Process Mining in a Line Production

Cristina Santos[1], Joana Fialho[1,3], Jorge Silva[1,2] and Teresa Neto[1]

[1] Polytechnique Institute of Viseu, Portugal
[2] Huf-Group
[3]CI&DEI
jfialho@estgv.ipv.pt

Abstract. The search for more efficient strategies, cost savings, time optimization and productivity are the main goals of any successful company. Process mining arises in this context and, although it is not a new concept, its expansion and applicability in the market has recently become notorious. Through an extensive set of data recorded over time, it is possible to determine the real state of a company's processes, allowing to diagnose failures and improve the efficiency of these processes.
This paper describes a project realized in the scope of process mining, developed in the company Huf Portuguesa. The machines of a production line record thousands of data. In a specific production line, data was collected in a time range, cleaned and processed. Process mining techniques allowed the discovery and analysis of the real state of the production line. All paths were detected, and each path was analyzed individually. The conformity was also analyzed.

Keywords: Production line, Process mining, Process Mining tools.

1 Introduction

The race for privileged positions in the business industry results in constant pressure on organizations. Authenticity, credibility, and leadership are the main goals and motivations of any successful company, leading to a pursuit of business process innovation. This innovation goes beyond simple data analysis or productivity levels; it requires a much deeper approach capable of acting in various areas, improving methods, and processes. The search for efficient solutions allows the rise of information systems (IS) that are increasingly tailored to reality.

We are in digital age, where the importance of social networks, rankings, data collection, and manipulation or access to algorithms for personal benefit, are becoming more and more prevalent. Currently, there are few activities that do not leave any kind of record. With such an extensive and comprehensive collection of data, there is the opportunity for processing them and drawing relevant conclusions.
The registration of events produced daily in organizations allows storing the necessary information to, after proper manipulation, seek to understand the actual behaviors of people and/or processes executed in those organizations. On the other hand, manipulating this information can also help identify potential points of failure and

deviations. This understanding is crucial so that, after a correct analysis, the best decisions can be made to lead to effective improvements.

Huf Portuguesa, a member of the German Huf Group, headquartered in Tondela, Portugal, is an automotive industry that produces components for automobiles, such as locks, keys, steering column locks, door handles, and rear door emblem handles with a vision camera.

The organization has a diversity of processes in different departments and various information systems (IS) that record data during the execution of these processes. This study focused on a specific process: the series production line of components manufactured by the company. Data recorded (event logs) can be studied and analyzed to detect routines, patterns, and preferences, allowing for the improvement or redesign processes. This technique of discovering, monitoring, and improving real processes by extracting knowledge from data available in information systems is known as process mining.

Process mining can be defined as a technique for extracting information from an event log containing relevant data about the activities performed by an organization or system in a business area. It involves the use of Data Mining combined with process modeling and analysis. The objective of process mining is, therefore, the automatic discovery of a process model by observing events recorded by certain corporate systems.

The paper is organized as follows: in addition to this introductory section, Section 2 briefly discusses the area of process mining, Section 3 describes the analyzed Huf process, and Section 4 concludes the paper.

2 Process Mining

It becomes necessary to organize information since, as cited by [1], "organizations have to confront uncertainty and disorderly events coming from both the interior and the exterior while still providing a clear, operational, and well-defined conceptual scheme for the participants."

The potential and importance of Information Technologies (IT) and Information Systems (IS) within organizations are more than proven. The development of IT allows introducing changes in business and responding to the evolution of markets. Companies face the challenge of analyzing massive amounts of data gathered daily (Big Data). Big Data are attributed to have the features Volume, Velocity, and Variety. Demchenko et al. (2013) proposed wider definition of Big Data as 5 Vs: Volume, Velocity, Variety and additionally Value and Veracity [2]. The volume of data, the speed at which it is recorded, the variety, its veracity, and value are a reflection of the scenario experienced by organizations [3].

Process Mining has revolutionized this area by enabling the fusion of machine learning algorithms and Big Data concepts. This concept is titled as "the Artificial Intelligence revolution that drives process excellence" [4].

