CHECKPOINT: IDENTIFYING A DRILLING STATE MACHINE

*FORMERLY SUPERVISED LEARNING METHODS FOR DRILLING STATE MACHINE PREDICTION

Prepared by Nicole McCarthy on June 11, 2025

Deere Development Company, LLC.

Contents

Abbreviations

DT	Decision Tree
GPM	Gallons per Minute
GMM	Gaussian Mixture Model
HDBSCAN	Hierarchy Density-Based Spatial Clustering of Applications with Noise
HMM	Hidden Markov Model
MFI	Mud Flow Index
MR	Motor (on a drill string)
PD	Pulser Device
RF	Random Forest
ROP	Rate of Penetration
RPM	Revolutions per Minute
RT	Real Time
WOB	Weight on Bit

Executive Summary

The client, Deere Development Company, has requested a method for predicting the drilling state machine (as defined by them) based on typical downhole drilling parameters for flow, rotation, vibration, and WOB for oil rig drilling. The drilling state machine defines five states: Pumps Off, Pumps On, Sliding Drilling, Rotating, and Rotational Drilling. By identifying the state correctly, the client can better implement other software for drilling optimization. This combination will make oil rig work more cost effective, time efficient, and safe for workers. The methods slated for development, in order of priority, are Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), and Hierarchy Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). The two clustering methods will be implemented as a Random Forest (RF) for real-time identification.

Introduction

This project focuses on using drilling parameters of oil rig drill strings to increase the knowledge and efficiency of runs. The client, Deere Development Company, has requested a method of estimating the drilling state machine, as defined by them, based on the drilling parameters sensed by the tool in real time. The knowledge of the drilling state machine can then be utilized to turn on and off other software that optimizes drilling efficiency and safety, critically in terms of energy, cost, and time.

The drilling state machine has been defined with five modes. High and Low are relative terms to describe how each parameter changes from the last drilling state. The drilling state machine can only move consecutively with the exception of (1) providing access to (2) and (3), but (2) not providing access to (3) (see Figure 1) (Deere, 2025).

There are no existing solutions for this specific problem, but similar work has used the same drilling parameters to predict geological formations and torque on the drill string. Therefore, this project aims to expand on the methods previously used for classification in real time and shift their focus towards better understanding the drilling itself.

Drilling Parameter	Flow	Rotation	Vibration	WOB
(0) Pumps Off	Low	Low	Low	Low
(1) Pumps On	High	Low	High	Low
(2) Sliding Drilling	High	Low	High	High
(3) Rotating	High	High	High	Low
(4) Rotational Drilling	High	High	High	High

