D13.2 Systematic Field Testing and Data-Driven Validation of the OptiDrill Drilling Advisory System: Objectives, Data, and Outcomes

Deliverable No	D13.2
Work package No. and Title	WP13 OPTIDRILL system field-scale test and validation in a borehole
Version - Status	V1.0 – Final
Date of Issue	16/12/2024
Dissemination Level	PUBLIC
Filename	D13.2 - Final

This project has received funding from the European Union's Horizon 2020 research and innovation action under grant agreement No 101006964

DOCUMENT INFO

Authors

Author	Organization	e-mail
Henning Knauer	IEG	henning.knauer@ieg.fraunhofer.de
Shahin Jamali	IEG	shahin.jamal@ieg.fraunhofer.de

Document History

Date	Version	Editor	Change	Status
16.12.2024	0.0	Henning Knauer	Finished first draft	draft
15.04.2025	1.0	Henning Knauer	Change of scope based on revision by PO	Draft
20/04/25	V1.0	Shahin Jamali	Final revision	Final

TABLE OF CONTENTS

TABLE OF CONTENTS 3
LIST OF FIGURES 4
LIST OF TABLES 4
EXECUTIVE SUMMARY 5
1. INTRODUCTION 6
2. FIELD TESTING 7
2.1 Setup and Procedure 7
2.2 Technical Objectives 8
2.3 Acquired Data 8
3. Expectations and Results 11
4. MACHINE LEARNING MODEL OPTIMIZATION 15
4.1 ROP Prediction Model 15
4.2 Drilled Lithology Prediction Model 16
5. CONCLUSION 18

LIST OF FIGURES

Figure 1: Locations of the drill sites of the development dataset (Geostar 2) and the field test (Push-It P1 wellbore) 7

Figure 2: Plot of the gathered and processed drilling process parameters over the measured depth for the first field test in Bochum 9

Figure 3: Actual ROP in blue and predicted ROP in red plotted over the measured depth for the P1 dataset processed with the new scaler. The MAE for a window of 100 data points is plotted in orange. 15

Figure 4: Confusion matrix of the model predictions for the Push-It P1 wellbore with a new scaler 16

LIST OF TABLES

Table 1: Sample of the drilling data recorded during the first field test. (Units: depth [m], RPM [rpm], WOB [kN], torque [kNm], pressure [kPa], mud flow [l/min], ROP [m/h]) 8

Table 2: Sample of the OptiDrill systems outputs generated during the drilling operation. (pre. = predicted, exp. = expected, after optimization, rec. = recommended based on optimization) 10

Table 3: Cross-validation results of the ROP prediction model on the Geostar 2 dataset from Bochum 12

Table 4: ROP prediction model metrics from the two field tests 12

Table 5: Cross-validation results of the drilled lithology prediction model on the Geostar 2 dataset from Bochum (Clst. = Claystone, Slst. = Siltstone, Sast. = Sandstone) 13

Table 6: Statistical overview over predicted and actual ROP values including the absolute error of the model 15

Table 7: Performance metrics for the reconstructed drilled lithology prediction model outputs 16

EXECUTIVE SUMMARY

This report deals with the discussion of the results of the two rounds of field testing of the OptiDrill drilling advisory system with regard to the technical objectives of the conducted tests and whether the results that could be achieved met the expectations of the project team. A list of technical test objectives is formulated aiming at the general functionality and reliability of the system, but more importantly at the predictive performances of the deployed artificial intelligence based software modules. Furthermore, samples of the datasets created during the system deployment are shown and described.

The main discussion focuses on how the expectations on the AI-based modules performances during deployment in a real drilling operation could be met. The expectations were quantified based on the results of the cross-validation of the models as part of the training and evaluation prior to the field tests. Concerning the ROP prediction and optimization module, the expectation could be met. The ROP prediction model was able to deliver promising results due to unfavorable circumstances during inference. The drilled lithology prediction module, however, could not meet the expectations and the results from the field tests, while an improvement could be observed, could not reach a level comparable to that achieved on the training data. The expectations towards the anomaly detection module were more difficult to formulate since the underlying algorithm is based on unsupervised learning and drilling problem are typically very scarce within datasets.

1. INTRODUCTION

Work package number 13 deals with the field testing of the OptiDrill system and the evaluation and validation of the system's outputs. Within deliverable 13.1 the first field test of the OptiDrill system including the drilling advisory system prototype and the sensors is reported on. Deliverable report 13.3 focuses on the second round of field testing and presents and discusses the respective results. Within this report we are going to describe the detailed technical objectives of the field tests that were conducted, the data that has been gathered during the deployment of the OptiDrill drilling advisory system and the results that could be achieved.