The activities performed by people, machines and software leave traces in the so-called event logs [5]. Process mining techniques use these logs to discover, analyze and improve business processes [6].
Process Mining focuses on acquiring specialized information regarding corporate processes. By using the available event logs, it is possible to analyze various factors, such as associating patterns, identifying the most common paths, exceptions, error-prone regions, what triggered a decision, and who made them [7].

Process Mining is a concept that can be interpreted as the intersection of two areas, namely, data science and process science, as shown in **Fig. *1***. Data science is a concept that encompasses the study and algorithms for problem-solving, including data extraction, preparation, and transformation. Process science "is the interdisciplinary study of processes aimed at understanding, influencing, and designing processes" [8]. In other words, it refers to a generic term regarding a broad discipline that combines IT with knowledge of management sciences, with its main focus on processes [9].

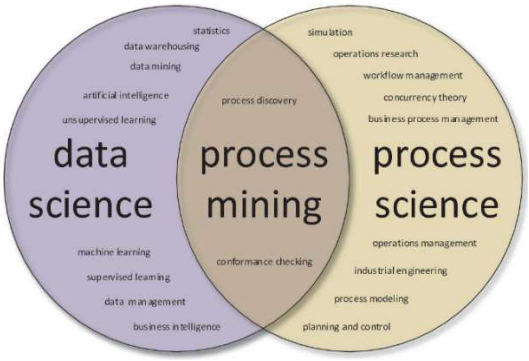

Fig. 1 Process Mining [9]

Fig. *2* highlights the phases of process mining. Initially, the extraction of data from Information Systems (IS) is crucial since it forms the basis of the entire process. However, it may be necessary to explore data before applying process mining.

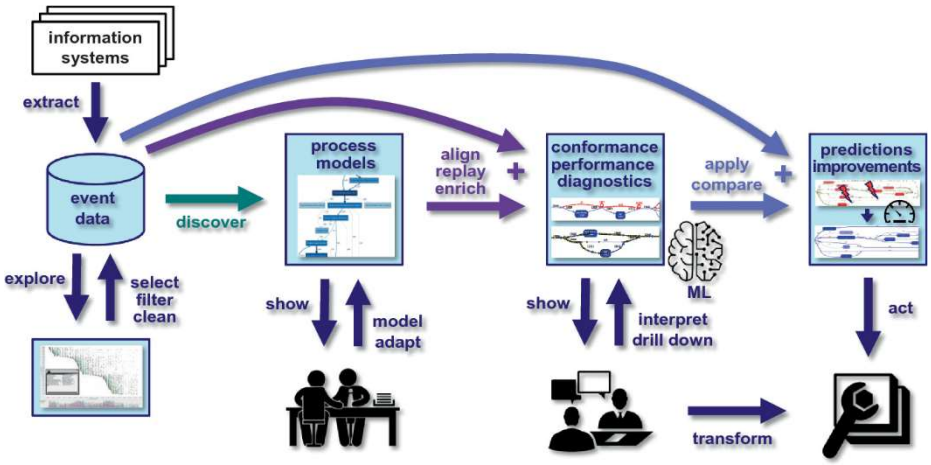

Fig. 2 Process Mining stages [9]

After the data collection, it is essential to explore its usefulness, select relevant information, and proceed with data treatment and cleansing. The data must contain three essential fields (case id, activity, timestamp) to enable its analysis. The case id is a unique field that identifies the event log, the activity represents the record of the activity, and finally, the timestamp allows for the identification of the timeline and detection of the sequence of activities (**Fig. *3***).

With the filtered data, the first of the three stages of process mining begins - discovery. This stage involves organizing all data and identifying the real paths of the process.

The second stage of process mining is the conformance/compliance verification. Here, it is verified whether the real events follow the predicted path/model. In this phase, a diagnosis can be drawn, leading to the final stage of process mining - applying improvements.

