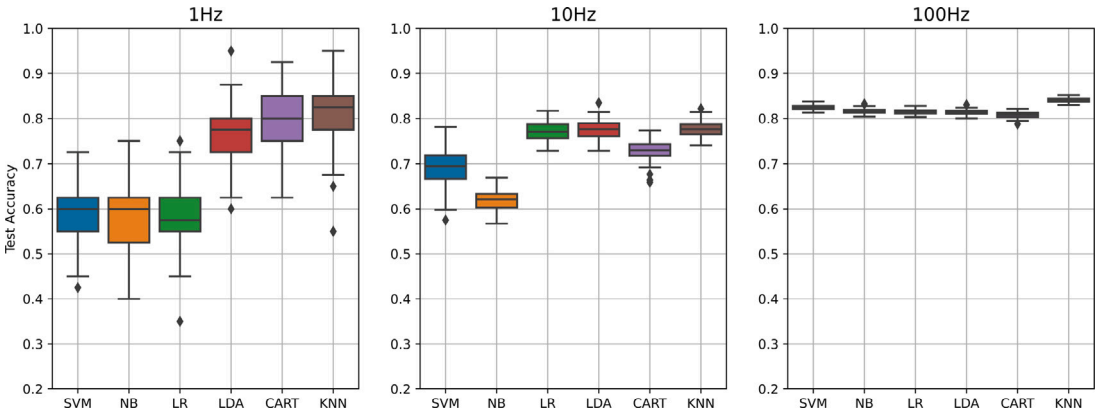

Fig. 10. Comparison of test accuracies of different classifiers trained on data collected at different sampling frequencies.

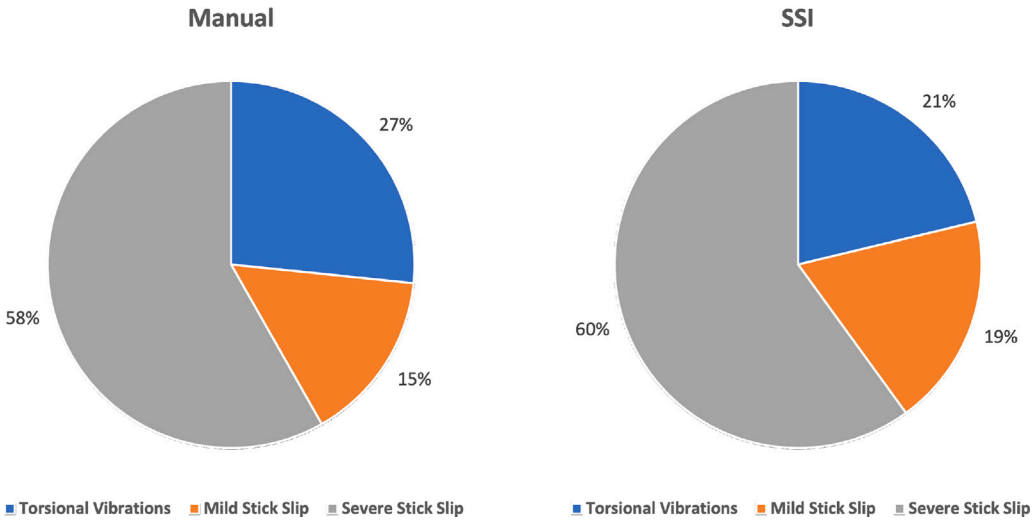

Fig. 11. Comparison of label distributions obtained using the two labeling techniques.