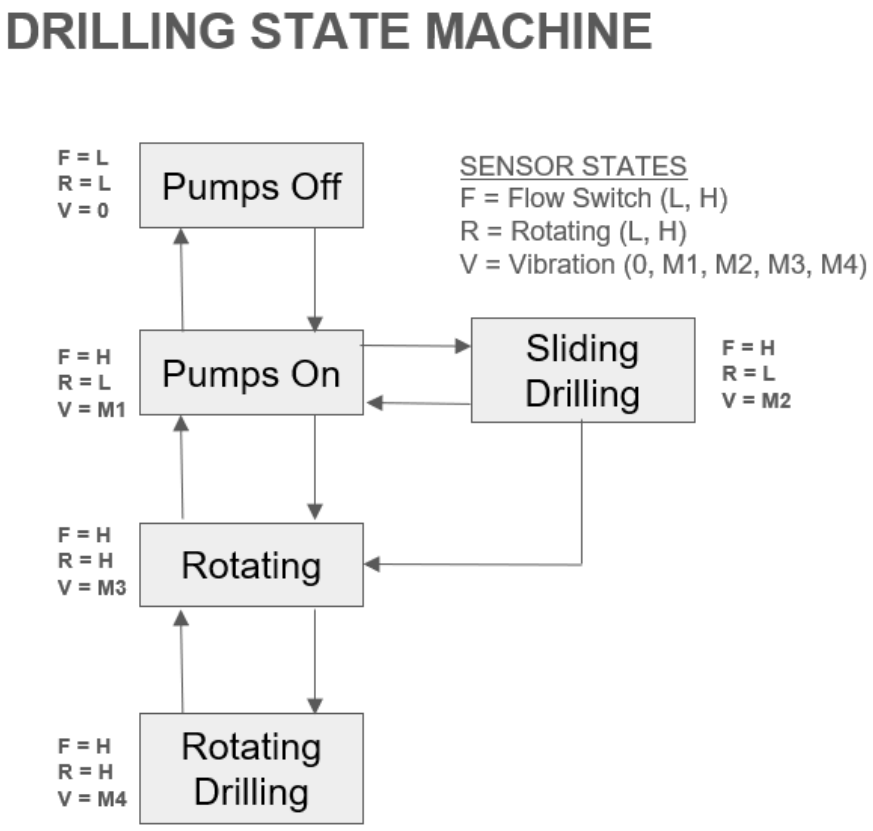

Figure 1 Deere Development Company's Drilling State Machine

Related Work

The problem of using drilling parameters to predict things about the downhole environment has become popular recently, especially for predicting geological formations. These projects have similar parameters to those proposed in this project to solve other problems. For instance, Hassan et al. (2024) tested three different models using ROP, GPM (flow), RPM, strokes per minute, torque, and WOB to predict the porosity and permeability of the geological formations being drilled. Of decision tree (DT), random forest (RF), and support vector machine (SVM) models, each had a correlation coefficient of above 0.8, but RF performed the best with 0.92. This is likely due to the robustness built into RF, since it uses the aggregate of multiple decision trees. A visualization on the prior work done on predicting geological formations from

drilling parameters can be found in Hassan et al. (2024, p. 17068). These methods range from supervised to unsupervised learning.

Unsupervised learning techniques are also relevant for a range of research topics, both within the oil and gas industry and outside of it. For instance, Tashnizi (2022) wrote his master's thesis on finding outliers in pump jack performance to identify failures and optimize future performance. During this research, he tested K-Means clustering, Gaussian Mixture Model (GMM), and Hidden Markov Model (HMM) with a GMM implementation. Both GMM and HMM outperformed K-Means in accuracy for cluster sets larger than four (Tashnizi, 2022, p. 67).

Furthermore, in a study on clustering time series data, Morad (2020) found that GMM alone was a viable candidate for accurate clustering this type of data. While the applications to oil and gas for GMM are not as prolific, one study used this method to successfully cluster images of oil spills and leaks from offshore drilling platforms (Weiwei et al., 2016).

No applications of Hierarchical Density-Based Spatial Clustering for Applications with Noise (HDBSCAN) were found in the oil and gas industry, but its potential applications for nonlinear data were promising in other transferable fields. For example, HDBSCAN has been applied in an industrial environment, that of welding robotic cells, to monitor work for anomalies or defects. This paper specifically noted its competence in dealing with noisy data and time-dependent parameters (Blachowicz et al., 2025). Another study noted its superior clustering efficiency and stability in comparison with K-Means and DBSCAN when classifying records of safety and environmental conservation management (Zhang et al., 2024).

This project will build on the findings of these experiments by focusing on unsupervised learning techniques for labeling data and applying the drilling parameters to predict the drilling state machine rather than the surrounding environment or torque. This technology will be more applicable to a wider range of associated technologies for optimizing drilling in terms of efficiency, safety, and accuracy.

Proposed Work

The datasets for this project have been provided by the client, Deere Development Company. They represent six runs from three different oil rigs, named Flybar 1WB, Flybar 1WC, and Flybar 2WC. These datasets include time and depth as well as the following attributes relevant to this project. Each item indicated below represents a column in the dataset that will be used for the project.