	CASE ID	ACTIVITY	TIMESTAMP
part_number	serial_number	operation	transaction_timestamp
123456	345622	A	17/08/2009 08:00:00
123456	345622	B	17/08/2009 08:00:23
123456	345622	C	17/08/2009 08:00:55
123456	345622	D	17/08/2009 08:01:04

Fig. 3 Data Collection example

The main advantage of implementing process mining techniques is the optimization of internal company processes and make data profitable. By using the data collected daily, it is possible to generate flowcharts that represent the processes exactly as they occur. Process mining ensures a comprehensive and informed view of the actual state of the company, which aids in identifying improvement opportunities, making decisions, increasing productivity, and saving time. Acting based on real data is more reliable than relying solely on studies and assumptions.

Process mining can be applied in any area, if there is a constant and temporal record of data. Its applicability provides better visibility into processes to meet the company's and its potential customers' expectations. Time optimization, cost reduction, improved outcomes, and increased productivity are particularly attractive in the industrial sector. In the financial services and banking industry, process mining is a great solution to gain transparency over processes and financial transactions, detecting issues that may cause losses. Additionally, the healthcare sector is equally enticed by the possibility of understanding the patient's journey, improving hospitals' operational efficiency, and enhancing diagnoses [4].

To apply process mining algorithms, there are several software tools, some open source, e.g. ProM and Apromore, other commercial tools, e.g. Disco, Celonis, ProcessGold, among others. In the literature, ProM [10], Disco [11] and Celonis [12] are mostly used. In this experiment, ProM and Celonis were used.

The development of Celonis began in 2011, and nowadays, it stands as one of the leading tools in the market, serving globally recognized clients such as Siemens, Uber, Cisco, Vodafone, among others. One of the significant advantages of Celonis is its real-time integration capability with Relational Database Management Systems (RDBMS). In addition, it also allows the importation of event logs from traditional file formats such as .csv and .xls.
Celonis is a cloud-based process mining tool, and its main benefit lies in not requiring any installation on the user's computer. It is a commercial tool that requires payment, but it provides some basic features available for free use, which was the version utilized for the development of this project.

ProM is an open-source extensible framework that supports a wide range of process mining algorithms as plug-ins. It is a tool very used [6]. The framework is flexible with respect to the input and output format, as it supports several formats, e.g. Petri nets, social networks [13], among others. Plug-ins can be used in a variety of ways and combined to be applied in real-life situations [10]. This software offers more than 1500 plug-ins [14]. It is an independent platform developed at Eindhoven Technical University by a research group led by Wil Van der Aalst [15]. The group actively invites investigators to contribute in the creation and development of new plug-ins, enriching the tool and maintaining the existing ones.

3 Production Line Process

The case study relates to a specific production line where the pieces of a product follows a sequential path from one machine to another. Each machine performs its operation and records numerous data regarding the passage of the piece. Ideally, the first record of the piece occurs at the first machine of the line, it goes through all the machines sequentially, and finally passes through the last machine for validation.

The BPMN model is used to represent processes in a standardized way through representative standard icons. BPMN diagrams use specific symbols and elements to represent different activities, events, gateways, and flows within a process, making it easier for stakeholders to understand and analyze complex processes.

Fig. *4* represents a BPMN (Business Process Model and Notation) model of the entire production line layout. The notation of this diagram identifies the machines and describes the process logic and the flow of activities.

