Process Mining of Mining Processes: Analyzing Longwall Coal Excavation Using Event Data

Edyta Brzychczy, Agnieszka Żuber, Wil van der Aalst, *Fellow, IEEE*

***Abstract*—The mining industry faces many challenges, prompting the adoption of new technologies and continuous improvement of processes to improve operational efficiency and personnel safety. Using data from information systems combined with novel process-mining techniques creates new possibilities for improving industrial processes. The paper presents a comprehensive method of modeling and analyzing the longwall process in underground mining based on event data using process mining (PM4LMP). The method comprises four basic steps: data gathering, data preprocessing, creation of event logs, and process mining tasks. In our method, we proposed, among all, case ID identification based on heuristics using context data and activity identification with supervised and unsupervised approaches, which provide complementary information about process execution. The method assumes an in-depth analysis of processes based on sensor data and knowledge gathered in IT systems, which can significantly improve the quality of information at managers' disposal when making decisions regarding the mining process.**

***Index Terms*—process mining, longwall mining, coal mining, sensor data, industrial processes**

I. Introduction

In dynamic and competitive market conditions, underground mining companies need to adapt and enhance their processes and activities. Achieving success in this area requires modern management aimed at creating value for these enterprises in the long run [1]. One of the fundamental concepts enabling the modern management of an organization is *Business Process Management* (BPM), which includes the identification of processes and the creation of process architecture as well as their modeling and analysis. Process management is aimed at the continuous improvement of processes whereby data available from various IT systems are used as a source of information and expertise.

Intensive development of virtualization and computerization in the mining industry enables monitoring of practically all events and activities involved in mining processes. Advances in sensor technology (*Industrial Internet of Things* - IoT) and enhanced capabilities of computer hardware and software are increasingly used in underground coal mines [2]; however, along with the increased amount of collected data, there is a growing need to develop tools for its efficient processing and analysis.

Edyta Brzychczy is with the Faculty of Mechanical Engineering and Robotics, AGH University of Krakow, Poland (e-mail: brzych3@agh.edu.pl); Agnieszka Żuber is with the Faculty of Civil Engineering and Resource Management, AGH University of Krakow, Poland (e-mail: toga@agh.edu.pl); Wil van der Aalst is with Process and Data Science Group, RWTH Aachen University, Germany (e-mail: wvdaalst@pads.rwth-aachen.de)

This is authors version - the final version is available at IEEE: 10.1109/TSMC.2023.3348496

The primary data analytic approaches used in underground mining are data-oriented and *not* process-centric; they focus on knowledge discovery from data through the use of traditional and advanced analytic techniques such as Machine Learning (ML), data mining, and statistics. However, process improvement activities require a process-oriented analysis of the available data. *Process Mining* (PM) techniques provide new opportunities for knowledge discovery from data about processes. However, process discovery from industrial raw sensor data has to meet several requirements imposed by traditional PM techniques, which may not be an easy task in such cases. Applying PM to raw mining sensor data requires customization and development of solutions tackling the following challenges [3]:

- There is no single, comprehensive dataset (event log) that could be directly used for the purpose of process mining analysis. Raw sensor data are stored in different systems, databases, or platforms;
- The event log generated from the sensor data can be affected by noise, inaccurate measurements, and ambiguous information, which may result in an unreliable event log;
- It is necessary to identify the case ID - a single process execution;
- The translation of low-level data into high-level data, known as event abstraction, requires appropriate data aggregation techniques.

Hence, the main factors limiting the applicability of PM in underground mining include the lack of suitable event logs for process modeling and analysis. It results mainly from the structure of IT systems monitoring underground processes and the granularity of collected data. Most of the data presents low-level readings coming directly from sensors. There are many different sensors, and their data form is not suitable for direct usage by PM techniques. This requires special preprocessing in order to create suitable event logs to model and analyze the underlying process.

With our work, we aim to address the gap related to creating higher-level event logs from sensor data collected in the monitoring system for the underground mining process and present an example of the usage of PM to model and analyze the longwall coal excavation process. We identify two important contributions to the fields:

1) *Mining industry* - we propose a sophisticated method named *Process Mining for Longwall Mining Process* (PM4LMP) to create an event log based on raw low-level data from IT monitoring system enabling in-depth process analysis of selected underground process and we

present an example of how PM can support the process improvement in the mining industry.
2) *Process mining* - we present challenges and solutions related to PM based on real-life industrial sensor data, including complementary usage of supervised and unsupervised techniques to increase knowledge about process execution, taking into consideration the mixed character of sensor data, still not being recognized fully by the PM community due to a lack of real-life examples and experiments with mixed types of variables (also of continuous type).

The paper is structured as follows. In Section 2, we introduce the preliminaries related to process mining artifacts. Section 3 summarises the related work. Section 4 presents the longwall mining process and data characteristics. A general description of our method considered in this paper is given in Section 5. Section 6 presents the results of the application of our method to model and analyze the longwall mining process. The final section summarizes our work and presents our future research plans.

II. Preliminaries

PM is a relatively new field of research bringing together capabilities of several known disciplines, e.g., data mining, machine learning, modeling, and analysis of processes [4]. The most important function of PM is to extract knowledge from recorded event log data for discovering, monitoring, and improving real processes.

The two main artifacts used in PM are the *event log* and *process model*. An event log is a collection of events. An event e can have any number of attributes, and most of the process mining techniques require the following three attributes to be present: case $\#_{case}(e)$, activity $\#_{act}(e)$, and timestamp $\#_{time}(e)$. For process mining techniques focusing on control flow, it often suffices to focus only on the activity attribute and the ordering within a case. This leads to a much simpler event log notion.

