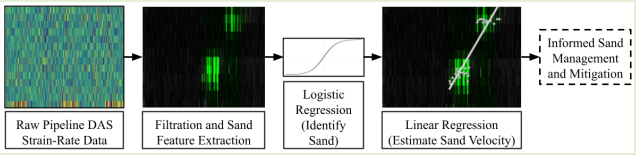
Machine Learning for Automated Sand Transport Monitoring in a Pipeline Using Distributed Acoustic Sensor Data

Harrison Gietz, Jyotsna Sharma, and Mayank Tyagi

***Abstract*—Uncontrolled sand production presents a substantial challenge to wellbore and pipeline integrity and efficiency of hydrocarbon production operations, often leading to equipment damage and compromised productivity. Traditional sand detection methods on the surface alert operators to sanding issues, but they are often a lagging indicator of downhole sanding events and do not provide precise identification of the problematic reservoir zones. Addressing this limitation, this study harnesses a combination of efficient signal processing and machine learning (ML) to analyze data from optical-fiber-based distributed acoustic sensors (DASs), thus serving as the first instance (to the authors' knowledge) of an automated and real-time approach to monitoring sand migration patterns and velocity estimation along a pipeline. The DAS data acquired from an experimental flow loop were analyzed using the developed algorithms, and the performance was evaluated for different flow speeds and sand ingress scenarios. The model training only required roughly 25% of the total data, and the remaining data were used to demonstrate the generalizability of the proposed ML models, through blind testing. Analysis of eight distinct experimental datasets provided a credible approximation of sand velocities, corroborating previous studies and theoretical expectations. Using the best-performing trained models, sand detection accuracies attained an average of 93.4% on blind testing data, along with sand velocity estimates with an average error of 10.1% from analytical results. The results from this study validate the use of DAS combined with ML for autonomous sand monitoring and flow characterization, both for boosting well performance and concurrently mitigating environmental hazards.**



***Index Terms*— Distributed acoustic sensing, distributed fiber-optic sensing, machine learning (ML), sand detection.**

I. Introduction

SAND production poses a significant asset integrity challenge in the realm of oil and gas extraction [1], [2]. These concerns are far-reaching and large in scale, costing the oil and gas industry millions of dollars annually in sand-related expenses such as production choke-backs, infrastructure cleaning, and equipment repair [3]. The negative side effects that arise from sand production are not purely economic, however; sand production also entails risks of significant potential environmental harm and contamination that can result from infrastructure erosion and failure [4], [5], [6]. For instance, well casing plays a role in well control and safety, as it helps in containing pressure and fluids within the well, reducing the risk of blowouts or uncontrolled releases of oil, gas, or other substances. However, unchecked sanding can lead to loss of well casing integrity, and as a result, monitoring and controlling sand production is of crucial importance for mitigating environmental harms. In addition, excessive sand production can prematurely curtail the production life of a reservoir, necessitating remedial operations, including the drilling of new wells, and resulting in the underutilization of irreplaceable oil and gas resources [3].

In addressing these challenges, monitoring the ingress location and velocity of sand is crucial. It enables operators to modulate fluid velocities below the erosional velocity limits for the pipeline material, thus maintaining safe production without compromising the integrity of the well and associated equipment, as well as providing insights on sanding zones for targeted mitigation. As a result, past work has explored methods of measuring and understanding the sand flow behavior and detection of sand inside of pipelines. However, previous work has used techniques which rely primarily on surface sensors [7], [8], [9], thus offering delayed and sometimes inadequate information of downhole sand ingress and movement further down the pipe.

The use of distributed acoustic sensors (DASs) allows for significant advancement in this arena. DAS data can provide

Manuscript received 10 May 2024; accepted 17 May 2024. Date of publication 6 June 2024; date of current version 16 July 2024. This work was supported by Louisiana Board of Regents and Shell Explorations and Production Company. The associate editor coordinating the review of this article and approving it for publication was Dr. Antreas Theodosiou. *(Corresponding author: Jyotsna Sharma.)*

Harrison Gietz is with the Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 USA.

Jyotsna Sharma and Mayank Tyagi are with the Department of Petroleum Engineering, Louisiana State University, Baton Rouge, LA 70803 USA (e-mail: JSharma@LSU.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSEN.2024.3408140, provided by the authors.

Digital Object Identifier 10.1109/JSEN.2024.3408140

real-time spatiotemporal information along the entire length of the installed fiber in the wellbore or pipeline [10], presenting a rich source of information, ripe for exploration through machine learning (ML). Previous attempts to monitor sand velocity or to leverage DAS data for mitigating the harms of sand production have had some success [1], [11], but fall short along various aspects. For example, no past work has jointly leveraged DAS data for measuring sand velocity, and previous work has required manual analysis of the data [12], an arduous process which also makes the analysis prone to human error, especially given the voluminous nature of distributed fiber-optic sensors due to the high spatial and temporal data resolutions [13], [14].

The proposed approach innovates in this space using ML to automate sand monitoring, requiring minimal data input for training the model to provide reliable sand detection and sand velocity estimates. The success of this method is demonstrated using experimental data that represent various flow rates and sand ingress locations, underscoring the generalizability of the approach. The strong performance of the proposed methodology can be attributed in part to the preprocessing techniques that are leveraged, which involve converting the signal into frequency band energy (FBE) data [15], [16], [17], [18]. A significant benefit of data preprocessing using the FBE approach is the large reduction in DAS data size without losing useful signal information, which helps with computational efficiency and data management, as demonstrated by Tabjula and Sharma [19] and Sharma et al. [18]. FBE processing and feature extraction reduces data size by two orders of magnitude, making this method much more applicable to real-time sand monitoring applications.

The generalizability of the proposed ML models is demonstrated on flow rates that are not encountered during ML model training. Analysis on eight experimental datasets provides a reliable approximation of sand velocities, corroborating previous studies and theoretical expectations. The average sand detection accuracies are as high as 96% (average 93.4%) across the blind testing datasets; the average F_1 score on blind testing (a common metric for ML problems that use imbalanced data) is 0.87; and the average error in sand velocity estimates is within 10.1% of theoretical expectations. The results show higher prediction accuracy when using the sand's characteristic frequency fingerprint, which aligns with the previous manual analysis presented by Shetty et al. [12]. By streamlining this process, this approach of sand detection and monitoring not only has the capacity to bolster production efficiency but also significantly mitigates environmental risks associated with sand production in oil and gas extraction.

This article is arranged as follows. Section I-A summarizes the novel contributions of this study on automated sand monitoring, vis-à-vis the published literature. Section II presents details about the experimental data collection and processing methods used before implementing the ML algorithm. In Section III, a formulation of the ML approaches used for sand detection and velocity estimation is given, followed by details on hyperparameter tuning and model selection. Section IV discusses the performance of the proposed ML algorithms on two blind test datasets, demonstrating the generalizability of the method. Finally, Section V presents a summary of the study's results and outlines potential future research directions for the work.

A. Literature Review and Novelties of This Study

In the realm of downhole sand management, the evolution of detection and monitoring technologies and real-time data analysis methodology is crucial for operational efficiency. This study enhances the past research in this area by integrating ML with DAS measurements for real-time, automated sand monitoring and analysis.

Fiber-optic sensing technology has widespread use in a variety of domains, such as telecommunication, healthcare, aerospace, and environmental monitoring [20], [21], [22], [23]. One such application of fiber optics is that of DASs [24], which provides acoustic monitoring that can be used for a variety of ways, including (in the present study) pipeline monitoring.

Several past studies have demonstrated the successful implementation of ML on DAS datasets for a variety of unique monitoring applications [25], [26], [27], [28], [29]. Although the focus of these previous works was not on solid particulates such as sand, the success of their ML applications in interpreting distributed sensor data sets a precedent for continued work along these lines for the application of sand monitoring.

More specifically, some previous works have demonstrated the application of DAS for downhole sand monitoring, a concept that was first introduced by the case study of Mullens et al. [30] in Azerbaijan. Their work used a combination of distributed temperature logs and DAS data to qualitatively estimate a suspected sand-entry point and produce (unverifiable) estimates of gas phase slip velocity. Given the specific nature of their study (experiments conducted on only three specifically chosen wells in Azerbaijan), the results could not generalize, nor could they be verified or easily expanded upon by future researchers, which is something that the current study addresses. Following Mullens et al. [30], DAS technology was then used for sand monitoring and sand production management by Thiruvenkatanathan et al. [11], who built upon the previous work using a signal processing technique which significantly reduced the computational requirements of using DAS data. Thiruvenkatanathan et al. [11] also introduced the concept of empirically uncovering a sand "acoustic fingerprint": a range of frequencies in DAS data that could be used to distinguish sand from other particulates or fluid signals. Their study uses this knowledge, accrued from DAS data in an experimental flow loop, to form logs of downhole sand ingress location in various wells. A case study by Hasanov et al. [31] found success with a comparable version of this approach; both the works used this technology for further production management, hence highlighting the practical applications of such research.

Although these works were strong steps forward in the realm of sand production management using DAS, their approaches fall short in that they do not extend to calculating verifiable sand velocity, a variable that is often of crucial consideration for optimal sand management practice and operating below the erosional velocity limits [32].

Moreover, they require manual analysis of data and sand logs, a time-consuming process susceptible to human error. This is especially true given the voluminous nature of distributed sensor data, which are often of the order of terabytes per hour for long-term wellbore monitoring [13], [14]. The method presented in this study automates this process, addressing both sand detection and the measurement of sand velocity, hence providing an innovative contribution to the problem of oilfield sand management.

Shetty et al. [12] laid important groundwork for this study by analyzing DAS data for sand detection and monitoring in multiphase flow within horizontal pipes in an experimental flow loop. The present study is in direct conversation with this study, using the same data, which allows for a robust comparison of methodologies. Where Shetty et al. [12] concentrated on manually determining the acoustic fingerprint and frequency analysis, this study expands on their approach using ML to refine and automate sand monitoring and the estimation of sand flow velocities. This progression underscores the potential of the method to enhance real-time decision-making in sand management. In addition, as previously mentioned, this work takes after Shetty et al. [12] in that it uses state-of-the-art signal processing: converting the distributed acoustic signal into FBE data [15], [16], [17] prior to use within the ML algorithms. This significantly amplifies the signal of sand while efficiently reducing the data size; despite training on under 2 min of signal from distributed sensors (a tiny "drop in the bucket" compared with the terabytes of data that distributed sensors typically produce for downhole sensing operations), the proposed ML algorithm attains exceptional performance, as discussed in the results section.

