http://dx.doi.org/10.5755/j02.eie.31126

# Creating a Data Generator and Implementing Algorithms in Process Analysis

**Cigdem Bakir[1, *], Mecit Yuzkat[1, 2], Fatih Karabiber[1]**
[1]*Department of Computer Engineering, Yildiz Technical University,*
*Davutpasa Campus, 34220 Istanbul, Turkey*
[2]*Department of Software Engineering, Faculty of Engineering and Architecture,*
*Mus Alpaslan University,*
*Mus, Turkey*
*cigdem.bakir@dpu.edu.tr*

*Abstract*—**Process mining is a new field of work that aims to meet the need of the business world to improve efficiency and productivity. This field focuses on analysing, discovering, managing, and improving business processes. Process mining uses event logs as a resource and works on this resource. Hence, the system is developed by analysing the event logs, including each step in the process model. Our study is made up of two significant stages: a data generator for processes and algorithms applied for discovering the created processes. In the first stage, the aim was to develop a simulator with the ability to generate data that could help process modelling and development. Within the framework of this study, a system was created that could work with various process models and extract meaningful information from these models. More productive and efficient processes can be developed as a result of his system. The simulator consists of three modules. The first module is the part where users create a process model. In this module, the user can create his own business process model in the system's interface or select from other registered models. In the second module, team-based data are simulated through these process models. These generated data are used in the third module, called "analysis", and meaningful information is extracted. In conclusion, the process can be improved considering the information about time, resource, and cost in the generated data. At the second stage, processes were discovered using alpha, heuristic, and genetic algorithms, which are process mining discovery algorithms and synthetic and real event logs. The discovered processes were demonstrated with Petri nets, and the algorithms' performances were compared using the fitness function, accuracy rates, and running times. In our study, the heuristic algorithm is more successful because it improves the noise in the data and incomplete processes, which are the disadvantages of the alpha algorithm. However, the genetic algorithm yielded more successful results than the alpha and heuristic algorithms due to its genetic operators.**

*Index Terms*—**Alpha algorithm; Data generator; Genetic algorithm; Heuristic algorithm; Process mining; Petri nets.**

## I. INTRODUCTION

In our age, information processing systems play a crucial role in the business world. It is of great significance for businesses to make quick decisions and improve efficiency. At this stage, results such as the performance of the employees and the cost of the steps in the process are obtained by reviewing the business processes. By reviewing the steps in the processes in detail, information such as the completion time and cost of the steps is reached, the problematic parts of the process are detected, and the efficiency of this process can be enhanced [1], [2]. Process mining is a comparatively new field of research that includes process modelling and process analysis, on the one hand, and business intelligence and data mining, on the other. The most important purpose of process mining is to extract business processes from event logs recorded in today's information systems and to discover, control, and improve them. Process discovery, model and log comparison and deviation control, social network/organizational mining, automation of simulation models, model development, model repair, status estimation, and history-based recommendations are among the areas covered by process mining [3]. Today, the recording of event logs is one of the most significant parts of business processes. In almost every software used in corporate organizations, event logs are recorded, which helps them be analysed and observed. Analyses can be performed on these recorded event logs regarding the process.

Our study aimed to create a data generator that would perform process modelling, simulation, and analysis. By simulating processes that had not been used or had been thought to be used before, it was aimed at extracting meaningful information such as what kind of results the process would produce, how much time it could be completed in, and what kind of results the performance distributions between teams would produce in the case of dividing the process into teams in parallel. There are three main modules in the system developed in this study. The initial step of these modules is to define the process in the system. In the second module, the simulation module, log files complying with the previously designed process model are created. Finally, in the analysis module, meaningful information regarding the process, such as individual and team performance, is extracted from the generated event logs. It was intended to discover processes using synthetic event logs created using the data generator and real event logs available in today's information systems. Several

Manuscript received 9 April, 2022; accepted 12 August, 2022.

process mining discovery algorithms, such as alpha, heuristic, and genetic algorithms, were used to discover the processes. The success rates of the algorithms were compared according to the fitness function, accuracy rates, and running times. Briefly, in this study, the aim was to develop a process mining system that could both generate process data and analyse and discover processes for synthetic or real process data. Scientific contribution of our study includes the production of synthetic data with different situations and scenarios for a process model, the visual examination of the flows of the processes, the implementation of widely used process mining algorithms, the development of a new software platform that can be used in the field of process mining.

The remainder of the article is organized as follows. Related and similar studies will be discussed in Section II. Process mining is presented in Section III. The proposed model is discussed in Section IV and its application with experimental study is detailed in Section V. The evaluation and conclusions are detailed in Section VI.

## II. RELATED WORKS

In recent years, the number of studies on process mining has increased, and the importance of this subject is increasing with each passing day. A brief summary of previous similar studies on process mining is given below.

Sun, Liu, Qi, Du, Ren, and Liu [1] proposed to discover multiple-concurrency short-loop structures via Petri nets. This study consisted of the artificial model and the actual model. The activities are matched with a triangular two-degree loop or a quadrilateral two-degree loop. Then, two types of short-loop structures are identified from the incomplete logs. But this work has limitations in that it does not consider the impact of duplicate activities or invisible activities and needs to study more complex structures.

One of the applications of process mining in the healthcare field has been process mining analysis with event logs obtained from four different hospitals in Australia, and vital information about patients was reached. The results of the studies were tried to be given comparatively. However, one of the challenges of process mining is that it is quite difficult to analyse, find, and improve data in more than two multiple data sets. Since all information about patients is present together, comparative analysis of big data, visualization of data, and division of data bring about some difficulties. In this study, these shortcomings were tried to be eliminated, although partially [4]. In other study, a methodology for the application of process mining in healthcare entitled “Process Mining Project Methodology in Healthcare” was developed (PM2 HC) [5]. Moreover, it was developed over a series of steps involving bibliographical reviews on the methodologies of application of process mining in the general and on applications of process mining in health case studies. However, these works were not achieved, as categorical information about patients was not taken into account.

Yasmin, Bemthuis, Elhagaly, Wijnhoven, and Bukhsh [6] developed ProM to collect all process mining algorithms on a single platform. With this tool, it is possible to discover processes, measure processes’ fitness, and improve processes. ProM offers a platform that shows that many process mining algorithms such as the genetic, fuzzy, heuristic, and alpha algorithms function in connection with each other. It was tried analyse data records of 569 events and 299 cases obtained from the records of students at Taylan University [7]. This study aims to rearrange the extraction of information from sequential events. However, the desired success could not be fully achieved due to the noisy data and incomplete cases.

Koonce and Tsai [8] tried to create software for business management in production using a genetic algorithm. The purpose of this study is to specify the optimum business process by finding the shortest processing time from various operations that can be carried out against various activities. In the study, different operations were created for different processing times by setting various rules. However, the performance problem encountered by the genetic algorithm could not be fully addressed in this study. Weijters, van der Aalst, and Alves performed process discovery and analysis from 12,000 different event logs by employing a heuristic algorithm in process mining [9]. In the study, they worked on noisy data and processes that were not clearly expressed.

In another reviewed study, stroke patient care applications with process mining techniques were developed [10]. In this study, they used process mining to discover the treatment processes of stroke patients in different hospitals. They examined the events that the patients were associated with and conducted analyses. This study indicates that process mining techniques can be used successfully in clinical data from different hospitals and different patients. However, real-world healthcare problems are not practical or applicable, and necessary healthcare specificities are taken into account when developing new process mining techniques. Goedertier, De Weerdt, Martens, Vanthienen, and Baesens [11] discovered processes from the event logs of information systems in the telecom sector. Owing to the analyses of the event logs, it was concluded that processes could be created automatically and that meaningful data could be extracted.

Liu, Zhang, Li, and Jiao [12] integrated related processes using event graphics for operational decision support systems. They conducted analyses from the event logs generated for different scenarios. Their results demonstrate that event log analyses play a significant role in process management and development.

Dogan, Bayo-Monton, Fernandez-Llatas, and Oztaysi [13] aimed to investigate customer behaviors about gender by using their paths with process mining. They can determine the gender of customers if they go to the men’s or women’s bathroom on the behaviors of male and female customers. Furthermore, the study shows that the process mining technique is a viable way to analyse customer behavior using bluetooth-based technology. However, personalized guidance can be developed for customers. Customer-specific suggestions or future state prediction may be considered. And this study was performed on a single data set.

