Journal of Theoretical and Applied Information Technology

31st August 2023. Vol.101. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PROCESS MINING APPROACH FOR DISCOVERING AND ANALYZING THE HEALTHCARE PROCESSES IN PYTHON

ABDEL-HAMED MOHAMED RASHED[1], NOHA E. EL-ATTAR[2], DIAA SALAMA ABDELMINAAM[3], MOHAMED ABDELFATAH[4]

[1,2,3,4] Information System Department, Faculty of Computers and Artificial Intelligence, Benha University, Egypt

[2]Department of Computer Science, Faculty of Computers and Informatics, Misr International University, Egypt

E-mail: [1]hameederasheede@yahoo.com, [2]noha.ezzat@fci.bu.edu.eg, [3]diaa.salama@fci.bu.edu.eg, [4]Mohamed.abdo@fci.bu.edu.edu,

ABSTRACT

As healthcare expands, there is an increasing need for innovative approaches to improve healthcare quality. One such approach is process mining which involves data analytics to extract insights from careflows and patient data in real-time. The process mining techniques can support business processes based on event log from information systems to model those processes and uncover challenges and bottlenecks; it supports the professionals of the field with a view of the problems that are currently occurring in the field. This paper applied process mining methods and techniques in the healthcare field to analyze the careflows and gains meaningful insights. The proposed approach has three main steps; preprocessing the dataset, applying miner algorithms to discover the process model, and applying the performance analysis to discover the existence deviation in the discovered model. This work can contribute to bring a Python (PM4Py framework) as a process mining tool rather than traditional tools that typically neglect to help process mining methods in large-scale analysis. This proposed method can be applied across several hospitals; it analyzes the patient careflows from the control-flow and the performance perspectives. It provides accurate analysis and insights to hospital administrators.

It furthermore, it provides accurate analysis and insights to hospital administrators. The paper used a dataset of cardiology patients in an Egyptian hospital. The results of the applied approach based on PM4Py framework are efficient and satisfactory; it gives objective analysis and insights to hospital managers to improve performance care processes.

Keywords: *Process mining, Healthcare process, Cardiology, PM4Py, Event log.*

1. INTRODUCTION

The business analysis process is essential for any business organization for many different reasons; including evaluating an organization's internal performance, increasing awareness of how people work and interact, and recognizing opportunities for resource and efficiency improvements. In addition, organizational information systems can record activities that produce frequent event logs. These event logs that were produced in the past were mainly employed for tracking, accounting, auditing, etc.

Organizations have recently started using this data for business analytics and improving service quality. Process mining is the approach that employs the event logs to discover the actual process model to assist the professionals and managers in mining profoundly the big data and extracting meaningful insights, and making business better. However, some challenges occurred during the execution of the business as delays and

bottlenecks due to complex rules when modeling the business that was no longer applicable or irrelevant, so the process mining comprises finding knowledge from event data that is accessible in modern business applications, IT framework and security frameworks to monitor, and enhance process models. For process mining purpose, every event that is kept in an event log contains the accompanying data: a case id, which recognizes the process instance; an activity or task name, which recognizes the action that has been performed; a user name or originator name, which distinguishes member who played out the task; a timestamp that shows the date and time the task was finished [1]. For example, a sample event log from one instance of the medical process is depicted in Figure 1.

ActivityNo	CaseNo	ActivityName	StartDate	FinishDate	lifecycle:transition	org:resource
.....	.....	.....	.....	.....	.....	.....
A201707	14019304	First Admission	3/30/2022 9:00	3/30/2022 9:30	complete	Receptionist...
A201708	14019304	Initial Checkup	3/30/2022 9:30	3/30/2022 10:00	complete	General Practitioner...
A201709	14019304	Lab tests	3/30/2022 10:00	3/30/2022 13:00	complete	Laboratory specialist...
A201710	14019304	Radio tests	3/31/2022 1:00	3/31/2022 1:20	complete	Radiology specialist...
A201711	14019304	Cardiac Stent	4/1/2022 10:00	4/1/2022 11:30	complete	Catheter and stent ...
A201712	14019304	Consultant Checkup	4/4/2022 10:00	4/4/2022 10:40	complete	cardiology consultant...
A201713	14019304	Decide the surgery	4/5/2022 22:40	4/5/2022 23:00	complete	Administration official...
A201714	14019304	Admission&Request	4/9/2022 23:00	4/10/2022 23:00	complete	Receptionist...
A201715	14019304	Discharge &Leave	4/20/2022 20:05	4/20/2022 20:20	complete	Receptionist...
.....	.....	.....	.....	.....	.....	.....

Figure1: An example of a sample event log

The organization that wishes to use process mining in their business processes will get the advantage of decreasing process time; once identifying the issues and solving them; also the business will be more efficient. But many challenges will face the organization that implements process mining. From these challenges the data quality; when data is suitable and fit, the outcomes of the process mining are good too. Another challenge when no enough data is being recorded by process systems.

There are many obstacles to manage healthcare processes, including the fact that they are increasingly multidisciplinary, highly dynamic, complex, ad hoc, and affect both the quality and cost of care [2]. Noteworthy, the complexity of the processes can be up to several months. Process mining can be a good solution by redesigning the business processes to be compatible with the standard model. The process mining can reveal insights to prove that the predesigned blueprint model (based on beliefs and opinions) contradicts reality. Based on event logs extracted from business information systems, business process mining employs methods and tools to discover, monitor, and enhance actual processes.

The proposed method in this paper uses different process mining techniques to discover the process models from historical event data; in order to obtain a closer reflection of reality. The process discovery algorithms aim to discover patient careflows from the event logs extracted from the hospital's information system. The discovered model investigates the patient's journey throughout the hospital, from admission to discharge. The best-discovered model will be chosen to check the consensus with the reference model to overcome any bottleneck or deviation. Furthermore, selected most properly discovered model and event log were analyzed under the performance perspective. The analysis results are significant for any recommended improvements in hospital processes. Still, there a rare of Python based process mining applications in healthcare as it is studied and reviewed in the next subsection 1.3 "related works", so this work wants to explore the powers of PM4PY framework in analyzing the business processes of patients.

The rest of the paper is organized as follows; the first section briefly overviews process mining in healthcare, its tools, and related works. Then, section 2 describes the research methodology and the case study, while the results obtained are discussed in Section 3. Finally, a short discussion concludes the paper and future works.