The use of ML for automating sand detection necessitates the consideration of certain common shortcomings of ML. While using this methodology, the present study is careful to consider potential pitfalls of ML such as overfitting, model generalization to new data, model sensitivity, and model interpretability [33]. These potential concerns are discussed and addressed throughout the article; for instance, via careful selection of interpretable models, and intentional preprocessing and augmentation of the training and testing data which encourage generalization and discourage overfitting, as discussed later in Sections II and III.

Overall, the approach adeptly demonstrates the power of using limited data streams to address sand monitoring challenges effectively; it makes efficient, real-time downhole insights a more actionable reality for sand control.

II. Experimental Setup and Data

A. Experimental Setup and Data Collection

This study leverages DAS data from the earlier study by Shetty et al. [12], to allow for a fair comparison between the present automated approach and the past manual approach. Distributed acoustic sensing involves sending laser pulses through an optical fiber, using the backscattered light in the Rayleigh spectrum to measure the vibrations (acoustic signal) throughout the length of the fiber [24]. As a result, the DAS data correspond to phase shift of the backscattered light, where large shifts indicate more significant vibrations at any given time and depth. The specifications for the DAS fiber and acquisition method for the experimental data analyzed are given in Table I.

TABLE I
DAS Acquisition Parameters [12]

Gauge Length	0.82 m
Spatial Resolution	0.82 m
Sampling Frequency	10 kHz
Fiber Type	Single mode (9μm core, 125 μm cladding, tight buffered with PVC jacket)

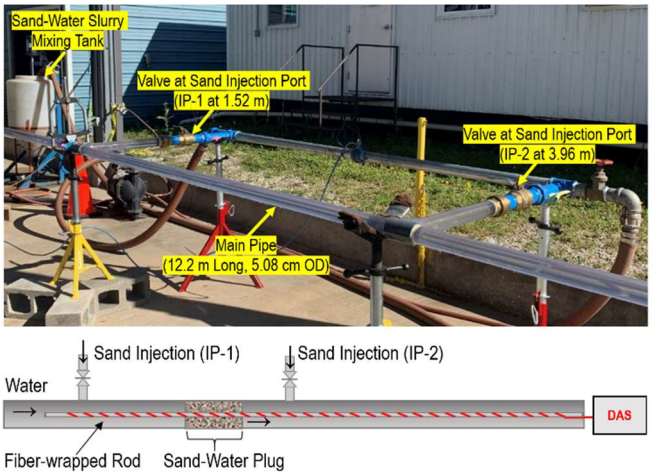


Fig. 1. Visual depiction of the experimental setup used for data collection.

As in Shetty et al. [12], the collected data are partitioned into various experimental trials; each trial consists of water flow, along with the injection of a sand–water slurry of concentration 0.001 volumetric sand concentration in water (0.001 v/v). In oilfield units, this equates to 1000 pounds of 300-μm sand per thousand barrels of water. The slurry is injected using valves located at two injection ports (IPs) at 1.52 m (IP-1) and 3.96 m (IP-2) along the 12.2-m-long horizontal PVC pipeline, with 5.08-cm outer diameter (OD), as shown in Fig. 1. To emulate various possible downhole sanding conditions, each trial has some differing parameters; these include the location and method of sand injection, as well as the fluid flow rate. The various fluid (water) flow rates in the main pipe used in this study included 1.77, 2.02, and 2.27 L/s. In oilfield units, these flow rates translate to 28, 32, and 36 gal/min, respectively. DAS data are acquired using a single-mode fiber which was wrapped helically on a 0.95-cm OD steel rod, sometimes referred to as a stinger in oilfield operations, which was inserted inside the main pipe. Additional details of the experimental setup can be found in Shetty et al. [12].

In total, about 40 GB of DAS data were used for the training, validation, and testing processes, with roughly 10 GB of this used for training and the remaining left for validation and testing. Table II describes the experimental conditions and specifies the data used for training, validation, and testing. All the three flow rates analyzed in this study represent flow above critical settling conditions where sand remains suspended in the carrier fluid (water), thus representing similar governing

TABLE II
EXPERIMENTAL CONDITIONS AND DATA USED FOR TRAINING, TESTING, AND VALIDATION

Trial	IP	Description	Main Pipe Flow (L/s) (Train/Test/Val)		
			1.77	2.02	2.27
A	IP-1	Sand slurry injection followed by fluid flow in main pipe.	Train	Test	-
B	IP-1	Sand slurry injection occurs while the fluid is already flowing in main pipe.	Val	Test	Test
C	IP-2	Sand slurry injection occurs while the fluid is already flowing in main pipe.	Train	Test	Test

physics. Even though the flow rates are close in magnitude, given the relatively small dimensions of the experimental flow loop, they represent distinct sand flow behavior (in terms of sand distribution across the pipe and the relative magnitudes of the governing viscous, drag, lift, and gravitational forces). Thus, they provide distinct datasets for ML model training and testing.

The unique features of the training and testing datasets (varying between two sand injection points and three unique fluid flow rates) make them conducive to testing the robustness of the ML models across a variety of conditions. This train–test splitting schema was used as an alternative approach to taking 25% of the data from each experimental setting for training. Splitting the training and testing data by "experimental setting" more adequately mimics the reality of varying operating conditions; each of the varying conditions corresponds to different potential operational settings common in real well-bore operations. Hence, by training on one set of conditions while validating and testing on others, the ML methods used are less sensitive to environmental variations, and the strong performance of the models is more significant, since overfitting to the full dataset is less likely.

B. Data and Preprocessing

The ML algorithms used take advantage of processing the raw DAS vibration data in the frequency domain using FBE. This preprocessing takes after the same methodology used by Shetty et al. [12]. The DAS signal is first processed into FBE data for various frequency ranges. This study sets out to empirically validate (by comparing ML model accuracies) the conclusions of Shetty et al. [12]: that certain frequency bands (e.g., 200–800 Hz) are particularly associated with the presence of sand in the experimental setup used. FBE data were generated for a variety of frequency ranges, including data for 0–200, 200–400, 400–600, 600–800, 800–1000, 1000–2000, 200–800, and 0–5000 Hz. Here, 0–5000 Hz represents the DAS data in the entire acquisition frequency range (up to the Nyquist frequency), implying no specific signal extraction from the acquired data; the other frequency bands demonstrate signal extraction corresponding to their frequency ranges.

An example of the normalized acoustic data for the 200–800-Hz FBE plot is shown in Fig. 2. As the valve is opened to introduce the sand slurry into the pipeline, a high acoustic signal is sensed at all depths, as shown in Fig. 2.

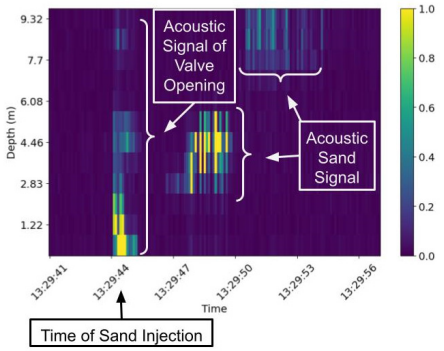


Fig. 2. Example of annotated spatiotemporal FBE data (200–800 Hz) for trial B (validation).

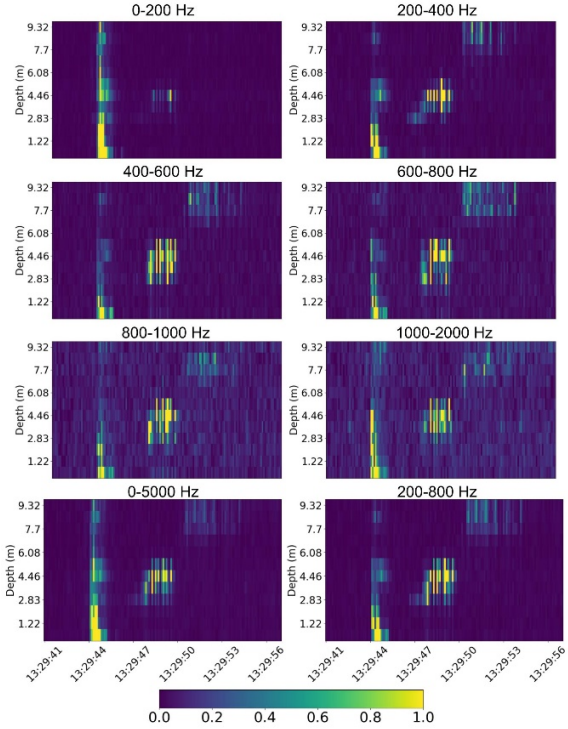


Fig. 3. Example of spatiotemporal FBE data for various frequency bands (after normalization).

The movement of the injected sand slurry, which travels as a sand–water plug along the pipeline, is subsequently observed. It is noted that the sand signal is not continuously observed across the depths which is due to the bending of the stinger rod, resulting in a nonuniform coupling of the fiber with the fluid flow (as described in detail by Shetty et al. [12]). The acoustic sand signal is the data used for the velocity estimation of the sand slurry. The valve opening signal is an example of the signal that is filtered out prior to ML, as is discussed later in this section.

In Fig. 3, it can be visually verified that the FBE signal in the range of 200–800 Hz is among the most effective for isolating the presence of sand in the work of Shetty et al. [12]; when compared with other ranges of frequencies, the afore-mentioned FBE data allow for the clear visualization of the sand slug over time, whereas some of the alternatives present more noisy or hard-to-interpret sand signature.

For instance, in Fig. 3 the data for trial B with the 1.77-L/s flow rate present notably more background noise in the FBE ranges of 400–600, 600–800, 800–1000, and 1000–2000 Hz compared with the 0–5000- or 200–800-Hz data; this makes the sand signal harder to distinguish in the former datasets. In addition, the data generated from 0 to 200 Hz have far less visual sand signal in the latter half of the depicted trial (B).

In the experimental setup used, a strong, noisy acoustic signal occurs at the time of the valve opening to allow the injection of sand slurry at IP-1 and IP-2 (shown in Fig. 1). This interferes with the analysis of the sand signature and mandates a data filtration process to clean the signal before passing it to the ML algorithm. The valve acoustic signal has a specific appearance following FBE processing, whereby the vibrational intensity along the whole length of the pipe is significantly increased. Hence, it is possible to isolate this signal with a simple filtration algorithm. As such, the data processing pipeline consists of the following ordered steps.