In section two of the report the field test procedure and setup is described and a list of technical objectives is given that were to be examined during the two iterations of field testing. The list comprises objectives focusing on the general functionality, practicability and reliability of the drilling advisory system on the one hand and the predictive capabilities of the AI-based software modules on the other hand.
Furthermore, some examples of the data that has been gathered during the field test are given. These examples comprise samples from two datasets, one containing the processed drilling parameters that were gathered and used as inputs to the different software modules and a second one consisting of the outputs that were generated during the AI-based models' inference.
Both field tests have been completed successfully providing valuable insights into the OptiDrill system's performance and its strengths and weaknesses. Based on the results from the field tests within section three we are going to discuss and compare the expectations regarding the different AI modules' performances. The expectations in terms of measurable metric values are derived from the k-fold-cross validation results from the model training and validation.
As an addition, within section four the optimization of the ROP and drilled lithology prediction models used during the first field test is discussed. In order to adapt to the differences of the drilling process parameters values that were encountered during the first field test in comparison to the values from the training data, a different scaler was used to process the input data and the model outputs could be reproduced, leading to improved performance in both cases.

2. FIELD TESTING

As already described in deliverable report 13.1, the first round of field testing of the OptiDrill system was conducted in late September at a drill site in Bochum. The drilling operation of the field test took four full days and the OptiDrill system was run for several hours on each day of the drilling operation. During the drilling operation a total of around 127m of drilled depth was monitored and analysed using the OptiDrill software. The project partner Oliveris (formerly PVI) also joined the drilling operation for one day to test the OptiDrill sensor system.

Figure 1: Locations of the drill sites of the development dataset (Geostar 2) and the field test (Push-It P1 wellbore)

The second round of field testing was conducted in November 2024 at the same location, however due to some issues encountered by the drilling team, with a change of drilling equipment and drilling strategy. The drilling operation took 8 days in total, with a overall run-time of the OptiDrill system of over 40 hours while around 100 metres of borehole were monitored and analysed. PVI joined again for the testing of the sensor system with the same setup as during the first field test.

2.1 Setup and Procedure

During the whole drilling operation, which was reported on in the previous deliverable 13.1, the OptiDrill Drilling Advisory system was run on an outdoor laptop that was connected to the drill rig. The laptop was set up in a save distance away from the drill rig and other machinery that was operated at the drill site.

The data received from the drill rig was in parallel captured and saved using the *Wireshark*[1] software in case it would be needed for further analysis or troubleshooting at a later point. Since, not technical issues were encountered during the field test, the data was not needed at a later point since the system's logs files contained all data from the drilling operation. Additionally, screen recordings of the OptiDrill software have been taken regularly when a new drill pipe was added to the drill string. The outputs of all three machine learning based modules have been saved together with the respective time stamps and drilling process parameters in a csv file. The complete drilling process parameter dataset that has been created during the software runtime has been continuously saved while running the system and can be used to reconstruct the OptiDrill systems outputs.

---

[1] Wireshark · Go Deep

The testing of the OptiDrill system was started after the standpipe was set at around 13 metres of measured depth. In total, over the four days of the drilling operation, the OptiDrill system ran for over 17 hours, while making around 8000 unique predictions. In some cases, the drilling process parameters used as input to the machine learning models did not change or were invalid, especially during flushing or pipe changes, leading to duplicate outputs of the models.

2.2 Technical Objectives

The technical objectives of the field testing of the OptiDrill system include the testing and evaluation of the overall functionality of the developed prototype in a real drilling environment on the one hand and the evaluation of the AI modules performances on completely new, unseen data from a different location on the other hand.

The test objectives can be summarized as follows:

1. Test the prototypes overall functionality and technical reliability in a real drilling environment (system runs smoothly without unexpected crashes or malfunctions)
2. Test the system's data processing and gathering functionalities
3. Test the practicality and user friendliness of the graphical user interface (GUI) designed for the visualization of the system's outputs
4. Test and evaluate the system's predictive capabilities based on the three AI-based software modules deployed during the field testing
    a. ROP prediction and optimization
    b. Drilled lithology predictions
    c. Anomaly detection

2.3 Acquired Data

The following tables and graphs show some excerpts and give examples of the data that has been gathered during the first field test of the OptiDrill drilling advisory system. The gathered data includes the processed, in this case meaning filtered, cleaned and merged, surface drilling process data that was received from the drill rig during the field testing. This data was visualized on the system's GUI and used as input data for the three AI-based software modules.

