Teaching and Research Area of Business Process Management Foundations and Engineering

RWTH AACHEN UNIVERSITY

This work was submitted to:

Business Process Management Foundations and Engineering (Informatik 9), RWTH Aachen University

Comparative Process Mining In Healthcare

Master's Thesis

Author: **Asad Tariq**

Student ID: **403326**

Supervisor: Univ-Prof. Dr. Ir. Sander J.J. Leemans

Examiners: Univ-Prof. Dr. Ir. Sander J.J. Leemans
Prof. Dr. Wil M. P. van der Aalst

Registration Date: 2022-10-24

Submission Date: 2023-04-24

Abstract

In the pursuit of optimising healthcare services, evidence-based innovations are deemed essential. Process mining is a technique that uses historical data from hospital information systems to provide valuable insights to healthcare stakeholders to enhance healthcare procedures. Healthcare procedures are flexible and complex, making it difficult to fully utilise traditional process mining tools. This highlights the need for enhanced methodologies to study less structured processes in healthcare and other fields. This thesis introduces a novel stochastic-aware approach for grouping process behaviour into cohorts based on trace attributes or other trace features, intending to improve process mining results in complex environments by dividing a large event log into smaller homogeneous subsets. The proposed approach is based on log profiles, which enable the identification of relevant attributes and the creation of new attributes to form homogeneous subsets. Through the evaluation of the study, we demonstrate that our approach is effective in discovering patterns and analysing processes in complex environments and has the potential for application in various domains. The methodology is evaluated using two real-world healthcare datasets.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the modern world, almost all steps in a business or organisation's procedures are recorded digitally as a routine practice. These recorded event data, which can range from selecting the payment plan for your health insurance on a website to booking a flight with an airline, provide the foundation for process mining algorithms, a group of analytical tools that look through a lot of event data to find insights into a process. Organisations use these insights to enhance business operations, reduce costs, optimise resource allocation, avoid potential bottlenecks and forecast future behaviour to achieve a competitive advantage. However, connecting the fields of process science with data science is essential if this is to be accomplished. Process mining may bridge the gap by giving visibility into the real-life processes consumers or employees use within a business and spotting bottlenecks and variances that must be addressed [1].

1.1 Background

The healthcare sector has several difficulties that need the systemic incorporation of process improvement. Data-driven innovations have become more crucial to increase healthcare's efficacy and efficiency [2, 3]. Thanks to these new strategies, healthcare firms may quickly modify their operations to meet changing patient needs. Healthcare businesses all over the world recognise the need to constantly enhance both their clinical and administrative procedures. Hospital information systems are heavily relied upon by healthcare organisations to facilitate various activities, including the recording of process execution data [4], which supports both clinical and administrative procedures. Sequences of actions taken during patient care, hospital stays and other situations are included in this data. Healthcare firms can improve their operations by using this data to analyse their procedures and make data-driven decisions.

Process mining is a fast-expanding field that has attracted interest recently because of its capacity to shed light on insights into complex processes in organisations. It includes a variety of analytical methodologies that use event data captured throughout different process phases, such as ordering an item from an online marketplace to booking a ticket for a football match, to produce process insights. The process discovery, conformance checking and process enhancement are the three main sub-fields of process mining [5]. The first subarea, process discovery, focuses on building appropriate process models from

event data that has been captured. This subarea uses an event log as its input and its objective is to identify a process model that faithfully captures the organisation's current state of the process. This stage aids in determining the process's control flow and building a model that encapsulates its structure [5]. Process mining's second sub-field, conformance checking, involves contrasting recorded event data with a suitable process model. This subarea enables the detection and measurement of differences between the process model and the process's actual execution in practice. For example, this subarea aids in identifying where people stray, where tasks are completed at the incorrect time and where tasks are switched. Organisations can increase their understanding of process models and pinpoint areas that need work through this subarea [5]. Process enhancement, which leverages the understanding gained from the first two subareas to find bottlenecks and strengthen the current process, is the third and final subarea of process mining. Organisations can streamline operations and gain a competitive advantage by determining where to improve. In general, process mining is a vital tool that aids businesses in understanding their procedures, spotting errors and making data-driven decisions to enhance their operations [5].

Cohort comparison is a process mining technique that compares cohorts, or groupings of cases, within a process [6]. It is possible to learn a lot about the distinctions between different cohorts' processes and how they differ. Such analysis can pinpoint optimum practices, for instance, if the processing is anticipated to be similar. However, if differences are anticipated, the analysis can draw attention to process-based commonalities [6, 7]. Such a comparative process analysis can then improve the related processes. Various perspectives can be used to compare processes. For example, the **control-flow** perspective discusses the activities that can be carried out in a process and how they are organised into pathways. In contrast, the **stochastic** perspective describes how likely activities, pathways and behaviour in processes are [8].

Considering both the control flow and stochastic perspectives in comparative process mining analysis is advisable. For example, although a rework loop may be possible, the control flow perspective may not be able to describe its likelihood or its effects on the process. Conversely, determining the importance of process-based variations between cohorts depends critically on understanding the stochastic perspective. For instance, a seldom-used rework loop may be a component of standard operating procedures. In contrast, a regularly used loop may have a considerable impact on the performance of the operation. As a result, healthcare businesses can discover possible bottlenecks and streamline their procedures to better serve the requirements of their patients and staff by contrasting the two points of view [2].