not impact classification performance significantly, since all labeling techniques should ideally agree on the ground truth for them to be considered effective. From the techniques listed under 'Labeling' in Section 2, we compare manual and SSI-based labeling. For our experimental data, manual labeling was performed using subject matter supervision. SSI-based labeling is automated and was performed using a sliding time window method, where the Nth time window w_N is labeled as below

$$\text{Label}(w_N) = \begin{cases} \text{Torsional Vibration} & \text{SSI}(w_N) < 1.0 \\ \text{Mild Stick–Slip} & 1.0 \leq \text{SSI}(w_N) < 3.0 \\ \text{Severe Stick–Slip} & \text{SSI}(w_N) \geq 3.0 \end{cases} \quad (6)$$

where $\text{SSI}(w_N)$ is computed as described in Section 2 using a siding time window method. Further, having already shown that a 100 Hz sampling rate is more effective than 1 Hz and 10 Hz, we only present results for the 100 Hz data, thus choosing to restrict the comparison to the choice of labeling, and not that of sampling frequency. This is also done for sake of clarity. In Fig. 11 we compare the label distributions for 100 Hz data generated by both labeling techniques that we employ, and the corresponding results for this experiment are presented in Fig. 12.

The results suggest that SSI-based labeling yield consistently lower test accuracies for all tested classifiers than even data labeled by a subject matter expert. This is likely due to the fact that the SSI, while easy-to-interpret, is actually an empirical metric and may not always be accurate, and as a result yields lower classification performance than manual labeling for all classifiers.

Thus, we draw the following conclusions from this experiment:

- Manual labeling carried out by a subject matter expert, while costly, is more accurate.
- Automated labeling techniques using empirical metrics such as SSI, while easy-to-interpret and time-efficient, may actually be erroneous. We recommend an additional 'label inspection' step carried out by a subject matter expert when using such labeling techniques.

5.3. Feature extraction

The goal of this experiment is to investigate the impact of different feature extraction techniques on the performance of different classifiers. We compare two feature extraction techniques – statistical and spectral feature extraction, listed in Section 2. Both approaches involve using a sliding time window. We extract the following statistical features from each time window – mean μ, standard deviation σ, skewness γ, kurtosis κ, minimum, and maximum. Further, these statistical features are extracted for the following sensor data – Torque (mNm), RPM T (top drive), RPM B (bit), and WOB (Weight on bit), resulting in a total of 24 features (6 statistical features × 4 sensors). Given the variety of approaches available for spectral feature extraction, we adopt a straightforward approach due to the simple nature of patterns in our data. We employ the fact that the torsional vibrations and the stick–slip phenomenon both result in periodic patterns in the RPM B data which occur under 1 Hz. Thus, we extract three spectral features, each of which denotes the strength of the periodicity in one of three frequency ranges - ≤0.5 Hz, 0.5–1 Hz, ≥1 Hz. The strength of different

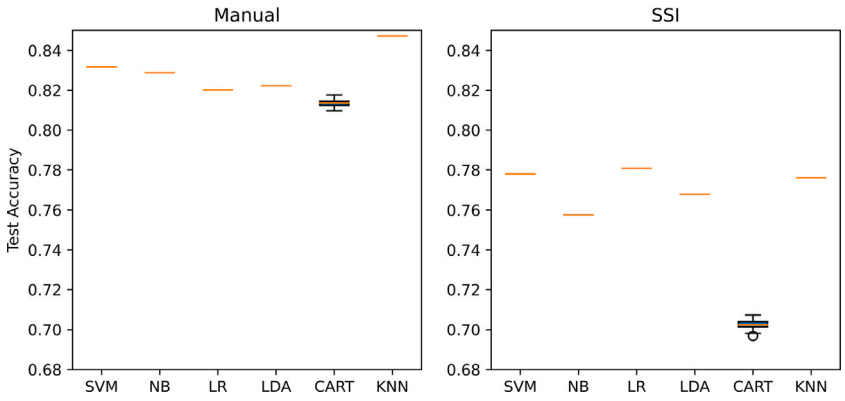

Fig. 12. Comparison of test accuracies of different classifiers trained on data labeled using two different techniques.