1.1 Healthcare Process Mining

Process mining includes more than automated process discovery, conformance checking, and organizational mining; it can provides more functions as constructing a simulated models, model extension, model repair, predictive monitoring, and process-based recommendations.

Event data may be extracted from various specialized systems in healthcare environments, including those based on electronic patient records (EPR). Another example is the ability of radiology information systems (RIS) to document the examination request to report workflow for patients. Likewise, the emergency information system records the patients' medical processes of, and billing information system generally consolidates information from other medical systems about the processes performed on a patient [3]. Figure 2 illustrates the execution of the process mining in a medical care center. Event logs can be used for various process mining roles once extracted from the hospital information system, these can be classified into three main roles: discovery, conformance checking and enhancement. Firstly; the discovery process aims to

create process models from the extracted event logs automatically. Secondly; the conformance checking begins by contrasting the event logs and predesigned blueprint models to recognize conformance and settle bottlenecks. At long last, the upgrade process gives the found model the bits of knowledge extricated from the event logs, for example, utilizing execution data or timing information on a model to show the bottlenecks, throughputs, and frequencies [4].

Event data logs can be investigated from various perspectives: (1) the control-flow perspective; (2) the performance perspective; (3) the organizational perspective; and (4) the data or case perspective. The control-flow perspective is focused on the behavior of the process, precisely the steps in the process and the order in which they are carried out. The relationships between the users who carried out the activities are the primary focus of the organizational perspective. This includes whether the users are members of the same group or different groups or organizational units. The performance perspective tries to figure out performance indicators like throughput times and sojourn times or find bottlenecks. Finally, the case/data perspective focuses on how the data serve the activities. The case can be identified by its path or the people working on it.

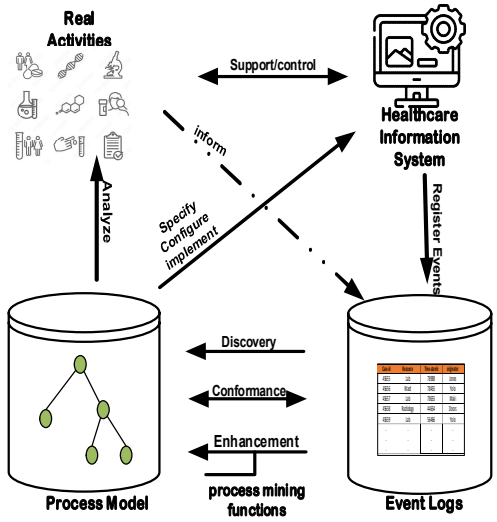


Figure 2: Process mining functions: process discovery, conformance checking and enhancements for thehospital [4].

1.2 Process Mining Tools

Process mining received a lot of attention from academia and industry, which resulted in the creation of several open-source and commercial process mining tools. Several open-source tools that support process mining techniques such as ProM [5], RapidProM [6], Apromore [7], bupaR [8], PM4Py [9], and PMLAB [10]. Also, there are famous commercial tools in process mining, to name a few, such as Disco [11], Celonis [12], Uipath [13], and QPR ProcessAnalyzer [14]. In addition, most of the open-source projects give a standalone tool that permits importing an event log and performing process mining analysis on it, providing a graphical user interface that is simple to use to engage users who are not experts and showcase process mining to a broader audience.

Process Mining for Python (PM4Py) [9] aims to close the gap between data science and process mining; as a library of process mining features, PM4Py integrates with Python and a set of process mining features. The PM4Py library offers several well-known methods for process mining, including: Process discovery, Conformance Checking, Filtering, Graphs, and Social Network Analysis. The PM4Py library provides the Process discovery algorithms such as: α- miner [15], Inductive miner [16], and Heuristic miner [17].

The PM4Py library provides the Conformance Checking algorithms such as token-based replay and alignments [18]. In addition, the PM4Py measures the fitness, precision, generalization, and simplicity of process models. Also the PM4Py provides: Python visualization libraries such as NetworkX, GraphViz, and Pyvis. The PM4Py library has the main advantages of making algorithmic development and customization more accessible; allowing process mining algorithms to be easily integrated with algorithms from other data science areas; carried out in many state-of-the-art Python packages; providing conversion capabilities for converting event data objects between formats; PM4Py provides analyzing big datasets due to it is using pandas data frames; also many process model notations, such as Petri networks, heuristic networks, process trees, Petri networks, and transition systems, can be customized by PM4py.

Several process mining applications were implemented in several business environments such as education, industry, banking and etc. Also, hundreds of academic or commercial tools were utilized to implement the process. According to the article [19], ProM (43 %) and Disco (20 %) were the process mining tools that were utilized the most, accounting for more than half of all options. And many programming languages that used for developing process mining algorithms such as Python, C#, R, and C. Still, there a rare of Python based process mining applications in healthcare.

Table 1 many features of two famous process mining tools (Prom and Disco) and the PM4Py.

1.3 Related Works

This section presents various related process mining applications utilizing the PM4Py framework. Study [20] used the PM4Py framework to demonstrate a novel modeling approach that can calculate a graph showing the relationships between activities without requiring the user to specify a case notion. The resulting models are called Multiple Viewpoint (MVP) models in which classes and objects serve as links between activities and events; what sets MVP models apart from the nearest similar approaches (OpenSLEX and OCBC models) is fast execution time and usability. The search [21] has aimed to demonstrate the effectiveness and suitability of Interactive Process Discovery (IPD) to model healthcare processes. Using the PM4Py framework, IPD lets the user discover the process model interactively, taking advantage of the domain knowledge alongside the event log; the work is performed by utilizing a dataset from an Italian Clinic, the study has limitations of that methodology is executed on the logs of single hospital, so it is better to apply the methodology in different healthcare context.