Several methods, including visual comparisons and analytical approaches, can be used to compare the two processes. Before comparing a single process, splitting it into variations or cohorts based on shared traits is commonly recommended. Cohorts can be found in event logs using various methods, including decision trees, rule-based approaches and clustering algorithms. The cohorts can be compared once they have been found using different process mining techniques, including process discovery, conformity testing and performance analysis. Process optimisation efforts can be concentrated on the areas with the most influential impact by comparing cohorts to acquire insights into the differences and similarities between the process variants. An example of cohort identification for a

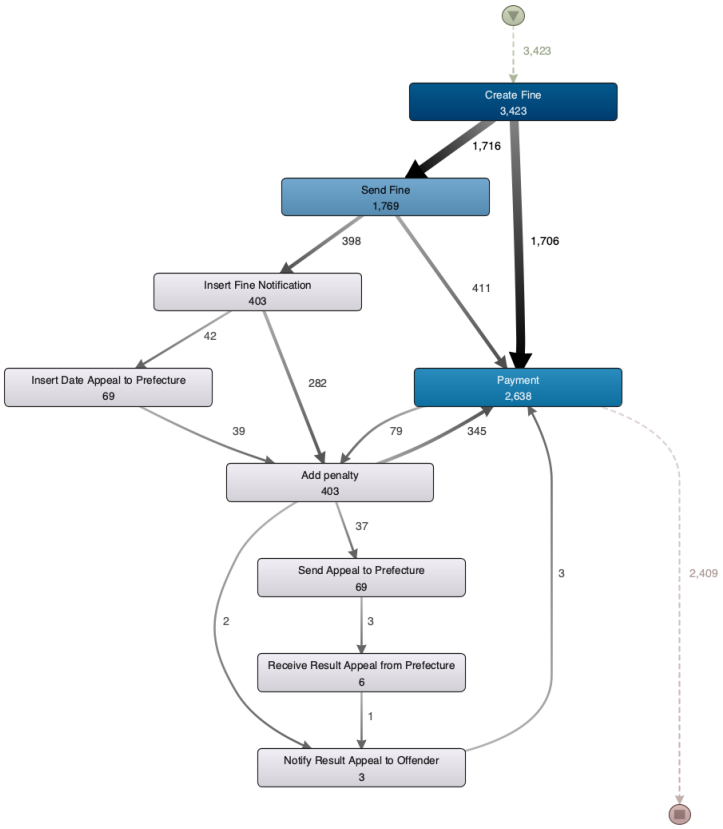


Figure 1.1: Road Traffic Fine Management Process

complex Road Traffic Fine Management Process[1] is shown in Figure 1.1. The process map has payment and appeal modules and possible cohort divisions could be two smaller processes, each focusing on a specific module, as shown in Figures 1.2 and 1.3.

1.2 Motivation and Problem Statement

The Process Comparison Methodology (PCM), a general approach to combining several process comparison methodologies, was proposed in prior work [7]. However, PCM does not consider the stochastic perspective and necessitates much manual labour without much-automated support, as mentioned in Section 2. So far, there needs to be a way to automatically input an event log file to identify cohorts. We suggest enhancing the PCM method with stochastic awareness to close this gap. Our objective is to give a systematic method for operationalising PCM and provide advice on effectively using the approaches.

1.3 Research Question and Goal

Identifying cohorts in the process is essential for understanding the behaviour of groups of entities following a similar path. However, manually identifying cohorts can take time and

[1] 10.4121/uuid:270fd440-1057-4fb9-89a9-b699b47990f5

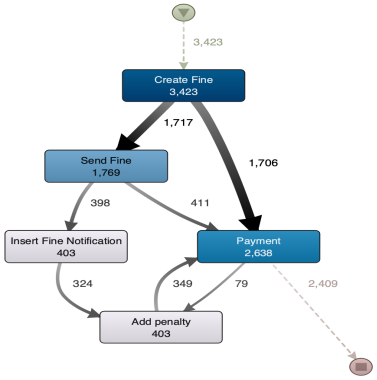


Figure 1.2: Payments

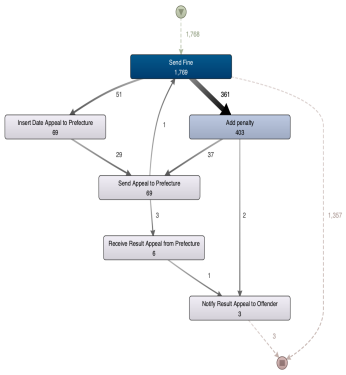


Figure 1.3: Appeals

effort. Identifying cohorts can now be automated using data mining and process mining techniques. Therefore, the research question for this study is:

How can stochastic-aware process mining and data mining techniques be utilised to identify cohorts in the process?

1.4 Contributions

The thesis proposes a semi-automated approach for identifying and instantiating alpha attributes in process mining, resolving the domain expert's dependency issue. The thesis also contributes to process mining by demonstrating how it can identify cohorts, providing organisations with deeper insights into group behaviour and process optimisation. Overall, the contributions of this thesis can help organisations leverage process mining to improve their processes and gain a competitive edge without the involvement of domain experts.