Nafasa, Waspada, Bahtiar, and Wibowo [14] developed the application of the alpha algorithm in the process mining application for online Learning Activities Based on Moodle event log data. The experimental conclusion proved that the implementation of the alpha algorithm can work correctly on the Moodle event log data. However, this study is needed for data analysis and pre-processing that is provided in Moodle event log data easily in integrated environment.

In our study, a new approach is presented to obtain information about activities and relations between them according to Petri nets obtained from event logs. Also, process mining discovery algorithms were used in addition to traditional algorithms used to measure the suitability of business processes. Since process mining is a new field in the world and there are few scientific studies in this field, a process mining system that performs process modelling, simulator, analysis, and discovery has been developed in this study to eliminate this deficiency.

The model we created has a great importance in the field of process mining in terms of having a unique data generator compared to previous studies. With the model we developed, unlike synthetic data used in the literature [15]–[17], it can perform a comprehensive analysis for the data sets and real data we have developed with its own data generator. Furthermore, our system analyses incomplete and noisy data sets apart from the completed data sets.

Augusto, Mendling, Vidgof, and Wurm [18] provide two major contributions to the measurement of process complexity and to the evaluation of process mining algorithms. First, they analysed existing measures for process complexity based on event logs. Second, they evaluated the identified set of process complexity measures, including our novel measures, using a benchmark collection of event logs and their corresponding automatically discovered process models [18].

In another article in [19], the authors focus on the study of automated process discovery using the Inductive visual Miner (IvM) and Directly Follows visual Miner (DFvM) algorithms to produce a valid process model for educational process mining to understand and predict the learning behavior of students. However, the performance of process mining algorithms in improving quality education and other benchmarking data sets obtained from alternate data should be explored [19].

Fauzi and Andreswari [20] used process mining techniques to extract streams from the event log. In the study, the discovery process was carried out to determine the flow carried out by the programmers in the software development team20. The authors only mentioned the software team discovery processes in their studies. In our study, many discovery processes such as health, education, and business processes were analysed. This system is aimed to be a reference for future studies in this field in all the world. In addition, the production of synthetic data with different situations and scenarios for an exemplary process model, the visual examination of processes’ flows, the implementation of commonly used process mining algorithms and the development of a software platform that can be used in the field of process mining constitute the original contribution of our study.

## III. PROCESS MINING

Process mining is a new field of study that aims to meet the need of the business world to increase productivity and efficiency. This field carries out studies on the analysis, discovery, management, and improvement of work processes. Process mining uses event logs as a resource and studies on this resource. Thus, the system is developed by analysing event logs that comprise every step of the process model [21].

Process mining is a new scientific discipline that analyses and models processes, utilizes data mining methods, and has become popular in recent years. This discipline aims to determine processes using different algorithms and to calculate the performance of processes. One of the reasons for the popularity of process mining nowadays is finding useful information in big data. It also contributes to the development of processes in the business world [22]. The purpose of this field is to determine processes and to calculate their performance employing different algorithms. The widespread use of big data in information technologies is bringing some problems with it to the entities. Process mining enables effective and efficient use of data by reanalysing the processes, discovery, modelling, and improvement of processes by employing process mining techniques. Thus, the use of process mining techniques by large entities becomes mandatory in a competitive business environment.

The general structure of the process mining is shown in Fig. 1. Each process in the real world is kept in event logs as a sequence of events using software systems. Thereafter, information in such event logs is analysed by discovery, fitness check, and by modelling with improvement, which are stages of process mining. Thus, meaningful data are offered to people and systems.

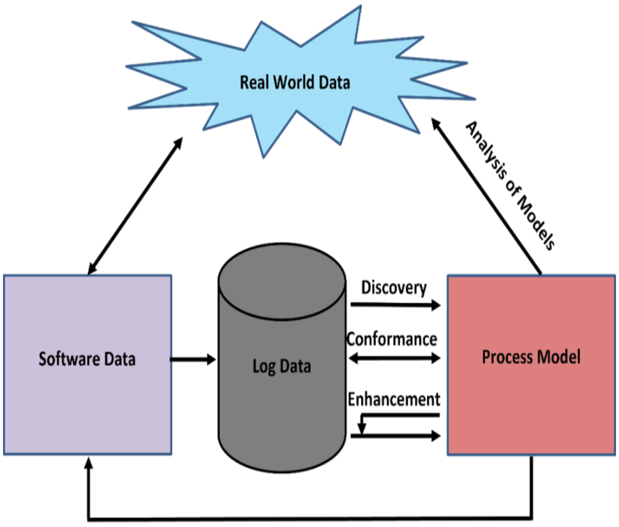


Fig. 1. General structure of process mining.

In our study, although it contains a structure similar to the example process mining structure seen in Fig. 1, it provides the production of synthetic data with different situations and scenarios, the visual examination of the flows of the processes, the implementation of commonly used process mining algorithms, and the development of a software platform that can be used in the field of process mining.

An event is any recorded movement. An event can be started and completed, or it can be canceled after it has been started. The event log is the collection of events accepted as input by a particular source in the process mining. Sequential states with a start and end point in a given time period represent an event log. An example order event log for process mining is created in Fig. 2. Detailed information about the order is given by showing it with a Petri net.

Here, order taking, bill payment, and delivery to the customer are carried out according to a certain time. The event log table for this business process example is shown in Table I.

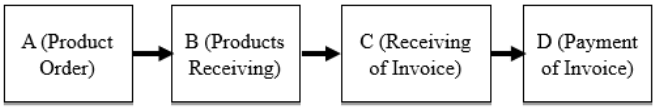


Fig. 2. Sample order event log [23].

TABLE I. EVENT LOG TABLE OF ORDER SAMPLE.

| State ID | Event ID | Time | Activity | Person |
|---|---|---|---|---|
| 1 | 1000 | 01.01.2013 | A (Product Order) | Peter |
| 1 | 1001 | 10.01.2013 | B (Products Receiving) | Micheal |
| 1 | 1002 | 13.01.2013 | C (Receiving Invoice) | Frank |
| 1 | 1003 | 20.01.2013 | D (Payment of Invoice) | Tanja |
| 2 | 1004 | 02.01.2013 | A (Product Order) | Peter |
| 2 | 1005 | 03.02.2013 | B (Products Receiving) | Micheal |
| … | … | … | … | … |

## IV. OUR MODEL

As seen in Fig. 3, our study consists of two main stages, namely, data generator and process mining discovery algorithms. In the first stage, there is a structure that can generate data with an automatic data generator. It consists of three submodules: process model creation, simulation, and analysis. In the second main stage, the performance of the processes in terms of accuracy, fitness, and time is evaluated by applying process mining discovery algorithms (alpha, heuristic, and genetic) for event logs, real data, and other synthetic data of the data produced by the data generator. Therefore, meaningful data are presented to individuals and systems.

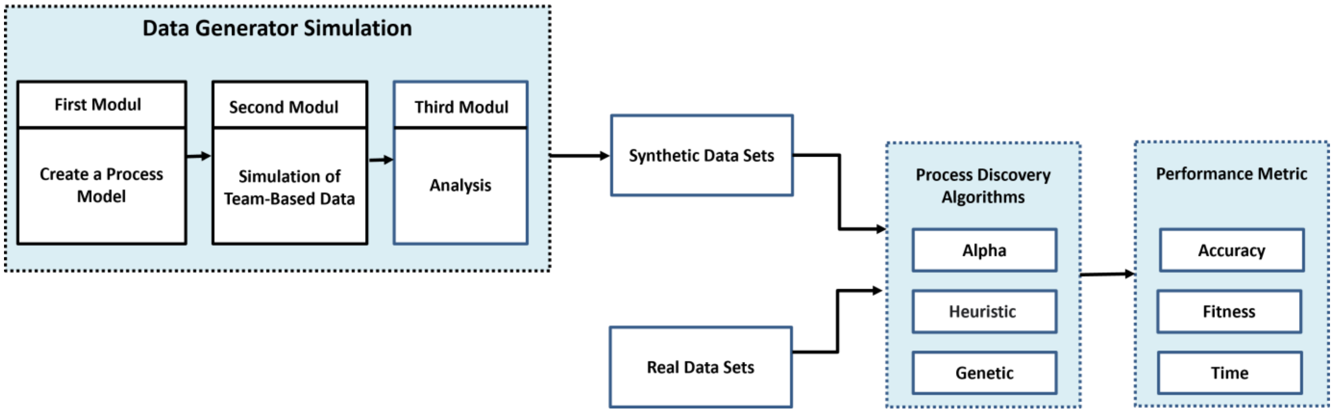


Fig. 3. The steps of our model.