Table 1. Comparison of two famous process mining tools versus the PM4Py

Features	ProM (6.5.1)	Disco (1.9.5)	PM4Py
License	Open source	Evaluation, Academic, Commercial	Open source
Output model notation	BPMN, WF, Petri nets, ECPs, transition heuristics	Fuzzy model	BPMN, WF, Petri nets
Supported platform	Standalone desktop version	Standalone desktop version	Standalone And Web based
Import type support	MXML, XES	CSV, XLS, MXML, XES, FXL	CSV, XLS, MXML, XES , FXL
Import log size capacity	unlimited	Up to 5 million events	unlimited
Filtering data	Yes	Yes	Yes
Process discovery	Yes	Yes	Yes
Conformance checking	Yes	No	Yes
Social network mining	Yes	No	Yes
Decision rule mining	Yes	No	Yes
Process visualization	Yes	Yes	Yes
Performance reporting	Yes	Yes	Yes
Discriminative rule mining	Yes	No	Yes
Trace clustering	Yes	No	Yes
Delta analysis	Yes	Yes	Yes

Another study in [22] proposed a model based on the PM4Py, a process mining framework for Python. The model has been wholly modified in which the user specifies a sporadic number of ranges for the process's activities and traces that the user needs to analyze; also, using an example from the real-world event log, this study used a method to reveal new insights into the process, The used technique in this study has limitation; it has not worked in real world scenarios where multi-level filtering can reveal a possible pattern of fraud or

non-compliance. Another study [23] used the Python PM4Py library to propose a process mining method. The proposed approach comprises data preparation, process discovery, analysis, and modification of the discovered process model. An algorithm for the automatic transformation of the process model into a verification model is presented in this paper; process model executions based on actual event logs are used to simulate the process model executions that are used in the process mining process, the main advantage of this study; providing the process mining results with automated analysis by executing the simulation and verification of process model.

An approach to process discovery for uncertain event data was the paper's goal of [24]; it has introduced a method for generating Uncertain Directly-Follows Graphs (UDFGs), directed graph-based models that combine information about the process's uncertainty, the limitation of this approach is the quantitative evaluation of the quality of the discovery algorithm, in addition to the process mining methods over uncertain logs. The proposed approach has been executed utilizing the PM4Py, the limitation of this works appears when evaluating the quality of the discovery algorithm, where it must define the metrics and measures over uncertain event logs and process models. This paper [25] presents the event logs of semi-automated production processes of e.GO is the subject of a comprehensive application and investigation of process mining techniques where HR is an active, indispensable part of the production processes. The paper used the Prom tool to implement the dotted chart and the PM4Py to identify the bottlenecks. There are challenges while prepressing the data such as improper logging and data quality Issues. Also the study addresses problem to uncover which activities cause a production delay. The study [26] proposed an investigation tool based on process mining using the PM4Py framework. The proposed approach aimed at discovering a process model for exploring and mining malignant authentication events across client accounts. Process mining depends on the concurrence of process modeling and data mining. It mines and extracts process models from the event logs of businesses. These event logs can be analyzed to identify conformance issues and workflow bottlenecks. The limitation of this paper when investigating the attacks, the difficulty is to analyze each authentication user and events triggering an alert. The paper [27] proposed an approach to discover process models incrementally. By adding trace by trace to an existing process model, the user can to incrementally discover a process model using the algorithm. Consequently, the process model under development gets steadily expanded. The proposed approach was implemented the extending PM4Py on dataset of road fine management process. The limitation of is the impact of ordering of traces incrementally on discovering process model, so it has been included as a future work to this research.

from discussion aforementioned of process mining using different fields event logs , and after searching the studies that had been applied process mining techniques in several event logs such as[x28-30x], it is observed the methods of analysis of business processes contains two main steps: discovering the process model and then applied the performance analysis and organizational analysis. In the step of discovering the process model, some papers applied one miner algorithm; others applied more than one miner algorithms so these papers chose the best resulted model based on four quality metrics (fitness, precision, generalization, and simplicity). This paper differs from other process mining papers; it is evaluated the resulted models from the miner algorithms based on five quality metrics not four, it is used the soundness metric to support the evaluation process.

2. THE PROPOSED METHODOLOGY

Discovering an accurate and structured process model is important point for follow-up steps to analyze and improve the business processes; accurate modeling of patients' careflows may help hospital management to identify the main problems as bottlenecks, deviations, etc. In this part, the methodology proposed in this study is introduced in Figure 3. The approach that's implemented in this study consists of three main steps, (1) data preparation, (2) model discovery, (3) conformance checking.

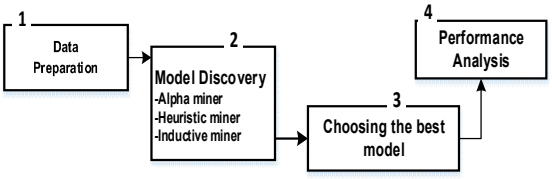


Figure 3: The proposed methodology

2.1 Data Preparation

The primary step in this phase is to extract events from the data set and determine the correct attributes for the event log. After that, the extracted log is cleaned of unwanted issues like irrelevant data, outlier data, missing data, and noise data, and lastly filtering the event log from the issues related to process instances such as incomplete instances or outrange time events.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.2 Model Discovering

The primary process mining function that automatically generates process models from the event logs is model discovery. The actual process as seen through actual process executions is reflected in the produced process model. Process mining views the discovery of a model as the foundation upon which to construct a model, such as a graphical structure based on actual events. A process model is created by process discovery using the event log. Several discovery model techniques support the process mining; the PM4Py framework includes three miner algorithms only (α, heuristic, inductive), the three discovery model algorithms used throughout the paper will be described in this section:

-**α-algorithm** [15] produced a Petri net representing the event log, the first process discovery algorithm. Most process discovery miners made enhancements to it. After analyzing the event log, the algorithm creates various task-dependent relationships. Relationships among tasks that are regarded as casual and descriptions of their order. But it is sensitive to noise and cannot discover the invisible and duplicated tasks. The fact that it does not guarantee soundness and does not take into account event frequencies is its primary limitation. As a result, this algorithm does not work well with real-world data.

-**Heuristic miner** addresses numerous issues with the α-algorithm [17]; Event logs are analyzed using the activities' dependency values to create a model. The heuristics miner builds model by creating a causal matrix and a dependency graph. The dependence (or causation) of events is depicted on a dependency graph. The dependency matrix and the length-one loop dependency are built to generate a dependency graph. Then, it constructs the straightforwardly follows network by utilizing the frequencies between the exercises. A heuristics net, or object containing the activities and their relationships, is the heuristics miner's output. After that, Petri net can be made from the heuristics net. It outperforms the α-miner takes into account event frequencies and ignores unusual behavior, single events, and short loops. It allows inferring a slight model from noisy real-world data.

-**Inductive miner** [16] incorporates two steps toward accomplishing its work. First and foremost, it makes a Stochastic Task Graph (SAG) from the event log, and by then, it synchronizes the designs of event log cases, to deliver the process model. It finds a principal split in the event log (there are various parts: sequential, parallel, concurrent and loop). The algorithm iterates over the sub-logs found by applying the split once the split has been found until a base case is found. Like heuristics miner, an inductive miner doesn't consider low-frequency events or isolated events/event loops. There are two ways to derive process models: Process Tree and the Petri Net it likewise, ensures sufficiency.