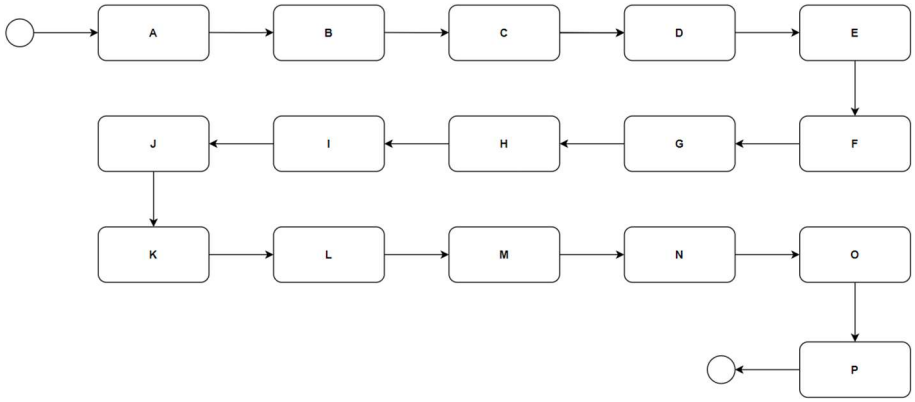

Fig. 4 Production Line layout

This is an older production line, and despite being composed of a larger number of machines, it was not possible to collect data from all of them. The line includes machines from other suppliers, and their data is not accessible. However, the BPMN model presented below (**Fig. *5***) corresponds to the sequence of machines that provide records.

.

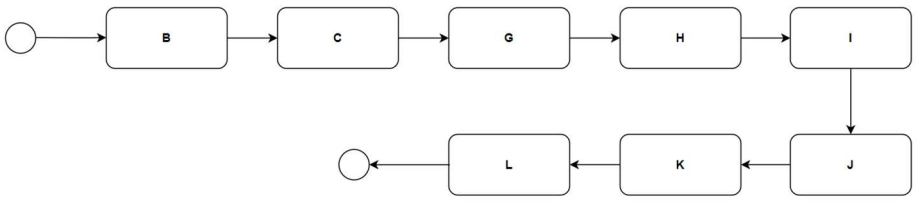

Fig. 5 Layout used.

The data was provided using CSV files. Its cleansing was performed using Python language and functions from the Pandas library. To ensure a more realistic analysis, all records that did not contain data from the first or second machine and the last machine were excluded. In other words, only nearly complete records of the production line were considered for the study.

3.1 Event Logs

Event logs typically correspond to an extensive database with a vast amount of data. In this study, the final dataset consists of 25 595 cases (number of case ids under analysis) and 181 461 activities (number of records). The data was collected in two periods: from May 22nd (06:02:57) to May 27th (05:53:38) and from May 29th (05:59:59) to May 30th (16:41:34).

The information and fields provided by the information systems are generally more extensive than necessary for a particular analysis. In this specific project, the following fields were considered:

- part_number: identifies the reference of the produced part. Each reference is composed of multiple records of serial_number.

- serial_number: represents the serial number of each part, a unique value that corresponds to the case id of the event log.

- operation: Identifies the machine, a unique value that serves as the machine's case id.

- transaction_timestamp: records the date and time when the activity was executed, essentially serving as the timestamp of the event log.

3.2 Discovery

The discovery phase is responsible for identifying the actual behavior of the process. It involves identifying the most frequent paths and unusual sequences. This phase was executed using the Celonis tool. The event logs were imported and the discovery process was performed, resulting in the real mapping of the process, as shown in **Fig. *6***. For each path, it was possible to associate the corresponding serial_number. The diagnostics and conclusions herein presented are described in order to safeguard any confidential information related to the company.