Table 1: Sample of the drilling data recorded during the first field test.
(Units: depth [m], RPM [rpm], WOB [kN], torque [kNm], pressure [kPa], mud flow [l/min], ROP [m/h])

Timestamp	Depth	RPM	WOB	Torque	Pressure	Mud_Flow	ROP
25.09.2024 13:54:33	13.67	36.00	16.06	3.51	1000.00	1735.00	11.33
25.09.2024 13:54:37	13.68	36.33	15.79	3.55	1000.00	1737.00	11.00
25.09.2024 13:54:40	13.69	36.00	15.48	3.48	1000.00	1738.00	10.75
25.09.2024 13:54:44	13.70	36.00	15.34	3.50	1000.00	1741.00	11.00
25.09.2024 13:54:47	13.71	36.00	15.79	3.53	1000.00	1728.67	11.33
25.09.2024 13:54:50	13.72	36.00	16.35	3.56	1000.00	1733.00	12.67
25.09.2024 13:54:53	13.73	36.00	16.33	3.56	1000.00	1735.50	14.50
25.09.2024 13:54:55	13.74	36.00	16.51	3.63	1000.00	1733.67	14.00
25.09.2024 13:54:59	13.75	36.00	16.74	3.58	1000.00	1733.33	12.67
25.09.2024 13:55:02	13.76	36.00	16.87	3.61	1000.00	1733.00	13.00
25.09.2024 13:55:04	13.77	36.00	16.68	3.71	1000.00	1733.00	13.67
25.09.2024 13:55:07	13.78	36.00	16.74	3.63	1000.00	1738.67	13.00
25.09.2024 13:55:10	13.79	36.00	16.35	3.58	1000.00	1731.67	12.00

25.09.2024 13:55:13	13.80	36.00	16.55	3.60	1000.00	1734.00	12.67
25.09.2024 13:55:17	13.81	36.00	16.63	3.65	1000.00	1735.00	15.50
25.09.2024 13:55:19	13.82	36.00	16.53	3.63	1000.00	1740.50	15.50
25.09.2024 13:55:21	13.83	36.00	16.68	3.65	1000.00	1738.00	14.33
25.09.2024 13:55:24	13.84	36.00	16.48	3.63	1000.00	1735.00	13.00
25.09.2024 13:55:27	13.85	36.00	16.28	3.78	1000.00	1735.00	15.00

The table above shows a sample of the surface drilling data that was recorded and processed during the first field test. Several filters are applied to the real-time data that is received from the drill rig during the drilling operation. These filters are used for cleaning the data and discarding any data points that are invalid, duplicate, or erroneous. The filtering includes the removal of data points that contain negative or otherwise implausible measurements for certain parameters, e.g. very high rate of penetration (ROP, drill speed) values or negative values. Furthermore, data points containing declining depth values are discarded, for instance indicating tripping or pipe change operations that do not deliver valuable data for the optimization of the drilling operation. After inspecting the complete dataset gathered and processed during the first field testing run of the OptiDrill system the proper function of all filters and additional processing operations could be verified.

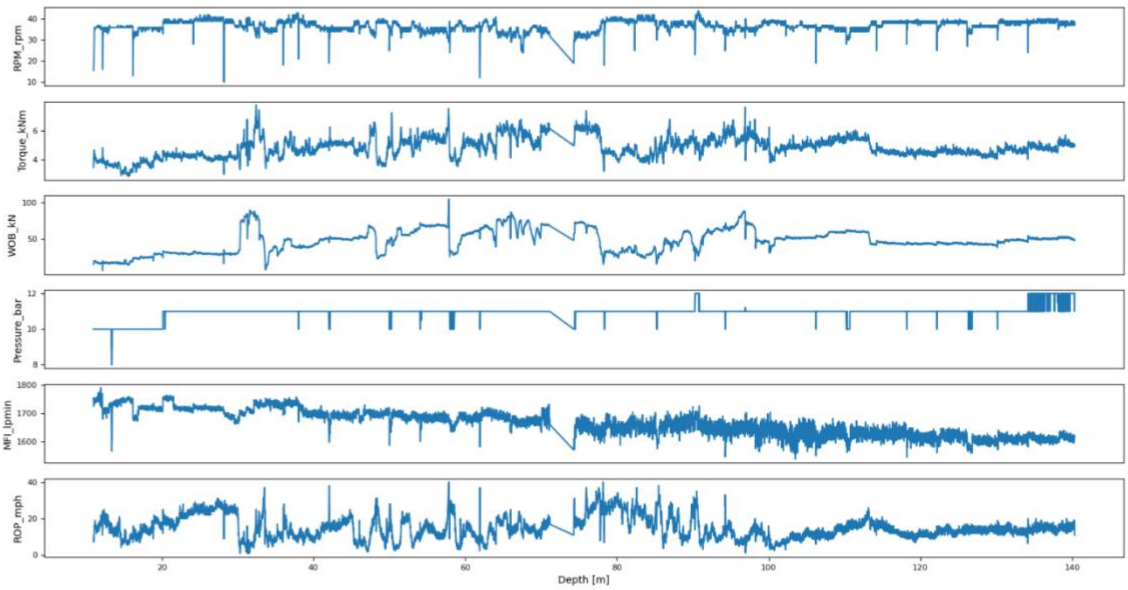

Figure 2: Plot of the gathered and processed drilling process parameters over the measured depth for the first field test in Bochum

The graphs above show the individual drilling process parameters that were gathered and processed during the first field test plotted over the measured depth of the wellbore. When looking at the plots one anomaly, in this case being a gap with no recorded data, can be observed at around 75 metres depth. This gap in the data was caused by human error. During the pipe change at this depth an operating error on the drill rig caused the weight on bit (WOB) measurements recorded to be negative resulting in the respective data points from this drill pipes to be discarded. The error was fixed after a new drill pipe was attached.