2.3 Choosing the Best Model

Four quality dimensions of the discovered process model were proposed in [31]: fitness and accuracy, generalization, and simplicity. Also, another dimension is essential to detect the usability of the model; Soundness. In the following line the fifth dimension will be explained:

a) Fitness indicates that the process model can display every trace, beginning to end, in logs. When a model can reproduce the observed behavior, it is considered fit. Alignments and token replay are the two primary methods for measuring fitness metrics. Finding the best alignment between the observed trace and the process model is necessary for the alignment-based method. As a result, it is guaranteed to return the model run closest to the trace [32]. Based on a specific process model, the token-based method calculates the fitness of an observation trace using four counters; produced tokens, consumed tokens, missing tokens, and remaining tokens [31].

b) Precision refers to avoiding under-fitting; it implies the model's capability, which disallows unlikable behavior. If all of the model's actions are observed, precision is determined by aligning the logs with the model [33].

c) Generalization or "avoiding over-fitting" refers to the process model's tendency to display other behaviors not present in the event logs displayed in the model. The method by which the "generalization" is measured is comparable to that used for precision measurements [33].

d) Soundness: Van der Aalst in [34] states three terms for any process model to be a sound (safeness, proper completion and option to complete). The safeness concept indicates that all of the events in the model are permissible. Proper completion says that no events can be executed when a process is complete. The option to complete means that the process model's final state can be reached from any state. No dead transition means that each transition in the model can be enabled. The inductive miner distinguishes from α- miner and heuristic miner that it guarantees the soundness.

e) Simplicity implies the most manageable model that best depicts the model's behavior. The simplicity is similar to Occam's Razor principle

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

which states that "one should not increase, beyond what is necessary, the number of entities required to explain anything". Because it will be challenging to comprehend and demonstrate complex process models, this has implications for human comprehension. The complexity of the model is measured by "how many arcs and nodes in the model" using the simplicity metric. The model itself, not the event log, is the sole focus of the concept of simplicity. Size, diameter, density, connectivity, node degree, separability, structuredness, sequentiality, depth, gateway mismatch, gateway heterogeneity, control-flow complexity (Cardoso), cyclicity, token splits, and soundness are some of the metrics introduced by [35] to gauge the model's simplicity or complexity. The simplicity cannot be quantified using a single perfect metric. Instead, each metric has advantages and disadvantages.

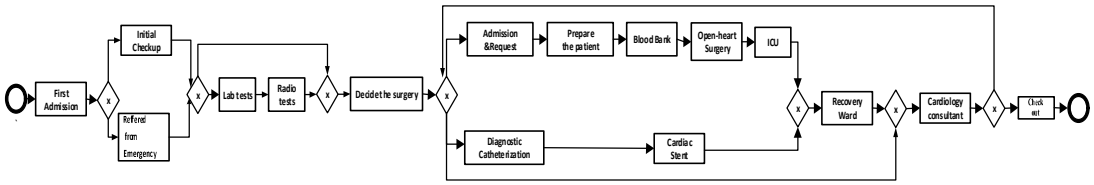


Figure 4: BPMN of Standard business model of the cardiac diseases made by Domain Knowledge.

2.4 Performance Analysis

Performance analysis techniques are conducted by statistical analysis or by conformance checking.

a) Conformance checking, or conformance analysis; conformance checking is measured as the fitness between the log and the predesigned blueprint model is measured, by aligning the process model and an event log. Managers and professionals tend to assume that their business works according to the predesigned process blueprint; most of the time, they expect only a few variations that deviate from the predesigned blueprint to exist. These deviations can cause unforeseen issues and delays in the process. As a result, it's critical to get rid of them. Conformance analysis is an incredible method for disposing of process deviations and ensuring the business runs according to the blueprint model. The conformance analysis in process mining can be used to analyze and improve business processes in different ways such as to detect if the real life events follow the rules and regulations of what the managers planned, and how the effects of not following the predesigned blueprints.

b) Statistics analysis: Statistical analysis is used by modern businesses to organize data better and make business decisions. These analyses help organizations reduce costs and improve the processes' efficiency. PM4Py includes many statistics to help decision-makers that are conducted on event logs. These statistics include the duration of the case process, the duration of each activity; the resources used with each activity, what kinds of activities take a lot of time, and how long waiting times are before they start.

3. DISCUSSION RESULTS ANALYSIS

3.1The Case Study

The paper used real log data; was collected from the cardiac diseases center in Egypt, where the unit received more than 1154 heart treatment cases and just 1028 that began and finished their treatment during the half year from January 2022 to July 2022. The log data is extracted from the hospital's information system. The managers and the professionals plan the predesigned process blueprint in the undertaken hospital, shown in Figure 4, the patient referring to the hospital based on advice from out doctor or as an emergency case from the emergency depart; the registration is the first activity must be taken, and when he finished his treatment, he must pay for the services and leave. Approximately 15 activities; the patient can perform after the registration. The patient executes all or some of them related to his state. Letters will be used to represent the activities' names for illustration propose, the activities and its symbols as follows: "First Admission" as A, "Referred from Emergency" as E, "Initial Checkup" as B, "Lab tests" as C, "Radio tests" as L, "Diagnostic Catheterization" as D, "Cardiac Stent" as F, "Decide the surgery" as G, "Open-heart Surgery" as H, "Blood Bank" as K, "Prepare the patient" as P, "Admission& Request" as R, "Cardiology consultant" as S, "ICU" as U. "Recovery Ward" as V, and "Check out" as Z. the undertaken hospital wants to hide the names of patients, doctors, nurses, and other work staff as privacy issues.

3.2 Data Preparation Results

The primary step in this phase is to extract the events from the data set then clean and prepare the extracted log, followed by the extraction of events from the data set and the selection of appropriate attributes for the event log, then the step of event data filtering for using with model discovery algorithms.

Figure 5: The sub steps of preprocessing stage.