1.5 Thesis Structure

The remainder of this thesis is structured as follows. In Chapter 2, we discuss related work. In Chapter 3, we present basic preliminaries used throughout the thesis. In Chapter 4, we discuss the methods proposed to find alpha attributes. In Chapter 5, we evaluate the applicability and usefulness of our proposed methods. In Chapter 6, we analyse the results we extracted and the challenges in the proposed methods. And in Chapter 7, we conclude the work and discuss future work and prospects.

Chapter 2

Related Work

In this section, related literature is discussed. It starts with the case studies in the process mining, after which related work for process mining methods, comparative process mining in healthcare, the PCM framework and trace clustering methods are discussed.

2.1 Case Studies in Process Mining

Over the past few years, the process mining techniques have been leveraged in many case studies in various sectors, including finance [9], [10], education [11], [12], information technology [13], [14], the automotive industry [15], ERP system [16], warehouse management [17], manufacturing [18], block-chain [19], etc. The healthcare industry has conducted a sizable number of case studies. A couple of case studies have been conducted in a Dutch hospital [20], [21]. A case study regarding tele-consultation Healthcare has been carried out in an Italian Hospital [22]. Recently, pandemic-related research was carried out for Covid-19 in [23]. In a Chicago outpatient clinic, a case study was conducted to analyse workflows in Clinical Care [24]. A case study related to Sepsis management is also performed for EMR Data [25]. Process mining project methodology for a tertiary hospital is carried out in [26]. Interesting research about cancer registry data is discussed in [27]. The case studies suggest that process mining can provide valuable insights and help organisations achieve their goals by optimising processes and improving outcomes.

2.2 Process Mining Methods

Only a few methods, meanwhile, have been published for structured process mining initiatives as well. In [28], the research proposes that with prior domain expertise, the Process Diagnostics Method is suggested to gain a comprehensive picture of the process swiftly. In [29], a methodology for applying process mining techniques focused on identifying habitual behaviour, process variants and exceptional medical cases are proposed. Finally, the L* life-cycle methodology has been proposed for mining process models in [30]. The L* life-cycle methodology summarises the life cycle of a typical process mining project that aims to optimise processes and encompasses a wide range of methodologies. Moreover, the researchers in [31] proposed PM2: a Process Mining Project Methodology since the former proposed methods have limitations, including limited applicability for well-

structured processes, resource-intensive, lack of flexibility and dependency on technical expertise, which may be challenging for organisations with limited resources. The Process Diagnostics Method limitedly concentrates on presenting a comprehensive perspective utilising a constrained range of process mining tools, and the L* is limited to analysing structured processes. An iterative analysis is the main emphasis of PM2[31], intended to help initiatives that strive to optimise process performance or compliance. The case study demonstrated the application of PM2 on data provided by IBM. PM2, like L*, covers many process mining methods. However, PM2 is more suitable for structured and unstructured process analysis than L*.

2.3 Comparative Process Mining

The approaches provided valuable insights but still need improvement as they primarily focus on analysing a single process. Some published research has focused on comparing processes based on event logs. A visual approach to spot statistically significant differences in event logs based on process metrics is proposed in [32]. The approach allows users to explore and discover interesting patterns, enabling a better understanding of business processes and improvement opportunities. In [33], the authors compared business process variants using process models and event logs. The method utilises techniques from process mining to discover and differentiate different variants of a process, providing insights into how processes are executed and improvement opportunities. In [34], the authors calculated and visualised the difference between process models in multi-dimensional process mining. The approach focuses on generating a compact representation of the differences between the models, allowing users to quickly identify the most significant differences. In [35], a framework in which the stochastic distance between cohorts defined by sets of these features is measured by a system that extracts features from trace attributes and then displays to users the landscape of sets of features and their impact on process behaviour is proposed. The method aims to help analysts identify data patterns, enabling a better understanding of the process and supporting decision-making processes. The approach is evaluated on real-life event logs, demonstrating its effectiveness in identifying relevant cohorts and suggesting actionable insights. The authors in [36] have devised a new method for discovering and checking how well a process matches its intended design. The method analyses large amounts of data quickly, and they tested it on a dataset with over 4 million events. The results showed that the method is much faster than existing methods, and they incorporated it into a popular open-source tool called ProM. Various research papers have proposed several approaches to analyse business processes using process mining techniques, including comparing different process variants, visual approaches, and frameworks to identify patterns and cohorts, enabling a better understanding of the process. However, there is still room for improvement.

2.4 Comparative Process Mining in Healthcare

Comparative process mining techniques have yet to be exposed to healthcare extensively. Nevertheless, this field has seen some fascinating studies. In [37], the authors propose a multi-level approach for identifying process changes in cancer pathways by combining process mining techniques with expert knowledge. The approach involves identifying and analysing the process models at different levels of granularity and comparing them to detect

changes. In [38], a method for improving the comparison of structural medical processes by combining domain knowledge and mined information is proposed. It involves extracting domain-specific knowledge from medical guidelines and combining it with mined process models to identify their differences. A method to enhance the medical process mining and trace comparison by leveraging semantic labels for multi-level abstraction is proposed in [39]. In [40], the authors present an analysis comparing process mining applied to clinical processes in four Australian hospitals. The authors evaluate the potential of process mining to improve healthcare processes, identifying process variations and providing insights into how process mining can enhance patient outcomes and reduce costs. In [41], cross-organisational process comparison in collections of process models and their executions is proposed. Nevertheless, most process comparison methods only consider the control-flow component (i.e., the presence, routing, and frequency of activities), ignoring other dimensions.