Table 2: Sample of the OptiDrill systems outputs generated during the drilling operation.
(pre. = predicted, exp. = expected, after optimization, rec. = recommended based on optimization)

Time	Pre. Lithology	Anomalous	Pre. ROP	Exp. ROP	Rec. RPM	Rec. WOB	Rec. Mud_Flow
25.09.2024 14:58:59	Claystone	TRUE	23.35	23.49	31.2	23.15	2047.8
25.09.2024 14:59:03	Claystone	TRUE	21	23.08	31.2	23.84	2052
25.09.2024 14:59:08	Claystone	TRUE	25.23	25.2	31.2	23.54	2061.6
25.09.2024 14:59:12	Claystone	TRUE	21.29	23.8	31.2	22.76	2050.8
25.09.2024 14:59:16	Claystone	TRUE	21.85	23.04	31.2	23.15	2052
25.09.2024 14:59:20	Claystone	FALSE	24.44	25.33	31.2	23.35	2059.2
25.09.2024 14:59:25	Claystone	TRUE	22.63	23.52	31.2	23.35	2052
25.09.2024 14:59:29	Claystone	TRUE	21.86	23.1	30.4	23.05	2056.8
25.09.2024 14:59:34	Claystone	FALSE	23.79	24.96	31.2	23.25	2048.4
25.09.2024 14:59:38	Claystone	TRUE	21.34	23.66	30.4	23.25	2050.8
25.09.2024 14:59:43	Claystone	TRUE	21.6	23.56	30.8	23.35	2040
25.09.2024 14:59:47	Claystone	FALSE	22.64	24.13	31.2	22.96	2050.8
25.09.2024 14:59:52	Claystone	TRUE	22.87	23.09	31.2	23.35	2053.8
25.09.2024 14:59:56	Claystone	TRUE	21.27	23.41	31.2	22.86	2059.2
25.09.2024 15:00:00	Claystone	TRUE	22.12	24.09	30.4	22.96	2053.2
25.09.2024 15:00:04	Claystone	TRUE	22.12	24.09	30.4	22.96	2053.2
25.09.2024 15:00:09	Claystone	TRUE	22.12	24.09	30.4	22.96	2053.2
25.09.2024 15:00:13	Claystone	TRUE	22.12	24.09	30.4	22.96	2053.2
25.09.2024 15:00:17	Claystone	FALSE	22.12	24.09	30.4	22.96	2053.2

The table above shows a sample of the OptiDrill system's outputs from the first field tests. For better visibility some additional columns, e.g. containing the respective input data, have been excluded from the sample. The table contains the timestamp at which the predictions were made alongside the outputs of the respective software modules for the drilled lithology prediction, the anomaly detection, and the ROP prediction and optimization. The column "Pre. Lithology" contains string values representing the class of lithology that was predicted. The column "Anomalous" contains the outputs of the anomaly detection module, which is a Boolean that assumes the value *True* in case an anomaly is detected. The two columns "Pre. ROP", predicted rate of penetration, and "Exp. ROP", expected rate of penetration, contain the outputs of the ROP prediction model. While the value in the column "Pre. ROP" represents the ROP predicted based on the current drilling process parameter values, which can be easily compared to the real, recorded ROP, the column "Exp. ROP" contains the ROP values from the ROP optimization framework that are based on fictitious scenarios of parameter settings with altered values for the controllable drilling process parameters. In this case the controllable drilling process parameters comprise the RPM, the WOB, and the mud flow. The values of these parameters that were determined during the optimization are shown in the columns "Rec. RPM", "Rec. WOB", and "Rec. Mud_Flow" and give the set points that were recommended to the drilling operator over the GUI during the field test.

3. Expectations and Results

The results from the field tests are discussed in the respective deliverables 13.1 for the first field test and 13.3 for the second field test. Within this section we are going to compare the expectations, especially for the performance of the AI-based software modules to the actual results acquired during the drilling operations of the field tests. For this purpose, we are going to refer to the list of test objectives given in section 2.2 of this report.