1) Converting the raw DAS strain rate data into FBE.
2) Normalizing the FBE data to values between 0 and 1.
3) Removing the manual valve-opening signature; this means removing data at times when a brief, strong acoustic signal is present throughout most of the pipeline (see Fig. 2 for an example of this valve signal).
4) Input into ML model (described in Section III).

This preprocessing workflow is depicted for trial B of the 1.77-L/s flow rate in Fig. 4. As can be seen in the figure, the initial raw DAS strain rate data are largely uninterpretable, further demonstrating the need for the signal processing and normalization steps.

In this study, the initial training data exhibited a significant imbalance, with the ratio of sand to nonsand signal classes being approximately 1:7. To rectify this disparity and enhance the model's learning capability, the present study adopted an oversampling strategy for the sand class, achieving a balanced 1:1 ratio [34]. This was accomplished by replicating the existing sand samples and introducing a subtle variance. Specifically, each duplicated entry was modified with a 2% Gaussian noise, characterized by a mean of 0 and a variance of 0.02. This method effectively prevented the exact replication of data, while maintaining the integrity of the original samples in the new, balanced dataset. Extensive experimentation was conducted on the validation dataset to determine the impact of varying noise levels on the validation accuracy when oversampling. The study explored noise variations ranging from 1% to 15% yet observed that the validation accuracy remained consistent across these different levels. Consequently, the 2% noise level was selected, though not due to any specific advantageous outcome.

Concerns about the model's generalizability, particularly regarding the consistent location of sand in most training and validation datasets, led to the implementation of data flipping techniques, a common practice in ML settings [35], [36], [37], [38]. This approach involved spatially distorting a certain fraction (ranging from 20% to 50%) of the oversampled data, which included both sand and nonsand entries to alter the sand signal location. This method aligns with the physical dynamics of the system, as the flipped data vectors represent plausible variations in sand location that could occur under different experimental conditions. Three distinct flipping methods were

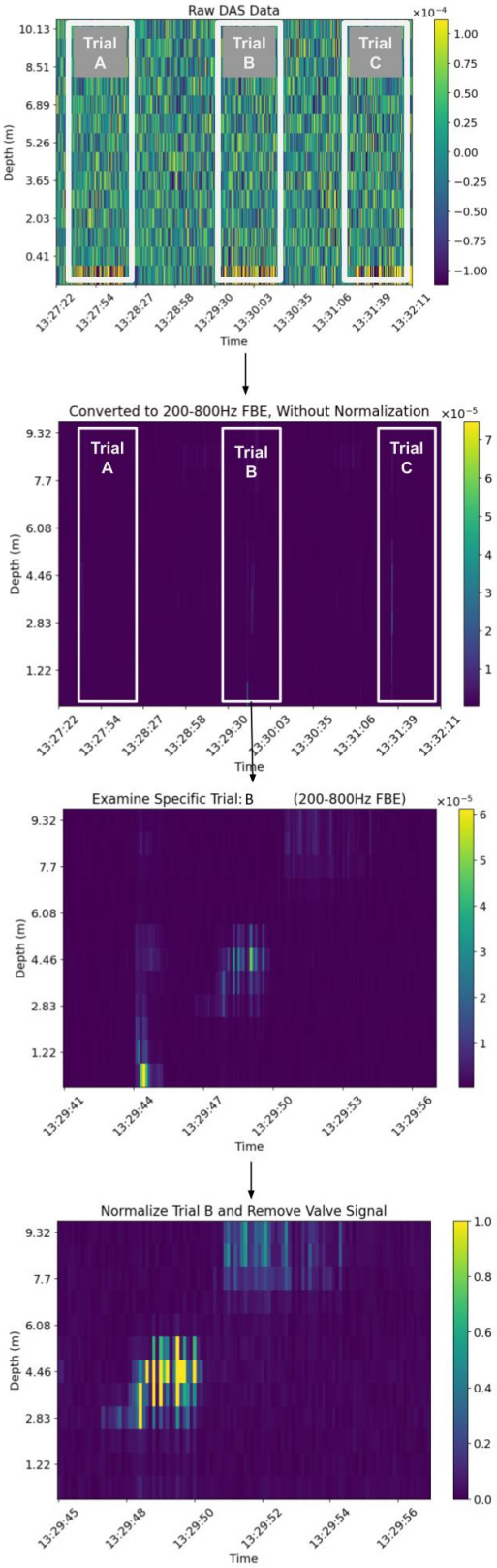


Fig. 4. Example of the processing pipeline performed on the DAS data, starting with the raw strain-rate data to FBE estimation, valve signal removal, and data normalization.

used: reversing the entire data vector, rolling the vector upward (where the top 25% of the data vector shifts to the bottom and the remaining data move up), and rolling downward (where the bottom 25% of the data vector shifts to the top, with the rest moving down accordingly). The choice of flipping method for each data point was randomized from these three options, providing a diverse range of data alterations to enhance the robustness and adaptability of the model. For each of the models obtained after hyperparameter tuning, the amount of flipped data varied based on the performance of the model along the F_1 and F_2 metrics (see Section III-B for details).

III. ML Methodology

A. Algorithm for Sand Detection

Past work which analyzes DAS data for detecting sand ingress and velocity has required human interpretation of data [12], which is prone to potential issues. The automation approach in the present study, described in the section below, has multiple advantages compared with this previous standard, which are detailed below. In real-time monitoring applications, using the method by Shetty et al. [12] would require manual interpretation and calculation of velocity, which is costly given the vast quantities of pipeline DAS data that are recorded on a day-to-day basis. The automated classification approach is superior in that it requires less costly labor, it can be scaled to large quantities of data with much greater ease, and it can avoid human bias in manual estimations.

In addition, the velocity estimation and sand detection method used by Shetty et al. [12] requires preexisting knowledge of the acoustic sand fingerprint, which they obtain from manual spectrogram analysis using several frequency ranges. The present proposed ML method avoids this human-intensive and time-consuming step, instead verifying the frequency fingerprint of the sand through comparison of different ML models' performances. Some details of this comparison are provided in Section III-D, as well as in supplementary documentation.

To perform supervised ML on the normalized FBE data, data labels (chosen from two classes: positive, or "sand," and negative, or "no sand") are manually interpreted by a human labeler. This is based on the visual presence of strong FBE signal at pipe depths and times where the presence of sand is already known.

The goal of the initial model is to predict the presence of sand at a given timestep across a column of data representing the full fiber length on the experimental pipe. Initially, FBE data from the test trials with flow rates of 1.77 L/s (28 gal/min) are used to train the models. As in Shetty et al. [12], these data consisted of three trials (A, B, and C). Two of the three 1.77-L/s trials, trial A with sand injection at IP-1 (at 1.52 m along the pipe) and trial C with sand injection at IP-2 (at 3.96 m along the pipe), were used for model training. The data from trial B, an experiment where the sand injection was also at IP-1 but injected into already-flowing water, were labeled and saved for validation of the model. This is also summarized in Table II presented earlier.

Let $X = \{\bar{x_1}, \bar{x_2}, \ldots, \bar{x_n}\}$ represent a sequence of normalized FBE data from a selected frequency band, with values in each vector $\bar{x}_i$ ranging from 0 to 1. Here, the length of any one vector $\bar{x}_i$ is based on the length of the pipeline, and the length of the sequence X is based on the amount of time that the data span (up to time n). Based on the strength and location of the signal for any one $\bar{x}_i$, the data are labeled as either containing sand or not containing sand at time i.

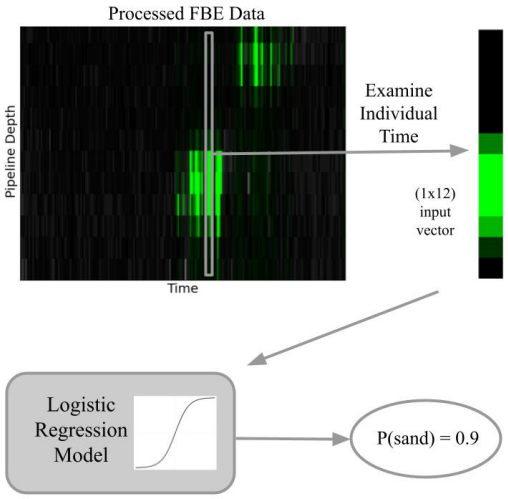


Fig. 5. Depiction of the approach used for sand classification.

Using the data and labels from the two trials mentioned above, a probabilistic linear model (logistic regression [39]) is trained to assign a probability of sand at each timestep. The approach was chosen as a natural and simple approach to binary classification (sand versus nonsand), and the rationale for the choice is elaborated on in Section III-B. At time i, the vector of acoustic signal throughout the pipe can be represented as $\bar{x}_i = \langle x_{i1}, x_{i2}, \ldots, x_{im} \rangle$, where each index (1 through m) represents a different depth along the flow loop. The logistic regression model finds an optimal coefficient vector $\bar{\alpha} = \langle \alpha_0, \alpha_1, \ldots, \alpha_m \rangle$ based on the data and labels of the training data, which allows for calculations of sand probability using the following equation:

$$P\left(\text{Sand}_i \mid \bar{x}_i\right) = \sigma(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_m x_{im}). \quad (1)$$

Here, σ is the sigmoid function defined by

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (2)$$

and is used to convert a value to a probability of sand between 0 and 1.

This sand detection workflow is depicted schematically in Fig. 5 using example DAS FBE data from trial A for 2.02 L/s for the 200–800-Hz frequency band.

B. Choice of ML Algorithm

For the ML-assisted sand classification, the selection of logistic regression as the model-of-choice was based on a variety of important considerations.

Empirically, four different common ML classification methods were used on the training and validation data to find a classification method, taking into account the model performance, complexity, and computational time. The methods tested were support vector machines, decision trees, random

forests, and logistic regression models [40]. The results of these tests are included in the supplementary document (see Section II) which show that the logistic regression performed strongest among the four tested classifiers.

In addition, theoretical considerations contributed to the choice of logistic regression for sand classification, including the below.

1) Ease of Training and Deployment: Logistic regression models are computationally inexpensive, making them better suited for real-time analysis and decision-making. Of note, the required time for ML processing of certain testing data (including velocity estimation, described later) was recorded: for an examined segment of 130 s of DAS data, the total required processing time was (on average) under 10 s when performed on a single CPU. Hence, since only a fraction of the data's total time is required to perform sand monitoring, the proposed method is well-suited for real-time settings.