2.5 Trace Clustering in Process Mining

Finding groupings of related objects in a data set using the well-known data mining approach called cluster analysis [41]. Trace clustering techniques are methods for grouping comparable traces or process instances. This division is frequently accomplished by finding a process model for each trace cluster, relying on patterns, similarities between the traces, or both. Typically, it falls under the unsupervised machine learning category [42]. For example, sequence analysis methods in bioinformatics, sequential pattern mining for consumer purchasing habits, and sequence labelling for part of speech tagging are all typical applications of clustering techniques [43]. For more than a decade [44], trace clustering has been studied as a preprocessing method for process discovery. The approach, similarity metrics, computing cost, and maturity have significantly progressed. The techniques published before 2015 are well-reviewed in the morphological box [45]. A context-aware technique to group traces based on generic edit distance [46]. With an unstructured process and problems with production performance indicators, the case study [47] offers an applied situation in industrial manufacturing production. Several methodologies were applied to comprehend how the process works, how many trace clusters should be recognised as homogeneous process variants, and what leads to production inefficiencies. The concept of trace profiling for clustering is proposed in[48]. Similarly, Comparative Trace Clustering has been applied to detect process changes, leading to valuable insights and process improvement[49]. Trace clustering methods have significantly progressed over the last decade and have been studied as a preprocessing method for process discovery. With various methodologies, trace clustering has been used to comprehend process variants and identify process changes, leading to valuable insights and process improvement.

2.6 Process Comparison Methodology

In light of the increasing interest in comparing processes from viewpoints other than control flow and the need for more methodological support for applying process comparison in a process mining project practically, the authors propose "Process Comparison Methodology (PCM)" in [50]. They provided an innovative methodology by considering multiple aspects, including the organisation, data, performance, etc., in contrast to previous process mining methodologies. In the case study, they used real-life data supplied by Xerox Services to

confirm their methods.

The following is the paper's standout contributions:

- Outline an approach for comparing processes that focuses on analysing several processes. Several factors, including control flow, organisation, data, performance, etc., are considered by this methodology.
- Utilise actual data to validate the methodology in a case study.

The alpha attribute in [7] is chosen in the initial phases. However, more guidance is needed on how this attribute can be chosen. The automated extraction of the alpha attribute is a research area that has received limited attention in the academic literature. Mostly the alpha attribute is selected by the domain expert. For example, in [7], the batch number of the insurance is selected as the alpha attribute based on domain knowledge. For a non-expert, finding the alpha attribute to divide the big event into smaller sub-logs is a challenge. To overcome this dependence on a domain expert to pick out an alpha attribute, data mining techniques can help us retrieve the alpha attribute to divide the event log into smaller but influential groups. In event logs, we have information related to the case and trace attributes. From these attributes, we can mine our alpha attribute. Machine learning methods have played a keen role in pre-processing of event logs in Process mining. In [51], the demo shows the capabilities of the Proactive Insights engine, which uses machine learning techniques to detect deficiencies in business processes, find their underlying reasons, and provide information regarding guidance to enhance the efficiency of the process. In [52], a log-lifting framework based on machine learning is proposed to bridge the gap between process model activities and event log discrepancies. The framework involves two phases, log segmentation and classification based on machine learning. In [53], to improve the efficiency of process mining, context awareness is added to unstructured event logs in logistics using machine learning. The study focuses on applying context awareness based on machine learning in process mining for logistic processes and depicts how it works in a logistical situation. This paper provides several automated techniques that assist analysts

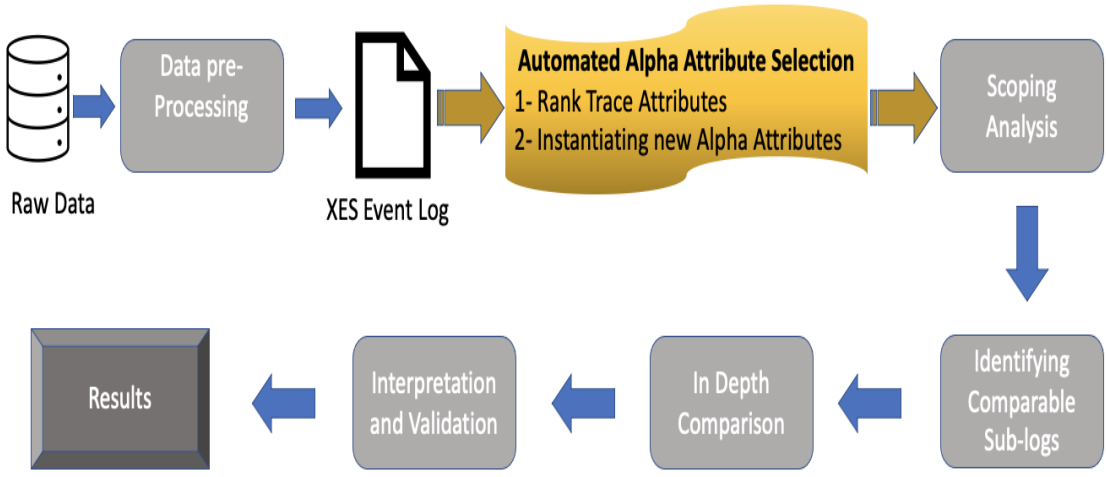


Figure 2.1: PCM

in selecting alpha attributes or creating new ones. Additionally, Figure 2.1 graphically

demonstrates the PCM technique, emphasising the phase on which we focus specifically, denoted by the golden tile.