### A. *Creation of the Data Generator*

In this study, a suitable data generator for real life was designed to develop process mining algorithms. This design consists of three main modules: process creation, data generator, and data analysis.

### B. *Process Creation*

Any process model can be reviewed as an example while creating a process model. The credit card application process, the steps of which are shown in Fig. 4, was examined.

In the created system, processes with start and end points are entered, as stated in Fig. 5. The steps in the process entered in this study can be created as many times as desired and can form cycles within themselves in addition to branching. The user can give percentiles for the steps subject to branching, enabling the generated data to produce data close to the rates desired by the user. After determining the path of the process, the user sets the cost and time level for each step. Meanwhile, the positions of the person who performs the steps can be specified. By assigning the specified positions to the selected steps, it will be possible to analyse the people in that position in the data to be generated later.

There is also a section that allows the user to use the existing processes before moving to the data generation stage to make this step easier for the user and create different but similar processes and see the differences between them. When the user creates a process, this process is added to the existing process page in the system and is ready to generate data.

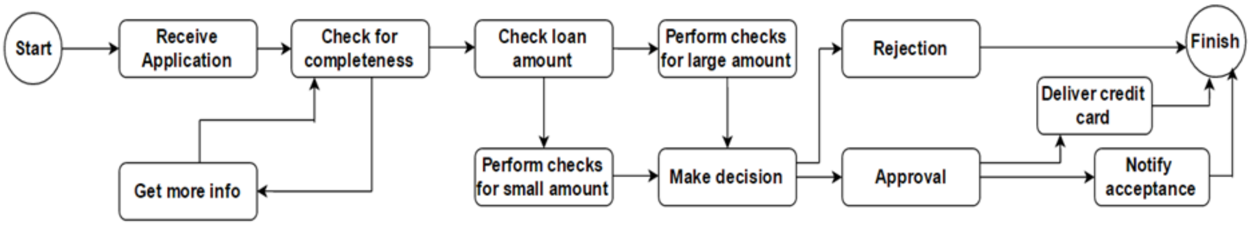
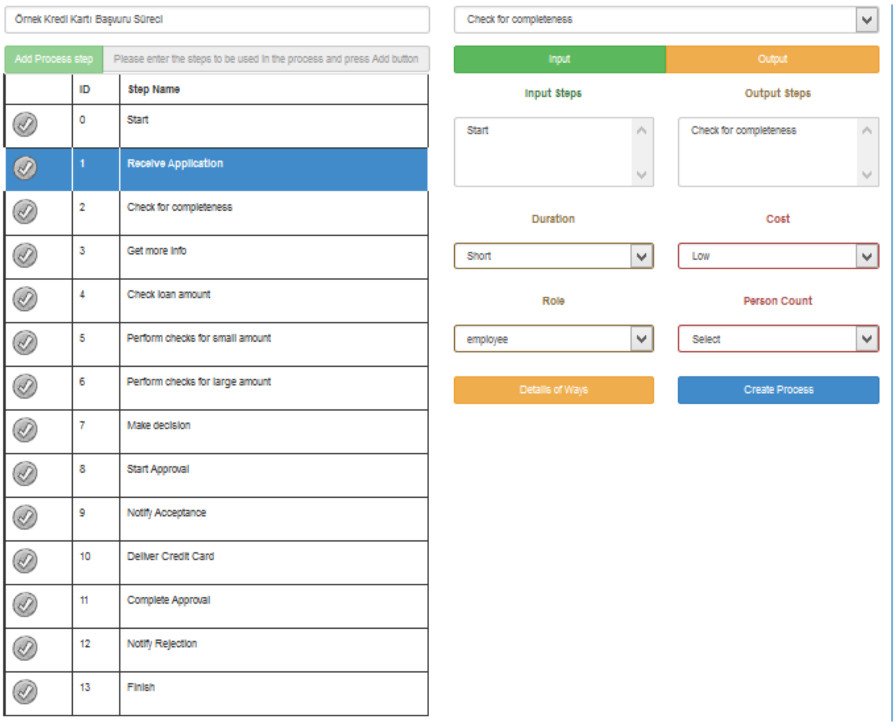


Fig. 4. Example credit card application process.



Örnek Kredi Kartı Başvuru Süreci

Add Process step | Please enter the steps to be used in the process and press Add button

| | ID | Step Name |
|---|---|---|
| ✓ | 0 | Start |
| ✓ | 1 | Receive Application |
| ✓ | 2 | Check for completeness |
| ✓ | 3 | Get more info |
| ✓ | 4 | Check loan amount |
| ✓ | 5 | Perform checks for small amount |
| ✓ | 6 | Perform checks for large amount |
| ✓ | 7 | Make decision |
| ✓ | 8 | Start Approval |
| ✓ | 9 | Notify Acceptance |
| ✓ | 10 | Deliver Credit Card |
| ✓ | 11 | Complete Approval |
| ✓ | 12 | Notify Rejection |
| ✓ | 13 | Finish |

Check for completeness

Input | Output

Input Steps: Start

Output Steps: Check for completeness

Duration: Short

Cost: Low

Role: employee

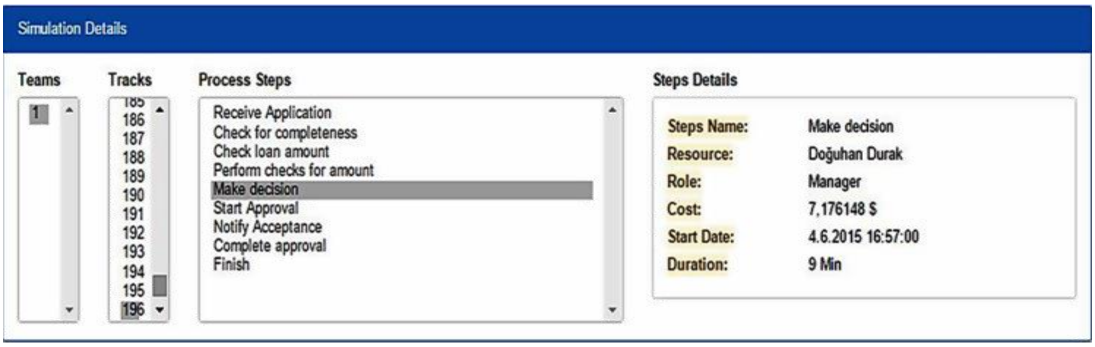
Person Count: Select

Fig. 5. Process creation screen.

### C. Simulation

The simulation module is the section where event logs are created. While event logs are created, data are generated for the process model selected by the user. On the simulation screen, the user is asked to define some parameters. These parameters include the starting date of the process, the starting time of the workday, the ending time of the workday, and the area where the process will be applied during the day. The created system is flexible enough to allow the user to create a model as he desires, and Track Count and Team Count can be defined as desired. When the team count is entered, event logs related to the process are generated in parallel, assuming that each team starts the process on the same date and at the same time in the data generated with respect to these teams. Since the need for generating data in a different process completion count and team count in a certain process is inevitable for the user, he/she is requested to name each simulation he/she produces. Accordingly, the user was given the opportunity to decide on the optimum team count by performing different analyses for different team counts for the process he/she chose.

An interface was designed, as in Fig. 6, to examine the data generated after the simulation.