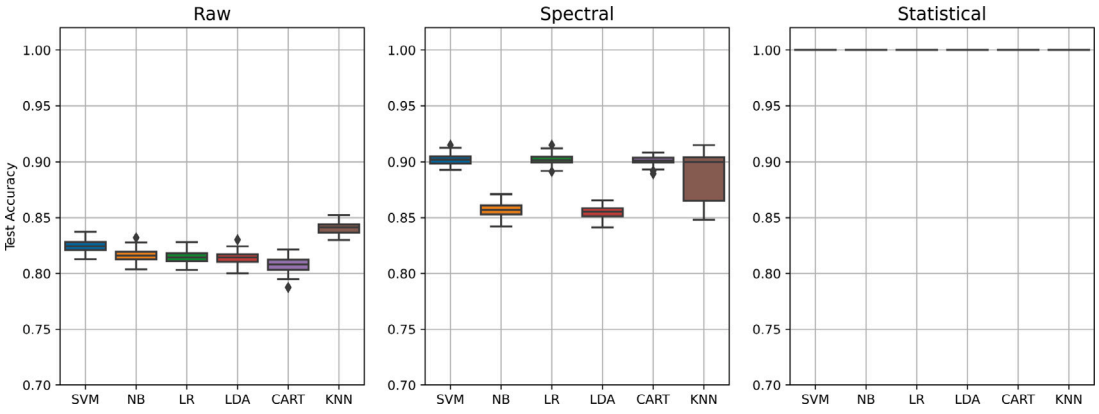

Fig. 13. Comparison of test accuracies of different classifiers trained on data obtained from different feature extraction techniques.

frequency components is computed using FFT. Our hypothesis is that the presence of frequency components in these ranges can help distinguish between torsional vibrations, mild stick–slip and severe stick–slip. We also present results obtained using raw features, without performing feature extraction for comparison. Note, we use data sampled at 100 Hz and labeled manually by a subject matter expert.

The results for this experiment are shown in Fig. 13. In general, both techniques for feature extraction result in higher classification performance than using raw features. It is worth noting that using even simple, FFT-based spectral features can result in classification performance gains. This shows that the extracted featured successfully captured the phenomenon manifesting in the data. More interestingly, using statistical features results in perfect classification performance on even the test data, across all classifiers. While such a result is typically indicative of model overfitting, that is not the case here. Upon investigating the pairwise scatterplots for the statistical features we find that certain features such as minimum, maximum, and standard deviation of RPM B (bit) are good enough to make all three classes of data perfectly linearly-separable. A subset of all pairwise scatterplots is shown in Fig. 14.

5.4. Class imbalance

The goal of this experiment is to investigate the impact of imbalanced data on the classification performance of different classifiers. To reiterate, the imbalance ratio of a data set is the ratio of the number of samples in the majority class to those in the minority class. In drilling applications, it is expected that the machinery will operate under normal operating conditions (torsional vibrations) more frequently, than it will under extreme operating conditions (mild and severe stick–slip), hence data sets obtained from drilling operations are expected to be naturally imbalanced. Thus, for our experiments we consider torsional vibrations to be the majority class, and mild and severe stick–slip to be minority classes.

Further, we consider three levels of imbalance for this experiment – 1:1, 10:1, and 100:1. To generate data sets with these imbalance rations, we fix the size of the minority class (mild and severe stick–slip), and generate increasing number of samples for the minority class (torsional vibrations). In general, to compute the required number of majority samples for each level of imbalance we employ the following relationship derived from the definition of imbalance ratio

$$N_{\text{torsional}} = \text{IR} \times (N_{\text{mild SS}} + N_{\text{severe SS}}) \tag{7}$$

where $N_{\text{torsional}}$ is the number of samples in the majority class, IR is the desired imbalance ratio, and $N_{\text{mild SS}}$ and $N_{\text{severe SS}}$ are the number of samples in each of the minority classes.

For this experiment, we continue to use only the data set sampled at 100 Hz since that was deemed to contain most information. However, despite statistical features yielding perfect classification performance, we restrict our experiments to raw features. We do this to prevent the perfect classification accuracies from masking the impact of different levels of class imbalance in our experimental results.