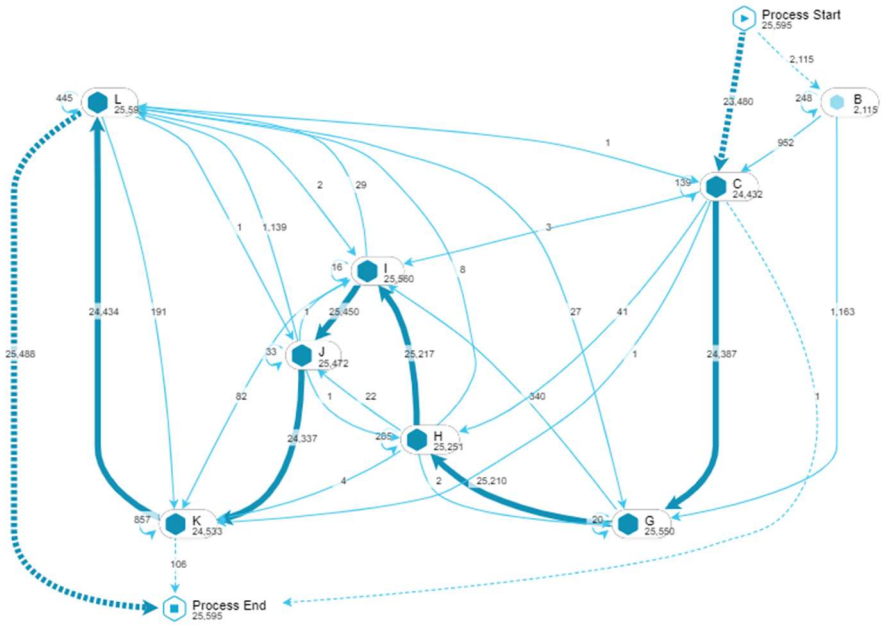

Fig. 6 All paths discovered

The numbers on the lines indicate the frequency of occurrences and the thicker the line, the higher is the frequency/probability of following that path/passing through that machine.

Fig. *7* shows the most frequent path, which corresponds to about 72,7% of all paths. Notice that the event logs do not initiate their process at the first machine of the line (B). This phenomenon occurs because the first machine is for validation, and it can be done manually by an operator. In these cases, the machine doesn't register anything, but this path is expected.

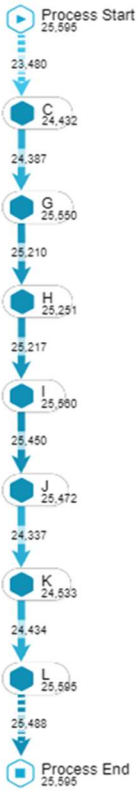

Fig. 7 More frequent path

The remaining paths will be presented in descending order of occurrence percentage and divided into three groups. Group A consists of cases that are also expected to occur in the production line. Group B records cases with a low percentage of occurrence, happening unexpectedly, but they are only process issues that do not interfere with the quality. Finally, Group C corresponds to other identified cases that, despite representing low percentages, are significant to analyze because the company evaluates its quality rate in units per million.

3.2.1 Group A

Table *1* displays the cases of Group A, with the frequency of occurrences and the percentage they represent.

Table 1 Variants of Group A

Cases	Frequency	Percentage
Case 1	2115	8,26%
Case 2	1163	4,54%
Case 3	952	3,71%

Case 1 in Group A represents the event logs that initiate the process at the first machine of the production line (B). Although this is the desired path, out of the 25 595 cases, only 2 115 start the process at the first machine of the layout, due to the reasons explained in the most frequent case. Additionally, as observed in **Fig. *8***, even though they start at the first machine, there is a division in the paths, which will be further analyzed in cases 2 and 3.

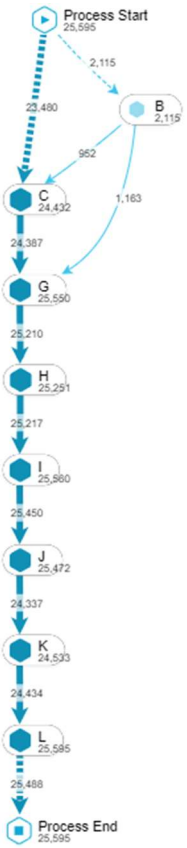

Fig. 8 Case 1 (Group A)

Case 2 corresponds to the scenario where the pieces go directly to the 3rd machine (G) (**Fig. *9***). This situation is explained by the existence of pieces in the production line that do not utilize this machine (G).

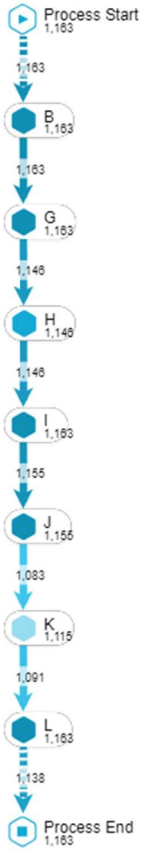

Fig. 9 Case 2 (Group A)

Case 3 (**Fig. *10***) represents the expected path, where the parts follow the production line sequentially, starting at the first machine of the layout, with no jumps. However, out of all the records analyzed, only 952 paths, approximately 3,71%, behave as the process was designed.