Chapter 3

Preliminaries

This chapter introduces vital fundamental ideas essential to comprehending the proposed study. Although we generally assume that the reader is already familiar with these fundamental ideas, this chapter aims to condense all the themes discussed in this work into one compact section. We start by outlining the fundamental mathematical definitions and notations we will use. The second section presents definitions and ideas related to process mining. Then, ideas for Machine learning methods and their evaluation are presented.

3.1 Basic Mathematical Concepts

This section introduces the basic mathematical concepts we use to formally describe our technique.

Sets

Mathematically, a set is a fundamental concept representing a collection of well-defined objects used to group certain elements.

Definition 3.1 (Set). A set is defined as a collection of unique elements. The elements of a set can be of an arbitrary type. For example, {1, 2, 3, 4, 5} is a set consisting of elements 1, 2, 3,4 and 5. If x is an element of a set X, it is written as $x \in X$; if x does not belong to X, it is written as $x \notin X$.

Given the two sets X and Y,

- **Cardinality of the set:** If the set $X = \{x_1, x_2, x_3, x_4, \ldots, x_n\}$ has n elements, then the cardinality of the set X is shown as $|X|$, therefore $|X| = n$.
- **Empty set:** The empty set is denoted by $\varnothing$.
- **Union of sets:** The union of sets X and Y is denoted by $X \cup Y$, i.e., $X \cup Y = \{x | x \in X \vee x \in Y\}$.
- **Intersection of sets:** The intersection of the sets X and Y is denoted by $X \cap Y$, i.e., $X \cap Y = \{x | x \in X \wedge x \in Y\}$.
- **Subset:** $X \subseteq Y$ represents that the set X is a subset of the set Y, i.e., $X \cap Y = X$.

- **Powerset:** The power set of a set X, i.e., $P(X) = \{Y|Y \subseteq X\}$ is denoted by $P(X)$. For example, given the set $X = \{1, 2, 3\}$, then the power set of X, i.e., $P(X) = \{\varnothing, \{1\}, \{2\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$.

3.2 Event Logs

Most process mining programs begin with event logs as their foundation. For example, information systems used by large institutions depict their daily operations. These processes' execution information is recorded in event logs that include all relevant information for a particular type of process. Depending on the specific use case, many types of information may be stored regarding a process. However, some generalisations apply to event logs: An event log comprises various traces. Each trace is a series of activities that describes a particular instance of the process execution. Each event, frequently represented as a row in an event log, includes details on a specific executed activity. In addition, each event includes details on the executed activity and the execution time, or timestamp, of the event. Each event in a trace is sorted sequentially according to its unique timestamp because every event has one. Each event has additional properties that can include the resource (user) that carried out a specific action or other details about the process.

For example, The table 3.1 displays two cases, A and B, together with the transactions, timestamps and activities connected to each event in each case. The events are shown in chronological sequence for case A and are identifiable by their event IDs.

Case ID	**Event ID**	**Transaction**	**Timestamp**	**Activity**
A	111	start	2/11/20 1:00	Register Patient
	111	complete	2/11/20 1:02	Register Patient
	123	start	2/11/20 1:03	Transfer Patient
	124	start	2/11/20 1:04	Call Nurse
	124	complete	2/11/20 1:06	Call Nurse
	123	complete	2/11/20 2:02	Transfer Patient
B	211	start	3/11/20 1:00	Call Doctor
	211	complete	3/11/20 1:02	Call Doctor
	223	start	3/11/20 1:03	Examine Patient
	224	start	3/11/20 1:04	Diagnose
	223	complete	3/11/20 1:06	Diagnose
	224	complete	3/11/20 2:02	Examine Patient
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Table 3.1: An example event log from the case study of hospital data.

Following, we provide formal definitions for the attributes of the event log elements that were previously discussed.

Definition 3.4 (Event). Let E represent the set of all potential event IDs or the event universe. Each event may contain several attributes, such as the timestamp, an activity that describes it, a resource that uses it, etc. We specify the following properties required for our method for each event e in E:

- **Unique Identifier of Event:** The unique identifier of an event e is denoted by $\pi_{id}(e)$.
- **Timestamp of event:** The timestamp of event e is denoted by $\pi_{time}(e)$.
- **Transaction type of event:** The transaction type of event e is denoted by $\pi_{trans}(e)$.
- **Instance id of event:** The instance id of an event is denoted by $\pi_{instance}(e)$.
- **Case identifier of event:** The case/trace identifier of event e is denoted by $\pi_{case}(e)$.

Definition 3.5 (Trace). Let E represent the universe of events. We define σ as the trace of an event log consisting of multiple events with the same case id, given the sequence $\sigma \in \mathcal{P}(E)$. Therefore, requiring $\forall \sigma \in \mathcal{P}(E)$ and $\forall 1 \leq i < j \leq |\sigma| : \pi_{case}(\sigma(i)) = \pi_{case}(\sigma(j))$.