Results are shown in Fig. 15 (top). At first glance, it appears that the test accuracy of all classifiers improves as the imbalance ratio increases. However, one of the known pitfalls of using test accuracy as an evaluation metric is that it does not take into account the imbalance in the training data. Thus, we also present the F1-scores for each case in Fig. 15 (bottom). Since F1-score is computed as the harmonic mean of the Precision and Recall, it does take into account the class imbalance. Consequently, the F1-score provides better insight into the

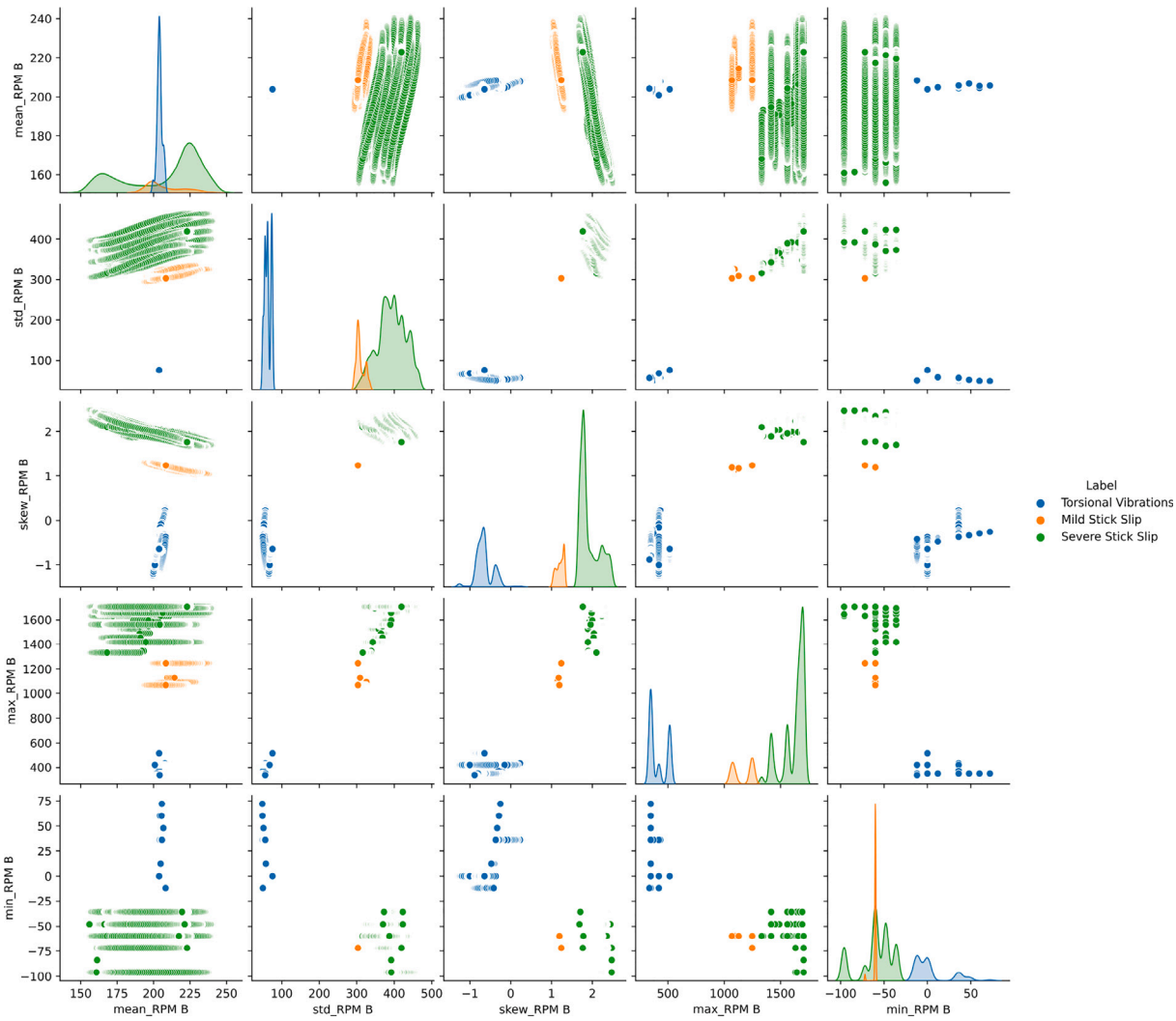

Fig. 14. Pairwise scatterplots for different statistical features using statistical features makes the data classes linearly-separable.