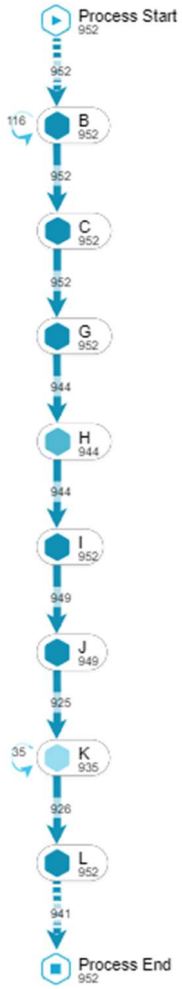

Fig. 10 Case 3 (Group A)

3.2.2 Group B

Table *2* displays the cases of Group B, their frequency of occurrences and percentage.

Table 2 Variants of Group B

Cases	Frequency	Percentage
Case 1	1139	4,45%
Case 2	857	3,35%

Case 1 represents a set of 1 139 paths that do not pass through machine K (**Fig. *11***). These cases are common and are related to a process issue associated with component detection. The operator removes the piece from the production line, validates it visually, and then places it back on the line to proceed to the next machine. This situation results in the absence of a record in that machine. It may be considered to implement a specific registration process for these cases.

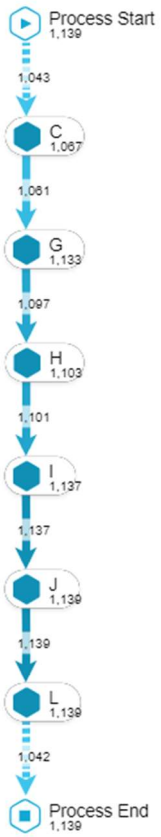

Fig. 11 Case 1 (Group B)

Case 2 (**Fig. *12***) corresponds to the loops detected at machine K. Approximately 857 cases (3,35%) are identified as instances where the piece passes through the machine twice, despite the first pass not encountering any errors. The machine issues 2 confirmation telegrams due to robot programming reasons.

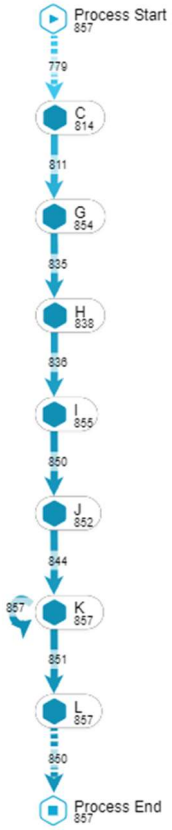

Fig. 12 Case 2 (Group B)

Similarly to the loops detected at machine K, several other cases were identified (**Fig. *13***) with a lower percentage, but for the same reason. The machines emit a double signal of OK confirmation, with a time difference of seconds, which is less than the machines' working period. However, two registers are emitted, resulting in false loops.

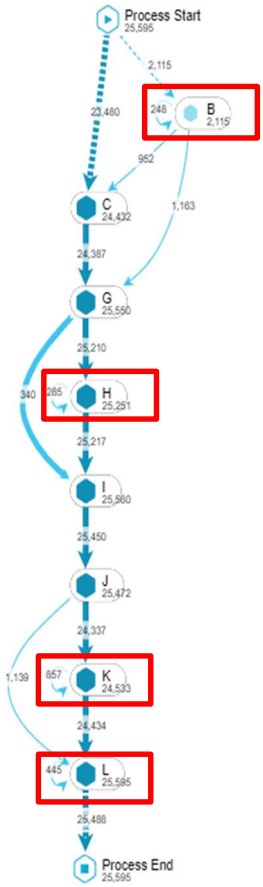

Fig. 13 Similar cases of Case 2

3.2.3 Group C

Table *3* displays the case information of Group C.