As a trace is a finite sequence of events $\sigma \in \mathcal{P}(E)$, each event appears only once in a trace, i.e., $\forall 1 \leq i < j \leq |\sigma| : \sigma(i) \neq \sigma(j)$.

Definition 3.6 (Event Log). An event log is a collection of cases L that belongs to the power set of C, denoted by $\mathcal{P}(C)$. Here, C represents the set of all possible case identifiers. It is important to note that in an event log, each event occurs at most once, meaning that for any $c, c' \in L$ where $c \neq c'$, the intersection between the set of all events associated with case c and the set of all events associated with case c' is empty, denoted by $\partial(c) \cap \partial(c') = \varnothing$.

3.3 K-means Clustering

This work's fundamental idea is the grouping of comparable objects into categories. Various clustering methods divide the initial data objects into several groups. K-means clustering is one approach that does this. Due to its simplicity, speed, scalability, clustering quality and robustness, K-means clustering is a preferred clustering technique above others. It is a quick and effective technique that can deal with noisy data and huge datasets with plenty of features. It also builds clusters with slight intra-cluster variance and is less susceptible to initialisation. Furthermore, k-means is an excellent method for exploratory data analysis since it generates clusters that are simple to understand. Also, it requires a clustering parameter, k, which allows the user to divide the data points into the desired number of clusters.

Given a set of n data points $X = \{x_1, x_2, x_3, x_4..., x_n\}$ and k representing the desired number of clusters:

1. Initialise k cluster centroids $\mu_1, \mu_2, \mu_3, \mu_4..., \mu_k$ randomly.
2. Using Euclidean distance metric, assign each data point x_i to the nearest centroid μ_j using the:

   $C(i) = \arg\min_{j=1,2,3,4,...,k} ||x_i - \mu_j||^2$,

   where $C(i)$ represents the assignment of a cluster of the ith data point.
3. For each cluster, re-calculate the centroids μ_j as the mean of all data points assigned to it: $\mu_j = \frac{1}{|S(j)|} \sum_{x_i \in S(j)} x_i$,

where the set of data points assigned to the cluster j is represented by $S(j)$.

4. Until the algorithm converges, repeat steps 2 and 3, i.e., either the maximum number of iterations is reached or the cluster assignment no longer changes.

The output of the K-means clustering algorithm is k clusters, each having a set of data points assigned to it.

3.4 Random Forest Classifier

A random forest classifier is an ensemble learning method for classification problems that integrates various decision trees into a single model. Each sample in a training batch of n samples contains a collection of m features and a class label y corresponding to it.

The random forest classification method operates as follows:

1. Based on sections of the initial training data randomly chosen with replacement, a user-defined number of decision trees (n_estimators) are generated.
2. A random subset of features (max_features) is chosen at each node of every decision tree to find the optimal split.
3. The decision trees are expanded as large as possible (until the bare minimum of samples needed to split a node is achieved), producing diverse trees.
4. A new sample is sent to each decision tree for classification, and the majority decision of the trees determines the class.
5. The mode of each decision tree prediction makes up the final class prediction.

The random forest algorithm aims to decrease overfitting and boost accuracy by adding randomisation to the decision tree building. The output of the random forest classifier is a predicted class label for the new sample and a measure of the prediction confidence. The technique also enables feature importance evaluation, which aids in feature selection and interpretability.

3.5 Trace Clustering

A process mining technique called trace clustering seeks to cluster related traces (sequences of events) in an event log based on distance or similarity metrics. The trace clustering algorithm operates as follows, given a log L of n traces. Each trace t represents a sequence of m events and a distance or similarity measure d (t1, t2) between any two traces t1 and t2 [54].

1. Create a n x n distance or similarity matrix by calculating the distance or similarity between each pair of traces in the log.
2. To group the traces into k clusters, where k is a user-defined value, use a clustering method (such as k-means, hierarchical clustering) to the distance or similarity matrix.
3. Based on the same distance or similarity metric, assign each trace to the cluster with the closest centroid, the cluster's mean or median trace.

4. Post-processing the generated clusters by splitting or merging them by specified criteria is optional. (e.g., minimum cluster size, maximum cluster diameter).

The trace clustering technique produces k clusters, each containing a collection of related traces. The clusters can be examined to learn more about how the underlying process behaves, spot trends and abnormalities, and enhance process efficiency. Process mining applications frequently employ trace clustering, including process identification, conformance testing, and performance analysis.

3.6 Silhouette Score

The silhouette score [55] is a metric that compares an object's similarity to its cluster to those of other clusters. A higher score suggests better-defined clusters and offers a mechanism to rate the quality of the clusters. A score of 1 means the object is well-matched to its cluster and poorly matched to others. The silhouette score can vary from -1 to 1. A score of -1, on the other hand, denotes an object that is well-matched to nearby clusters but poorly matched to its cluster. A silhouette score greater than 0.5 often denotes a successful clustering strategy, whereas a score lower than 0.5 may indicate that the clusters are ill-defined or overlap [56]. It is crucial to understand the score in the context of the issue because the best value of the silhouette score can vary depending on the dataset and the objectives of the investigation.