2) Limited Data Requirements: Logistic regression requires less data to train than other potential classification methods, such as Naïve Bayes classifiers or deep neural networks [41].

3) Probabilistic Modeling: Logistic regression allows for an output probability, rather than a binary yes or no output. This allows for tracking and quantification of uncertainty, which is helpful in scenarios with large economic stakes such as hydrocarbon production. For instance, it may be economical to only make operating decisions based on predictions of sand that are (for example) 90% certain; if this is the case, the decision boundary of the logistic regression can be adjusted to account for the risk tolerance of the operator.

4) Interpretability: Due to their simplicity, logistic regression models are more interpretable than larger, multilayered models such as deep neural networks.

5) Sensitivity: Logistic regression has lower outlier sensitivity than other potential classification methods such as support vector machines and decision trees [40].

C. Hyperparameter Selection

Model selection was based on a hyperparameter search conducted on the validation dataset, where the top performing models were selected based on weighted harmonic means of the precision and recall (F-scores) [42], specifically the F_1 and F_2 scores. These values can be calculated using the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN), as shown in the following equations:

$$\text{Precision(Prec)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

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

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{5}$$

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{\text{Precision} \times \text{Recall}}{\left(\beta^2 \cdot \text{Precision}\right) + \text{Recall}} \tag{6}$$

$$\text{Accuracy(Acc)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{7}$$

Note that F_2 score is equivalent to F_β score where $\beta = 2$.

An observation was made regarding the use of flipped training data, previously described in Section II: despite its potential to improve generalization, it appeared to adversely affect the validation results, possibly due to similarities between the validation and training datasets. Nevertheless, the decision was made to incorporate a certain percentage of flipped data in the training set, with the foresight that the model must generalize effectively to diverse data scenarios during blind testing (representing different sand ingress locations), which might not be perfectly represented in the training set. To determine the optimal extent of data flipping, the fraction of total training data subjected to flipping was varied (this was performed following oversampling, which was done to allow for a balanced data size for the two classes, as discussed in Section II) and rigorously evaluated for its impact on model performance on the validation dataset.

Concurrently, the study investigated multiple probabilistic linear models, each differing in their loss functions [43], [44], [45]. A critical adjustment was made in the loss ratio attributed to positive (sand) versus negative (no sand) classes during training. This ratio was experimented with, ranging from a balanced 1:1 to a more skewed 10:1, the combination of loss weights summing to 1. Empirically, this adjustment proved to be significantly beneficial in enhancing model accuracy, along with F_1 and F_2 scores. The prioritization of recall in model selection, as evidenced by the use of F_2 score, was strategically chosen to improve the model's capability to accurately identify sand presence. This approach acknowledges that the negative impact of FPs is considerably less severe than the potential consequences of failing to detect sand, which could lead to hazardous and expensive outcomes. Hence, the model's design inherently favors a higher tolerance for FPs to ensure maximum reliability in sand detection.

A comprehensive grid search was conducted on the validation set to identify the most effective combination of data flipping fraction and loss ratio. The flip fraction was tested within a range of 20%–50% of the total oversampled data, while the loss ratio was varied from 1:1 to 10:1 (representing the ratio of loss for sand versus loss on nonsand data points). For each set of training data corresponding to different FBE frequency ranges, certain top-performing models were selected to be used in testing (see Section IV). In summary, the hyperparameter grid search was conducted over two dimensions: fraction of flipped data in the training set, "flip," with search values including {0.2, 0.3, 0.4, 0.5}, along with sand-loss ratio, "weight," where the values were taken from {1:1, 2:1, 3:1, 4:1, 5:1, 6:1, 7:1, 8:1, 9:1, 10:1}.

In Tables II–VI, the "Top F_1" model refers to the model trained with the loss and flipping fraction that obtained the strongest F_1 score on the validation data; the "Top F_2" model refers to the model trained with the loss and flipping fraction that obtained the strongest F_2 score on the validation data; and the Base Case model uses a fixed loss ratio of 1:1 and flipping fraction of 20% (denoted with 0.2 in Table III), for comparison of model performance without hyperparameter tuning.

In practical applications of this workflow, the combination of hyperparameters, including the decision on how the loss function penalizes the different class predictions, will be a

TABLE III
PERFORMANCE ON THE CLASSIFICATION TASK USING THE VALIDATION DATA, FOR THE THREE SELECTED MODELS CHOSEN AFTER HYPERPARAMETER SEARCH. USES THE FREQUENCY BAND OF 200–800 Hz

Model	**Base Case**	**Top F_1**	**Top F_2**
Parameters	Weight = 1:1 Flip = 0.2	Weight = 6:1 Flip = 0.3	Weights = 9:1 Flip = 0.2
Acc	0.90	0.96	0.96
Precision	0.94	0.94	0.87
Recall	0.60	0.89	0.96
F_1	0.73	0.92	0.92
F_2	0.64	0.90	0.95

TABLE IV
PERFORMANCE OF DIFFERENT ML MODELS, AFTER HYPERPARAMETER SEARCH, ON THEIR RESPECTIVE VALIDATION SETS

FBE Frequency (Hz)	**Acc | Prec | Recall | F_1 | F_2**
Train and Validation	*Base Case*
0-200	0.79 | 1.00 | 0.09 | 0.16 | 0.11
200-400	0.85 | 0.92 | 0.40 | 0.56 | 0.45
400-600	0.89 | 0.92 | 0.58 | 0.71 | 0.62
600-800	0.86 | 0.87 | 0.46 | 0.60 | 0.50
800-1000	0.82 | 1.00 | 0.20 | 0.33 | 0.23
1000-2000	0.81 | 1.00 | 0.12 | 0.21 | 0.14
200-800	0.90 | 0.94 | 0.60 | 0.73 | 0.64
0-5000	0.89 | 1.00 | 0.43 | 0.60 | 0.48
	Top F_1
0-200	0.82 | 0.70 | 0.34 | 0.46 | 0.38
200-400	0.94 | 0.89 | 0.84 | 0.86 | 0.85
400-600	0.93 | 0.85 | 0.82 | 0.84 | 0.83
600-800	0.95 | 0.91 | 0.86 | 0.88 | 0.87
800-1000	0.89 | 0.76 | 0.73 | 0.75 | 0.74
1000-2000	0.86 | 0.65 | 0.73 | 0.69 | 0.71
200-800	0.96 | 0.94 | 0.89 | 0.92 | 0.90
0-5000	0.96 | 0.98 | 0.82 | 0.89 | 0.84
	Top F_2
0-200	0.23 | 0.23 | 1.00 | 0.37 | 0.60
200-400	0.94 | 0.82 | 0.93 | 0.87 | 0.90
400-600	0.92 | 0.83 | 0.84 | 0.83 | 0.84
600-800	0.94 | 0.89 | 0.86 | 0.88 | 0.87
800-1000	0.80 | 0.55 | 0.86 | 0.67 | 0.77
1000-2000	0.75 | 0.45 | 0.84 | 0.59 | 0.72
200-800	0.93 | 0.79 | 0.96 | 0.87 | 0.92
0-5000	0.95 | 0.84 | 0.96 | 0.90 | 0.93

techno-economic decision to be made by the oilfield operators depending on their sand-handling capability and production objectives. The goal of selecting three different models during the validation stage (as opposed to one) was to demonstrate the different performances achieved and the sensitivity of the hyperparameter selection criteria.

D. Evaluating Sand Acoustic Fingerprint

Based on the findings of the past work [12], this study sought to confirm which frequency bands (e.g., 200–800 Hz) correspond strongly to sand acoustic fingerprint in the experimental flow loop setup. To do so, multiple different frequency bands were used when processing the raw DAS data, and independent ML models were trained on each version of the data.

Table IV shows the results of the performance of the top three models selected on the validation data for DAS FBE in different frequency ranges. It can be observed that the frequency band 200–800 Hz and the bands contained within that range (200–400, 400–600, and 600–800 Hz) perform notably better than the alternative bands, indicating the relative "ease" with which the models can distinguish sand once DAS data is intelligently processed using the acoustic sand fingerprint.

In practical settings involving different flow loop conditions, this process of training and validating models on different frequency band ranges can be used to verify the acoustic fingerprint of the flow loop in question. In other words, by obtaining accuracy and F-scores of ML models from different frequency bands, the operator can compare the scores to make informed inferences about which acoustic fingerprint enables the best sand detection.

Based on the results and performance on the models in Table IV, two main frequency bands of data were used for creating the tested models (see Section IV): 200–800 Hz (the acoustic sand fingerprint), which has the strongest overall classification and F_1 performance in Table IV, and 0–5000 Hz, which represents the DAS data in the entire acquisition frequency, implying no specific signal extraction. In a supplementary document, further analysis results are presented for the performance of models trained on data from other frequency bands.

E. Estimating Sand Velocity

Following sand detection and classification, the second component of ML automation involves using information about detected sand to estimate sand velocity. To do this, the algorithm first records all the timesteps where the DAS data were classified as containing sand, based on the logistic regression discussed previously. Since the injected sand slurry is traveling in the pipeline as a sand–water plug (as shown in Fig. 1), a "center of mass" calculation is performed to determine an approximate spatial location (s_i) of the sand acoustic signal along the pipeline at any one given time, i. This is done by taking a weighted average of the normalized acoustic strength values along the length of the flow loop, at each time (using the same vector that was used for sand classification). For instance, if a vector of acoustic values, $\bar{x}_i$, is classified as sand, then the approximate spatial "center of mass" of the sand is calculated by finding the weighted average of the values of $\bar{x}_i$. Then, the point (i, s_i) is one of many that represents the approximate spatiotemporal location of sand.