Table 3 Variants of Group C

Cases	Frequency	Percentage
Case 1	340	1,33%
Case 2	191	0,75%
Case 3	106	0,41%
Case 4	82	0,32%
Case 5	41	0,16%

Case 1 identifies a set of 340 paths that do not pass through machine H. This is due to operator intervention, where the operator performs the task manually while the machine is performing its function on another piece.

Case 2 represents a set of 191 cases, approximately 1,33%, where paths skip machines and do not follow the natural sequence of the production line. This is due to piece revalidation. The piece is rejected by machine L (responsible for quality assurance) and is returned to the production line for quality parameter confirmation.

Case 3 consists of 106 records (0,41%) where the pieces do not pass through machine K and backtrack in the production line sequence. These are rare cases, and there is currently no explanation; it could be a reading issue.

Case 4 represents a set of 82 cases (0,32%) where there is no record at machine J. This may occur due to a machine reset, and the operator performs the assembly manually.

Case 5 corresponds to 41 records, approximately 0,16% of the cases, where the pieces do not pass through machine G. After analysis, it is concluded that these are isolated cases resulting from lack of experience of the production line operators.

4 Compliance Verification

The conformance/compliance analysis was performed in ProM and requires two inputs: a Petri net and the log file of the process model. The Petri net is defined graphically as a directed graph. It consists of two types of nodes: places, represented as circles, and transitions, represented as squares. The connecting arcs are represented by arrows, used to link two nodes of different types.

The CSV data file was converted to XES (a process mining format adopted by the IEEE Task Force and the latest version of this tool). Using the Inductive Miner algorithm plugin, responsible for extracting the process model from the logs, a Petri net was generated. In **Fig. *14***, it is possible to observe the data flow and the machines on the production line; the black box indicating an alternative flow [14].

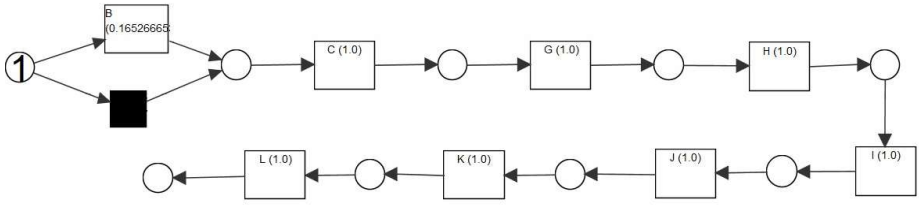

Fig. 14 Rede Petri net

To analyze conformance, the plugin "Replay a Log on Petri net for Conformance Analysis" was applied, resulting in the diagram shown in **Fig. *15***.

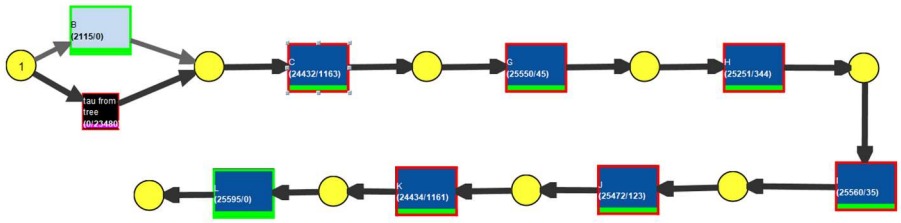

Fig. 15 Compliance verification

The yellow circles indicate movements outside the model, i.e., non-conformant behavior. When analyzing the diagram, several details are noteworthy:

- The green transitions show that the actual logs followed all the expected transitions.
- The red outline represents activities that are not in compliance.
- The black rectangles indicate decision points for trajectories to be followed after the execution of an activity, with the pink bottom bar indicating cases with divergent executions relative to the model.
- The activities with dark blue boxes represent frequent activities in the process execution.
- The value inside the rectangles indicates the conformance frequency/non-conformance frequency ratio.

In conclusion, the conformance analysis performed reinforces the results obtained by the Celonis tool. It was detected a majority conformance, which starts at the second machine on the production line, with only a small percentage of cases not conforming to the expected behavior.