Preparation of a proper event log is a complex task that can be accomplished in many different ways depending on anticipated results. Techniques dedicated to event log preparation can be organized in three main types [5]:

1) Event data extraction – identifying data elements that characterize events coming from diversified data sources.
2) Event correlation - grouping the data elements related to a single process instance.
3) Event abstraction - mapping data elements to events that correspond to activity executions in a business process.

Each of the mentioned tasks, especially in terms of industrial processes, is challenging to perform. First of all, in industrial databases, most information is stored as raw sensor data. It could be treated as event data but on a very low level of abstraction. Actually, it cannot be used in this form in further process mining analysis. Transforming input raw event data from industrial databases into the event log represents a challenge related to data quality. Currently, no fully automated approach is available for extracting event data or event logs from databases. Although the databases are loaded with data, there is no direct reference to events, cases, and activities. Thus, it brings additional challenges to event correlation and event abstraction.

Event abstraction approaches are a bridge between raw data gathered in information systems and a format of events that represent the execution of an activity and are correlated to traces and, as such, could be analyzed in process mining.

In the literature, various techniques for event abstraction are categorized into three major groups: (i) supervised, (ii) semi-supervised, and (iii) unsupervised learning strategies. Their detailed taxonomy is presented in [6]. It is worth emphasizing that event abstraction based on sensor data is recognized as one of the main challenges in BPM and PM domains [7].

As mentioned before, the second important artifact used by PM is the process model. The main objective of the process model is to reflect the execution of activities in the proper order. For process model creation, different process modeling languages (of imperative or declarative type) could be applied. Mathematical concepts and notations reported in the literature on the subject differ in their expressive power and formal semantics for process modeling, e.g., Transition Systems (TS), Petri Nets (PN), Workflow Nets (WF-nets), Business Process Modeling and Notation (BPMN), Process Trees (PT) or Directly Follows Graphs (DFG), Declare (DeC), Dynamic Condition Response Graphs (DCR) [8], [9].

The most popular process modeling languages used in the context of process mining are still PN and DFG models [8]. Hence, we briefly introduce both [4].

Definition 1 (Directly-Follows Graph): A Directly-Follows Graph (DFG) is a pair $G = (A, F)$ where $A \subseteq \mathcal{U}_{act}$ is a set of activities and $F \in \mathcal{B}((A \times A) \cup (\{\blacktriangleright\} \times A) \cup (A \times \{\blacksquare\}) \cup (\{\blacktriangleright\} \times \{\blacksquare\}))$ is a multiset of arcs. $\blacktriangleright$ is the start node and $\blacksquare$ is the end node ($\{\blacktriangleright, \blacksquare\} \cap \mathcal{U}_{act} = \emptyset$), where $\mathcal{U}_{act}$ is the universe of activities.

$\blacktriangleright$ and $\blacksquare$ can be viewed as artificially added activities to clearly indicate the start and end of the process. The other nodes in the graph denote the activities. In Definition 1, F is a multiset of arcs to be able to capture frequencies. A DFG $G = (A, F)$ represents all activity sequences corresponding to paths starting in $\blacktriangleright$ and ending in $\blacksquare$.

The main drawback of using DFGs is that they cannot express concurrency. If activities do not happen in a fixed sequence, immediately loops are created. A DFG is like a Markov chain, i.e., the state is determined by the last activity, and there is no memory. Therefore, we need notations like Petri nets.

Definition 2 (Labeled Accepting Petri Net): A labeled accepting Petri net is a tuple $PN = (P, T, F, l, M_{init}, M_{final})$ with P the set of places, T the set of transitions, $P \cap T = \emptyset$, $F \subseteq (P \times T) \cup (T \times P)$ the flow relation, $l \in T \not\rightarrow \mathcal{U}_{act}$ a labeling function, $M_{init} \in \mathcal{B}(P)$ is the initial marking, and $M_{final} \in \mathcal{B}(P)$ is the final marking.

A complete explanation of Petri nets is out of the scope of the paper, and we assume that the reader is familiar with the basics [8], [10], [11].

Workflow nets (WF-nets) form a subclass of Petri nets having precisely one source place p_{start} and one sink place p_{end}, and all other nodes on a path from source to sink [12].

The source place defines the $M_{init} = [p_{start}]$ is the initial marking and $M_{final} = [p_{end}]$ is the final marking. Hence, a WF-net defines an accepting Petri net modeling cases that can move from source place p_{start} to sink place p_{end}. A WF-net is *sound* if and only if the following requirements are satisfied: (1) option to complete: for each case, it is always still possible to reach the state which just marks p_{end}, (2) proper completion: if place p_{end} is marked all other places are empty for a given case, and (3) no dead transitions: it should be possible to execute an arbitrary activity by following the appropriate route through the WF-net, and (4) safeness: it is not possible to put two tokens in place at any point in time [12]. Techniques like the inductive mining approach guarantee to produce sound WF-nets [13].

Possessing the suitable event log (and process model if required), one can perform one or several PM tasks. Major process mining tasks include [4]:

- *Process model discovery* - the discovery of real process models. Transformation of input data from event logs into a process model without a *priori* information. The process model can be expressed in different notations (Petri nets, BPMN, process tree).
- *Conformance checking* - comparing an existing, formal model with an event log to check whether the process recorded in the event log is consistent with the formal models (procedural, organizational, declarative, business rules).
- *Enhancement*, also called performance analysis, includes in-depth analysis of the process through the use of contextual information stored in the event log, which is used to extend and refine the existing process model (e.g., by indicating process bottlenecks, bandwidths, and frequencies).
- *Comparative process mining* - uses as input multiple event logs from different periods, locations, or case categories to compare process executions and to search differences.
- *Predictive process mining* - introduces various Machine Learning techniques to better diagnostics and explanation of process behavior as well as process outcome prediction.
- *Action-oriented process mining* - aims to turn diagnostics into actions using low-code automation platforms for triggering the workflows.