Because this estimated location of sand signal can be produced for every timestep where sand was detected, a collection of these values can be used to garner an estimate of the movement of the sand–water plug over time. Hence, the collected points (i, s_i) can be used to inform a linear regression model, of which the slope corresponds to the sand velocity within the flow loop. This is depicted in Fig. 6, where the white points represent the "center of mass" of the sand signal at the corresponding timesteps. Though it is not used for prediction of additional data in this application, the use of linear regression to form a velocity estimation slope was largely inspired by Shetty et al. [12]. To estimate sand velocity, those authors manually annotated FBE images with lines that

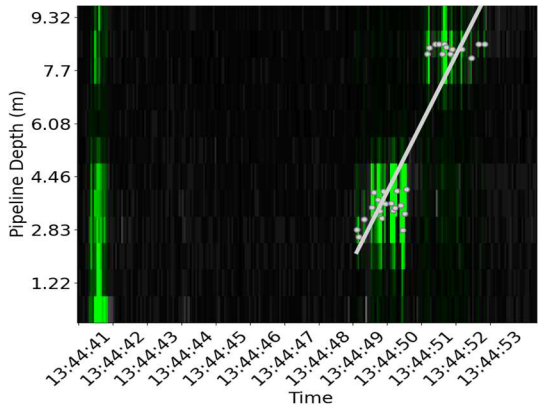


Fig. 6. Example of linear regression for automated sand velocity calculation. Data are from trial A of 2.02 L/s, using the FBE for 200–800 Hz.

followed the trend of the sand plug; in the same way, the use of the linear regression slope in this study allows this to happen automatically. Collectively, they show the trend of sand movement along the flow loop, as captured by the linear regression line (indicated in white). On the left is the valve signal, which is filtered out as a preprocessing step. Since the fiber is helically wrapped, a measured length correction is also applied to estimate the velocity in the horizontal frame of reference. As such, the initial velocity estimates from the regression are divided by 1.13, which is the helical correction reported in the previous work by Shetty et al. [12].

To corroborate the results of this ML approach, and following the methodology of Shetty et al. [12], an analytical model of fluid flow is used to calculate the theoretical fluid velocity and sand slip velocity [46], [47]. Given that this is only the second work (to the authors' knowledge) of experimentally calculating sand velocity with DAS data, the use of the same analytical models as Shetty et al. [12] allows for consistent comparison across the literature.

The equation for sand slip velocity can be expressed as

$$\frac{v_{\text{sl}}}{V} = \frac{v_t^2}{4gd_p}\left[1 - \frac{C}{q}\right]^{2.5}\left[\frac{V_c}{V}\right]^4 \tag{8}$$

where V_c is the critical settling velocity (0.728 m/s), C is the delivered solids (sand) concentration (2850 kg/m^3), q is the spatial solids concentration (3343, 3300, and 3269 kg/m^3, for fluid flow rates of 1.77, 2.02, and 2.27 L/s, respectively), g is the gravitational acceleration, v_t is the terminal settling velocity of a sand particle (0.033 m/s), d_p is the sand particle size (300 μm), and V is the mean fluid flow velocity (which varies depending on the trial). The above numeric values are based on those presented by Shetty et al. [12].

For the flow rates of 1.77, 2.02, and 2.27 L/s, the analytical slip velocities are thus derived as 0.059, 0.067, and 0.074 m/s, respectively, using (8) adopted from Shetty et al. [12]. Using these values, and the known carrier fluid (water) flow rate in the main pipe, the analytical sand velocities can be calculated and compared with the velocities found in this study (by subtracting the slip velocity from the fluid velocity). This gives the sand velocities corresponding to 1.77, 2.02, and 2.27 L/s, as 0.87, 0.99, and 1.12 m/s, respectively. The percentage errors between the ML-estimated velocity and the analytical velocity are summarized and discussed in Section IV.

TABLE V
PERFORMANCE OF THE CLASSIFICATION AND REGRESSION MODELS ALONG VARIOUS METRICS, FOR 1.77-L/s FLOW RATE (TRAINING AND VALIDATION DATA)

1.77 L/s	Vel (m/s) | Vel Err (%) | R^2 *Acc* | *Prec* | *Recall* | F_1 | F_2	
	200-800 Hz	**0-5000 Hz**
Trial	*Base Case*	
A	0.91 | 4.80 | **0.84** *0.96* | *0.83* | *0.96* | *0.89* | *0.93*	0.78 | 10.80 | **0.78** *0.76* | *0.44* | *1.00* | *0.61* | *0.80*
B	1.12 | 28.60 | **0.82** *0.90* | *0.94* | *0.60* | *0.73* | *0.64*	1.37 | 57.70 | **0.84** *0.89* | *1.00* | *0.43* | *0.60* | *0.48*
C	0.88 | 0.80 | **0.88** *0.97* | *1.00* | *0.53* | *0.70* | *0.59*	-0.85 | 197.7 | **0.34** *0.94* | *1.00* | *0.25* | *0.40* | *0.29*
	Top F_1	
A	0.88 | 0.60 | **0.84** *0.37* | *0.23* | *1.00* | *0.38* | *0.60*	0.78 | 10.90 | **0.78** *0.35* | *0.23* | *1.00* | *0.37* | *0.59*
B	1.07 | 22.80 | **0.78** *0.96* | *0.94* | *0.89* | *0.92* | *0.90*	1.16 | 32.7 | **0.82** *0.96* | *0.98* | *0.82* | *0.89* | *0.84*
C	0.84 | 3.90 | **0.88** *0.95* | *0.58* | *1.00* | *0.73* | *0.87*	0.80 | 8.50 | **0.85** 0.95 | 0.90 | 0.45 | 0.60 | 0.50
	Top F_2	
A	0.88 | 0.60 | **0.84** *0.21* | *0.19* | *1.00* | *0.32* | *0.55*	0.78 | 10.9 | **0.78** *0.19* | *0.19* | *1.00* | *0.32* | *0.54*
B	1.06 | 21.90 | **0.80** *0.93* | *0.79* | *0.96* | *0.87* | *0.92*	1.16 | 33.8 | **0.81** *0.95* | *0.84* | *0.96* | *0.90* | *0.93*
C	0.84 | 3.90 | **0.88** *0.72* | *0.18* | *1.00* | *0.31* | *0.52*	0.80 | 8.50 | **0.85** *0.95* | *0.90* | *0.45* | *0.60* | *0.50*

IV. RESULTS AND DISCUSSION

In the tables below, the results for both classification accuracy (denoted as "Acc") and estimated velocity (denoted as "Vel") are shown, using FBE data from 200 to 800 and 0 to 5000 Hz (the full available frequency range for the acquired DAS). These two FBE bands were specifically selected to investigate the performance of the frequency range corresponding to the sand signal (200–800 Hz) when compared with the full frequency range without any sand signal extraction (0–5000 Hz). For completeness, aggregate results for the other examined FBE frequency ranges are included in the supplementary file that further demonstrates 200–800 Hz to be the best performing band. The velocity error reported in Table V (as "Vel Err") is with reference to the analytical velocity for the respective cases, which is considered as the ground-truth value. As shown in Table V, a variety of metrics are evaluated to measure the model performance, including Precision (Prec), Recall, F_1, and F_2 of the sand classifier models and the R^2 value representing the "goodness of fit" of the regression models used for the velocity estimate.

Along with the tables, bar plots (see Figs. 7–9) are provided to visualize the average values of metrics across the three trials. In the bar plots, sand velocity errors are provided in a normalized form (rather than percentage) to allow them to be plotted side-by-side the other metrics. Hence, in the bar plots, a velocity error value of 1 corresponds to a percentage error of 100%, meaning the error between the estimated velocity and true sand velocity is 100%. Accordingly, a value of 0.5 corresponds to 50% error, 0.1 corresponds to 10% error, and so on. Note that a frequency band is considered to perform more strongly if the model has higher F_1, F_2, Acc, and R^2 scores while also giving a lower velocity error.

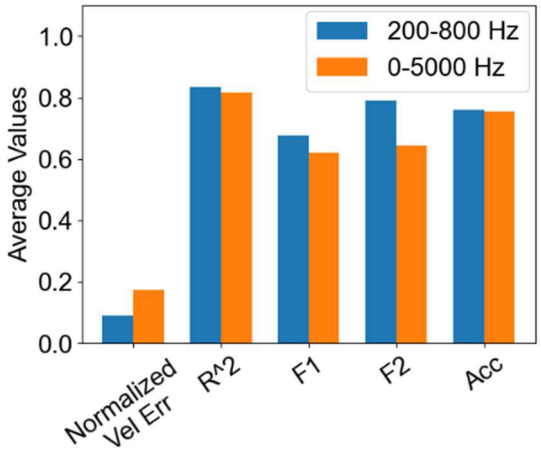


Fig. 7. Bar plot comparing the average performance across trials A–C (using the flow rate of 1.77 L/s) of the Top F_1 models, which were, respectively, trained on the data from the 200–800- and 0–5000-Hz frequency bands.

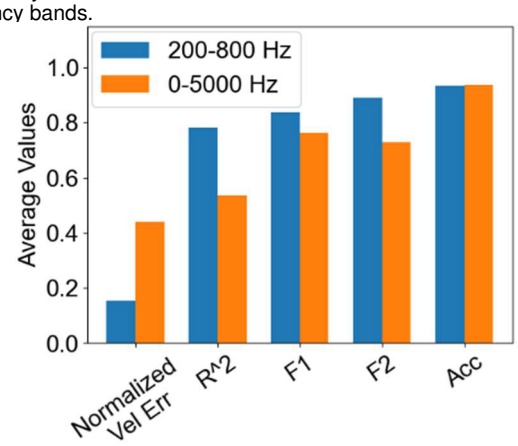


Fig. 8. Bar plot comparing the average performance across trials A–C (using the flow rate of 2.02 L/s) of the Top F_1 models, which were, respectively, trained on the data from the 200–800- and 0–5000-Hz frequency bands.

Examples of the regression analyses are depicted for trials A–C in Figs. 10–12, along with the slope, y-intercept, and R^2 values in the corresponding captions. Note that the y-intercepts, expressed in meters, do not have a conceptual meaning in this analysis. This is because the time values on the x-axis which correspond to the intercept merely mark when data were truncated in preparation for the classification and regression analysis; in practice, this was often a few seconds before the sand ingress time. Consequently, the intercepts lack a direct link to the actual ingress depth, and they are only included as a technical artifact of the linear regression analysis. Given this and the emphasis of this study on sand velocity estimation, the slope values (also recorded in the figures' captions) are the focus of attention from the linear regression.

A. Training and Validation Results (1.77-L/s Flow Rate)

Using the 1.77-L/s data, the sand-detection algorithm was trained on trials A and C, while being validated on data from trial B. Sand velocity results were calculated after the completion of training the detection models, to further verify the velocity estimation method.

