**Name of student:** Mickael LIU

**Name of your Level:** L0

**Source (e.g. scholars.google.com):** [**https://www.mdpi.com/2078-2489/13/5/234**](https://www.mdpi.com/2078-2489/13/5/234)

**Paper title:** A Deep Learning Approach for Repairing Missing Activity Labels in Event Logs for Process Mining

**Keywords specific to the paper:**

process mining; business process management; incomplete event logs; data quality; data management

**Summary of the main contributions:**

Business process management techniques are widely applied in modern information systems, such as financial, production, and hospital systems. Traditionally, process analysts model business processes through knowledge gained from interviews, workshops, or documents. On the one hand, modeling business processes by hand can be cost ineffective and time consuming. On the other hand, involving human beings to model processes can introduce unavoidable biases. Thanks to the large-scale deployment of computer systems, enterprise data have become more accessible. Process mining, a relatively new subject, was introduced to fill the gap between data mining and traditional process modeling. The goal of process mining techniques is to discover process insights directly from the data collected from target organizations.

One of the most critical parts of process mining is called process discovery, which aims at discovering a business process model automatically from process data. The datasets used to discover process models are called event logs. Each event log is a collection of traces, and each trace is an ordered sequence of events. Each event contains an activity, timestamp, and other attributes.

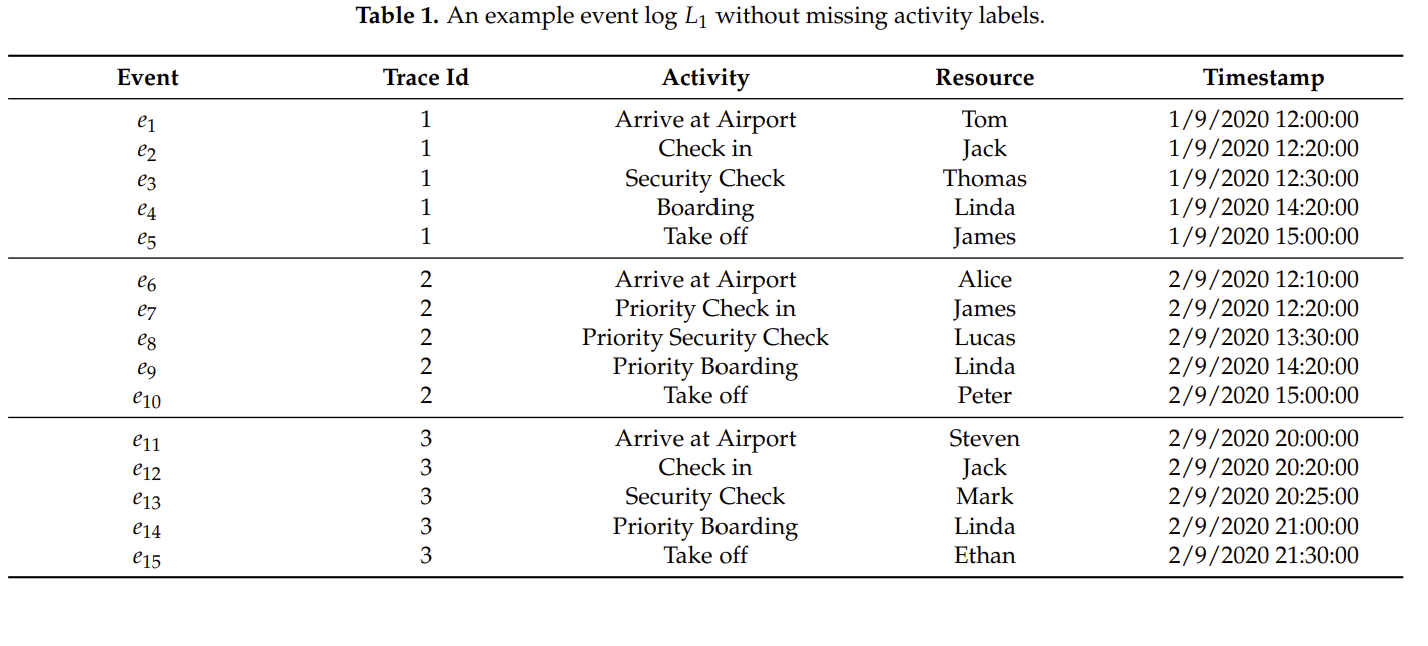
Different process discovery algorithms have been proposed in the last decade, and some of them can guarantee the production of accurate process models under certain circumstances. However, like other data mining techniques, the analysis results are heavily related to the quality of the input datasets. Most existing process discovery algorithms assume the event log to be complete, and they may not be able to discover accurate process models when some data in the input event log are missing. Missing data in event logs has been defined as one of the major data quality issues in process mining. Several methods were proposed in the field of process mining to repair event logs with missing data. However, none of these methods can accurately repair missing activity labels in event logs when a large number of activity labels are missing.