Our work focuses on process discovery, conformance checking, and performance analysis tasks to model and analyze the longwall mining process described in the following sections.

III. Related work

The usage of PM techniques in the mining industry is still limited. Only a few papers present its utilization in underground process modeling and analysis. The first scope of applications is related to the analysis of working machinery and installations in the primary process; the other area encompasses non-industrial processes, i.e., emergency rescue.

In terms of analysis of working machinery, PM applications were related to mechanized roof support operations [14], roof bolter operation [15] as well as LHD (Load Haul Dump) machine [16].

The application of PM in underground mining was also related to higher-level processes, such as emergency rescue processes after fatal accidents due to gas explosions in China [17].

PM on sensor data is currently one of the research directions aiming to introduce IoT data into the BPM domain more widely for process analysis and improvement. The event log creation from low-level events has been undertaken by many researchers, e.g. [6], [18].

Only a few examples of industrial use cases using sensor data for PM tasks can be found in the literature.

A general proposal to deal with sensor data is presented in [19]. The authors provide an interactive method for the process analyst to conduct an initial analysis of data sets from IoT for activity executions. The method comprises visualizations and filtering features to find patterns in data, enabling activity signatures identification and labeling of similar activities; however, most of the sensor readings used in the smart factory example are discrete and binary.

The approach using the continuous sensor data readings for PM was presented in [20]. In the paper, authors use window-based segmentation of the sensor measurements for the creation of event logs and discover processes of using a smart baby bottle. After data segmentation, they applied cluster analysis, in which results were labeled by a domain expert. The following steps comprised the grouping of segments into activities and process instance identification. Finally, they used created event logs to process discovery.

In [21] authors propose a similar window-based segmentation approach, including time-series-based process and product state data as additional attributes in the event log in the context of Digital control-flow Twin. Authors applied unsupervised techniques (cluster analysis) based on statistics in sized time-series segments to event abstraction. Obtained clusters were mapped to process states using domain knowledge and used for event log creation and process discovery.

A different approach is proposed in [22]. The authors present a supervised approach for analyzing the dependency relationships between events to generate multi-level models. In the example from a pulp mill, the authors defined the so-called operating regions of the process based on KPIs, aiming to transform data from continuous values to discrete ones. It was done based on domain expert knowledge. In the next step, a classifier (Decision Tree) was applied, with input variables coming from process monitoring and the operating region as a target variable, to extract meaningful patterns explaining the relation between observations and their assigned label. Discovered patterns are further used as events, and Case ID is identified as one day of operation. Finally, the created event log is used to process discovery.

In the mentioned works, various supervised and unsupervised techniques are presented. Our method assumes parallel usage of supervised and unsupervised approaches to activity identification, which reveals the potential of such complementary usage for in-depth process understanding and analysis.

The content of the presented references shows that PM on the industrial sensors is a promising direction of research in terms of event abstraction, case identification, and event log requirements, especially when working with data of a continuous type.

IV. Longwall mining process

Longwall mining is a technique whereby coal is mined by sliced blocks (from 0.6 to 1.2 meters in thickness), usually 100 to 300 meters wide and 1,000 to 3,000 meters long.

The process involves coal cutting along the width of the panel (longwall face length) to the depth of the intended slice (having the width of the longwall panel) by a longwall shearer - Fig. 1. Next, crushed coal is loaded onto a conveyor (Armored Face Conveyor). The sequence of operations continues after the roof support (Powered Roof Supports) is propped at the front, and the conveyor advances forward. The crushed coal is transported to the breaker feeder (Crusher) via a chain conveyor (Beam Stage Loader).

Most often, the shearer's front organ cuts the coal body's upper layer under the slant (clockwise rotation). The rear organ works in overflow, cutting the lower part of the coal body, and loads the hitherto unloaded material onto the scraper conveyor through the advanced organ.

The slotting phase (changing the organ position) occurs at the end and the beginning of the longwall face. The working shearer moves in two directions: back (Along) and forth (Return) alongside the length of the longwall.

The steps described above define a common cutting method involving operations repeated in a specified sequence and cyclical implementation of operations (activities) in a longwall face. The actual selection of operations depends on deployed technology, equipment, and work organization [23].

The cutting cycle could be divided into three main parts: main cut, shuffle, and turnaround [24]. Fig. 2 shows the model of the bi-directional cutting cycle of the shearer.

The shearer cycle consists of the following stages (marked in Fig. 2):

Direction Along:

1) Cutting at the beginning of the longwall face,
2) Stoppage in ON mode at the beginning of the longwall (30-40m from the minimum value),
3) Return to the drive I,
4) Stoppage in ON mode at the beginning of the longwall face,
5) Cutting at the middle section of the longwall face,
6) Cutting at the end of the longwall face, and
7) Stoppage in ON mode at the end of the longwall,

Direction Return:

8) Cutting at the end of the longwall face,
9) Stoppage in ON mode at the end of the longwall face,
10) Return to the drive II,
11) Stoppage in ON mode at the end of the longwall face,
12) Cutting at the middle section of the longwall face,
13) Cutting at the beginning of the longwall face, and
14) Stoppage in ON mode at the beginning of the longwall face.

The described process is remotely monitored from the surface by the mine's dispatcher, who possesses limited detailed information about the ongoing processes. The dispatcher's observations are primarily based on the real-time tracking of the shearer's position and sensor data collected from the machine. Consequently, our research focuses on the application of process mining techniques to enhance the monitoring process, facilitating in-depth process-oriented analysis and, subsequently, opportunities for process improvement [3].