Overall, with a few exceptions, the models using 200–800-Hz FBE data tend to outperform those using the full

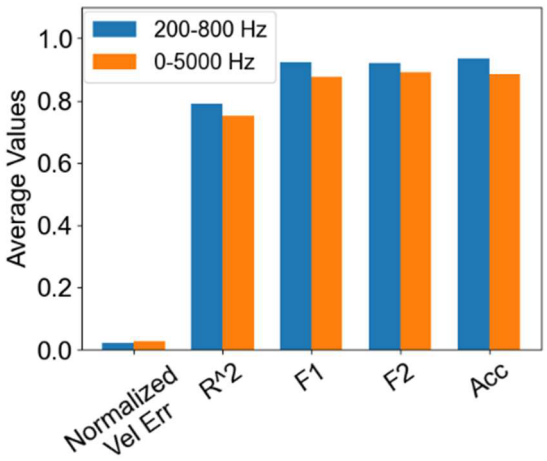


Fig. 9. Bar plot comparing the average performance across trials B and C (using the flow rate of 2.27 L/s) of the "Top F_1" models, which were, respectively, trained on the data from the 200–800- and 0–5000-Hz frequency bands.

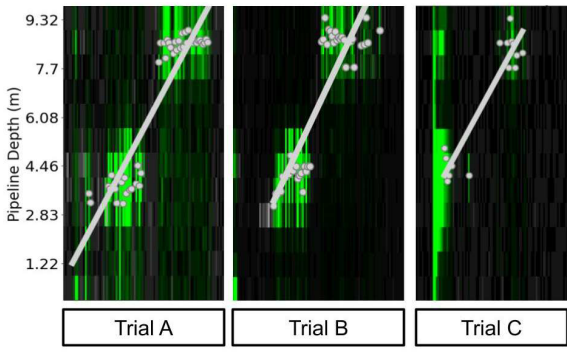


Fig. 10. Examples of the linear regression for automated sand velocity calculation for the 1.77-L/s flow rate, using the 200–800-Hz frequency fingerprint. The depicted trial A regression line, from the Base Case model, has a slope of 0.91 m/s, an intercept of −3.87 m, and an R^2 value of 0.84. The depicted trial B regression line, from the Top F_1 model, has a slope of 1.07 m/s, an intercept of −1.45 m, and an R^2 value of 0.78. The depicted trial C regression line, also from the Top F_1 model, has a slope of 0.84 m/s, an intercept of 1.49 m, and an R^2 value of 0.88.

frequency range of 0–5000 Hz, which further demonstrates the applicability of the acoustic sand fingerprint. This result is aligned with the conclusion from Shetty et al. [12]. Notably, the Top F_1 and Top F_2 models produce high detection accuracy and reasonable estimates of velocity that correspond with the analytical values. For trial A, the lower accuracy of the Top F_1 and Top F_2 models here is thought to be due to the increased sensitivity to sand (which is itself due to the higher loss weight for the sand class), leading to many FPs. This is evidenced by the low precision and high recall of the two models on trial A. Despite this, the two tuned models (Top F_1 and Top F_2) perform exceptionally well compared with o the Base Case on the blind testing data (i.e., the 2.02- and 2.27-L/s flow rates), as shown in Sections IV-B and IV-C. This highlights that the models generalize well because of this specific form of hyperparameter tuning.

It should also be noted that for the velocity estimation linear regression models, the relatively low R^2 value is somewhat expected due to the nature of the sand traveling as a plug (see Figs. 1 and 10–12), which results in spatially distributed

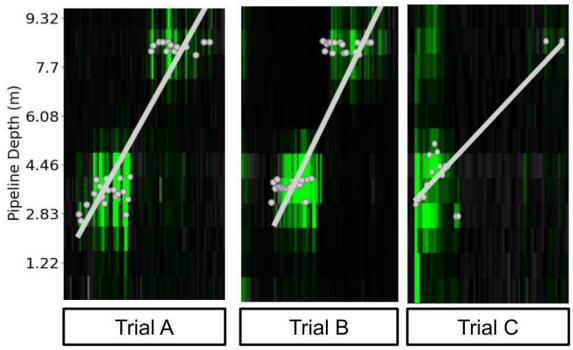


Fig. 11. Examples of the linear regression for automated sand velocity calculation for the 2.02-L/s flow rate, using the 200–800-Hz frequency fingerprint. All depicted regression lines are generated based on the points classified as sand by the Top F_1 model. The trial A regression line has a slope of 1.07 m/s, an intercept of −11.90 m, and an R^2 value of 0.82. The trial B regression line has a slope of 1.20 m/s, an intercept of −0.05 m, and an R^2 value of 0.79. The trial C regression line has a slope of 0.82 m/s, an intercept 1.69 m, and an R^2 value of 0.74.

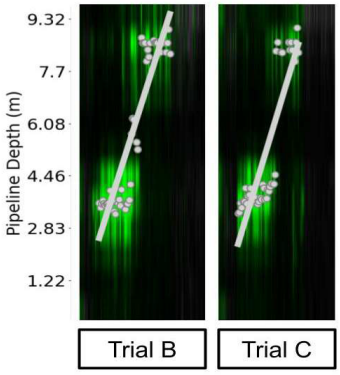


Fig. 12. Examples of the linear regression for automated sand velocity calculation for the 2.27-L/s flow rate, using the 200–800-Hz frequency fingerprint. Both the depicted regression lines are generated based on the points classified as sand by the Top F_1 model. The trial B regression line has a slope of 1.07 m/s, an intercept −9.26 m, and an R^2 value of 0.83. The trial C regression line has a slope of 1.12 m/s, an intercept of −11.52 m, and an R^2 value of 0.75.

sand detection across multiple locations at any given time. The spatial resolution of the DAS data also contributed to this.

The results for the linear regression models are pictorially illustrated for trials A–C of the 1.77-L/s flow rate data in Fig. 10. Recall that the slope of the regression line in the figure represents the estimated sand velocity. For each of the three trials, these values closely corroborate the theoretical expectations, as demonstrated in Table V.

For the 1.77-L/s flow rate, Fig. 7 provided a visual aggregation of the performance for the Top F_1 models across trials A–C. Notably, all four of the averages for accuracy, F_1-score, F_2-score, and R^2 are higher for the 200–800 = Hz data compared with the 0–5000-Hz data; at the same time, the error in estimated sand velocity is lower when using the 200–800-Hz frequency band. These outcomes demonstrate the power of using the acoustic sand fingerprint for sand classification and velocity estimation, specifically for the case of the 1.77-L/s flow rate.

TABLE VI
PERFORMANCE OF THE CLASSIFICATION AND REGRESSION MODELS ALONG VARIOUS METRICS, FOR 2.02-L/s FLOW RATE

2.02 L/s	Vel (m/s) | Vel Err (%) | R^2 *Acc* | *Prec* | *Recall* | F_1 | F_2	
	200-800 Hz	**0-5000 Hz**
Trial	*Base Case*	
A	1.32 | 33.40 | **0.82** *0.94* | *1.00* | *0.64* | *0.78* | *0.69*	1.18 | 18.90 | **0.82** *0.93* | *1.00* | *0.56* | *0.71* | *0.61*
B	1.44 | 44.80 | **0.75** *0.86* | *1.00* | *0.54* | *0.70* | *0.59*	1.42 | 42.90 | **0.75** *0.85* | *1.00* | *0.52* | *0.68* | *0.57*
C	1.41 | 42.1 | **0.58** *0.84* | *0.36* | *0.27* | *0.31* | *0.28*	0.11 | 89.30 | **0.00** *0.91* | *0.50* | *0.60* | *0.55* | *0.58*
	Top F_1	
A	1.07 | 8.10 | **0.82** *0.96* | *0.82* | *1.00* | *0.90* | *0.96*	1.19 | 20.30 | **0.84** *0.95* | *1.00* | *0.67* | *0.80* | *0.71*
B	1.20 | 21.10 | **0.79** *0.95* | *1.00* | *0.83* | *0.91* | *0.86*	1.21 | 22.00 | **0.77** *0.94* | *1.00* | *0.78* | *0.88* | *0.82*
C	0.82 | 17.00 | **0.74** *0.89* | *0.54* | *1.00* | *0.70* | *0.85*	0.11 | 89.3 | **0.00** *0.92* | *0.54* | *0.70* | *0.61* | *0.66*
	Top F_2	
A	1.08 | 9.10 | **0.82** *0.82* | *0.47* | *1.00* | *0.64* | *0.81*	1.14 | 14.50 | **0.85** *0.90* | *0.62* | *0.92* | *0.74* | *0.84*
B	1.18 | 18.90 | **0.80** *0.97* | *0.96* | *0.93* | *0.94* | *0.93*	1.14 | 15.10 | **0.77** *0.98* | *0.96* | *0.96* | *0.96* | *0.96*
C	0.82 | 17.00 | **0.74** *0.59* | *0.24* | *1.00* | *0.38* | *0.61*	0.11 | 89.30 | **0.00** *0.89* | *0.42* | *0.80* | *0.55* | *0.68*

B. Testing Results (2.02-L/s Flow Rate)

For the flow rate of 2.02 L/s, the analytical sand velocity is 0.99 m/s. Using this value as the ground truth, the percentage error between the estimated velocity and the analytical velocity is depicted in Table VI, along with all the other performance indicators (Precision, Recall, F_1, and F_2) for the three trained models. Notably, the 200–800-Hz Top F_1 model has the strongest consistent performance across the three trials along the F_1 and F_2 metrics, further corroborating the utility of the acoustic sand fingerprint for velocity and detection applications. For all the three trials, the Top F_1 model for both the 200–800-Hz FBE and 0–5000-Hz FBE data consistently identified the labeled sand with over 89% accuracy. Across all the models, the average sand detection accuracy for the 2.02-L/s trials is also 89%, with an average F_1 score of 0.71 and average F_2 score of 0.72. This a promising sign for sand detection using the proposed ML method, especially considering that the differing velocity of the 2.02 L-s flow rate compared with the training data implies the type of model can perform well out-of-distribution.

In addition, the sand signal extracted from the 200–800-Hz FBE data at this flow rate produces reasonable estimates of the velocity (with an average percentage error of 23.5% across all the three 200–800-Hz models). The failure of the 0–5000-Hz models to obtain an estimate of the velocity for trial C (see Table VI) is expected because of the higher environmental noise observed during this trial. Without any frequency filtering and by considering all the frequency content, the signal-to-noise ratio of sand was observed to be weak. For real field applications, this further demonstrates the need for preliminary characterization of characteristic sand frequency using data from some known sanding events, as was done with the validation data in this study.