The first three test objectives being the testing of the OptiDrill prototype's overall functionality and reliability in a drilling environment, the testing of the data processing and gathering functionalities, and the practicality and user friendliness of the GUI, could all be successfully tested during the two round of field testing and met the expectations.
The system ran throughout the whole time of the drilling operations for a total of around 58 hours over 12 days of drilling without any errors or software crashes. The system ran very smoothly all components worked as expected from a technical perspective. The outdoor laptop used as the central device for running the software proved to be a reliable and performant piece of equipment very well suited for the purpose of testing the software prototype in a drilling environment.
The data processing and gathering functionalities of the system delivered the required results which could be verified by analysing the created drilling process parameter datasets after the field tests were finished. All filters and additional processing tasks were performed correctly resulting in a processed, clean and ready to use development dataset for further development or the reproduction of the results obtained during the field tests.
The GUI developed by the project partner TVS proved to be a good and reliable solution for the real-time display of the OptiDrill system's outputs. The interface to the python software performing the data processing, model inference, and outputs processing worked without any issues. The design of the GUI is very simple, using the available screen space efficiently and providing all necessary information regarding the process monitoring and model outputs at one glance. This impression was shared by the colleagues at site that were not involved in the OptiDrill project, and no major negative feedback was received. The GUI did completely fulfil its purpose during the field testing campaigns however one major downside of the current implementation is that it does not offer any interactivity or adaptability through the software, e.g. for adjusting the plots. An iOS app run on an iPad providing these features was developed for demonstration purposes and was presented during the project review meeting in Manchester. A video showing the app with all of its functionalities will be published on the project website in the near future.

The remaining test objectives focusing on the performance of the AI-based software modules that were deployed during the field tests will be discussed in the following paragraphs. For the supervised learning-based modules, being the ROP prediction and optimization as well as the drilled lithology prediction, we will use the results that were achieved during the model training and evaluation on the data from the Geostar 2 project from Bochum as a guideline for the quantification of the expectations related to the predictive performances. The models used during the field tests were trained based on this dataset, which was created based on logging data from drilling operations that were conducted were closely to the actual field test site. Regarding the unsupervised learning-based module for the drilling problem/anomaly detection the comparison of the results from the field test will be less straightforward, since no actual drilling problems were encountered however many anomalies were detected. A more detailed analysis on all three modules will be presented in deliverable report 13.4.

Concerning the ROP prediction model that was deployed during the field testing the following results could be achieved during the k-fold-cross-validation on the Geostar 2 dataset from Bochum:

Table 3: Cross-validation results of the ROP prediction model on the Geostar 2 dataset from Bochum

Well	MAE [m/h]	RMSE [m/h]	R2
H01	1.38	2.10	0.93
H02	1.22	1.65	0.92
H03	1.31	1.89	0.91
H04	1.42	1.97	0.96
H05	1.27	1.82	0.96
H06	1.50	2.24	0.95
H07	1.41	2.13	0.93
H08	1.38	2.36	0.92
H09	1.34	1.88	0.94
H10	1.68	2.39	0.91
H11	1.54	2.20	0.94
H12	1.43	1.94	0.90
Avg.	1.406	2.048	0.930

The table above shows the error metric values for the MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and the R^2-Score for the Geostar 2 dataset. All these metrics are commonly used for the performance rating of AI-based regression models. In the course of the cross-validation, during 12 iterations the model was trained on 11 of the 12 available well and then validated on the remaining well. It can be seen that the predictive performance differs from well to well with fluctuations of up to around 30% for the MAE and RMSE values. The average metrics scores that were obtained based on the cross-validation are 1.41 m/h for the MAE, 2.05 m/h for the RMSE and 0.93 for the R^2-Score. Using these very good metrics, achieved on a dataset with well from the very same location, drilled with the exact same equipment as basis for the expectation regarding the performance of the system at a new location with different equipment would not have been fair. However, the results can be used as an estimate for maximum performance that could be achieved under ideal circumstances, while the realistic expected performance can be assumed to be significantly lower, given the fact that the equipment used for the field testing differed from that used for the drilling operation of the Geostar 2 wellbores as well as the statistical properties of the data recorded in both cases. Based on the evaluation of the ROP prediction system's outputs from the two rounds of field testing the following performance metrics values were calculated:

Table 4: ROP prediction model metrics from the two field tests

Wellbore	MAE	RMSE	R^2-Score
P1	2.24	3.11	0.72
P2	1.89	2.47	0.78

The results achieved during the field tests do fulfil the expectations prior to the field test. Although there are significantly lower than the results from the model evaluation, they are still in within an acceptable range and show that the model was able to predict the ROP with a reasonable accuracy despite the significant differences in the input data.