Although the theoretical model is relatively simple, certain complications might arise in real-life conditions. In real processes, unexpected and unpredictable events occur quite frequently due to different geological and mining conditions or organizational settings [25], [26]. These issues related to the specificity of the mining process will be addressed in more detail in the following sections.

V. Our method

In this paper, we propose the Process Mining for Longwall Mining Process (PM4LMP) method. The PM4LMP method is employed to support the analysis and improvement of underground mining processes in the longwall face with the use of PM techniques to answer the following research questions:

RQ1: What is the empirical representation of the mining process carried out in the longwall face, considering the integration of sensor data from the longwall shearer?
RQ2: What specific operational states can be discerned during the execution of the mining process?
RQ3: To what extent do deviations or discrepancies manifest in the actual execution of the mining process when compared to the underlying theoretical process model?
RQ4: What are the primary bottlenecks within the longwall mining process?
RQ5: How effectively does the process address the specific requirements of business users?

The method involves four main steps:

1) Gathering of input data, including raw sensor data, a theoretical model of the process, and business questions formulated by users;
2) Data preprocessing;
3) Event log creation;
4) PM tasks, namely process discovery, conformance checking, and process enhancement.

The raw data from the underground mine that we used in our work was prepared for monitoring, visualization, and simple statistics but not for process mining. Data in the original format and structure cannot be used directly to analyze or improve process effectiveness due to the specific nature of the data but also due to the complexity of the analyzed process. Another aspect that obstructs the analysis of mining data is that real industrial data may be incomplete and involve noise. Thus, the preprocessing step is required before the final data set can be prepared in the format adequate for process mining. An important feature of sensor data in the mining industry is their mixed character, including categorical data (nominal, binary) and numerical data (discrete and continuous).

Fig. 1: Scheme of longwall face.

Fig. 2: Model of the bi-directional cutting cycle of the shearer.

The main aim of the data preprocessing step is data cleaning, dimensionality reduction, and selection of the most accurate variables describing the work of the shearer in a pertinent way.

Dimensionality reduction starts from the data cleaning with the imputation of missing values and outlier identification. This step also contains correlation analysis for numerical data and crosstables for categorical data to exclude dependent variables, as well as Principal Component Analysis (PCA). An additional step in data preprocessing is feature engineering, including the discretization of continuous variables (if needed).

There are two main challenges related to the usage of sensor data from longwall monitoring systems in process mining important for the creation of event logs: (i) they usually do not contain case IDs, and (ii) activity names are not given due to the low-level character of data.

While analyzing and monitoring the industrial machinery and equipment, a single, full-duty cycle should be identified. Due to the specificity of sensor data from a longwall shearer, such identification does not exist and requires manual preprocessing and case identification. Therefore, we proposed the heuristic procedure [27] enabling the cycle identification in data based on the shearer location variable.

Since the shearer activities' names are not available in the data, two approaches for activity identification are proposed:

1) Supervised - based on a theoretical model and expert rules,
2) Unsupervised - based on clustering techniques and incorporating domain knowledge,

In the first approach, based on the domain expert's knowledge and a longwall working technology manual, the longwall shearer operation process's theoretical stages were defined as a hierarchical state model[1].

The hierarchical model covers the main prescribed states in the longwall shearer operation process mentioned in Section IV. In line with the theoretical process model, 14 main stages were thus defined; however, some undesirable states can also occur during the process, namely: Moving and Reversion. All identified activities are general and do not cover the real diversity of working conditions.

When the supervised approach is employed, some actual non-typical behavior of the shearer can be omitted; thus, unsupervised activity identification with clustering methods is proposed as a complementary approach.

Clustering is an unsupervised data mining technique typically requiring a predetermined number of clusters. According to a specified number of clusters and the final statistical description, the optimal number of clusters can be determined based, e.g., on Silhouette information.

The last essential step of the unsupervised approach is assigning unique, understandable labels to discovered clusters based on cluster summary statistics and visualizations, consulting with domain experts, and verifying using the technical manuals.

Based on case ID identification and activity labeling, we create raw event logs with columns: timestamp ("TIMESTAMP"), case ID ("CASE_ID"), and activity ("ACTIVITY").

Created event logs are used in the selected PM tasks. Our goal in the process discovery is to analyze in general process execution based on specific activities and behavior recorded in data. For this purpose, we use DFG models and event logs created with an unsupervised approach.

[1]Due to the NDA of the company providing data for analysis, we cannot show the details of the used expert rules.

In the conformance checking procedure, we apply an event log labeled with expert rules and a theoretical process model. The main objective of this task is to reveal the discrepancies between the actual process execution and the prescribed process model so as to identify deviations and failures.

Finally, in the process enhancement, the time perspective of the shearer operation process is analyzed on the basis of created process models with Inductive Miner. We also analyze process execution in relation to the following business questions raised by practitioners interested in the occurrence and frequencies of the following events:

1) Reverse movements in specific parts of the longwall face (beginning and end),
2) More than two stoppages in operation at the beginning and the end of the longwall,
3) Moving in the middle section of the longwall face.

The implementation of our method on real-life examples is presented in the next section. In analyses, we used ProM [28] and Disco software [29].

VI. Implementation

The presented approach was evaluated on a data set containing raw sensor data from the monitoring system of underground longwall machinery. An analytical sample covering a period of one month. The data contains 460,000 records and 147 variables (continuous and binary type) related to the shearer operation process.