1- **Gathering data.** The hospital of interest for the study uses the database system to manage its business and data. This dataset focuses on tables that describe the services introduced to the patient, like registration, labs, radiology, wards, surgery, and medication departments. SQL statements filter the final dataset depending on these tables from the hospital information system. As a result of a large number of activity sequences within low-level abstraction, the initial event log contains a lot of data, which generates unstructured model and difficult to understand. To assist process discovery methods in discovering a process model that stakeholders cannot understand, the paper applies grouping of low-level events to recognizable activities on a higher abstraction level based on domain knowledge. The processes of treating patients from January 2022 to July 2022 were included in the collected data. These processes included the number of service users, cases, 1028 providers, and 12477 events from a running process to record activities from 16 activities; the initial data was collected and stored in the .CSV file format. The event log is often converted in an IEEE standard as XES (eXtensible Event Stream), supported by most process mining tools.

2-Data cleaning. Deleting any duplicate records. For improved outcomes, it is essential to eliminate irrelevant data from the event logs, such as noise and outlier data. Likewise, there are 18 empty values or noise data in the event logs in columns "StartDate" and "FinishDate". The appropriate approach is to use the mean values of the same event from other cases to approximate the actual values in the empty cells.

3-Event Data Filtering; PM4Py has several of pre-built filter functions that simplify standard process mining filtering tasks. The paper used the following PM4py functions to filter the instance case:

a) *filter_start_activities():* the traces containing the specified activity as the start event are retained (or eliminated).

b) *filter_end_activities():* traces that contain the given activity as the final event are retained (or dropped).

c) *filter_event_attribute_values();* uses a specific set of values to filter event attributes.

d) *filter_trace_attribute_values();* Only the traces that have an attribute value for the provided attribute key and are listed in the collection of corresponding values are kept (or removed if retain is set to False).

e) *filter_variants();* preserves the traces that correspond to a particular order of activity execution.

f) *filter_time_range();* filters the event log based on a two-period time range.

Table 2 shows a statistic for the event log data before and after applying preprocessing step. There is removing of 165 traces after applying the preprocessing step. Also the number of variants or the execution sequence is decreased to 8 variants from 16 variants.

3.3 Model Discovery Phase Results

The management specialists and doctors in the hospital have arranged a predesigned blueprint model or reference process model for planning the patient's processes during his medical services journey when the patient is admitted to the hospital toward the finish of the journey. The knowledge-based workers in the hospital anticipate that their standard model is ideal for process planning, but frequently it is far from actual processes. Process mining views the discovery of a model as the foundation upon which to construct a model, such as a graphical structure based on actual events. Table 3 displays the event log most common and complete traces, which are 93% identical to the original event log and do not contain any anomalies. A process model is created by process discovery using the event log. The output of applying the discovery miner algorithms, as

Table 2. Event data before and after applying preprocessing step.

	No. of cases	No. of events	No. of variants	No. of processes
Before applying filters	1028	12477	16	16
After applying filters	863	8566	8	16

Table 3: Variants frequency

Variant	Freq %
A,B,C,L,S,G,R,P,K,H,U,V,S,Z	23%
A,E,C,L,G,D,F,V,S,Z	22%
A,E,C,L,G,S,Z	17%
A,B,C,L,G,D,F,V,S,Z	16%
A,E,G,S,Z	8.5%
A,E,C,L,F,S,G,R,Z	2.5%
A,S,G,R,P,K,H,U,V,S,Z	2.5%
A,E,F,L,C,S,G,R,Z	2.0%

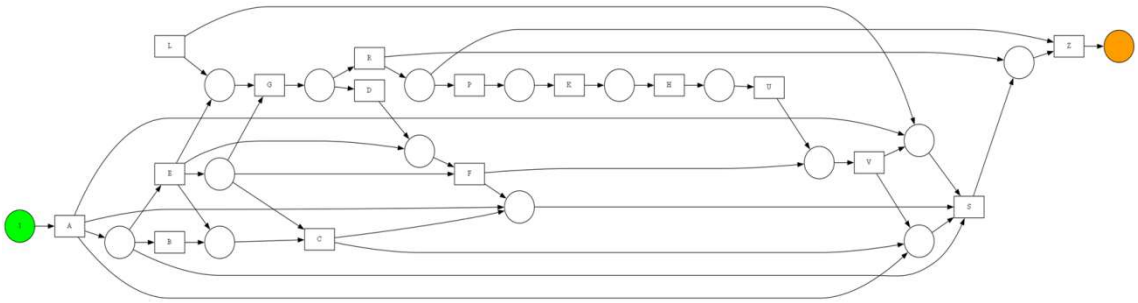


Figure 6: Petrinet discovered by the α-miner

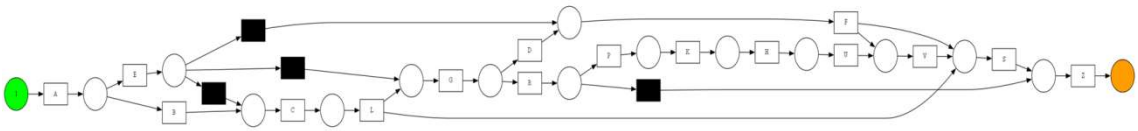


Figure 7: Petrinet discovered by the heuristic miner

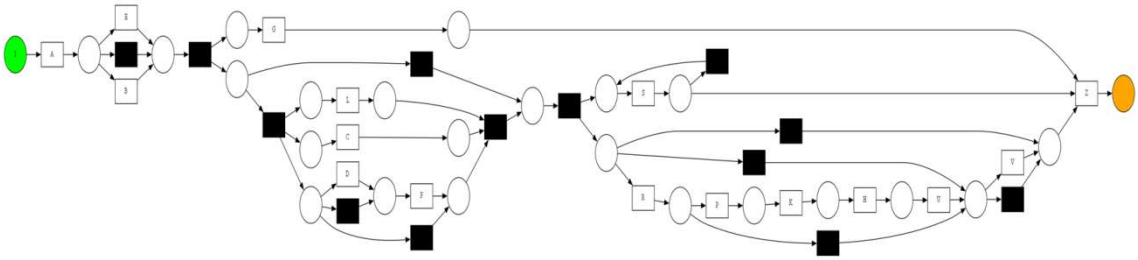


Figure 8: Petrinet discovered by the inductive miner

3.3 The Results of Choosing Between the Discovered Models

All through this section, the paper presents how it estimated the five quality measurements (fitness, precision, simplicity, generalization, and soundness) to assess the performance of the discovered process models and then choose the best model to apply the after-steps:

a. ***The fitness metric*** aims to determine the process model's acceptance of the log's behavior. The PM4Py framework proposes two different methods for computing replay fitness in light of token-based replay and alignments. Token-based replay returns a fitness value calculated in accordance with the scientific contribution [36] and the proportion of wholly fitted traces. A fitness value that is determined as the average of the fitness values of single traces is returned, as well as the proportion of fully fitted traces.