performance of different classifiers, and hence we consider it for the remainder of this discussion. NB, LR, and LDA yield generally lower F1 scores for higher levels of imbalance. This is expected for the NB classifier whose performance strongly depends on the prior probability distribution which, for imbalanced data sets, assigns lower probabilities to the minority class data, and thereby results in lower detections of the minority class. While the LDA model does not depend on the prior probabilities, unlike NB, it does not explicitly account for class imbalance either. This combined with the use of raw features for this experiment results in lower F1-scores for LDA. In fact, even the lower performance of LR at higher imbalance ratios can be attributed to the fact that LR does not explicitly account for class imbalance. Specifically, an imbalanced data set can artificially skew the classification threshold learned by the LR model, towards the majority class.

The performance of CART clearly stands-out in these results with test accuracies and F1-scores greater than 0.80 for all levels of imbalance. The CART algorithm is not known for its robustness to imbalanced data. CART has been found to fail occasionally when trained using imbalanced data (Liu et al., 2010). Hence, in this case, we may attribute this competitive performance to the fact that the data can be easily classified into one of three classes using only a few splits in CART's decision tree. This is also in line with the previous experiments where CART outperformed other classification models for different sampling frequencies, labeling techniques, and feature representations.

6. Discussion and conclusion

This paper explores the implications of drilling data quality on efficient application of ML techniques. The following conclusions can be made:

- High resolution data is data sampled at high sampling frequency. Through extensive testing, we have determined that high resolution data (minimum of 100 Hz) is crucial to better performing classifiers with high accuracy and low variance.
- Data labeling is an important step for machine learning models. Manual labeling techniques provide accurate labeling while being time and resource intensive. While automated techniques are faster to implement and provide higher testing accuracies, they are prone to errors and need a 'label verification' step, as suggested in Section 5.2. In either case, involvement of a subject matter expertise is crucial.
- Adding features to raw data improves the model's robustness and its classification performance across all classifiers used. Spectral features contribute to an average 5% increase in performance whereas statistical features account for greater than 15% model improvement. Adding a good mix of statistical and spectral features to time series data is highly recommended.
- Handling data imbalance helps improve classification accuracy, especially when dealing with low frequency surface and downhole measurements.