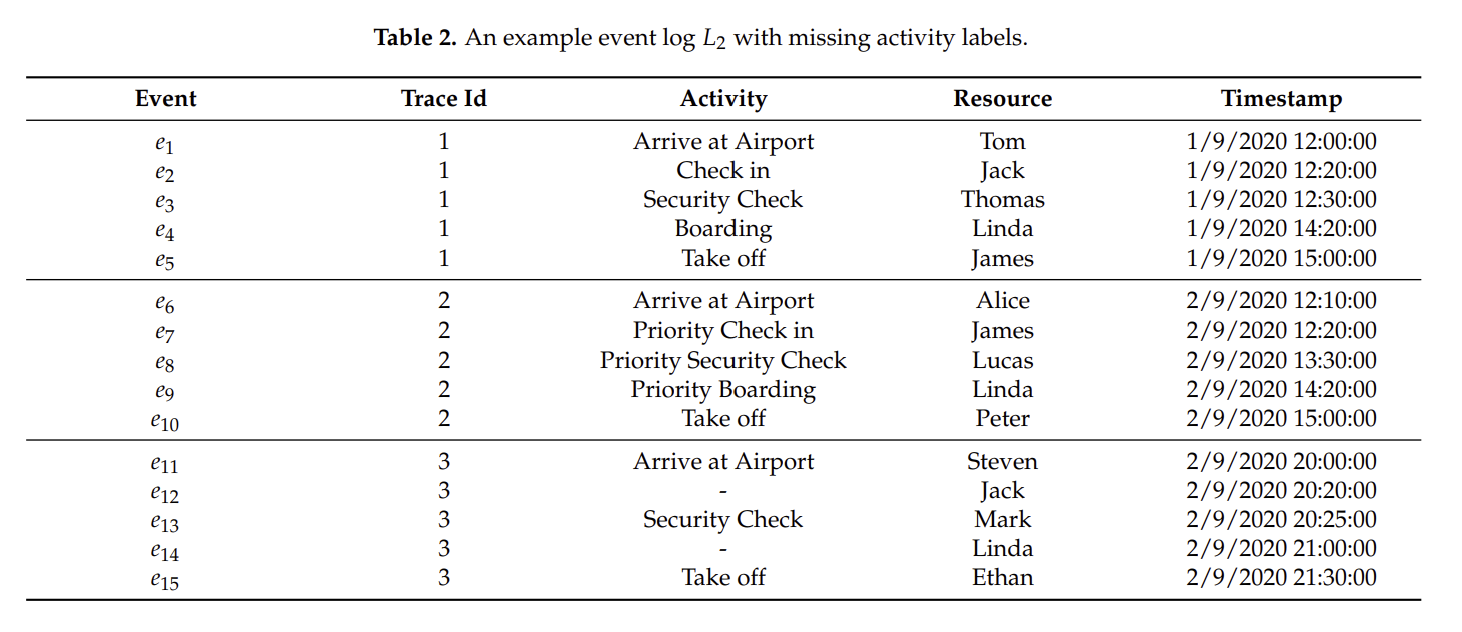
Inspired by recent research papers that successfully applied deep learning methods to predict the next activities in ongoing traces, they propose an LSTM-based prediction model to predict the missing activity labels in event logs. The prediction model takes both the prefix and suffix sequences of the events with missing activity labels as input. In addition, additional attributes of event logs are also utilized to improve the performance.

Table 1 shows an example event log L1, which describes a simple airport process. Each row is an event, which is an execution record of an activity. An event can have multiple attributes. In the example log, each event has an activity label, a resource, and a timestamp. The activity label describes which activity the event recorded, the resource describes the person who performed the event, and the timestamp describes the time when the event was recorded. The event log contains three traces, and each trace is a sequence of events ordered by timestamps. The goal of automated process discovery algorithms is to construct a process model that can accurately describe the process behaviors.

The goal of this paper is to propose a method as a data preprocessing tool that can accurately repair the missing activity labels. The repaired event logs can then be used by process discovery algorithms to discover accurate process models.

Without the labels we can think that the passengers can go to the security check without checking in.





However, since the proposed method uses neural networks to predict missing activity labels, there can be higher requirements for computer hardware to execute the method. For example, a GPU could be required to achieve good performance. In addition, the method requires many parameters from users, which may impact the performance of the method.

It has to be noted that, like other methods used to repair missing activity labels in event logs, it is assumed that we know the exact locations of the missing values. Although our method cannot be applied directly to event logs when the locations of missing values are unknown, the method can be used together with anomaly detection algorithms. For example, a missing event may exist between two events when the direct succession relation between two consecutive events is identified as an anomaly.