b.***The precision metric*** is how much behavior allowed by the discovered model is not allowed by the manually defined model. The PM4Py framework used two approaches for the measurement of precision: ETConformance (using token-based replay), proposed in paper [37], and the second measurement is Align-ETConformance (using alignments) that proposed in paper [30]. While token-based replay is quicker and uses heuristics, the outcome may not be exact. On the other hand, alignments are precise, work with any relaxed sound net, and are fast if the state space is enormous.

c. ***The generalization metric*** looks at how the process model and the log match up. Using the generalization method outlined in [38], the PM4Py platform computes the generalization between an event log and a Petri net model. Where token-based replay operation is carried out and the generalization is determined using the formula in equation 1:

$$\boldsymbol{Gen} = 1 - avg\_t\left(\sqrt{\frac{1.0}{\text{freq(t)}}}\right) \qquad (1)$$

Where (***t***) is the frequency of ***t*** following the replay, ***avg_t*** is the average of the inner value across all transitions. The resulting value is a in the range of 0 and 1.

d. ***The soundness metric*****:** the PM4Py framework checked the soundness using the WOFLAN approach discussed in [39], which can provide the final user with relevant statistics. When WOFLAN is applied to a Petri net that accepts it, it determines whether the net is sound.

e. ***The simplicity metric*** evaluates the intricacy of the model as "how several bends and hubs in the model". PM4Py framework measures the simplicity based on the study [40]; it measures the simplicity as the inverse arc degree from equation 2 below:

$$\boldsymbol{Simplicity} = \frac{1.0}{1.0 + \max(\text{mean\_degree} - \text{k},0)} \qquad (2)$$

The sum of the number of input arcs and output arcs is the average degree for a Petri net place or transition. The number must be at least two if each location has one input and output arc. ***k*** is a number among 0 and infinity. The resulting value of the simplicity is a number between 0 and 1.

Table 4. Event log Statistic before and after the preparation phase.

	Fitness		Precision	Generalization	Soundness	Simplicity
The miner	Token-based	Alignment				
α	0.73	-	-	0.95	No	0.5
Heuristic	0.92	0.62	0.80	0.90	No	0.74
Inductive	1.0	0.99	0.45	0.93	Ok	0.67

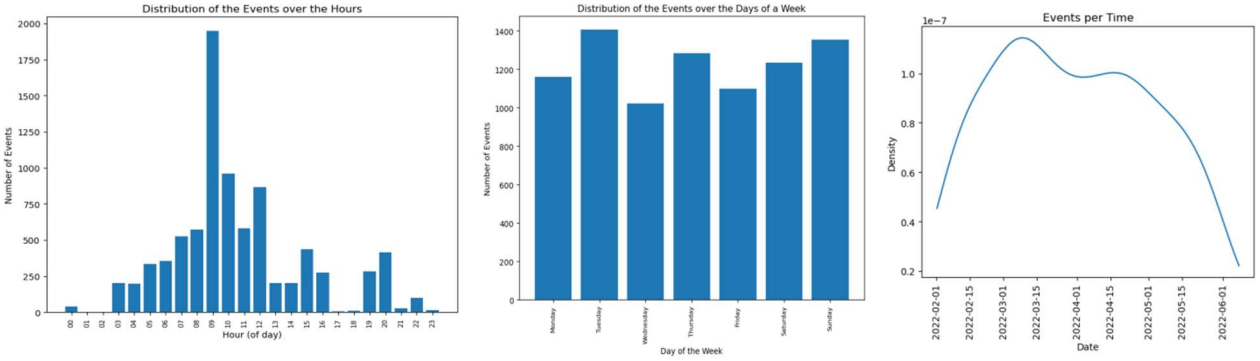
Table 4 shows the measurement results (fitness, precision, generalization, soundness, and simplicity) for the different process models generated from the three miners in Figures 6, 7, and 8. It is shown that the inductive miner produces a model with perfect fitness, followed by a model produced by the heuristic miner. However, the precision of the model produced by the inductive miner is lower than that of the heuristic miner. It is noted that the models produced from α-miner and heuristic miner are not sound; the precision of the model produced from α-miner is unbounded. Also, resulting model from the inductive miner is sound, but the simplicity of produced model from the heuristic miner is the highest. Considering this data, the ideal decision is to rely upon the inductive miner algorithm for modeling the processes for this situation study.

3.4 The results of the Performance Analysis and Evaluation

When the process mining method is implemented in the hospital, the performance analysis stage plays a significant role; when the model discovery phase is complete, it is put into action to use the results. There are many performance analysis indicators are crucial as measuring the patient's time spent in the hospital since admission, the time spent on each activity, the resources used for each activity, which activities occupy a lot of time with the patient, and the patient's lengthy waiting times for a specific operation. Additionally, it is essential to recognize the deviation in the discovered care-flow model from what was planned through the blueprint-designed model, and find the occasion of bottlenecks in executing specific activities. Two main analysis methods are used to measure the performance analysis in this case study:

a) Conformance checking analysis

A process model and an event log of the same process are compared using conformance checking techniques. The objective is to determine whether the event log follows the model and vice versa. Among procedural process models, PM4Py currently supports two techniques, i.e., token-based-replay and alignments. Conformance checking; result from mapping extracted event log from information system on Petri net of the predesigned blueprint process model "standard model", the standard model was handmade by hospital's domain experts. When complying between the reference model and the event logs



(a) Distribution of the events over the hours *(b) Distribution of the events over the days of a week* *(c) Distribution of the events over the four months*

Figure 9: Event distribution over time

from the hospital information system, it is found about a 88.5% deviation. It implies that the extracted event log has a deviation from the model with of 11.5%, there are many deviated activities when aligned the traces with the standard model as activities of "Prepare the patient", "Cardiology consultant", "Admission&Request", and "Diagnostic Catheterization". So it's recommended that the hospital management improve the predesigned blueprint model to include these traces.

b) Statistical analysis

The PM4Py framework can support the hospital knowledge domain using process mining with different statistic metrics that conducted on the event logs as follows:

-***Events distribution***; Observing how events are distributed over time provides valuable information about work shifts, working days, and the busiest and busiest times of the year.