A. Data preprocessing

Real-world industrial datasets frequently exhibit incompleteness and noise. In the dataset under investigation, approximately 30% of columns were entirely devoid of data, and nearly 50% of variables exhibited more than a 50% incidence of missing values. Consequently, only 57% of all variables proved suitable for subsequent analysis. Furthermore, it was noted that among the 44 logical variables associated with safety sensors, a single logical value was present. This observation implies that these safety sensors had not been activated at any point, thus warranting their exclusion from further analytical considerations.

Missing numerical values were imputed with Multivariate Imputation by Chained Equations (MICE) with the unconditional mean method, as well as interpolation and categorical features with mode value. The detection of outliers was carried out utilizing the Interquartile Range (IQR) method, and subsequently, any identified outliers were substituted with null values. In cases where the actual position of the shearer was unrecorded, a variable extrapolation technique was employed by estimating values between the two nearest data points. The theoretical cycles are thoroughly documented, allowing for the validation of extrapolation results based on other variables.

The subsequent phase involved correlation analysis on numerical data and generating cross-tabulations for logical data to identify and eliminate interdependent variables. Numerical attributes exhibiting a correlation coefficient exceeding 0.6 were excluded from further analysis.

TABLE I: Summary of selected variables

Variable	Description	Type	Range
SM_Daily RouteOfTheShearer	Daily route of the shearer [m]	Numerical	0-29904
SM_ShearerLocation	Shearer location [m]	Numerical	0-152
SM_ShearerSpeed	Shearer speed [m/min]	Numerical	0-25
SM_TotalRoute	Total route [km]	Numerical	0-100
SM_ShearerMoveInLeft	Shearer move to the left	Binary	0/1
SM_ShearerMoveInRight	Shearer move to the right	Binary	0/1
LCD_Average Three-phaseCurrent	Average three-phase current [A] left cutter drum (organ)	Numerical	0-769
RCD_Average Three-phaseCurrent	Average three-phase current [A] right cutter drum (implement)	Numerical	0-775
LHD_MotorCurrent	Motor current [A] left haulage drive (tractor)	Numerical	0
RHD_MotorCurrent	Motor current [A] right haulage drive (tractor)	Numerical	0-191
LP_Average Three-phaseCurrent	Average three-phase current [A] left pump	Numerical	0-171
RP_Average Three-phaseCurrent	Average three-phase current [A] right pump	Numerical	0-164

Fig. 3: Example of identified full cycles.

The next stage of dimensionality reduction was PCA. Initially, we selected the primary components, ranging from PC1 to PC9, guided by an assessment of cumulative variance (91%) as well as adherence to the Kaiser Criterion, which entailed the exclusion of components with eigenvalues exceeding one. To facilitate a more precise interpretation of these primary components, we scrutinized the relationship between the absolute values of their coefficients and the component calculations, opting for a threshold of 0.5 for the absolute coefficient value. Based on data quality evaluation and PCA results, the final list of variables related to the shearer operation, including the basic operating parameters (e.g., speed, currents in drums and tractors) and parameters related to its position with respect to the longwall face (e.g., location, the direction of movement) was created and confirmed by domain experts (Table. I).

The preprocessing step yields the final data set containing 12 original variables (10 numerical and two categorical) and 460,000 records.

B. Case ID Identification

The shearer duty cycles were identified based on the analysis of variables indicating the location and ride direction, namely SM_ShearerLocation, SM_ShearerMoveInLeft and SM_ShearerMoveInRight.

The result of the heuristic procedure implementation is presented in Fig. 3.

In most cases, the proposed heuristic approach enabled the correct identification of the cycle start and end. The errors in identification were caused mainly by data quality issues and

Fig. 4: Silhouette scoring for various numbers of clusters for each clustering task.

unpredictable situations responsible for abnormal operating conditions. Further analysis was restricted to cases with the correct ID.

C. Activity Identification

As it was mentioned in earlier sections, two approaches were adopted in activity identification. The first approach involves labeling stages with expert rules, and the second one uses clustering techniques.

In the practical implementation of activity identification based on domain expert rules the following variables were used: RCD_AverageThree-phaseCurrent, LCD_AverageThree-phaseCurrent, RHD_MotorCurrent, SM_ShearerMoveInLeft, SM_ShearerMoveInRight, SM_ShearerLocation.

In the unsupervised approach, we use hierarchical and non-hierarchical clustering techniques [30]. We divided clustering into the following tasks: clustering only numerical variables (RNV) with k-medoids algorithm, clustering discretized numerical and categorical variables (CV) with hierarchical clustering based on Gower's distance, clustering mixed type variables (MV) and clustering artificial numerical variables (based on mean values (ANV-AG) and standard deviations (ANV-SD) in time windows).

Results obtained with the use of respective methods were generated with a number of clusters in the range of 5 to 20. The Silhouette scores for different numbers of clusters obtained by various clustering methods are given in Fig. 4.

In further analysis, we used the results of numerical variables clustering (RNV) for which the Silhouette score obtained the highest value for 19 clusters.

In the final step, based on statistical and visualization methods, unique, understandable labels were assigned to discovered clusters, followed by the verification procedure with the domain expert.

The verification with the domain expert was carried out in the form of a direct interview, including the presentation of cluster statistics, cluster visualization (in the form of the cluster mapping into shearer cycle (e.g., Fig. 5), and initial defined cluster labels. After the interview, the labels of the obtained clusters were confirmed or revised by the domain expert.

Fig. 5: Defined states of the shearer operation obtained by each approach.