Chapter 4

Method

In addressing the challenge of choosing an alpha attribute from existing attributes (phase 1 of PCM), we propose four techniques that guide analysts to select one or more alpha attributes. Two new techniques rank existing trace attributes, while the other two perform trace clustering to instantiate a new alpha attribute each.

We propose two methods to select an existing alpha attribute: one based on an unsupervised learning algorithm, i.e. k-means clustering. We name this method ID-K. And another is based on a supervised machine learning algorithm, i.e. the Random Forest Classifier [1], we name this method ID-R. Each method returns a graph of the relative importance of each attribute. In addition, we rank the attributes by the feature importance as a guide to the user: higher values indicate more influence on the classification. Instead of selecting the alpha attribute from existing attributes, we introduce two methods that instantiate new alpha attributes to group the homogeneous traces into similar groups; we name these methods IN-V and IN-P, respectively.

Our proposed clustering methods, i.e. ID-k, IN-V and IN-P, depend on determining the optimal number of clusters in which the data points can be grouped. The Elbow method [57] is commonly used to identify the optimal number of clusters. It is based on the notion that the optimal number of clusters is one that, while maintaining a very low variability between clusters, minimises the within-cluster sum of squares (WCSS) [58]. The Within-Cluster Sum of Squares (WCSS) metric is used in clustering algorithms to quantify the distribution of data points within each cluster. To compute WCSS, the squared distances of each data point in a cluster to the centroid of that cluster are summed up. Clustering algorithms strive to accomplish tightly grouped clusters around their centroids by minimising WCSS. The elbow method plots the WCSS against the number of clusters and checks for the inflexion point or "elbow" in the plot. This elbow shows that increasing the number of clusters does not notably lower the WCSS, indicating a levelling down of the variability between clusters. For instance, Figure 5.5 shows the elbow graph for one of our evaluations. In this graph, the elbow is at the number of clusters = 2; after that, there is no notable decrease in the WCSS. Another measure used to evaluate the quality of clustering findings is the Silhouette score [55]. Like the Elbow method, the Silhouette score for various cluster sizes is calculated. The number of clusters that produce the highest

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`

overall Silhouette score is chosen to determine the ideal number of clusters. When trying to group data points into clusters, it's essential to know how well they work. That's where the Silhouette score comes in. It gives us a score between -1 and 1, which tells us how well each data point fits into its assigned cluster. The higher the score, the better the clustering result. However, a negative score means the data point has been put in the wrong cluster. We need to calculate the Silhouette score for different cluster numbers to determine the best number of clusters for a dataset. We want to find the number of clusters with the highest overall Silhouette score. But, the best number of clusters is not the same as the total number of data elements. The amount of clusters with the highest overall Silhouette score is what we are looking for. However, the optimal number of clusters might differ from the overall data points. The ideal number of clusters will rely on the distinct features of the dataset because clustering algorithms aim to group similar data points while reducing variability within clusters. When a data point is assigned to a cluster that does not match well, it is sometimes called an "incorrect" cluster assignment [59]. For instance, a negative Silhouette score indicates that the data point is more similar to other data points in a different cluster than those in the cluster to which it was assigned [60]. When this occurs, we may need to reconsider our clustering strategy for that dataset.

4.1 ID-K: k-means Clustering

Clustering is a technique to group data into clusters based on their similarity in certain features [61]. Our analysis starts with an XES event log and considers the event attributes. Numeric, boolean and categorical features are considered, while identifiers, timestamps, descriptions and comments are not considered. It is recommended to remove such attributes as they do not provide helpful information about the content of the data points and can introduce bias in the analysis [62]. This is because these features may comprise information that is not actual to the factual content of the data points, leading to skewed clustering algorithm results [63]. For instance, using identifiers or timestamps in clustering analysis may result in forming specific clusters based on the chronological order of events rather than the actual similarity of the data points. Furthermore, using these features can also impact the algorithm's scalability, making processing extensive datasets computationally expensive and challenging [64]. Additionally, these features may need to provide more valuable information about the content of the data points. As a result, they can yield noise in the data, thereby reducing the accuracy of clustering results. Conversely, eliminating these fields can enhance the accuracy of clustering results and augment the algorithm's scalability. This is because the clustering algorithm can concentrate on relevant characteristics that are more informative about the content of the data points. Furthermore, by transforming data into a format that machine learning algorithms can process, such as using factorisation[2] for categorical and boolean features, we can encode the variable as an enumerated type or categorical variable, which can then be utilised as input for clustering analysis [62].

Once the data has been transformed, the next step is determining the optimal number of clusters through the Elbow method. Then the k-means clustering algorithm is applied to the selected features, which groups traces in homogeneous groups based on the selected features. Finally, the contribution of each attribute in segregating the traces into different

[2]https://pandas.pydata.org/docs/reference/api/pandas.factorize.html

groups is studied and a plot of the relevant importance of each feature is plotted as the output of the method. This is done by calculating the relevance of each feature for each cluster based on the magnitude of the weight of the feature in the centroid vector [3]. This method allows for the identification of the most important features that contribute to the separation of the clusters.