Simulation Details

Teams: 1

Tracks: 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196

Process Steps:
Receive Application
Check for completeness
Check loan amount
Perform checks for amount
Make decision
Start Approval
Notify Acceptance
Complete approval
Finish

Steps Details

| | |
|---|---|
| Steps Name: | Make decision |
| Resource: | Doğuhan Durak |
| Role: | Manager |
| Cost: | 7,176148 $ |
| Start Date: | 4.6.2015 16:57:00 |
| Duration: | 9 Min |

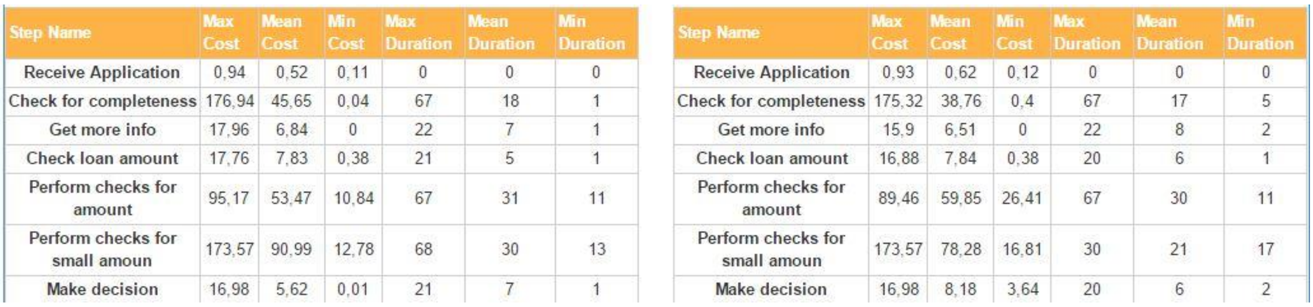
Fig. 6. Information on the simulation results.

In the system we have developed, the user can observe which team continues from which steps on which track in the generated data. There is information such as the start time of the selected step, which employee performed it, and duration and cost. Duration expresses the burden of the operation performed in the selected step in terms of time, while cost expresses the financial burden of the operation in the system. By extracting the event logs of the data he/she generated in XML format, the user can reach meaningful information about the fitness, accuracy, and completion times of the processes with process mining discovery algorithms in the discovery module of the system.

### D. *Analysis*

In the analysis module, the objective is to extract meaningful information from the event logs created in the simulation module.

On this screen, previous process simulations are listed, and the user selects the process to be analysed. After the selection of the model, the summary information is shown to the user, as given in Fig. 7.



| Step Name | Max Cost | Mean Cost | Min Cost | Max Duration | Mean Duration | Min Duration |
|---|---|---|---|---|---|---|
| Receive Application | 0,94 | 0,52 | 0,11 | 0 | 0 | 0 |
| Check for completeness | 176,94 | 45,65 | 0,04 | 67 | 18 | 1 |
| Get more info | 17,96 | 6,84 | 0 | 22 | 7 | 1 |
| Check loan amount | 17,76 | 7,83 | 0,38 | 21 | 5 | 1 |
| Perform checks for amount | 95,17 | 53,47 | 10,84 | 67 | 31 | 11 |
| Perform checks for small amoun | 173,57 | 90,99 | 12,78 | 68 | 30 | 13 |
| Make decision | 16,98 | 5,62 | 0,01 | 21 | 7 | 1 |

| Step Name | Max Cost | Mean Cost | Min Cost | Max Duration | Mean Duration | Min Duration |
|---|---|---|---|---|---|---|
| Receive Application | 0,93 | 0,62 | 0,12 | 0 | 0 | 0 |
| Check for completeness | 175,32 | 38,76 | 0,4 | 67 | 17 | 5 |
| Get more info | 15,9 | 6,51 | 0 | 22 | 8 | 2 |
| Check loan amount | 16,88 | 7,84 | 0,38 | 20 | 6 | 1 |
| Perform checks for amount | 89,46 | 59,85 | 26,41 | 67 | 30 | 11 |
| Perform checks for small amoun | 173,57 | 78,28 | 16,81 | 30 | 21 | 17 |
| Make decision | 16,98 | 8,18 | 3,64 | 20 | 6 | 2 |

Fig. 7. Summary information extracted from the analysis (left general and right team).

While the left table contains general information about the process model, the right table displays information about the selected team of the process model. The data from the team selected in the summary analysis are compared to the average data from the process, and a general performance comparison can be made between the teams.

### E. *Team Comparison*

As in Fig. 8 of the steps, graphs showing the time and cost distributions of general, team-based, and selected tracks are produced. Due to these graphics produced, the steps with the highest cost and time in the process can be determined and changes to be made in terms of improving the process can be determined. However, for tracks between teams, it is possible to identify the teams that finished in the shortest time and the longest time, and to see which team is working with which performance. It is possible to reach information about which team completed the process in which track and how long it took.

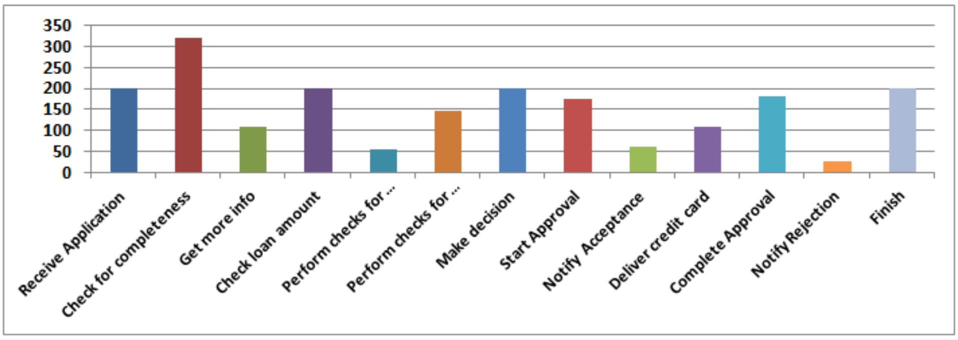


Fig. 8. Step frequency analysis data.

Thanks to the tool comparison structure, meaningful analyses can be made from the data produced in accordance with the multi-thread architecture. As a result of the flexibility of the system, comparisons can be made by generating data for the desired simulation at different times and for different tool numbers.

### F. *Process Mining Discovery Algorithms*

Every process in the real world is recorded in event logs as sequential events through software systems. Afterward, the information in these event logs is modeled and analysed with discovery, fitness checking, and development, which are process mining stages. Therefore, meaningful data are presented to individuals and systems. Many different algorithm approaches are used in process mining. These algorithms are Alpha, Multi-Phase Miner algorithm, Heuristic algorithm, Fuzzy Miner, and Genetic algorithm. In this study, performance comparisons were made using alpha and heuristic algorithms. As seen in Fig. 9, the most important stage of process mining is the discovery of processes from event logs and model creation. Various process mining algorithms, especially alpha and heuristic algorithms, can be used during process discovery.

Figure 10 shows the fitness checking stage of the processes. At this stage, processes are monitored, and the fitness of existing processes and processes obtained from the event logs is checked. Thus, it is used to detect differences between the event log and the model to determine their positions in the process and to measure their severity.

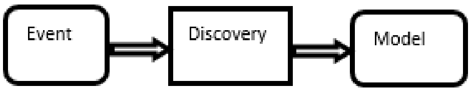


Fig. 9. Discovery phase of process mining.

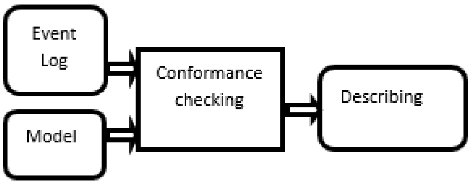


Fig. 10. Fitness checking phase of process mining.

Figure 11 shows the development stage of the processes. The aim of this stage is to determine and change the deficiencies of the existing model and to enable its redevelopment. For example, activities that are costly in terms of time are determined by reviewing the event logs, and obstructions between these activities are identified and remodeled.

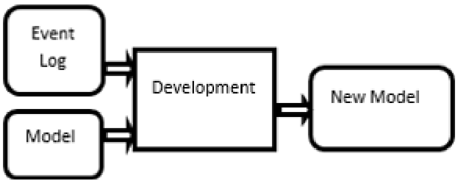


Fig. 11. Developing phase of process mining.

Today, the recording of event logs is one of the most significant parts of business processes. In almost all software used in corporate organizations, event logs are recorded, which helps them to be analysed and observed.