TABLE II: Excerpt from EVENT_LOG_EXPKNW

Timestamp	Case ID	Activity
05.06.2019 19:46:22	14	Stoppage_in_O_mode_beginning_along
05.06.2019 19:46:54	14	Cutting_beginning_along
05.06.2019 20:12:22	14	Return_to_drive_along
05.06.2019 20:12:38	14	Stoppage_in_O_mode_beginning_along
05.06.2019 20:12:48	14	Return_to_drive_along
05.06.2019 20:12:58	14	Stoppage_in_O_mode_beginning_along
05.06.2019 20:13:09	14	Cutting_beginning_along
05.06.2019 20:13:17	14	Stoppage_in_O_mode_beginning_along
05.06.2019 20:13:48	14	Cutting_beginning_along
05.06.2019 20:13:55	14	Stoppage_in_O_mode_beginning_along

Visualizations of defined states on the shearer cycle obtained by each approach are given in Fig. 5.

D. Event log characteristics

The event log preparation procedure yields two event logs:

1) EVENT_LOG_EXPKNW - Event log with activities identified with expert rules;
2) EVENT_LOG_NUMERIC - Event log with activities identified with clustering of numerical variables.

An excerpt from created event logs is presented in Tables II and III.

The EVENT_LOG_EXPKNW includes 44 full cases, and the total number of events is 4425. This event log includes 17 unique classes of events. In the event log EVENT_LOG_NUMERIC, there are 59 full cases and about

TABLE III: Excerpt from EVENT_LOG_NUMERIC.

Timestamp	Case ID	Activity
05.06.2019 19:46:26	14	Stoppage_Working_cutter_drum_2
05.06.2019 19:46:54	14	Moving_Beginning_Drive_from_start_longwall
05.06.2019 19:47:16	14	Stoppage_Working_cutter_drum_2
05.06.2019 19:47:26	14	Stoppage_Start_longwall
05.06.2019 20:11:15	14	Stoppage_Working_cutter_drum_2
05.06.2019 20:12:21	14	Moving_Beginning_Drive_from_start_longwall
05.06.2019 20:13:17	14	Stoppage_Working_cutter_drum_2
05.06.2019 20:19:50	14	Moving_Beginning_Drive_from_start_longwall
05.06.2019 20:20:17	14	Stoppage_Working_cutter_drum_2
05.06.2019 20:20:45	14	Moving_Middle_Crushing

7115 events, with 19 unique classes. The difference in case number between the event logs is the result of missing data in categorical data used for activity identification based on expert rules. Both event logs include only unique variants of the process, which proves that the analyzed process characterizes high variability. However, these two event logs have significant differences in relative frequencies. Certain tendencies in the occurrence of activities can be seen. In EVENT_LOG_NUMERIC, the events appear alternately whilst frequent and infrequent events are dispersed. A reverse situation is observed in EVENT_LOG_EXPKNW. In this case, there are clear sequences of very frequent and infrequent events. In EVENT_LOG_EXPKNW, specific sequences occur at similar places in the cases, e.g., the beginning of the case.

E. Process Discovery

In the process discovery phase, we used EVENT_LOG_NUMERIC to visualize the process flow of discovered specific states in the longwall shearer operation (RQ1). To model their relations in a general view, we used DFG (Fig. 6), created with Disco.

The discovered model shows the complexity of the real-life industrial process (although the presented view covers 65% of the paths in the event log). The most frequent states are mostly at the beginning and the end of the longwall and are related to frequent moving and stoppage operations (Stoppage_Working_cutter_drum_2 denotes stoppage with working-cutting drum, Stoppage_End_longwall_2 denotes stoppage at the end of the longwall without working drums). It can be caused by mining and geological conditions of the coal seam (e.g., problems with roof stability or rock parting in the coal seam). Also, very often, the working drum operations during stoppage can be noticed related to the start of cutting with many repetitions with moving, which can be found undesirable for process continuity and energy consumption (e.g., longwall shearers are machinery with high installed power, even over 2300 kW).

Interesting states, which can also be found in event data, are so-called quick stops, which denote sudden stoppage of the shearer, not recommended from the point of view of shearer operation (RQ2). That example of information can be missed relying only on an event log created with expert rules.

Models based on event logs created with the use of unsupervised techniques (after preprocessing and incorporating the domain knowledge) can provide insights into real process performance. The new knowledge acquired from their analyses could be used in updating:

1) Process and machinery maintenance – analysis of undesirable states related to excessive overload or non-obvious stoppages can significantly improve the operating conditions and control the wearing of machines and underground equipment, e.g., shearer overloading and its root cause analysis may provide additional information about typical coincidences with working conditions.
2) The safety management – through an in-depth analysis of deviations during the process, their causes can be better understood, and adequate protective measures are taken, e.g., a slower cutting rate could be an effect of methane occurrence, so its identification may help to improve air monitoring sensor systems.
3) Reporting and monitoring of work performance – a broad spectrum of analysis of the entire process enables the identification of bottlenecks in several perspectives, leading to improvement of time efficiency, which also translates into financial aspects, e.g., noticed problems in the longwall face due to bad organization of work resulting in time losses, could be resolved and improved with time and cost savings.

The performed analysis revealed the complexity of longwall shearer behavior. A comparison of recorded behavior in relation to the theoretical process model is presented in the next section.

F. Conformance checking

The main objective of the conformance checking task was to verify the discrepancy between the actual process execution and the theoretical process model (RQ3). Thus, we used the log labeled with expert rules (EVENT_LOG_EXPKNW) and the theoretical process model.

This task was supported by ProM. Replay a log on Petri net for Conformance Analysis plugin, yielding a Petri net with additional information such as specific colors and shapes. Figure 7 shows the results of conformance checking analysis – theoretical model as Petri net with colors indicating the deviations. The green bar at the bottom of the activities column indicates the frequency of times the log was synchronized with the model (the darker the color, the more frequent the synchronization). The purple bar at the bottom of the activity column indicates the frequency of cases where executions deviate from the model.