DRILLING PARAMETER	ORIGINAL DATA	PREPROCESSED DATA	PROCESSING CHANGES
TIME	RT_Time	Time	Rename, Fill Down
DEPTH	RT_Depth	Depth	Rename, Fill Down
FLOW	• RT_Pumps_Off • RT_Pumps_On	GPM	Aggregate, Fill Down
ROTATION	MR_RPM-AVG.MR	Motor_RPM	Rename, Fill Down
VIBRATION	• MR_VIBA, MR_VIBL • PD_Axial Vibration, PD_Lateral Vibration	• Motor_Axial_ Vibration • Motor_Lateral _Vibration • Pulser_Axial_ Vibration • Pulser_Lateral _Vibration	Rename, Fill Down
WOB	RT_WOB	WOB	Rename, Fill Down

This project initially intended to focus on using validation data to create labels and then use supervised learning methods, such as decision trees and support vector machines, to determine the drilling state machine at any given time during a run. However, after preprocessing the data, the validator columns showed signs of insufficient data for use as accurate labeling. Therefore, this project has had to be reformed as an unsupervised approach to labeling drilling parameter data as parts of a drilling state machine. The new plan is outlined here.

The primary goal is to test two clustering methods and one sequential modeling method for adding the drilling state parameter labels to the training data. Gaussian Mixture Models (GMM) and Hierarchy Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) are the clustering methods chosen in light of their ability to work with non-linear data. GMM relies on probabilistic methods while HDBSCAN creates a hierarchy of connected components sorted by distance (Carrasco and Whitfield, 2024; McInnes, 2016). Hidden Markov Model (HMM) has been added as a sequential modeling method to address the time-series nature of the data and potentially provide more accurate labels for the training data (Jurafsky and Martin, 2024). This is particularly well suited for this project as we already have a Markov chain, namely the prepared drilling state machine.

If there is time after these items have been completed, the project will move forward with training a Random Forest (RF) on the clustering methods to test on the testing dataset. In the case of the HMM, the model itself will be used for testing. The reason for choosing these two models is their purported efficiency and accuracy in real-time applications (Hassan, 2024).

Progress

Preprocessing

The data was cleaned such that it includes the parameters seen in the previous table with no missing values. To avoid null values, the last value was repeated until a new one was found. If no value was found in the first row, a 0 was inserted. After the data was cleaned, it was segmented into each run specified by well.

Warehousing

As a final step in the preprocessing code, the data frames for each rig (in its entirety) and run were saved to CSVs in a private folder so as to protect the information of both the client and their customers.

Evaluation

This project will utilize the data from Flybar 1WB and Flybar 1WC to train the three methods. These methods will then be tested on the singular, but lengthy, run conducted on Flybar 2WC. The clustering methods will be evaluated with the silhouette coefficient and Davies-Bouldin Index. The HMM will be evaluated in a similar manner. All three methods will also undergo a manual inspection that uses domain knowledge to determine how accurate the state labels are. Finally, the RF will be evaluated through testing on data clustered by the original method that was used in training the RF.

Discussion

The project is currently in the modeling stage. Understanding the data had already been completed before this project was initiated. With the data now preprocessed and stored in an

accessible format, the bulk of the remaining time will be spent on modeling. A rough timeline for the rest of the project is below.

The most significant challenge for this project is the condensed timeline. Should the scope of the project prove to be too large for the given time, the clustering methods will be eliminated and the HMM will be evaluated without comparisons to traditional clustering. Furthermore, the project may be able to gather clustering results from either GMM or HDBSCAN, or both. In this case, only the RF implementation and testing may be eliminated. At this time, there are no other challenges of significance anticipated. Although this was stated previously, the labeling data was found to be mute.