Analyses can be carried out on these recorded event logs regarding the process. The event log of the credit application process data exemplified for the process mining is given in Table II.

TABLE II. EVENT LOGS OF SAMPLE CREDIT APPLICATION PROCESS DATA.

| Status ID | Event ID | Time | Activity | Person |
|---|---|---|---|---|
| 1 | 1000 | 01.01.2019 | A | Mecit |
| 1 | 1001 | 10.01.2020 | B | Ali |
| 1 | 1002 | 13.06.2020 | C | Çiğdem |
| 1 | 1003 | 20.01.2021 | D | Furkan |
| 1 | 1004 | 02.02.2021 | G | Mecit |
| 2 | 1005 | 03.05.2021 | A | Çiğdem |
| …… | …… | …… | …… | …… |

### G. Alpha Algorithm

The purpose of the alpha algorithm is to form a work flow net (Petri net) from event logs. The steps of the alpha algorithm (α(W) = (**P**W, **T**W, **F**W)) are given below [24], [25]. The **P**W shown in this algorithm represents places of established (work flow steps) Petri net, **T**W transitions, and **F**W indicates combined status of all locations and transitions from input place to output place, i.e., established Petri net (work flow network). **W** refers to the log of a workflow on **T**. **T** stands for transitions. Using **T**W together refers to the transition of the workflow.

Algorithm 1. Alpha algorithm steps.

**Step 1:** $T_W$ all transitions of event logs are established.

$$T_W = \{t \in T \mid \exists_{\sigma \in W}\ t \in \sigma\},$$

**Step 2:** $T_I$ *input places in source node are formed*.

$$T_I = \{t \in T \mid \exists_{\sigma \in W}\ t = \text{first}(\sigma)\},$$

**Step 3:** $T_O$ output transitions of recipient are formed.

$$T_O = \{t \in T \mid \exists_{\sigma \in W}\ t = \text{last}(\sigma)\},$$

**Step 4:** Discovered places in Petri nets are defined ( $X_W$ and $Y_W$ ).

$$X_W = \{(A,B) \mid A \subseteq T_W \wedge B \subseteq T_W \wedge \forall_{a \in A} \forall_{b \in B}\ a \rightarrow W^{b} \wedge \forall_{a1,a2 \in A}\ a1\# \forall_{b1,b2 \in B}\ b1\# W^{b2}\},$$

**Step 5:** Discovered places in Petri nets are defined ($X_W$ ve $Y_W$).

$$Y_W \ \{(A,B) \in X_W \mid \forall_{(A',B') \in X_W}\ A \subseteq A' \wedge B \subseteq B' = \Rightarrow (A, B) = (A',B')\},$$

**Step 6:** Places of each input and output nets are formed separately.

$$P_W = \{\ p(A,B)\big|^{(A,B)} \mid \in Y_W\ \} \cup \{\ i_W,\ o_W\ \},$$

**Step 7:** By integrating places obtained at Step 6 Petri net started to be formed.

$$F_W = \{(a,\ p(A,B)\big|^{(A,B)} \in Y_W \wedge a \in A\} \cup \{(\ p(A,B)^{,b}\ ) \mid (A, B) \in Y_W \wedge b \in B\} \cup \{(\ i_W\ , t) \mid t \in T_I\ \} \cup \{(t,\ o_W\ ) \mid t \in T_O\ \}$$

As a consequence, the Petri net is obtained with the alpha algorithm.

In our previous study, a data generator simulation was performed [26]. In this simulation, many synthetic event logs were created.

In our study, a system was created that provides process discovery from event logs created with synthetic and real event logs, integrated with the data generator simulation carried out before. In this study, the Petri net created with the alpha algorithm for the credit application data set consisting of 7 activities, 227 statuses, and 856 event logs created by simulation is shown in Fig. 12.

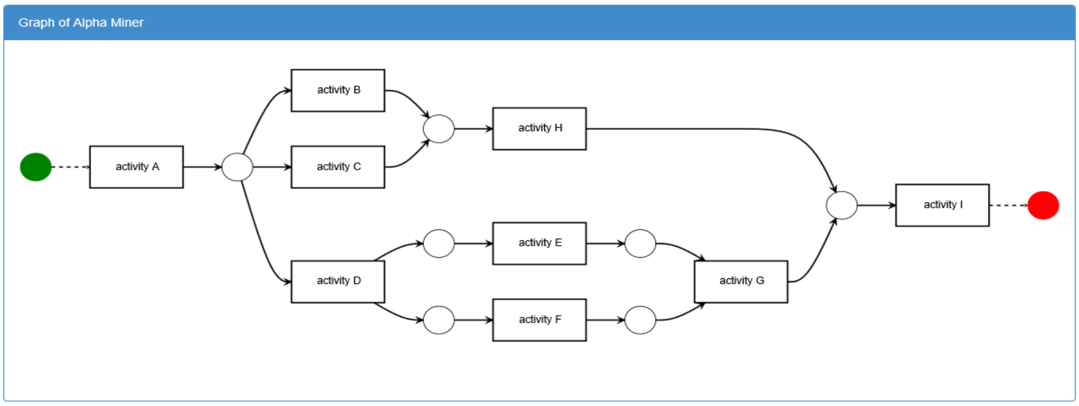


Fig. 12. Petri net created with the alpha algorithm of the sample credit data set.

### *H. Heuristic Algorithm*

The purpose of heuristic algorithm is to establish a Petri net that provides an optimum solution by using event logs [27]. The heuristic algorithm sequences transitions in the event log and removes less frequent transitions based on frequencies. The steps of the heuristic algorithm are as follows (see Algorithm 2).

Algorithm 2. Heuristic algorithm steps.

```
Step 1: All transitions from event logs are received.
    $T = \{t \mid \exists_{\sigma \in W} + [t \in \sigma]\},$
Step 2: Transitions calculated for a single loop length.
    $C_1 = \{(a, a) \in T \times T \mid a \Rightarrow W^{a} \geq \sigma L1L\},$
Step 3: Transitions calculated for length.
    $C_2 = \{(a, b) \in T \times T \mid (a, a) \not\exists C_1 \wedge (b, b) \not\exists C_1 \wedge a \Rightarrow 2 W^{b} \geq \sigma L2L\},$
Step 4: Strongest output value is determined.
    $C_{out} = \{(a, b) \in T \times T \mid b \neq End \wedge a \neq b \wedge \forall_{y \in T} [a \Rightarrow W^{b \geq a} \Rightarrow W^{y}]\}$
,
Step 5: Strongest input value is determined.
    $C_{in} = \{(a, b) \in T \times T \mid a \neq Start \wedge a \neq b \wedge \forall_{y \in T} [a \Rightarrow W^{b \geq x} \Rightarrow W^{b}]\},$
Step 6: Weakest output connections of two loops length are determined.
    $C'_{out} = \{(a, x) \in C_{out} \mid (a \Rightarrow W^{x}) < \sigma a \wedge \exists_{(b,y) \in C_{out}} [(a, b) \in C_2 \wedge ((b \Rightarrow W^{y}) - (a \Rightarrow W^{x})) > \sigma_{\gamma}]\}$
Step 7: Weak output connections are excluded.
    $C_{out} = C_{out} - C'_{out},$
Step 8: Weakest input connections of two loops length are determined.
    $C'_{in} = \{(x, a) \in C_{in} (x \Rightarrow W^{x}) < \sigma_{a} \wedge \exists_{(y,b) \in C_{in}} [(a, b) \in C_2 \wedge ((y \Rightarrow W^{b}) - (x \Rightarrow W^{a})) > \sigma_{\gamma}]\}$
Step 9: Weak input connections are excluded.
    $C_{in} = C_{in} - C'_{in}.$
Step 10: Output transitions calculation is determined according to threshold value.
    $C''_{out} = \{(a, b) \in T \times T \mid a \Rightarrow W^{b} \geq \sigma_{a} \vee \exists_{(a,c) \in C_{out}} [((a \Rightarrow W^{c}) - (a \Rightarrow W^{b})) < \sigma_{\gamma}]\},$
Step 11: Input transitions are determined according to threshold value.
    $C''_{in} = \{(b, a) \in T \times T \mid (b \Rightarrow W^{a}) \geq \sigma_{a} \vee \exists_{(b,c) \in C_{in}} [((b \Rightarrow W^{c}) - (b \Rightarrow W^{a} < \sigma_{\gamma}]))\},$
In consequence, linked graph is formed ($DG = C_1 \cup C_2 \cup C''_{out} \cup C''_{in}$).
```

In Fig. 13, the model created by the heuristic algorithm of the credit application process data that we used for the alpha algorithm above is shown. In this study, different models were obtained using alpha and heuristic algorithms and data from the credit application process.