The trace fitness to the theoretical process model is only 0.39. Some transitions are skipped, and 13 places indicate events in the log that could not be explained with the model. The white circles represent the paths followed according to the model, while the yellow circles indicate the occurrence of movements inconsistent with the model. Larger circles mean more frequent alternative movements. There are many places in the Petri net where activities occurred where they were not supposed to (e.g., the activity Cutting_end_along indicated as skipped in all cases after Cutting_middle_along, was performed 45 times after Stoppage_in_O_mode_end_along). This is also confirmation of difficult process performance in the end part of the longwall face (revealed during the process discovery task).

The event log also includes activities that do not occur in the theoretical model as Reversion and Moving, mainly in the

Fig. 6: Process map for EVENT_LOG_NUMERIC.

middle part of the longwall. They occur very often in the execution of the process: 193 and 68 times, respectively. It is also a sign of difficult conditions of the mining process, Reversiondenotes moving in the opposite direction to the cycle part and without cutting (e.g., to clean the conveyor path, among others, after roof collapsing), which is also not desirable from the process efficiency point of view.

Although the generalization of the theoretical model is quite high (0.9997), the performed analysis leads to the question of whether the theoretical process model is not too simple to imitate complex real-life behavior of the longwall shearer operation; maybe revision in this scope is needed. Some repetitions of activities, especially in difficult mining conditions, are normal daily basis practice and should not be pointed out as deviations from the process model.

G. Process enhancement

In the process enhancement, we mostly focused on the analysis of the time perspective (RQ4) and looked for the answers to business questions formulated by domain users (RQ5).

The time perspective analysis was obtained with ProM's plugin Replay a log on Petri net for Performance/Conformance Analysis This plugin allows for the identification of flows with the most frequently performed activities in the analyzed processes and time statistics.

The general statistics for process execution are as follows. The shortest cycle time was 1.75h, and the longest one was 1.09 days. The histogram of the cycle duration is presented in Fig. 8. A few cases can be seen that are too long (two cycles pass two days with night hours; however, taking this into account, the longest cycle time covers 16 hours). The average cycle duration is 6.5h with a standard deviation equal to 5.9h, confirming the high variability of the analyzed process. The four longest activities in the process are:

- Reversion_return with an average duration of 39.61 min and a standard deviation of 47.55min,
- Reversion_along with an average duration of 36.80 min and a standard deviation of 2.01h,
- Cutting_middle_return, with an average duration of 27.80 min and a standard deviation of 1.13h,
- Return_to_drive_return, with an average duration of 25.05 min and a standard deviation of 1.82h.

What is worth noting is that the first two activities are additional to the theoretical model; their presence may indicate problems in the process execution due to mining conditions. The existence of the activity Return_to_drive_return on the list confirms problems in the end part of the longwall face.

The performance of activity Cutting_middle_return requires a more in-depth analysis of process execution in the middle part of the longwall. For this purpose, we created a heat map of events based on EVENT_LOG_NUMERIC (Fig. 9).

The analysis of the heat map revealed the following specific states related to shearer overloading (Moving_Middle_Overloading Moving_Middle_Cooling_down, Stoppage_Quick_stop) confirming issues related to coal cutting activity in the middle of the longwall.

In response to queries from business users, process analysis reveals the type of changes and the places where the changes should be effected. The frequency analysis of activities involved in process executions in specific sections of the longwall face and comparison of the real process performance with the theoretical model leads us to the following conclusions:

Fig. 8: Histogram of shearer cycle duration.

1) Reverse movements in specific parts of the longwall: face at the beginning of the longwall face, reverse movement occurs 26 times, in the middle section—143 times, and at the end of the longwall face—24.
2) More than two stoppages in operation at the beginning and the end of the longwall: 410 stoppages occurred in the beginning section of the longwall face and 812 stoppages at the end of the longwall face.
3) Moving without mining in the middle section of the longwall face occurred two times.

The results obtained, as well as knowledge about events and their occurrence frequency in different longwall locations, can be used in process improvement.

Firstly, the presented analysis could be used to report the most problematic places regarding shearer operation in the longwall face. Secondly, knowledge about event frequency, e.g., undesirable stoppages, can be extended with context data (if available) and case perspective analysis with building classifiers on top of external factors, e.g., mining and geological conditions, labor force, or management conditions for searching causes of process failures.

VII. Conclusions

The underground mining process is a specific example of a business undertaking, which could be viewed as a complex process system enabling minerals extraction.

Presently, control and monitoring systems in underground mines allow the collection of data characterizing the operation of mining machinery and equipment. Sensors located in underground mines measure major operational and environmental parameters. However, integrated management systems implemented in underground mines according to ISO standards define the selected processes, but the underlying theoretical, general models are not confronted with the actual process performance data stored in the mine's IT systems. Such analyses are recommendable because they show the existing structure and method of process implementation, providing the mine operators with information on the deviations and anomalies in process execution [3].

Fig. 7: Conformance checking results on the theoretical model.