Figure 9(a) shows the distribution of events over the day hours; the range of day hours from 12 am to 3 am is less busy than other hours. The Figure 9(b), it is noted there is equal distribution of the events over the days of the week, also the Figure 9(c) shows the distribution of events over the four months; the range from the start of 2-2022 to the start of 6-2022; is equal distribution of events.

-***Throughput Time*** recovering the list of cases' durations (in days) from throughput time is feasible. The average throughput time for each case is 13 days.

-***Case Arrival/Dispersion Ratio:*** The average distance between two consecutive cases' arrivals in the log can be retrieved as the case arrival ratio. The average case takes 1 hour and 50 minutes to arrive.

-***Sojourn Time;*** The average sojourn time statistic permits us to know for each activity, how much

time was spent executing the activity. The average of between the activity's start and completion timestamps is used to calculate this. The duration of the activities' sojourn time is shown in the table below.

Table 5. The sojourn time for the activities

Activity	Sojourn time(Hr)
First Admission	00:30
Referred from Emergency	01:30
Initial Checkup	00:28
Lab tests	02:59
Radio tests	02:31
Decide the surgery	15:40
Diagnostic Catheterization	00:38
Cardiac Stent	00:43
Recovery Ward	57:24
Cardiology consultant	00:47
Admission& Request	10:02
Prepare the patient	87:46
Blood Bank	24:00
Open-heart Surgery	05:41
ICU	45:23
Check out	00:10

-The length of stay (LOS); analysis of the correlation between the length of hospitalization spent time and the disease prognosis. The experts in the hospital and doctors classified the patients upon on their disease prognosis into three groups (open heart surgery, stent and catheterization, and medication without any surgery). It was found that the group of patients who did open heart surgery had the highest LOS than other groups with a value near 23 day, and patients who took a treatment trip without surgery with a value close to the 8 days. The second group of the patients who did Cardiac Stent and Diagnostic Catheterization with LOS ≈7 days.

-Analysis the bottlenecks in the activities*;* based on the previous statistics of throughput, sojourn times for the activities, also based on the distribution of event as in Figure 9, and handover network as shows in Figure 10; the day hours that are over loaded from 7 am to 4 pm every day, so these hours of day need more resources to serve in. Also there several activities have the risk of the bottlenecks such as "Cardiology consultant", " Recovery Ward", " Blood Bank", and " Decide the surgery ", so it`s recommended to the hospital management to increase the resources that serves these activities.

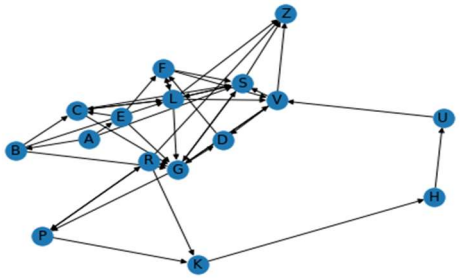


Figure 10: Resource handover network

analyze the hospital event log. The contributions of this work are as follows. First, we proposed methodology in four stages; first, extract the event log and prepare it for the following steps. Then discovered the model from the event log, three algorithm miners are applied (α, heuristic, and inductive) to produce three models. When choosing the best model from the three models based on evaluating the quality metrics (fitness, precision, generalization, soundness, and simplicity), the best model was chosen now is ready to apply the performance analysis to help decision-makers in the hospital with an extensive view and a several of insights to detect the inefficiency and lousy planning of processes. So they can develop solutions to existing problems. Second, the paper ensures to add a fifth metric; soundness, when evaluating the quality metric of extracted models. Also the paper concluded that the PM4Py framework has many power points as a process mining framework, as explained in section 1.2. The paper ensures and utilizes these features in the healthcare sector. It is noted that PM4Py has a wide range of ready statistics to be applied and enable the developer to customize any code easily to develop any new feature or algorithm. Also the PM4Py framework is efficient while dealing with large datasets as in this case study. But from the limitations of the PM4Py framework; it has only three miner algorithms (α, heuristic, and inductive). So as a future work it is good to embed others miner algorithms into the PM4Py framework. Based on the cardiac patients' event logs, a dataset from an Egyptian hospital was used in this paper. This work applied complete process mining method to analysis the healthcare careflows, but the method does not include trace clustering, and was choosing between three mining algorithms only. The limitation of this work; framework applied in this

paper; is applied to one hospital with one type of diseases "Cardiology diseases", hence it is necessary to apply this framework to different hospital event log to validate it with different type of patients. Future work will focus on improving the business process's performance by developing a decision mining model that can predict the next activities or traces. Also, it may be good to design a process-aware recommender system to advise the process mining users with best practices and alter them for any inefficient next activities.

REFERENCES

[1] Van Der AW, Adriansyah A, Medeiros, AKAD, Arcieri F, Baier T, Blickle T et al, "Process mining manifesto",International conference on business process management, Berlin (Heidelberg): Springer, 2011.

[2] Á Rebuge, DR Ferreira,"Business process analysis in healthcare environments: A methodology based on process mining", Inf. Syst. Elsevier, Vol. 37, No. 2, 2012, pp. 99–116.

[3] Mans, Ronny S., Wil MP Van der Aalst, and Rob JB Vanwersch, "Process mining in healthcare: evaluating and exploiting operational healthcare processes", Heidelberg: Springer International Publishing, 2015.

[4] Rashed A-HM, El-Attar NE, Abdelminaam DS, Abdelfatah M , "Analysis the patients' careflows using process mining", PLoS ONE, 2023, Vol. 18, No 2.

[5] Van Dongen, Boudewijn F., et al. "The ProM framework: A new era in process mining tool support", Applications and Theory of Petri Nets: 26th International Conference, ICATPN, Miami, USA, Vol. 26, 2005.

[6] Mans, Ronny, Wil MP van der Aalst, and H. M. W. Verbeek. "Supporting Process Mining Workflows with RapidProM.", BPM (Demos),Vol. 56 , 2014.

[7] La Rosa, Marcello, et al,"APROMORE: An advanced process model repository", Expert Systems with Applications, Vol.38, No.6, 2011, pp.7029-7040.

[8] Janssenswillen, Gert, et al,"bupaR: Enabling reproducible business process analysis." Knowledge-Based Systems, Vol.163, 2019, pp. 927-930.

[9] Berti, Alessandro, Sebastiaan J. Van Zelst, and Wil van der Aalst, "Process mining for python (PM4Py): bridging the gap between process-and data science", arXiv preprint arXiv, 2019, 1905.06169.