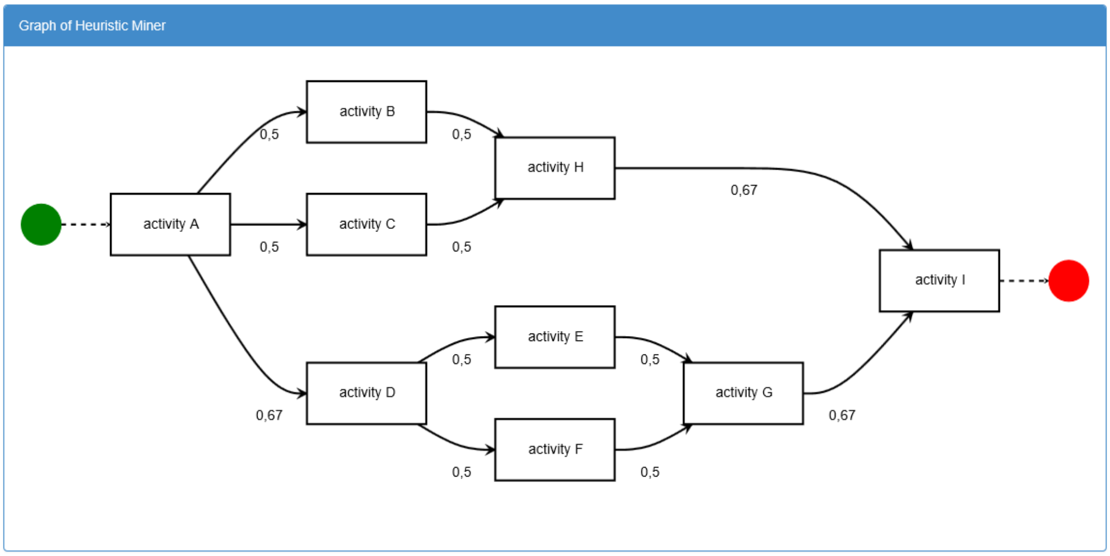


Fig. 13. Petri net created with the heuristic algorithm of the sample credit data set.

### *I. Genetic Algorithm*

Genetic algorithm is one of the methods based on biological processes used to perform modelling for the solution of algorithm searching and optimization problems. Genetic algorithms are used to solve problems that are difficult or impossible to solve with traditional methods [28].

Genetic algorithm is being used commonly in different areas, primarily for optimization, as well as automated programming and information systems, mechanical learning, finance, tabulation problems, facility layout problems, traveling salesman problems, vehicle routing problems [29]. The steps of the genetic algorithm are shown Fig. 14.

Algorithm 3. Genetic algorithm steps.

```
Step 1: Number of individuals determines mutation and crossover ratios.
Step 2: Random population is formed.
Step 3: Repeat Steps from Step 4 to Step 7 until new population is formed.
Step 4: Selection process is carried.
Step 5: Crossover applied.
Step 6: Mutation process applied.
Step 7: New baby individual is created.
Step 8: Best individuals solving the problem are generated.
```

The Petri net created with the genetic algorithm for the sample credit data set is shown in Fig. 15.

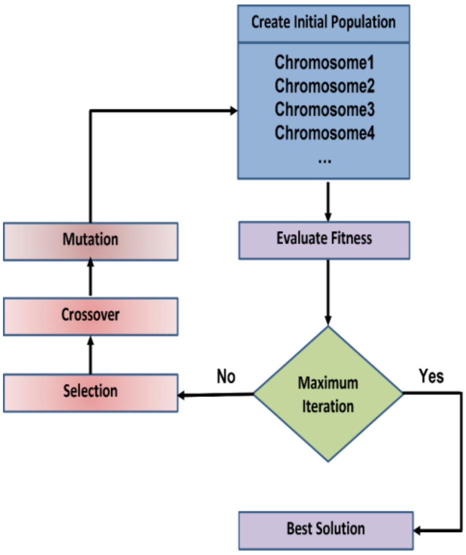


Fig. 14. Steps of the Petri genetic algorithm.

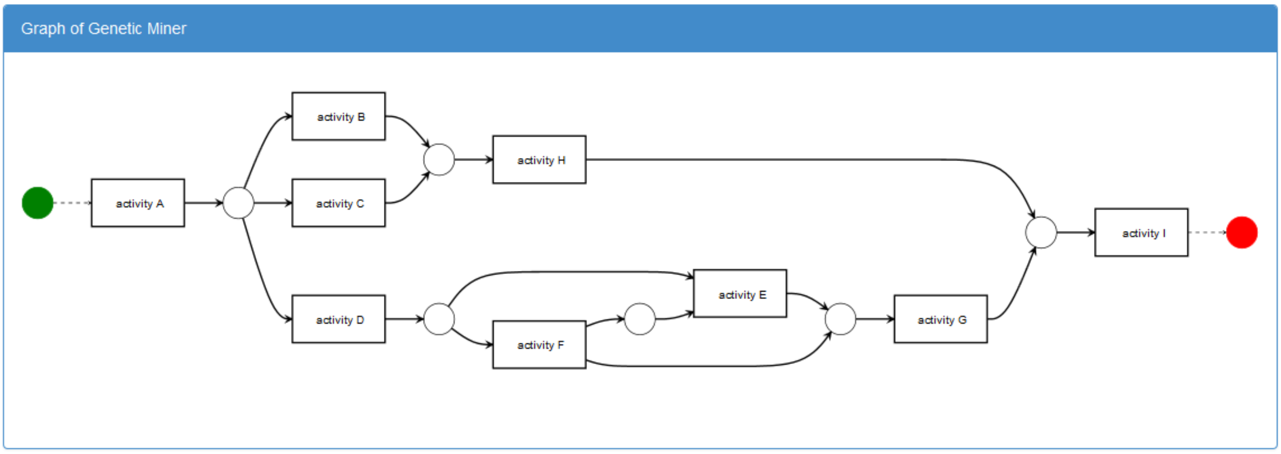


Fig. 15. Petri net created with genetic algorithm of the sample credit data set.

## V. Experimental Study

Unlike the alpha algorithm, the heuristic algorithm calculates the frequency values between the activity groups in the event logs. In the alpha algorithm, each event log is taken once. In the heuristic algorithm, frequency values are obtained by using all the event logs. The frequency calculation is made by taking the sum of the activity groups that are related to each other.

To measure the performance of algorithms, several different criteria are used. Accuracy and time criteria, primarily fitness function developed in this study, are compared for process mining discovery algorithms.

The fitness function is calculated to identify the most suitable individual to a solution to the problem. The fitness function for alpha and heuristic is shown as F. Equation (1) expresses the proposed fitness function for the alpha algorithm

$$\mathrm{F} = \frac{1}{2} * \left( \frac{(s-y)}{s} \right) + \frac{1}{2} * \left( \frac{(s-k)}{s} \right). \tag{1}$$

In (1), **s** is the total processes, **y** is the non-separable processes, and **k** is the remaining processes that can be separated. Equation (2) expresses the proposed fitness function for the heuristic algorithm

$$\mathrm{F} = \frac{3}{5} * \left( \frac{(s-(y+t))}{s} \right) + \frac{2}{5} * \left( \frac{(s-k)}{s} \right). \tag{2}$$

In (2), **s** is the total processes, **y** is the non-separable processes, **t** is the repeating processes, and **k** are the remaining processes that can be separated. The fitness function for the genetic algorithm is calculated in two ways, $\mathrm{F_S}$ and $\mathrm{F_C}$. These calculations are calculated as in (3) and (4):

$$F_S = 0.2\mathrm{x}\frac{PEAS}{TAS} + 0.3\mathrm{x}\frac{TTIS}{TIS} + 0.5\mathrm{x}\frac{TUOTS}{TIS}, \tag{3}$$

$$F_C = 0.4\mathrm{x}\frac{PEAS}{TAS} + 0.6\mathrm{x}\frac{TUOTS}{TIS}. \tag{4}$$

In (3) and (4), **PEAS** represents the number of activities parsed, **TAS** the number of all activities, **TTIS** the number of all completed tracks, **TIS** the number of all tracks, and **TUOTS** the number of all tracks suitable for completion.