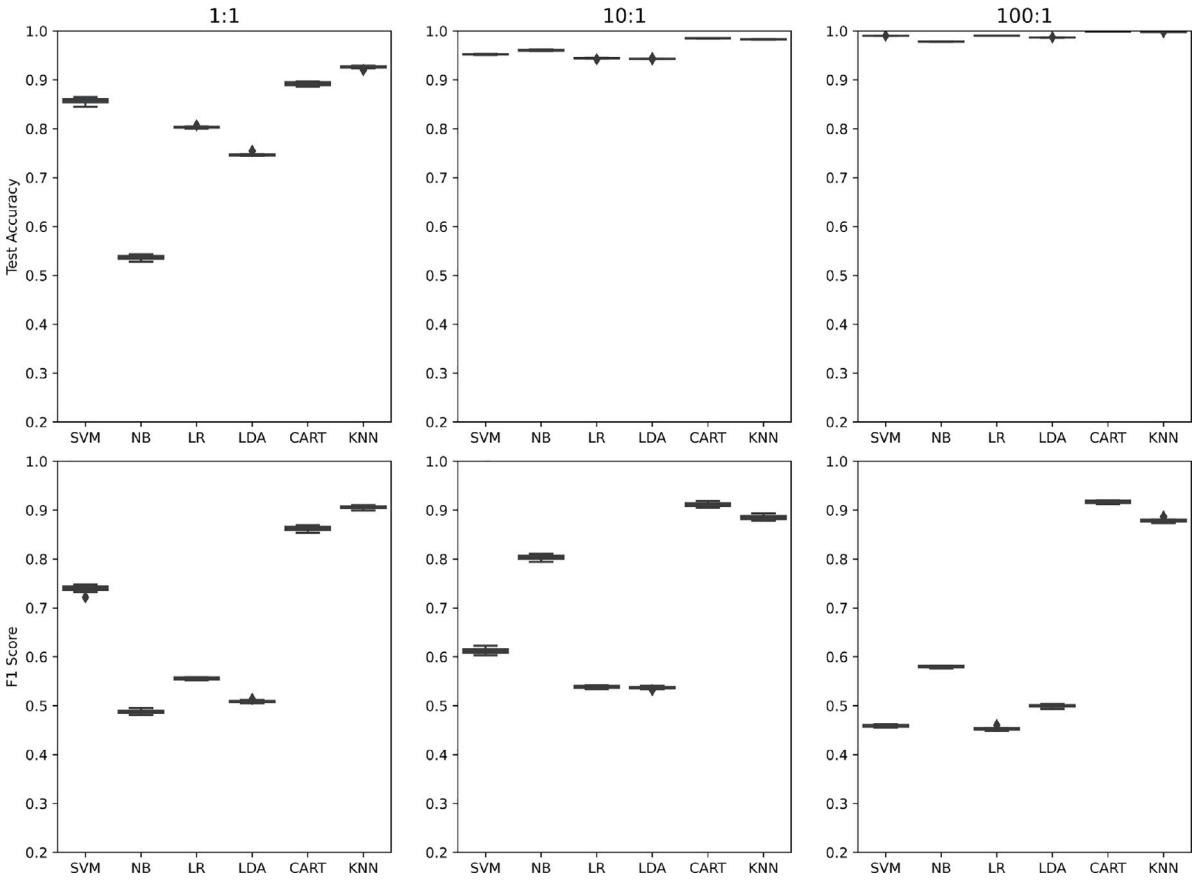

Fig. 15. Test accuracy and F1 scores for different classifiers trained on data containing different levels of class imbalance.

- Performance of CART, a decision tree model has been better than other classifiers for all testing scenarios.
- Using correct feature representation for each data set can yield competitive results with all classifiers including simple ones like LDA, so that should be the goal.
- Another idea is to use an ensemble of classifiers all of which have a different fundamental formulation. This can ensure that even if any one of them overfits to the data, not all do.

CRediT authorship contribution statement

Saket Srivastava: Conceptualization, Methodology, Experimental, Formal analysis, Writing – original draft, Writing – review & editing, Editing reviewer comments. **Rushit N. Shah:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Editing reviewer comments. **Catalin Teodoriu:** Writing – review & editing, Editing reviewer comments, Supervision of experiments, Project acquisition. **Aditya Sharma:** Experimental, Writing – review & editing, Editing reviewer comments, Supervision of experiments.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Catalin Teodoriu reports financial support was provided by Helmerich and Payne.

Data availability

The data that has been used is confidential.

Acknowledgments

We thank Helmerich and Payne Inc., USA for supporting and funding the experimental study in the Drilling Vibration Laboratory at the University of Oklahoma.

References

Abu-Mahfouz, I., 2003. Drilling wear detection and classification using vibration signals and artificial neural network. Int. J. Mach. Tools Manuf. 43 (7), 707–720.

Al Gharbi, S., Ahmed, M., ElKatatny, S., 2018. Use metaheuristics to improve the quality of drilling real-time data for advance artificial intelligent and machine learning modeling. Case study: cleanse hook-load real-time data. In: Abu Dhabi International Petroleum Exhibition & Conference. OnePetro.

Baumgartner, T., van Oort, E., 2014. Pure and coupled drill string vibration pattern recognition in high frequency downhole data. In: SPE Annual Technical Conference and Exhibition. OnePetro.