5 Conclusions and future work

After the discovery and conformance analysis, non-conformances that disrupt the expected process flow were identified. These cases need to be carefully analyzed, as they may be responsible for inefficiencies in the production line, performance losses, misutilization of available resources, or even productivity delays. Consequently, it becomes feasible to consider implementing improvements or devising a plan.

It is indeed relevant to consider reprogramming the machines for future implementation to extract more reliable, concise, and valuable data for process mining and, potentially, leading to predictive models. With optimized data extraction, the following could be achieved:

- Forecasting execution times based on shifts.
- Identifying the most efficient sequence of machines/production line.
- Assessing if the required number of operators remains constant concerning the references in production.

• Managing the distribution of the number of operators based on the references in production.
• Predicting potential machine failures and scheduling preventive maintenance.
• Identifying patterns of issues in the production lines.
• Detecting periods of profitability.
• Estimating the daily production numbers to meet customer demands.

Implementing these suggestions across various production lines in the company would not only increase productivity but also minimize errors and faults, leading to more efficient and streamlined processes.

Acknowledgement

This work is funded by National Funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project Refª UIDB/05507/2020. Furthermore, we would like to thank the Centre for Studies in Education and Innovation (CI&DEI) and the Polytechnic of Viseu for their support.

References

1. Daft, R., & Lengel, R: Information Richness: A New Approach to Managerial Behavior and Organizational Design. Greenwich: CT: JAI Press (1984).
2. Demchenko, Y., Membrey, P. et al.: Addressing Big Data Issues in Scientific Data Infrastructure. (2013) 10.1109/CTS.2013.6567203, Last accessed 2023/07/27.
3. Oliveira, R.: Mineração de Processo com Celonis Framework (2016). https://www.linkedin.com/pulse/minera%C3%A7%C3%A3o-de-processo-com-celonis-framework-rosangela-oliveira/?originalSubdomain=pt. Last accessed 2023/07/08
4. UpFlux Process Mining. https://upflux.net/pt/process-mining/, Last accessed 2023/07/02
5. Van Der Aalst, W.: Process mining: Overview and opportunities. ACM Transactions on Management Information Systems, 3 (2), pp. 1–17 (2012)
6. (Batista, E., Solanas, A.: Process mining in healthcare: A systematic review. In 9th International Conference on Information, Intelligence, Systems and Applications, pp. 1–6 (2018).
7. Iervolino, L.: Process Mining: Entenda a realidade dos seus processos. (2018) https://www.linkedin.com/in/luigi-iervolino-67981b/recent-activity/articles/, Last accessed 2023/06/20.
8. vom Brocke, J. and van der Aalst, W. et al.: Process Science: The Interdisciplinary Study of Continuous Change (2021). Available at SSRN: https://ssrn.com/abstract=3916817, last accessed 2023/30/06
9. van der Aalst, W.: Process Mining: A 360 Degree Overview. in W. C. van der Aalst (Ed.), Process Mining Handbook. Lecture Notes in Business Information Processing. Springer, (2022)
10. Van Dongen, BF, De Medeiros, AKA, et al.: The ProM framework: A new era in process mining tool support. Lecture Notes in Computer Science, 3536 (i), pp. 444–454 (2005).

11. Günther, C. W., & Rozinat, A.: Disco: discover your processes. In N. Lohmann, & S. Moser (Eds.), DEMONSTRATION TRACK OF THE 10TH INTERNATIONAL CONFERENCE ON BUSINESS PROCESS MANAGEMENT (2012)
12. Badakhshan, P., Geyer-Klingeberg, J. et al.: Celonis process repository: A bridge between business process management and process mining. In CEUR Workshop Proceedings, 2673, pp. 67–71 (2020)
13. Van Der Aalst, WMP, Song, M.: Mining Social Networks: Uncovering Interaction Patterns in Business Processes. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3080, pp. 244–260 (2004)
14. ProM Tools, ProM Documentation: https://promtools.org/prom-documentation/, last accessed 2023/07/07
15. Ailenei, I.: Process mining tools: A comparative analysis. Master Thesis. Eindhoven University of Technology (2011)