In our work, we presented the PM4LMP method with the aim of enabling the modeling and analysis of an underground

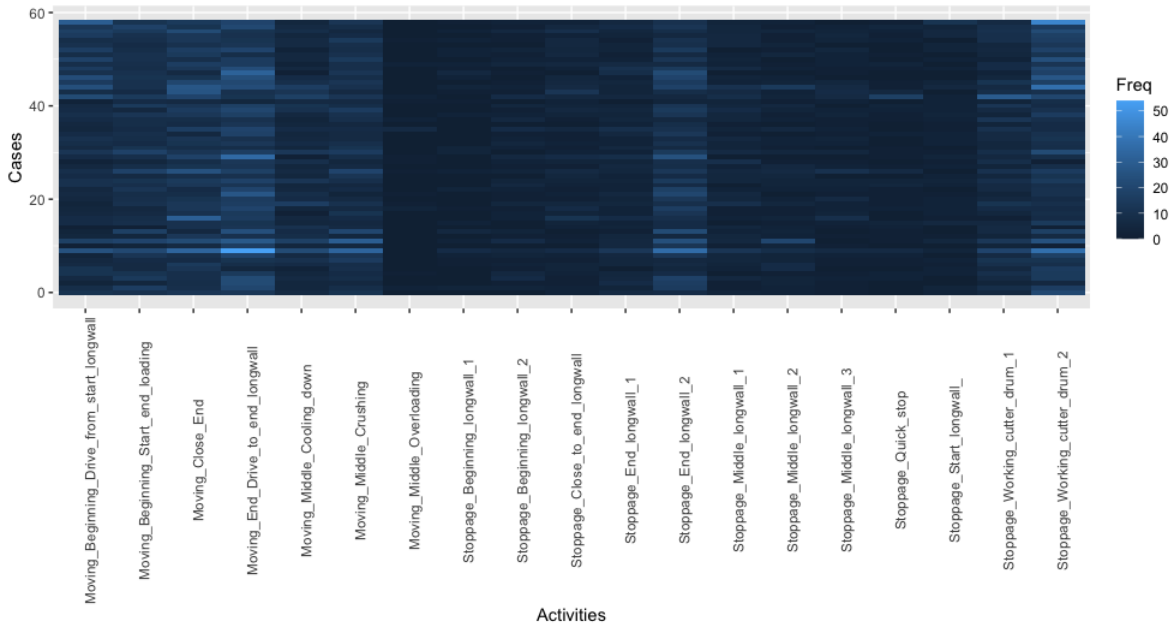

Fig. 9: Frequencies of activities performed in the longwall face.

mining longwall process based on event logs. We showed the potential of complementary usage of supervised and unsupervised approaches in event log creation, enabling the identification of important issues in the process execution.

The analysis of the mining process based on event data is far from a trivial task, presenting several challenges; thus, our method imposes certain limitations, e.g.:

- *Case ID identification* - Because of the data's low quality and numerous errors, the heuristic approach cannot be regarded as entirely reliable and can lead to unreliable results, requiring further event log cleaning.
- *Activity identification* - In the method, only k-medoids clustering was tested as the most common ones.

Considering the mentioned limitations, our future research directions will include exploring alternative ways to generate case ID relying on more variables and more complex rules for detecting the end and start of the cycle. Additionally, data providers will be advised to collect the identifications of such cycles or directly monitor the performance of the machine with incorporated dedicated sensors. Further research is also needed to extend already used clustering techniques with those based on density, e.g., DBSCAN [31] or the newest algorithms like [32].

We are aware that the proposed method is a new approach to the analysis of the longwall mining process and requires improvement, but still, it appears to be a viable alternative to traditional data-oriented analysis in the mining industry operating in the reality of Industry 4.0.

Acknowledgments

The authors thank Dr. Jacek Korski from ITG KOMAG Institute for his support and for sharing his unique expert knowledge about the mining process.

The part of research was done under the PACMEL project, the National Science Centre, Poland, under CHIST-ERA programme (NCN 2018/27/Z/ST6/03392) and statutory funds of AGH University of Krakow.

The part of research is funded by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy, Internet of Production (390621612)

References

[1] J. Bijanska and K. Wodarski, "Process management in a mining enterprise - basic areas and research problems," *Scientific journals of the Silesian University of Technology*, vol. 126, pp. 35–50, 2018.

[2] C. Zhou, N. Damiano, B. Whisner, and M. Reyes, "Industrial internet of things (iiot) applications in underground coal mines," *Mining Engineering*, vol. 69, pp. 50–56, 12 2017.

[3] A. Żuber, "Analyzing mining process based on event data," Ph.D. dissertation, AGH University of Science and Technology, Krakow, 2022, unpublished.

[4] W. M. P. van der Aalst, "Process mining: A 360 degree overview," in *Process Mining Handbook*, ser. Lecture Notes in Business Information Processing, W. M. P. van der Aalst and J. Carmona, Eds. Springer, 2022, vol. 448, pp. 3–34. [Online]. Available: https://doi.org/10.1007/978-3-031-08848-3_1

[5] K. Diba, K. Batoulis, M. Weidlich, and M. Weske, "Extraction, correlation, and abstraction of event data for process mining," *WIREs Data Mining Knowl. Discov.*, vol. 10, no. 3, 2020. [Online]. Available: https://doi.org/10.1002/widm.1346

[6] S. J. van Zelst, F. Mannhardt, M. de Leoni, and A. Koschmider, "Event abstraction in process mining: literature review and taxonomy," *Granular Computing*, vol. 6, no. 3, pp. 719–736, 2021.

[7] C. Janiesch, A. Koschmider, M. Mecella, B. Weber, A. Burattin, C. Di Ciccio, G. Fortino, A. Gal, U. Kannengiesser, F. Leotta, F. Mannhardt, A. Marrella, J. Mendling, A. Oberweis, M. Reichert, S. Rinderle-Ma, E. Serral, W. Song, J. Su, V. Torres, M. Weidlich, M. Weske, and L. Zhang, "The internet of things meets business process management: A manifesto," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 6, no. 4, pp. 34–44, 2020.

[8] W. M. P. van der Aalst, *Process Mining: Data Science in Action*. Springer, 2016.

[9] T. Slaats, "Declarative and hybrid process discovery: Recent advances and open challenges," *J. Data Semant.*, vol. 9, no. 1, pp. 3–20, 2020. [Online]. Available: https://doi.org/10.1007/s13740-020-00112-9