Regarding the drilled lithology prediction model that was used for the field tests a similar approach can be used for the formulation of the expectations of the model performance. The model was trained on the same data from Bochum, however the results that could be achieved during the model evaluation were not as good as the results achieved for the ROP prediction model. The following table shows the results of the cross-validation of the drilled lithology prediction model on the Geostar 2 dataset:

Table 5: Cross-validation results of the drilled lithology prediction model on the Geostar 2 dataset from Bochum (Clst. = Claystone, Slst. = Siltstone, Sast. = Sandstone)

			Individual F1-Scores			
ID	**Acc**	**F1_avg**	**Clst.**	**Slst.**	**Sast.**	**Clst./Slst.**
H01	0.67	0.60	0.72	0.40	0.49	0.78
H02	0.49	0.36	0.30	0.01	0.62	0.53
H03	0.53	0.41	0.26	0.20	0.49	0.70
H04	0.62	0.51	0.58	0.14	0.63	0.69
H05	0.64	0.59	0.70	0.38	0.59	0.67
H06	0.67	0.62	0.82	0.45	0.62	0.59
H07	0.58	0.40	0.52	0.13	0.20	0.74
H08	0.58	0.52	0.43	0.45	0.52	0.69
H09	0.55	0.51	0.55	0.40	0.66	0.44
H10	0.69	0.64	0.83	0.40	0.69	0.63
H11	0.61	0.47	0.30	0.54	0.74	0.30
H12	0.72	0.49	0.68	0.09	0.81	0.37
Avg.	0.61	0.51	0.56	0.3	0.59	0.59

The table shows the accuracy and F1-Score calculated on all classes and the individual F1-Scores for each class. It can be seen that again the results differ quite significantly from one well to another with the accuracy ranging from 49 to 72% and the F1-Score ranging from 0.36 to 0.64. The same fluctuations can be observed in the individual F1-Scores. All in all, calculated for the whole dataset an average accuracy of 61% and an average F1-Score of 0.51 could be achieved. Concerning the results that could be expected the same argumentation as for the expectations regarding the ROP prediction model is used here. The performance achieved during the cross-validation should be seen as the maximum reachable performance under ideal conditions. However, due to the beforementioned differences in equipment and data also the geological conditions at the drill site were different compared to those of the Geostar 2 site. One special feature of the geology at the field test site was that the encountered layers of lithologies were not oriented horizontally but rather vertically or at a steep angle.

During the first field test the drilled lithology prediction model almost exclusively predicted only one class, namely *Claystone. Claystone* was the predominant lithological class encountered with almost 50% of the data points of the dataset belonging to this class. However, due to the significantly different ranges of the process parameters encountered during the field test this behaviour was caused by incident and does in anyway show a malfunction or limitation of the model in this case. Using a more suitable scaler as described in section 3 of this report could fix this issue to some extent. Furthermore, there was a difference between the classes the model was trained on and the classes that were actually encountered at the drill site. Since these two groups of lithology classes did not match, the model was not able to predict instances of the unknown class and there predicted instances of a class that was not present in the actual dataset, leading to many errors in the predictions and at the same time showing an important limitation of the system. The model might not have fulfilled the expectations and was not able to reach a satisfactory performance, on the basis of the results achieved during the cross validation, but the model inference and evaluation helped to identify some important challenges and limitations of the approach followed.

For the second field test the drilled lithology prediction model was retrained on the data from the first field test, as reported on in deliverable 13.3. The very limited amount of data available for the retraining, consisting only of the data gathered from the first field test, caused some major limitations for the model training and validation. It was not possible to train the model in the same manner as it was originally, keeping wells separated for training and validation purposes. Instead, the model was trained using a stratified train and validation split. From experience this leads to good performance on both train and validation data, but worse performance on new, unseen data. Since this was the only way of training and validating the model with the data at hand to enable a prediction of all classes present at the validation site, it was decided to do so.
The results that were achieved during the model training could as expected not be reproduced on new unseen data from the second wellbore. The renewed change of equipment and drilling strategy, due to the strong deviation of the drilling path during the first borehole, led to additional difficulties for the model. At the end an improved overall predictive performance on the majority classes could be observed, while the minority class *Sandstone* was not predicted by the model during the second field test.

The anomaly detection module developed by BGS was mostly validated on UK onshore data, but since it is based on an unsupervised learning approach this does not matter much for the model inference. The main reasons for choosing this approach were data scarcity, for drilling problems data, and the high adaptability to new data without the need of retraining. For both field test the results of the anomaly detection inference were very similar. No actual drilling problems were encountered, while many anomalies were detected within the recorded data. On the UK dataset the model was validated on, around 9% of the anomalies detected could be matched to actual drilling problems. More insights into the anomaly detection frameworks outputs are given within deliverable report 13.4.