Bello, M., Nápoles, G., Sánchez, R., Bello, R., Vanhoof, K., 2020. Deep neural network to extract high-level features and labels in multi-label classification problems. Neurocomputing 413, 259–270.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. pp. 144–152.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees–crc press. Boca Raton, Florida.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2 (2), 121–167.

Chandel, A.K., Patel, R.K., 2013. Bearing fault classification based on wavelet transform and artificial neural network. IETE J. Res. 59 (3), 219–225.

Chandola, V., Banerjee, A., Kumar, V., 2007. Outlier detection: A survey. ACM Comput. Surv. 14, 15.

Domingues, R., Filippone, M., Michiardi, P., Zouaoui, J., 2018. A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognit. 74, 406–421.

Dong, G., Chen, P., 2016. A review of the evaluation, control, and application technologies for drill string vibrations and shocks in oil and gas well. Shock Vib. 2016.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7 (2), 179–188.

Fix, E., Hodges, J.L., 1989. Discriminatory analysis. Nonparametric discrimination: Consistency properties. Int. Statist. Rev./Revue Int. Statist. 57 (3), 238–247.

Ghojogh, B., Samad, M.N., Mashhadi, S.A., Kapoor, T., Ali, W., Karray, F., Crowley, M., 2019. Feature selection and feature extraction in pattern analysis: A literature review. arXiv preprint arXiv:1905.02845.

Hegde, C., Millwater, H., Gray, K., 2019. Classification of drilling stick slip severity using machine learning. J. Pet. Sci. Eng. 179, 1023–1036.

Hodge, V., Austin, J., 2004. A survey of outlier detection methodologies. Artif. Intell. Rev. 22 (2), 85–126.

Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression. Vol. 398, John Wiley & Sons.

Jansen, J.D., van den Steen, L., Zachariasen, E., 1995. Active damping of torsional drillstring vibrations with a hydraulic top drive. SPE Drill. Complet. 10 (04), 250–254.

Javanmardi, K., Gaspard, D., 1992. Application of soft-torque rotary table in mobile bay. In: IADC/SPE Drilling Conference. OnePetro.

Karkoub, M., Abdel-Magid, Y., Balachandran, B., 2009. Drill-string torsional vibration suppression using GA optimized controllers. J. Can. Pet. Technol. 48 (12), 32–38.

Krama, A., Gharib, M., Refaat, S.S., Palazzolo, A., 2021. Design and hardware in-the-loop validation of an effective super-twisting controller for stick-slip suppression in drill-string systems. J. Dyn. Syst. Meas. Control 143 (11), 111008.

Le Cam, L., 1990. Maximum likelihood: an introduction. Int. Statist. Rev./Revue Int. Statist. 153–171.

Liu, W., Chawla, S., Cieslak, D.A., Chawla, N.V., 2010. A robust decision tree algorithm for imbalanced data sets. In: Proceedings of the 2010 SIAM International Conference on Data Mining. SIAM, pp. 766–777.

Middleton, D., 1999. Non-Gaussian noise models in signal processing for telecommunications: new methods an results for class A and class B noise models. IEEE Trans. Inform. Theory 45 (4), 1129–1149.

Millan, E., Ringer, M., Boualleg, R., Li, D., 2019. Real-time drillstring vibration characterization using machine learning. In: SPE/IADC International Drilling Conference and Exhibition. OnePetro.

Mohan, J., Krishnaveni, V., Guo, Y., 2014. A survey on the magnetic resonance image denoising methods. Biomed. Signal Process. Control 9, 56–69.

Myung, I.J., 2003. Tutorial on maximum likelihood estimation. J. Math. Psych. 47 (1), 90–100.

Navamani, T., 2019. Efficient deep learning approaches for health informatics. In: Deep Learning and Parallel Computing Environment for Bioengineering Systems. Elsevier, pp. 123–137.

Noshi, C.I., Schubert, J.J., 2018. The role of machine learning in drilling operations; a review. In: SPE/AAPG Eastern Regional Meeting. OnePetro.