Here, $\mathrm{F_C}$ is more advantageous compared to $\mathrm{F_S}$. Because when there are uncompleted cases in $\mathrm{F_S}$, the fitness function does not provide effective results. Moreover, when there is a parsing fault in $\mathrm{F_S}$, the parsing ends.

*Accuracy rate* is found by the ratio of parsed processes to the number of total processes. It is calculated in the same manner for all process algorithms.

*Time criteria* means the average working period of algorithms.

### *A. Performance Analysis of Process Mining Discovery Algorithms for Simulation Data Sets*

Table III lists the results of the process mining algorithms on various synthetic data sets generated using simulation. Missing Synthetic Data6 (Patient Follow-up Data) were created by deleting some activities in Synthetic Data5 (Patient Follow-up Data).

As the change between the fitness function of these two data is shown in Table III, the fitness function of the alpha algorithm decreased by 0.16 (0.78–0.62), while the fitness function of the heuristic algorithm decreased by 0.15 (0.83–0.68), and the fitness function of the genetic algorithm decreased by 0.04 (0.87–0.83). In the missing data, the change in the fitness function of the genetic algorithm decreased less than in the other algorithms.

### *B. Performance Analysis of Process Mining Discovery Algorithms for Synthetic Data Sets*

Table IV presents the results of process mining algorithms in existing [30] various synthetic data sets.

Concerning the average fitness and accuracy values of the algorithms, the genetic algorithm yielded successful results compared to the heuristic and alpha algorithms. Moreover, it is observed that the heuristic algorithm yielded better results than the alpha algorithm.

### *C. Performance Analysis of Process Mining Discovery Algorithms for Real Data Sets*

The results of the process mining algorithms on real data sets are presented in Table V [31], [32].

In Table V, some events were randomly deleted from the Missing Real Data3 (Patient Information) event log. Some events were randomly added to and deleted from the Missing and Noisy Real Data5 (Traffic) event log.

For these data sets, the performance of the alpha algorithm is lower than that of the heuristic and genetic algorithms. However, the genetic algorithm yielded better results than the heuristic algorithm.

### *D. Performance Evaluations of Process Mining Algorithms for the Same Data Set*

In Table VI, the success results of the real data set related to a patient process obtained from a hospital are given.

We obtained data similar to this data set with our own data generator simulation. However, since we could not obtain a synthetic version of this data set, we only compared the performance of these two data sets with alpha, heuristic, and genetic process mining algorithms. Similar results were obtained as seen in Table VI.

TABLE III. RESULTS OF PROCESS MINING ALGORITHMS ON SIMULATION DATA SETS.

| **Simulation Data Sets** | **Fitness Value** | | | **Accuracy Rate (%)** | | |
|---|---|---|---|---|---|---|
| | **Alpha** | **Heuristic** | **Genetic** | **Alpha** | **Heuristic** | **Genetic** |
| Data1 (Daily activity) | 1 | 1 | 1 | 100 | 100 | 100 |
| Data2 (Work application) | 0.97 | 1 | 0.99 | 95.5 | 97.5 | 96 |
| Data3 (Master application) | 0.92 | 0.94 | 0.97 | 93 | 95 | 97.6 |
| Data4 (Credit application) | 0.91 | 0.93 | 0.96 | 87.4 | 89.5 | 92.5 |
| Data5 (Patient Chart Data) | 0.78 | 0.83 | 0.87 | 79 | 84.5 | 86.7 |
| Missing Data6 (Patient Chart Data) | 0.62 | 0.68 | 0.83 | 60 | 65.3 | 72.5 |
| Avg | 0.86 | 0.89 | 0.93 | 85.8 | 88.6 | 90.8 |

TABLE IV. RESULTS OF PROCESS MINING ALGORITHMS ON SYNTHETIC DATA SETS.

| **Synthetic Data Sets [26]** | **Fitness Value** | | | **Accuracy Rate (%)** | | |
|---|---|---|---|---|---|---|
| | **Alpha** | **Heuristic** | **Genetic** | **Alpha** | **Heuristic** | **Genetic** |
| Synthetic Data1 (Exercise1) | 0.97 | 1 | 1 | 97.5 | 100 | 100 |
| Synthetic Data2 (Exercise2) | 0.92 | 0.98 | 0.95 | 95.5 | 97.5 | 95 |
| Synthetic Data3 (Exercise3) | 0.87 | 0.73 | 0.81 | 80 | 70.5 | 77.5 |
| Synthetic Data4 (Exercise4) | 0.72 | 0.81 | 0.84 | 70 | 85 | 86 |
| Synthetic Data5 (Exercise5) | 0.67 | 0.73 | 0.81 | 70 | 80.5 | 85.5 |
| Missing Synthetic Data6 (Exercise1) | 0.52 | 0.69 | 0.74 | 66.4 | 73.5 | 78 |
| Avg | 0.77 | 0.82 | 0.85 | 79.9 | 84.5 | 87 |

TABLE V. RESULTS OF PROCESS MINING ALGORITHMS ON REAL DATA SETS.

| Real Data Sets [27], [28] | Fitness Value | | | Accuracy Rate (%) | | |
|---|---|---|---|---|---|---|
| | Alpha | Heuristic | Genetic | Alpha | Heuristic | Genetic |
| Real Data1 (Teleclaims) | 0.69 | 0.98 | 0.99 | 71 | 86 | 91 |
| Real Data2 (Patient Information) | 0.67 | 0.84 | 0.89 | 68 | 74 | 83 |
| Missing Real Data3 (Patient Information) | 0.61 | 0.78 | 0.82 | 0 | 40 | 52 |
| Real Data4 (Traffic) | 0.57 | 0.73 | 0.62 | 0 | 55 | 64 |
| Missing and Noisy Real Data5 (Traffic) | - | - | 0.54 | - | - | 56 |

TABLE VI. RESULTS OF PROCESS MINING ALGORITHMS FOR SAME DATA SETS.

| Algorithms | Fitness Value | | | Accuracy Rate (%) | | |
|---|---|---|---|---|---|---|
| | Alpha | Heuristic | Genetic | Alpha | Heuristic | Genetic |
| Simulation Patient Process Data [7] | 0.69 | 0.78 | 0.84 | 71 | 80 | 86 |
| Real Patient Process Data [6] | 0.7 | 0.77 | 0.84 | 70.5 | 79.6 | 85.3 |
| Synthetic Patient Process Data | - | - | - | - | - | - |

## VI. CONCLUSIONS

In today’s information systems, information is recorded on events that take place in business processes. By reviewing these event logs, meaningful information about time, cost, and process can be extracted. Due to this information, it is possible to improve processes or measure the performance of the process.

At the first stage of this study, a web-based data generator was developed. In this data generator consisting of three main modules, process models can be created arbitrarily, data suitable for processes can be generated, and analysis and team comparisons can be made regarding the generated data. In the first module, a sample data model is created. In the simulation module, the second module, event logs are produced in compliance with the previously entered model. Meaningful information about the processes is extracted from the event logs produced in the analysis module.

In the second stage of this study, a system was created, enabling process discovery from existing event logs or event logs created with a data generator. After the data were taken in the MXML and XES formats and uploaded to the system, various process mining algorithms such as alpha, heuristic, and genetic algorithms were applied, and the results were compared according to fitness, accuracy, and time criteria.

When comparing the results of alpha and heuristic algorithms, the heuristic algorithm yielded more successful results than the alpha algorithm in terms of fitness and accuracy rates. The genetic algorithm yields better results than the heuristic algorithm regarding incomplete processes due to the application of genetic procedures (selection, crossover, and mutation).

When the results are compared in terms of the time criterion, the alpha algorithm ends in a shorter time than the heuristic and genetic algorithms for simple data. However, the genetic algorithm ends in shorter times than the alpha and heuristic algorithms for complex data.