TABLE VII
PERFORMANCE OF THE CLASSIFICATION AND REGRESSION MODELS ALONG VARIOUS METRICS, FOR 2.27-L/s FLOW RATE

2.27 L/s	Vel (m/s) | Vel Err (%) | R^2 *Acc* | *Prec* | *Recall* | F_1 | F_2	
	200-800 Hz	**0-5000 Hz**
Trial	*Base Case*	
B	1.05 | 6.5 | **0.86** *0.85* | *1.0* | *0.59* | *0.74* | *0.64*	1.15 | 3.0 | **0.84** *0.83* | *0.89* | *0.64* | *0.74* | *0.68*
C	1.07 | 4.5 | **0.77** 0.78 | 1.0 | 0.59 | 0.74 | 0.64	1.07 | 4.6 | **0.69** *0.83* | *0.96* | *0.7* | *0.81* | *0.74*
	Top F_1	
B	1.07 | 4.2 | **0.83** *0.94* | *0.92* | *0.92* | *0.92* | *0.92*	1.11 | 1.1 | **0.81** *0.87* | *0.8* | *0.91* | *0.85* | *0.88*
C	1.12 | 0.1 | **0.75** *0.93* | *0.96* | *0.91* | *0.93* | *0.92*	1.07 | 4.6 | **0.69** *0.90* | *0.90* | *0.90* | *0.90* | *0.90*
	Top F_2	
B	1.07 | 4.2 | **0.83** *0.90* | *0.80* | *0.98* | *0.88* | *0.94*	1.11 | 1.1 | **0.81** *0.64* | *0.52* | *0.96* | *0.68* | *0.83*
C	1.12 | 0.1 | **0.75** *0.95* | *0.92* | *0.99* | *0.95* | *0.97*	1.07 | 4.6 | **0.69** *0.80* | *0.73* | *0.97* | *0.83* | *0.91*

This difficulty to obtain an estimate of the velocity can be noted in the section of Fig. 11 which corresponds to trial C; in trial C of the 2.02-L/s data, only a small number of points in time following the initial injection were classified as sand by the 200–800-Hz Top F_1 model. Despite this, as depicted for all the trials in Fig. 11, the estimated linear regression slopes follow the visual path of the sand signal quite closely. The strong performance of these velocities is further illustrated by the estimated values presented in Fig. 8 and Table VI.

For the 2.02-L/s flow rate, Fig. 8 provides a visual display of the performance for the Top F_1 models across trials A–C. For the 200–800-Hz models, three of the averages (F_1-score, F_2-score, and R^2) are higher than the 0–5000-Hz data. Besides, the error in estimated sand velocity is significantly lower on average for the 200–800-Hz frequency band, when compared with using the 0–5000-Hz band (15.4% versus 43.9%). While the average classification accuracy across trials is slightly higher for the 0–5000-Hz data, it should be noted that accuracy is the least-valuable metric to report on in the case of imbalanced data; when attention is brought to the more critical metrics of F_1-score and F_2-score, the 200–800-Hz data consistently outcompete the 0–5000-Hz data. Hence, these average results further demonstrate the power of using the acoustic sand fingerprint for sand classification and velocity estimation, specifically for the case of the 2.02-L/s flow rate.

C. Testing Results (2.27-L/s Flow Rate)

For the flow rate of 2.27 L/s, the analytical sand velocity is 1.12 m/s. Using this value, the percentage errors between the estimated velocity and the analytical velocity, in addition to all other performance metrics for the three models, are depicted in Table VII.

Similar to the testing on 2.02 L/s, the Top F_1 model using the 200–800-Hz data performs the strongest, with F_1 and F_2 scores consistently above 0.92, and overall accuracy at or above 93%. Across all the 200–800-Hz models, the average error in velocity was only 3.3%. The models trained with the 0–5000-Hz data do not perform as strongly on classification, though they still do give promising overall results (with an average accuracy of 82%, an average F_1 score of 0.80, an average F_2 score of 0.82, and an average velocity error of 3.2%).

The linear regression lines produced by the classified data from these models are depicted in Fig. 12. Here, the sand signal is seen to move across the pipeline across time for both trials B and C; the velocity estimates produced by these models for the 2.27-L/s data were some of the strongest found in this study, as demonstrated by the aforementioned average error (across all three of the 200–800-Hz models) of only 3.3%.

For the 2.27-L/s flow rate, Fig. 9 provided a visual aggregation of the performance for the Top F_1 models across trials B and C. Notably, all four of the averages for accuracy, F_1-score, F_2-score, and R^2 are higher for the 200–800-Hz data compared with the 0–5000-Hz data; at the same time, the error in estimated sand velocity is lower when using the 200–800-Hz frequency band. Hence, the averaged results further demonstrate the power of using the acoustic sand fingerprint for sand classification and velocity estimation in the case of the 2.27-L/s flow rate. Further comparison of the performance across a variety of frequency bands is provided in a supplementary document.

D. Discussion of Results

Ultimately, the used models perform well at both the classification and velocity estimation tasks, with the 200–800-Hz Top F_1 models having the best performance overall. The ML results using the selected 200–800-Hz frequency band and Top F_1 model show an average sand detection accuracy of 93.4%, an average F_1 score of 0.87, an average F_2 score of 0.85, along with an average velocity error of 10.1% across several gigabytes of blind testing data. Given the very limited training data and the use of out-of-distribution data for testing (the different flow rates and injection settings discussed in Table II of Section II), the performance of the models is noteworthy. These results demonstrate that the model provides reasonably accurate corroboration of sand detection and velocity estimation in the experimental flow loop.

V. FUTURE WORK

While this study marks a substantial step forward in the use of ML for sand monitoring, it also opens avenues for further refinement and exploration in this field.

For one, the current ML algorithms used have the potential for improvement and verification. Steps were taken throughout the training and tuning process to ensure the robustness of the ML models used. However, more investigation of model sensitivity could be conducted to uncover weaknesses and potential avenues of improvement. For instance, input perturbation and analysis of threshold sensitivity could be conducted to further explore the behaviors of the logistic regression models; on the other hand, analysis of the models' sensitivity to outlier points for linear regression, as introduced by Cook [48], could help improve the velocity estimation.

In addition, verification of the data labels and improvements of model performance, certainty, and robustness could be attained using various alternative sensors in addition to DAS. This could include distributed temperature logs as in

Mullens et al. [30], pressure sensors, or surface sensors to corroborate data labels as done by Thiruvenkatanathan et al. [11].

Besides, potential improvement can be made to the scalability and efficiency of the model. While logistic regression is an inherently computationally inexpensive classification method (see Section III-B), the proposed use of the algorithm is limited in that it can only process small inputs corresponding to any one timestep at once. As such, future work could involve modifying the ML model to detect sand over broader time horizons at once, which could improve overall processing speed. This could mandate small changes to the existing logistic regression architecture or could involve exploration of entirely new methods of ML for sand monitoring, such as convolutional networks used on DAS data [49]. The use of CNNs in future work may also hold promise for different reasons: if the input of a convolutional neural network is spatiotemporal in nature, there is possibility for the model to be trained to detect sand presence and velocity concurrently, using a single model. This could once again improve the efficiency of the real-time monitoring approach.

Finally, though it was outside of the scope of this initial investigation, future exploration could include implementing the proposed methodology using well-scale oilfield datasets, to test the method's performance and robustness in real-world operational situations.

VI. Conclusion

This study demonstrates the successful application of ML, combined with advanced signal processing, for automating the detection of sand, as well as estimation of sand velocities using DAS data. By offering a more streamlined and automated approach to sand monitoring, the study contributes to enhancing operational efficiency and reducing the risks associated with sand production in oil and gas operations, such as equipment damage and environmental hazards.

The ML approach was implemented on experimental datasets from eight distinct sand transport tests, representing different flow rates, sand ingress scenarios, and injection locations. A key novelty of the approach was to use only about 25% of the data for training and most of the data for blind testing and validation. The proposed method also stands out for its ability to process and analyze vast quantities of DAS data in real-time addressing a key challenge in using high-resolution distributed sensor information for real-time decision-making. This was enabled using FBE preprocessing for intelligent data compression and enhancement of the sand signature from background signals.

The results also corroborate key conclusions from previous studies and empirically show high performance for characteristic frequencies corresponding to the sand fingerprint. The ML results using the selected 200–800-Hz frequency band and Top F_1 model show an average sand detection accuracy of 93.4%, an average F_1 score of 0.87, and average F_2 score of 0.85 across several gigabytes of blind testing data, along with an average error in the sand velocity estimates of 10.1%, as compared to theoretical expectations. Different combinations of hyperparameters are considered for the ML models, to demonstrate their sensitivity and offer flexibility to end users to customize them based on techno-economic criteria.

Acknowledgment

The authors would also like to acknowledge the assistance from Rishikesh Shetty.

References

[1] J. Carlson, D. Gurley, G. King, C. Price-Smith, and F. Waters, "Sand control: Why and how," *Oilfield Rev.*, vol. 4, no. 4, pp. 41–53, 1992.

[2] F. Sanfilippo, M. Brignoli, D. Giacca, and F. Santarelli, "Sand production: From prediction to management," in *Proc. SPE Eur. Formation Damage Conf.*, The Hague, The Netherlands, 1997, pp. 389–398.

[3] H. Ben Mahmud, V. H. Leong, and Y. Lestariono, "Sand production: A smart control framework for risk mitigation," *Petroleum*, vol. 6, no. 1, pp. 1–13, Mar. 2020.

[4] R. Kiran et al., "Identification and evaluation of well integrity and causes of failure of well integrity barriers (A review)," *J. Natural Gas Sci. Eng.*, vol. 45, pp. 511–526, Sep. 2017.

[5] X. Q. Li, T. G. Zhu, L. T. Fang, and Y. Yuan, "Effect of sand production on casing integrity," *J. Can. Petroleum Technol.*, vol. 46, no. 12, pp. 22–26, Dec. 2007.

[6] U. I. Duru, P. M. Ikpeka, C. Ndukwe-Nwoke, A. Arinkoola, and S. I. Onwukwe, "Quantitative risk assessment of the effect of sand on multiphase flow in pipeline," *Rudarsko-Geološko-Naftni Zbornik*, vol. 37, no. 4, pp. 37–52, 2022.

[7] K. Wang, G. Liu, Z. Liu, and Y. Li, "Non-intrusive identification of offshore sand production in water-gas pipe flow via acoustic sensing method," in *Proc. 29th Int. Ocean Polar Eng. Conf.*, Honolulu, HI, USA, 2019, pp. 2023–2026.