Okoli, P., Cruz Vega, J., Shor, R., 2019. Estimating downhole vibration via machine learning techniques using only surface drilling parameters. In: SPE Western Regional Meeting. OnePetro.

Otalvora, W.C., AlKhudiri, M., Alsanie, F., Mathew, B., 2016. A comprehensive approach to measure the realtime data quality using key performance indicators. In: SPE Annual Technical Conference and Exhibition. OnePetro.

Patil, P.A., 2013. Investigation of Torsional Vibrations in a Drillstring using Modeling and Laboratory Experimentation. Papierflieger-Verlag.

Pavone, D., Desplans, J., 1994. Application of high sampling rate downhole measurements for analysis and cure of stick-slip in drilling. In: SPE Annual Technical Conference and Exhibition. OnePetro.

Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. 2 (11), 559–572.

Ramaswamy, S., Rastogi, R., Shim, K., 2000. Efficient algorithms for mining outliers from large data sets. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. pp. 427–438.

Rish, I., et al., 2001. An empirical study of the naive Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. Vol. 3, (22), pp. 41–46.

Sharma, A., Srivastava, S., Teodoriu, C., 2020. Experimental design, instrumentation, and testing of a laboratory-scale test rig for torsional vibrations—The next generation. Energies 13 (18), 4750.

Srivastava, S., Shah, R.N., Teodoriu, C., 2021. Natural language processing based information extraction from drilling reports to classify drilling dysfunction severity. Geothermal Resources Council Transactions 45, 1311–1323, URL https://www.geothermal-library.org/index.php?mode=pubs&action=view&record=1034454.

Srivastava, S., Teodoriu, C., 2019. An extensive review of laboratory scaled experimental setups for studying drill string vibrations and the way forward. J. Pet. Sci. Eng. 182, 106272. http://dx.doi.org/10.1016/j.petrol.2019.106272, URL https://www.sciencedirect.com/science/article/pii/S092041051930693X.

Srivastava, S., Teodoriu, C., 2020. Characterizing drilling vibrations by interlinking surface data, drillstring design and lithology of rock in utah FORGE deep test well 58-32. In: Stanford Geothermal Workshop.

Vaseghi, S.V., 2008. Advanced Digital Signal Processing and Noise Reduction. John Wiley & Sons.

Wang, H., Bah, M.J., Hammad, M., 2019. Progress in outlier detection techniques: A survey. Ieee Access 7, 107964–108000.

Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometr. Intell. Lab. Syst. 2 (1–3), 37–52.

Yang, H., Mathew, J., Ma, L., 2003. Vibration feature extraction techniques for fault diagnosis of rotating machinery: a literature survey. In: Asia-Pacific Vibration Conference. (42460), pp. 801–807.

Yen, G.G., Lin, K.-C., 2000. Wavelet packet feature extraction for vibration monitoring. IEEE Trans. Ind. Electron. 47 (3), 650–667.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., Saeed, J., 2020. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. J. Appl. Sci. Technol. Trends 1 (2), 56–70.

Zha, Y., Pham, S., 2018. Monitoring downhole drilling vibrations using surface data through deep learning. In: SEG Technical Program Expanded Abstracts 2018. Society of Exploration Geophysicists, pp. 2101–2105.

Zhang, H., Zeng, Y., Bao, H., Liao, L., Song, J., Huang, Z., Chen, X., Wang, Z., Xu, Y., Jin, X., 2020. Drilling and completion anomaly detection in daily reports by deep learning and natural language processing techniques. In: SPE/AAPG/SEG Unconventional Resources Technology Conference. OnePetro.

Zhao, Y., Nasrullah, Z., Li, Z., 2019. Pyod: A python toolbox for scalable outlier detection. arXiv preprint arXiv:1901.01588.

Zhu, X., Wu, X., 2004. Class noise vs. attribute noise: A quantitative study. Artif. Intell. Rev. 22 (3), 177–210.