DATE	ASSIGNMENT DUE	ASSIGNMENT TASKS	RESEARCH TASKS
JUNE 12			HMM Training
JUNE 13			HMM Testing
JUNE 16			UMAP Processing
JUNE 17			GMM + HDBSCAN
JUNE 18			Clustering Evaluation
JUNE 19			RF Training + Testing
JUNE 20		Methodology, Results, Discussion	RF Evaluation
JUNE 23	Final Report	Conclusion, Summary, Presentation Slides	
JUNE 24	Final Presentation	Record Presentation	

Conclusion

This is an ambitious project taking course over the month of June 2025. Its primary goal is to label the training data with Deere Development Company's drilling state machine using drilling parameters for time, flow, rotation, vibration, and WOB. Its secondary goal is implementing and testing methods for predicting a tool's drilling state in real time. By using the

most accurate methods for labeling the nonlinear, time-series data, further work on creating real-time drilling state prediction tools may be done. The final goal is for the client to be able to implement other software with higher effectiveness and significantly decrease costs, time, and safety risks for workers. The building and testing of methods will be prioritized as such:

1. Hidden Markov Model Training and Testing
2. Gaussian Mixture Model Training
3. HDBSCAN Training
4. Clustering Methods Evaluation
5. Random Forest Training and Testing

The methods outlined above will be trained on drilling data from five runs on two different wells. It will then be tested on one run from a third well. All drilling data was provided by the client from real oil rig work. Once the methods that were developed (based on time availability) have been tested, they'll be evaluated for performance based largely on domain knowledge and inter/intra cluster similarity indexes.

References

Blachowicz, T., Wylezek, J., Sokol, Z., and Bondel, M. (2025). "Real Time Analysis of Industrial Data Using the Unsupervised Hierarchical Density-Based Spatial Clustering for Applications with Noise in Monitoring the Welding Process in a Robotic Cell." *Information 16*(2). https://doi.org/10.3390/info16020079.

Carrasco, O. and Whitfield, B. (2024). "Gaussian Mixture Model Explained." *Built In.* https://builtin.com/articles/gaussian-mixture-model.

Deere, P. Personal Communication. March 8, 2025.

Hassaan, S., Mohamed, A., Ibrahim, A. F., and Elkatatny, S. (2024). "Real-Time Prediction of Petrophysical Properties Using Machine Learning Based on Drilling Parameters." *ACS Omega*, *9*(15), 17066–17075. https://doi.org/10.1021/acsomega.3c08795.

Jurafsky, D. and Martin, J. (2024). “Hidden Markov Models.” *Speech and Language Processing*. https://web.stanford.edu/~jurafsky/slp3/A.pdf.

Lee, S. (2025). “10 Statistical Methods to Boost Clustering Validity Results.” *Number Analytics*. https://www.numberanalytics.com/blog/10-statistical-methods-clustering-validity-results.

McInnes, L. (2016). “How HDBSCAN Works.” *HDBSCAN*. https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html.

McInnes, L. (2018). “How UMAP Works.” *Uniform Manifold Approximation & Projection*. https://umap-learn.readthedocs.io/en/latest/how_umap_works.html.

Morad, N. (2020). “Modeling Methods in Clustering Analysis for Time Series Data.” *Open Journal of Statistics 10*(3), 565-580. https://doi.org/10.4236/ojs.2020.103034.

Tashnizi, M. (2022). *Application of Hidden Markov Model in Production Data Analysis*. [Master’s thesis, Montan Universitat Loeben]. https://pureadmin.unileoben.ac.at/ws/portalfiles/portal/17548059/AC16839895.pdf.

Weiwei, J., Yupeng, Z., Wei, A., and Jianwei, L. (2016). “Application of Gaussian Mixture Model in Identification of Oil Spill on Sea.” *Sixth International Conference on Machinery, Materials, Environment, Biotechnology, and Computer*. https://www.atlantis-press.com/article/25858820.pdf.

Zhang, L., Su, X., Wang, Y., Wang, M., Yang, X., and Xu, Z. (2024). "HDBSCAN-Based Semantic Clustering Model in Classifying Incidents on Security and Environmental Conservation Management", *Ninth International Symposium on Advances in Electrical, Electronics, and Computer Engineering*. https://doi.org/10.1117/12.3033910.