[10] Carmona Vargas, Josep, and Marc Solé, "PMLAB: an scripting environment for process mining.", Proceedings of the BPM Demo Sessions 2014: Co-located with the 12th International Conference on Business Process Management (BPM 2014) Eindhoven, The Netherlands, 2014.

[11] Disco. Available [online]: https://fluxicon.com/disco

[12] celonis. Available [online]: https://www.celonis.com/solutions/

[13] Uipath. Available [online]: https://www.uipath.com/

[14] processanalyzer. Available [online]: https://www.qpr.com/process-mining/qpr-processanalyzer

[15] Van der Aalst, Wil, Ton Weijters, and Laura Maruster, "Workflow mining: Discovering process models from event logs." IEEE transactions on knowledge and data engineering, Vol. 16, No.9, 2004, pp. 1128-1142.

[16] Leemans, Sander JJ, Dirk Fahland, and Wil MP Van Der Aalst,"Discovering block-structured process models from event logs-a constructive approach", Application and Theory of Petri Nets and Concurrency: 34th International Conference, PETRI NETS 2013, Milan, Italy, 2013, Proceedings 34.

[17] Weijters, A. J. M. M., Wil MP van Der Aalst, and AK Alves De Medeiros, "Process mining with the heuristics miner-algorithm.", Technische Universiteit Eindhoven, Tech. Rep. WP 166, 2017 pp. 1-34.

[18] Adriansyah, Arya, Natalia Sidorova, and Boudewijn F. van Dongen, "Cost-based fitness in conformance checking", Eleventh International Conference on Application of Concurrency to System Design. IEEE, 2011.

[19] Guzzo, Antonella, Antonino Rullo, and Eugenio Vocaturo,"Process mining applications in the healthcare domain: A comprehensive review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2022.

[20] Berti, Alessandro, and Wil van Der Aalst. "Extracting multiple viewpoint models from relational databases.", Data-Driven Process Discovery and Analysis: 8th IFIP WG 2.6 International Symposium, SIMPDA 2018, Seville, Spain, 2018, and 9th International Symposium, SIMPDA 2019, Bled, Slovenia, 2019, Revised Selected Papers 8. Springer International Publishing, 2020.

[21] Benevento, Elisabetta, et al, "Evaluating the effectiveness of interactive process discovery in healthcare: a case study", Business Process Management Workshops: BPM International Workshops, Vienna, Austria, 2019, Revised Selected Papers 17. Springer International Publishing, 2019.

[22] Vidgof, Maxim, et al., "Cherry-picking from spaghetti: Multi-range filtering of event logs", Enterprise, Business- Process and Information Systems Modeling: 21st International Conference, BPMDS 2020, 25th International Conference, EMMSAD 2020, Held at CAiSE 2020, Grenoble, France, 2020, Proceedings 21. Springer International Publishing, 2020.

[23] Zakarija, Ivona, Frano Škopljanac-Mačina, and Bruno Blašković, "Automated simulation and verification of process models discovered by process mining", Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije, Vol.61, No.2, 2020, pp.312-324.

[24] Pegoraro, Marco, Merih Seran Uysal, and Wil MP van der Aalst, "Discovering process models from uncertain event data", Business Process Management Workshops: BPM 2019 International Workshops, Vienna, Austria, 2019, Revised Selected Papers 17. Springer International Publishing, 2019.

[25] Uysal, Merih Seran, et al,"Process mining for production processes in the automotive industry", Industry Forum at BPM, Vol. 20, 2020.

[26] Lagraa, Sofiane, and Radu State, "Process mining-based approach for investigating malicious login events", NOMS IEEE/IFIP Network Operations and Management Symposium, IEEE, 2020.

[27] Schuster, Daniel, Sebastiaan J. van Zelst, and Wil MP van der Aalst, "Incremental discovery of hierarchical process models", Research Challenges in Information Science: 14th International Conference, RCIS 2020, Limassol, Cyprus, 2020, Proceedings 14. Springer International Publishing, 2020.

[28] Mans RS, Schonenberg MH, Song M, van der Aalst WM, Bakker PJ. Application of process mining in healthcare–a case study in a dutch hospital. In Biomedical Engineering Systems and Technologies: International Joint Conference, BIOSTEC 2008 Funchal, Madeira, Portugal, pp. 425-438

[29] Cho M, Song M, Yoo S. A systematic methodology for outpatient process analysis based on process mining. In Asia Pacific Business Process Management: Second Asia Pacific Conference, Brisbane, QLD, Australia, July 3-4, 2014. pp. 31-42.

[30] Zhou Z, Wang Y, Li L."Process mining based modeling and analysis of workflows in clinical care-a case study in a Chicago outpatient clinic". In Proceedings of the 11th IEEE international conference on networking, sensing and control 2014 Apr 7, pp. 590-595.

[31] Rozinat, Anne, and Wil MP Van der Aalst. "Conformance checking of processes based on monitoring real behavior", Information Systems, Vol.33, No.1, 2008, pp. 64-95.

[32] Adriansyah, Arya, Boudewijn F. van Dongen, and Wil MP van der Aalst, "Conformance checking using cost- based fitness analysis", ieee 15th international enterprise distributed object computing conference. IEEE, 2011.

[33] Adriansyah, Arya, et al,"Measuring precision of modeled behavior", Information systems and e-Business Management,Vol.13, No.1, 2015,pp. 37-67

[34] Van Der Aalst, Wil,"Process mining: data science in action", Heidelberg: Springer, Vol. 2, 2016.

[35] Mendling, Jan, Gustaf Neumann, and Wil Van Der Aalst. "Understanding the occurrence of errors in process models based on metrics", Lecture notes in computer science 4803, 2007.

[36] Berti, Alessandro, and Wil MP van der Aalst. "Reviving Token-based Replay: Increasing Speed While Improving Diagnostics." ATAED@ Petri Nets/ACSD. 2019.

[37] Muñoz-Gama, Jorge, and Josep Carmona,"A fresh look at precision in process conformance", International Conference on Business Process Management, Springer, Berlin, Heidelberg, 2010.

[38] Buijs, Joos CAM, Boudewijn F. van Dongen, and Wil MP van der Aalst, "Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity", International Journal of Cooperative Information Systems, 2014.

[39] Verbeek, Henricus Marinus Wilhelmus, "Verification of WF-nets.", Ph.D. Thesis, Technische Universiteit Eindhoven, 2004.

[40] Blum, Fabian Rojas, "Metrics in process discovery", Technical Report TR/DCC-2015-6, Computer Science Department, University of Chile, 2015.