[8] I. Palmer, D. G. Vorpahl, J. M. Glenn, H. Vaziri, and J. McLennan, "A recent Gulf of Mexico cavity completion," *SPE Drilling Completion*, vol. 20, no. 3, pp. 219–223, Sep. 2005.

[9] C. McPhee, C. Farrow, and P. McCurdy, "Challenging convention in sand control: Southern North Sea examples," *SPE Prod. Oper.*, vol. 22, no. 2, pp. 223–230, May 2007.

[10] A. H. Hartog, *An Introduction to Distributed Optical Fibre Sensors*. Boca Raton, FL, USA: CRC Press, 2017.

[11] P. Thiruvenkatanathan, T. Langnes, P. Beaumont, D. White, and M. Webster, "Downhole sand ingress detection using fibre-optic distributed acoustic sensors," in *Proc. Abu Dhabi Int. Petroleum Exhib. Conf.*, Abu Dhabi, United Arab Emirates, 2016.

[12] R. Shetty, J. Sharma, and M. Tyagi, "Experimental study on sand detection and monitoring using distributed acoustic sensing for multiphase flow in horizontal pipes," *SPE J.*, vol. 29, no. 2, pp. 1045–1060, Feb. 2024.

[13] H. S. Bakka, S. Dümmong, K. E. Haavk, Q. Li, and P. A. Olsen, "Real-time fiber optics: A door opener for the wellbore environment," in *Proc. SPE Norway Subsurface Conf.*, Bergen, Norway, 2022.

[14] M. G. Schuberth et al., "A real-time fiber optical system for wellbore monitoring: A Johan Sverdrup case study," in *Proc. SPE Offshore Eur. Conf. Exhib.*, 2021.

[15] T. Park, R. Paleja, and M. Wojtaszek, "Robust regression and band switching to improve DAS flow estimates," in *Proc. SPE Annu. Tech. Conf. Exhib.*, Dallas, TX, USA, 2018.

[16] P. Kasouvi et al., "Correlating distributed acoustic sensing (DAS) to natural fracture intensity for the Marcellus Shale," in *Proc. SEG Tech. Program Expanded Abstr.*, Houston, TX, USA, 2017, pp. 5386–5390.

[17] I. Pakhotina, S. Sakaida, D. Zhu, and A. D. Hill, "Diagnosing multistage fracture treatments with distributed fiber-optic sensors," *SPE Prod. Oper.*, vol. 35, no. 4, pp. 0852–0864, Nov. 2020.

[18] J. Sharma, T. Cuny, O. Ogunsanwo, and O. Santos, "Low-frequency distributed acoustic sensing for early gas detection in a wellbore," *IEEE Sensors J.*, vol. 21, no. 5, pp. 6158–6169, Mar. 2021.

[19] J. Tabjula and J. Sharma, "Feature extraction techniques for noisy distributed acoustic sensor data acquired in a wellbore," *Appl. Opt.*, vol. 62, no. 16, pp. E51–E61, 2023.

[20] J. Selker et al., "Distributed fiber-optic temperature sensing for hydrologic systems," *Water Resour. Res.*, vol. 42, no. 12, 2006.

[21] C. K. Kao and G. A. Hockham, "Dielectric-fibre surface waveguides for optical frequencies," *Proc. Inst. Elect. Eng.*, vol. 113, no. 7, pp. 1151–1158, 1966.

[22] X. Lang, R. Singh, B. Zhang, and S. Kumar, "Highly sensitive TIT4T fiber-based waveflex biosensors functionalized with MXene-QDs for xanthine detection," *IEEE Sensors J.*, vol. 24, no. 2, pp. 1564–1571, Dec. 2024.

[23] X. Liu et al., "SFFO cortisol biosensor: Highly sensitive S-flex fiber optic plasmonic biosensor for label-free cortisol detection," *IEEE Sensors J.*, vol. 24, no. 2, pp. 1494–1501, Jan. 2024.

[24] H. F. Taylor and C. E. Lee, "Apparatus and method for fiber optic intrusion sensing," U.S. Patent 5 194 847, Mar. 16, 1993.

[25] T. Sadigov et al., "Real-time water injection monitoring with distributed fiber optics using physics-informed machine learning," in *Proc. Offshore Technol. Conf.*, Houston, TX, USA, 2021.

[26] C. Cerrahoglu et al., "Real-time applications of sensor analytics for production and injection profiling," in *Proc. ADIPEC*, Abu Dhabi, United Arab Emirates, 2022.

[27] M. Alkhalaf, F. Hveding, and M. Arsalan, "Machine learning approach to classify water cut measurements using DAS fiber optic data," in *Proc. Abu Dhabi Int. Petroleum Exhib. Conf.*, Abu Dhabi, United Arab Emirates, 2019.

[28] J. Tejedor, J. Macias-Guarasa, H. Martins, J. Pastor-Graells, P. Corredera, and S. Martin-Lopez, "Machine learning methods for pipeline surveillance systems based on distributed acoustic sensing: A review," *Appl. Sci.*, vol. 7, no. 8, p. 841, Aug. 2017.

[29] D. Miklashevskiy et al., "Approach for wellbore production monitoring using distributed acoustic noise measurements," in *Proc. Int. Petroleum Technol. Conf.*, Dhahran, Saudi Arabia, 2020.

[30] S. Mullens, G. Lees, and G. Duvivier, "Fiber-optic distributed vibration sensing provides technique for detecting sand production," in *Proc. Offshore Technol. Conf.*, Houston, TX, USA, 2010, pp. 432–444.

[31] Z. Hasanov et al., "Production optimization of sanding horizontal wells using a distributed acoustic sensing (DAS) sand monitoring system: A case study from the ACG field in Azerbaijan," in *Proc. SPWLA 62nd Annu. Logging Symp.*, 2021.

[32] C. N. Emiliani et al., "Improved sand management strategy: Testing of sand monitors under controlled conditions," in *Proc. SPE Annu. Tech. Conf. Exhib.*, Denver, CO, USA, 2011.

[33] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2022.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1–9.

[36] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020.

[37] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. 7th Int. Conf. Document Anal. Recognit.*, Edinburgh, U.K., 2003, pp. 1–6.

[38] A. J. Ratner, H. R. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré, "Learning to compose domain-specific transformations for data augmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 1–11.

[39] D. R. Cox, "The regression analysis of binary sequences," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 20, no. 2, pp. 215–232, 1958.

[40] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[41] Y. Bengio, I. Goodfellow, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[42] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[43] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," in *Proc. 3rd IEEE Int. Conf. Data Mining*, Melbourne, FL, USA, Nov. 2003, pp. 435–442.

[44] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[45] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2001, pp. 1–6.

[46] V. Matoušek, "Solids transportation in a long pipeline connected with a dredge," in *Terra et Aqua*. Voorburg, The Netherlands: International Association of Dredging Companies (IADC), 1996.

[47] T. Yagi, T. Okude, S. Miyazaki, and A. Koreishi, "Analysis of hydraulic transport of solids in horizontal pipes," in *Port & Harbour Research Institute*. Yokosuka, Japan: Nagase, 1972.

[48] R. D. Cook, "Detection of influential observation in linear regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, 1977.

[49] H. Gemeinhardt and J. Sharma, "Machine-learning-assisted leak detection using distributed temperature and acoustic sensors," *IEEE Sensors J.*, vol. 24, no. 2, pp. 1520–1531, Jan. 2024.

[50] K. Rehman and F. Nawaz, "Remote pipeline monitoring using wireless sensor networks," in *Proc. Int. Conf. Commun., Comput. Digit. Syst. (C-CODE)*, Islamabad, Pakistan, Mar. 2017, pp. 32–37.

[51] J. L. Tabjula and J. Sharma, "Comparison of DAS and FBG sensitivity for detecting and quantifying small pipeline leaks," in *Proc. SPIE Defense + Commercial Sens.*, Orlando, FL, USA, 2023.

[52] J. Tabjula, R. Shetty, T. Adeyemi, and J. Sharma, "Empirical correlations for predicting flow rates using distributed acoustic sensor measurements, validated with wellbore and flow loop data sets," *SPE Prod. Oper.*, vol. 38, no. 4, pp. 678–693, Nov. 2023.

[53] A. H. Hartog, "Optical fibre sensors in the oil, gas, and geothermal energy extraction," in *Light, Energy Environ., OSA Tech. Dig.* Washington, DC, USA: Optica Publishing Group, 2014.

[54] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.

Harrison Gietz is pursuing the B.S. degree in mathematics with Louisiana State University, Baton Rouge, LA, USA.

He plans to pursue graduate studies in computer science, with research interests in machine learning (ML) for engineering applications, natural language processing, and artificial intelligence (AI).

Jyotsna Sharma received the B.Tech. degree in electrical engineering from Indian Institute of Technology, Delhi, India, in 2006, and the Ph.D. degree in petroleum engineering from the University of Calgary, Calgary, AB, Canada, in 2012.

She joined the Department of Petroleum Engineering, Louisiana State University (LSU), Baton Rouge, LA, USA, in 2019, after working in the oil industry for over eight years at Chevron, Schlumberger, and Shell, Canada. At LSU, she serves as an Assistant Professor with the Department of Petroleum Engineering and as an Adjunct Professor with the Department of Electrical Engineering. As a Research Engineer at Chevron, she led projects in USA, Indonesia, Canada, and Venezuela. At Schlumberger, she worked as a Field Engineer in Canada conducting wireline well logging. She worked as a summer researcher at Shell Canada, Stanford University, Stanford, CA, USA, and Technical University of Dresden, Dresden, Germany. Her research interests include application of distributed fiber optic sensing and machine learning in the energy industry.

Mayank Tyagi received the B.Tech. degree in mechanical engineering from Indian Institute of Technology (IIT), Kanpur, Kanpur, India, in 1995, and the Ph.D. degree in mechanical engineering from Louisiana State University (LSU), Baton Rouge, LA, USA, in 2003.

He serves as a Professor with the Department of Petroleum Engineering, LSU. He also holds a joint faculty appointment at the Center for Computation and Technology (CCT), LSU, since 2007. His current research interests span across physics-based simulations using high-performance computing, data analytics, and machine learning for interdisciplinary petroleum engineering applications such as image-based pore-scale modeling using lattice Boltzmann method, multiscale multiphase computational fluid dynamics, geothermal reservoir engineering, and unconventional reservoir simulations.