4. MACHINE LEARNING MODEL OPTIMIZATION

Based on the insights from section 3 of deliverable report 13.1 dealing with the results from the first field test of the OptiDrill software system, within this chapter we will reconstruct the outputs of the ROP prediction and the lithology prediction models using a different scaler than during the field test that is more suitable. The scaler used during this investigation is fit to the actual drilling data that was recorded during the drilling operation instead of being fit to the development data that was used for model training and validation. The purpose of this section is to give an insight into how the systems performance, more exact that of the ROP and drilled lithology prediction would have changed using a different scaler for the preprocessing of the input data.

4.1 ROP Prediction Model

In this section we will take a look at the reconstructed outputs of the ROP prediction model with a min-max-scaler fitted to the actual drilling process parameter ranges that were determined during the field test. The scaler is fitted to the dataset that was created by the OptiDrill system during the field test. Afterwards the same model that has been used during the field test is applied to the data recorded by the OptiDrill system during the field test scaled with the new scaler. The following tables shows a statistical overview of the ROP predictions made by the model using the new scaler.

Table 6: Statistical overview over predicted and actual ROP values including the absolute error of the model

	Act. ROP [m/h]	Pred. ROP [m/h]	Abs. Error ROP [m/h]
mean	14.90	14.82	1.44
std	6.34	5.75	1.58
min	1.00	3.97	0.00
50%	14.00	13.73	1.02
75%	18.00	17.29	1.88
90%	24.50	23.95	3.03
95%	27	26.57	4.05
99%	32	30.63	7.52
max	40	39.31	29.92

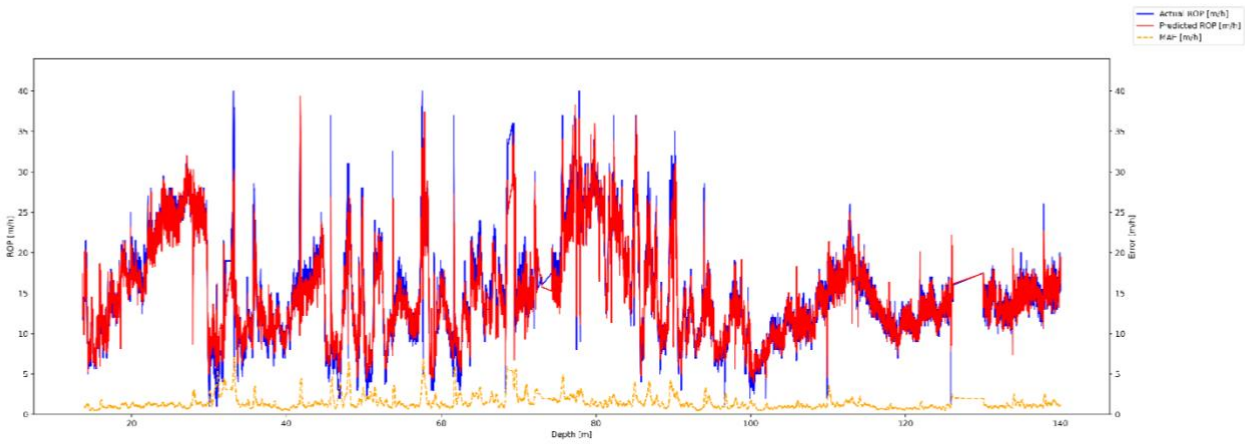

Figure 3: Actual ROP in blue and predicted ROP in red plotted over the measured depth for the P1 dataset processed with the new scaler. The MAE for a window of 100 data points is plotted in orange.

Looking at the table it can be seen that the predictions of the model are significantly better after applying the new scaler to the data preprocessing. The mean and the standard deviation, as well as the percentiles values of the predicted ROP are way closer to the values of the actual ROP. The error metrics values could also be improved significantly. The MAE could be improved by 36% from 2.24 m/h to 1.44 m/h, the RMSE could be improved by 31% from 3.11 m/h to 2.14 m/h, and the R2-score could be improved by 24% from 0.72 to 0.89. The following figure shows the plot of the new ROP prediction model outputs.

4.2 Drilled Lithology Prediction Model

Just as for the ROP prediction model, the outputs of the drilled lithology prediction model were reconstructed based on the processed drilling parameter dataset from the P1 wellbore and a new scaler. The following confusion matrix and the table with the performance metrics show the results of this evaluation run.