In the future, the aim is to improve the system by integrating different algorithms into this study. Furthermore, it is recommended to develop studies to deal with process mining problems such as noisy data and incomplete processes, which are the disadvantages of our study.

## CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

## REFERENCES

[1] H. Sun, W. Liu, L. Qi, Y. Du, X. Ren, and X. Liu, “A process mining algorithm to mixed multiple-concurrency short-loop structures”, *Information Sciences*, vol. 542, pp. 453–475, 2021. DOI: 10.1016/j.ins.2020.07.003.

[2] P. Zerbino, A. Stefanini, and D. Aloini, “Process science in action: A literature review on process mining in business management”, *Technological Forecasting and Social Change*, vol. 172, art. 121021, 2021. DOI: 10.1016/j.techfore.2021.121021.

[3] A. Pika, M. T. Wynn, S. Budiono, A. H M Ter Hofstede, W. M P van der Aalst, and H. A Reijers, “Privacy-preserving process mining in healthcare”, *International Journal of Environmental Research and Public Health*, vol. 17, no. 5, p. 1612, 2020. DOI: 10.3390/ijerph17051612.

[4] A. Partington, M. Wynn, S. Suriadi, C. Ouyang, and J. Karnon, “Process mining for clinical process: A comparative analysis of four Australian hospitals”, *ACM Transactions on Management Information System*, vol. 5, no. 4, pp. 19-1–19-18, 2015. DOI: 10.1145/2629446.

[5] G. B. Pereira, E. A. P. Santos, and M. M. C. Maceno, “Process mining project methodology in healthcare: A case study in a tertiary hospital”, *Network Modelling Analysis in Health Informatics and Bioinformatics*, vol. 9, art. no. 28, 2020. DOI: 10.1007/s13721-020-00227-w.

[6] F. A. Yasmin, R. Bemthuis, M. Elhagaly, F. Wijnhoven, and F. A. Bukhsh, “A process mining starting guideline for process analysts and process owners: A practical process analytics guide using ProM”, *DSI Technical Report Series*, pp. 1–18, 2020.

[7] S. Weerapong, P. Porouhan, and W. Premchaiswadi, “Process mining using σ-algorithm as a tool (A case study of student registration)”, in *Proc. of 2012 Tenth International Conference on ICT and Knowledge Engineering*, 2012, pp. 213–220. DOI: 10.1109/ICTKE.2012.6408558.

[8] D. A. Koonce and S.-C. Tsai, “Using data mining to find patterns in genetic algorithm solutions to a job shop schedule”, *Computers &*

*Industrial Engineering*, vol. 38, no. 3, pp. 361–374, 2000. DOI: 10.1016/S0360-8352(00)00050-4.

[9] A. J. M. M. Weijters, W. M. P. van der Aalst, and A. K. Alves, "Process mining with the HeuristicsMiner algorithm", *BETA publicatie: working papers*, vol. 166, 2006.

[10] N. Martin *et al.*, "Recommendations for enhancing the usability and understandability of process mining in healthcare", *Artificial Intelligence in Medicine*, vol. 109, art. 101962, 2020. DOI: 10.1016/j.artmed.2020.101962.

[11] S. Goedertier, J. De Weerdt, D. Martens, J. Vanthienen, and B. Baesens, "Process discovery in event logs: An application in the telecom industry", *Applied Soft Computing*, vol. 11, no. 2, pp. 1697–1710, 2011. DOI: 10.1016/j.asoc.2010.04.025.

[12] Y. Liu, H. Zhang, C. Li, and R. J. Jiao, "Workflow simulation for operational decision support using event graph through process mining", *Decision Support Systems*, vol. 52, no. 3, pp. 685–697, 2012. DOI: 10.1016/j.dss.2011.11.003.

[13] O. Dogan, J.-L. Bayo-Monton, C. Fernandez-Llatas, and B. Oztaysi, "Analyzing of gender behaviors from paths using process mining: A shopping mall application", *Sensors*, vol. 19, no. 3, p. 557, 2019. DOI: 10.3390/s19030557.

[14] P. Nafasa, I. Waspada, N. Bahtiar, and A. Wibowo, "Implementation of alpha miner algorithm in process mining application development for online learning activities based on MOODLE event log data", in *Proc. of 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, 2019, pp. 1–6. DOI: 10.1109/ICICoS48119.2019.8982384.

[15] G. P. Kusuma, S. Sykes, C. McInerney, and O. Johnson, "Process mining of disease trajectories: A feasibility study", in *Proc. of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies*, 2020, pp. 705–712, vol. 5. DOI: 10.5220/0009166607050712.

[16] A. K. A. de Medeiros, A. J. M. M. Weijters, and W. M. P. van der Aalst, "Genetic process mining: An experimental evaluation", *Data Mining and Knowledge Discovery*, vol. 14, no. 2, pp. 245–304, 2007. DOI: 10.1007/s10618-006-0061-7.

[17] Z. Huang and A. Kumar, "A study of quality and accuracy trade-offs in process mining", *INFORMS Journal on Computing*, vol. 24, no. 2, pp. 311–327, 2012. DOI: 10.1287/ijoc.1100.0444.

[18] M. Vidgof and B. Wurm, "The connection between process complexity of event sequences and models discovered by process mining", *Information Sciences*, vol. 598, pp. 196–215, 2022. DOI: 10.1016/j.ins.2022.03.072.

[19] H. AlQaheri and M. Panda, "An education process mining framework: Unveiling meaningful information for understanding students' learning behavior and improving teaching quality", *Information*, vol. 13, no. 1, p. 29, 2022. DOI: 10.3390/info13010029.

[20] R. Fauzi and R. Andreswari, "Business process analysis of programmer job role in software development using process mining", *Procedia Computer Science*, vol. 197, pp. 701–708, 2022. DOI: 10.1016/j.procs.2021.12.191.

[21] 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, pp. 719–736, 2021. DOI: 10.1007/s41066-020-00226-2.

[22] L. Reinkemeyer, *Process Mining in Action: Principles, Use cases and Outlook*, 1st ed. Springer, 2020. DOI: 10.1007/978-3-030-40172-6.

[23] N. Gehrke and M. Werner, "Process Mining", WISU - die Zeitschrift für den Wirtschaftsstudenten 7/13, pp. 1–16, 2012.

[24] W. van der Aalst *et al.*, "Process mining manifesto", in *Business Process Management Workshops. BPM 2011. Lecture Notes in Business Information Processing*, vol. 99. Springer, Berlin, Heidelberg, 2012, pp. 169–194. DOI: 10.1007/978-3-642-28108-2_19.

[25] A. K. A. de Medeiros, B. F. van Dongen, W. M. P. van der Aalst, and A. J. M. M. Weijters "Process mining: Extending the α-algorithm to mine short loops", *BETA publicatie: working papers*, vol. 113, 2004.

[26] M. Yüzkat, B. Şen, H. K. Caymaz, and F. Karabiber, "Implementation of data generator for process mining applications", in *Proc of 2015 23nd Signal Processing and Communications Applications Conference (SIU)*, 2015, pp. 1405–1408. DOI: 10.1109/SIU.2015.7130105.

[27] A. J. M. M. Weijters and J. T. S. Ribeiro, "Flexible Heuristics Miner (FHM)", in *Proc. of 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2011, pp. 310–317, 2011. DOI: 10.1109/CIDM.2011.5949453.

[28] A. Bogarín, R. Cerezo, and C. Romero, "A survey on educational process mining", *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 1, p. e1230, 2018. DOI: 10.1002/widm.1230.

[29] A. Lambora, K. Gupta, and K. Chopra, "Genetic algorithm - A literature review", in *Proc. of 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 380–384. DOI: 10.1109/COMITCon.2019.8862255.

[30] ProM Tools. [Online]. Available: http://www.promtools.org/doku.php?id=tutorial:introduction.

[31] Process Mining. [Online]. Available: http://www.processmining.org/event_logs_and_models_used_in_book (accessed 2021).

[32] 3TU Datacentrum. [Online]. Available: http://data.3tu.nl/repository/collection:event_logs_real (accessed 18 August 2021).