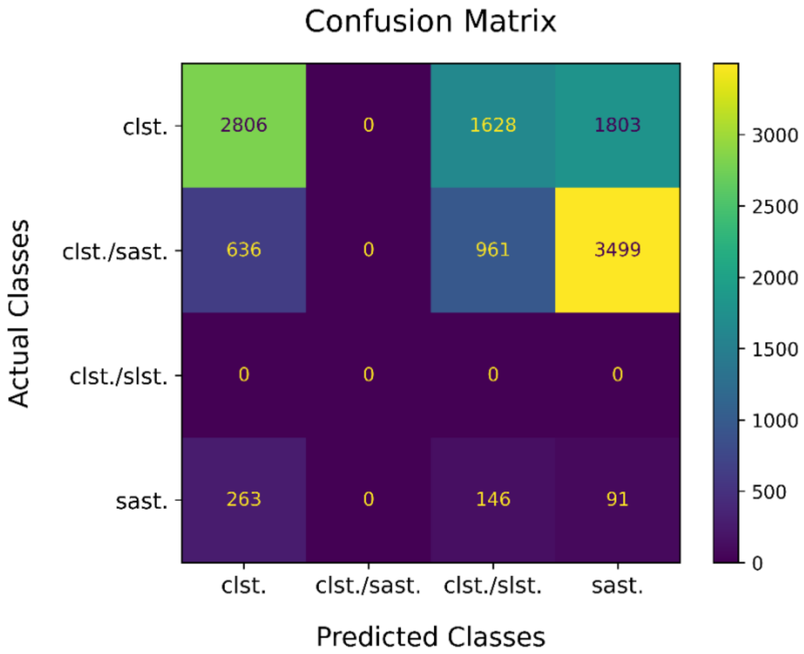

Figure 4: Confusion matrix of the model predictions for the Push-It P1 wellbore with a new scaler

Table 7: Performance metrics for the reconstructed drilled lithology prediction model outputs

	Claystone	Claystone/Sandstone	Claystone/Siltstone	Sandstone
Precision	0.76	0	0	0.02
Recall	0.45	0	0	0.18
F1-Score	0.56	0	0	0.03
Support	6237	5096	0	500

Looking at the confusion matrix above we can see that there is one row and one column with no actual predictions. The row corresponds to the class *Claystone/Siltstone* which is not represented in the actual classes from the P1 dataset, therefore this row has to remain empty. The column corresponds to the class *Claystone/Sandstone*, which is not represented in the Geostar 2 dataset on which the drilled lithology prediction model was trained. Since the model was not trained on this class, no prediction can be made leading to this column remaining empty. If we look at the individual classes row by row that were present in both the development and the inference dataset, we can see that around 45% of the instances of the class *Claystone* are predicted correctly, while in the remaining instances it is equally often confused with the classes *Claystone/Siltstone* and *Sandstone*. 69% of the instances from the class Claystone/Sandstone, on which the model was not trained on, were predicted as *Sandstone*, which does make sense. The remaining instance are predicted as *Claystone* in around 12% of the cases and as *Claystone/Siltstone* in around 19% of the cases. The class *Sandstone* is only predicted correctly in around 18% of the instances from this class. The remaining instances are predicted as Claystone in 53% and as Claystone/Siltstone in 29% of the cases. The table with the individual performance metrics of the classes shows a similar picture. All in all, it can be said that the predictions with the new scaler are more diverse and make more sense than those with the scaler used during the field test. One shortcoming here is that with the approach followed in the OptiDrill project, the classes that the model is able to predict are predefined by the classes present in the development dataset. If there are differences between the actual lithology classes in a new wellbore from a new location and the historical training dataset, these cannot be handled directly. However, if the classes have similar geological properties or share the same components, e.g. in classes that are a mix of two lithologies, the results are of course still useful and can be interpreted accordingly.

5. CONCLUSION

The evaluation of the field tests of the OptiDrill system in Bochum has provided valuable insights into its functionality, reliability, and predictive capabilities in a real drilling environment. Over the course of the testing, the system demonstrated robust performance, successfully running for a total of approximately 58 hours without any technical issues. The data processing and gathering functionalities were effective, resulting in a clean and usable dataset for further analysis.

The performance of the AI-based software modules, particularly the ROP prediction model, met expectations, achieving a Mean Absolute Error (MAE) of 2.24 m/h in the first test and improving to 1.89 m/h in subsequent evaluations. This indicates the system's ability to adapt and optimize drilling parameters effectively, even in varying conditions. However, the drilled lithology prediction model faced challenges, primarily due to overfitting and discrepancies between training and field data, highlighting the need for further refinement and retraining to enhance its accuracy. In both rounds of field testing the drilled lithology prediction module could not meet the expectation in terms of predictive capabilities. The anomaly detection module proved to be sensitive, identifying numerous anomalies during the tests, although no significant drilling problems were encountered. This sensitivity underscores the module's potential as a tool for monitoring drilling dynamics and identifying operational risks.
The other test objectives focusing on the functionality, reliability, and practicability of the drilling advisory systems were met without any mayor shortcomings.

Overall, the field tests have validated the OptiDrill system's capabilities and provided a foundation for future improvements. The insights gained will inform the next phases of development, including the planned UK field test, where direct downhole validation will further refine the system's predictive accuracy and enhance its applicability across diverse drilling environments. The findings from this testing phase represent a critical step toward realizing the full potential of the OptiDrill system in optimizing drilling performance and improving energy efficiency in the industry.