4.2 ID-R: Random Forest Classifier

We leverage a classification algorithm to investigate the influence of each trace attribute on a specific process-based feature: the length of a trace (activity count), as choosing the number of activities as a target feature can provide meaningful insights into the behaviour and performance of processes, anomaly detection and process prediction. We chose this feature to categorise the data into various categories. We want the outcome to factor in the effect of the trace length as a different number of activities carried out for a case may have interesting reasons, so we extract the count of activities per case and add it as the target feature in the data set. For example, analysing the activity count may help us identify possible process bottlenecks, compare the efficiency of different process variants, detect anomalous process behaviours and predict the outcome of the process based on trends and patterns of past behaviours. This enables us to classify the information into different categories depending on the number of activities executed in each case and detect the outliers with notably higher or lower activity counts. We then train a random forest classifier on all the features other than the target feature (activity count). After training, the classification model categorises the data points into the targeted classes. We choose to utilise an ensemble classifier, specifically the Random Forest classifier. The Random Forest classifier is an ensemble learning method that combines multiple decision trees to enhance the overall classification performance of the algorithm [65]. It operates by training multiple decision trees on randomly selected subsets of the data and then taking the average of the predictions of all the trees to make a final prediction. This approach helps to minimise over-fitting and improves the model's accuracy. Over-fitting may occur in process mining when the model is too complex and fits the training data too closely, which results in poor generalisation of new data [66]. The Random Forest classifier is also robust to the outliers, which are extreme data values that vary considerably from the anticipated pattern [67]. When there are unusual process behaviours, such as deviations from the anticipated order of activities or unexpected activity counts, outliers can appear in the setting of process mining [5]. In terms of the degree of variability or spread of the values in the data, the Random Forest classifier also manages high variance. When there are numerous potential process paths or variations that result in a broad range of activity counts and other process attributes, high variance can happen in process mining. Combining multiple decision trees, the Random Forest technique can capture complex relationships between process attributes and activity count, even in high variance [67].

The Random Forest classifier selects a feature that provides the maximum information gain or the minor impurity to divide the data into more homogeneous groups. This selection procedure is repeated until no more examples of the same class label exist in the leaf nodes. We can determine the feature utilised at each decision node and assess how well it distinguishes between meaningful classes by taking the root node, which has the maximum

[3] https://towardsdatascience.com/interpretable-k-means-clusters-feature-importances-7e516eeb8d3c

information gain. In a random forest classifier, the importance of features is calculated using the mean decrease impurity (MDI) method [68], which calculates the total reduction of Gini impurity that each feature provides across all the trees in the forest. The feature importance is then determined by averaging the reduction in impurity across all trees that use the feature. However, a feature's relative importance might still be influenced by where it is in the decision tree and how it affects how instances are categorised. As a result, it's crucial to assess the random forest classifier's performance on a validation set and interpret the outcomes in the context of the particular process mining issue.

4.3 Instantiating a new alpha attribute

Apart from finding the alpha attribute from the existing set of case attributes, we propose another activity-based approach to divide the main event log into sub-event logs using a trace clustering method. Trace clustering splits the event log into homogeneous sub-logs based on the pathway followed through the process by the trace [44]. To enable clustering, a collection of characteristics can be used to describe the trace: events, the frequency of events, length, order and other pertinent characteristics. We enhanced the trace clustering using the activity profiling approach proposed by Minseok Song and Wil Van der Aalst in [54]. Trace profiling is a method for turning a trace into a vector. However, the set-based depiction of activity profiles considers the presence of activity in a trace rather than how frequently it occurs. To get around this, a hybrid technique that combines a feature vector that provides the frequency of common activities conducted per trace with another vector that reveals the presence of activities in a trace is employed in this research. This helps to improve the robustness of the clustering algorithm. Based on this trace profile vector, we use a clustering method, i.e. k-means clustering, to group the traces into homogeneous sub-groups. The number of clusters, i.e., k, is calculated by the Elbow method or the Silhouette score. The optimal number of clusters, k, maximises the average Silhouette over a range of possible values for k.

An example of vector representation of an event log with eight activities where activities' A' and 'B' may be repeated in the same trace is shown in Figure 4.1. We create a trace profiling vector for each trace. This means that a vector of 10 elements is generated for each trace. The first eight elements represent one of the eight activities in the event log and the last 2 represent the count of the repetitive activities. This vector can provide information about the patterns or sequences of activities in the trace. After converting the event log into a vector representation, the clustering algorithm can be applied to the vectorised data. The algorithm subsequently produces cluster labels for each trace, which are added to each trace as a new attribute. This attribute is then returned as the alpha attribute.

We propose two methods to instantiate a new alpha attribute based on the trace clustering technique mentioned earlier. The first method involves utilising the vectorised data generated through the trace profiling approach, which we refer to as IN-V. The second method involves processing the vectorised data using Principal Component Analysis (PCA) to reduce its dimensionality, and we name this method IN-P. Principal component analysis, or PCA, is a mathematical method that facilitates complex data analysis by minimising the number of variables that must be considered. For example, imagine having a variety of cars, each with a different engine power, paint, seating capacity and speed. It would help

Traces	Vector (Set + Frequent Activity count)
A -> B -> B -> C ->H	< 1,1,0,0,1,0,0,1,1,2 >
B -> B -> B -> D -> F	< 0,1,0,1,0,1,0,0,0,3>
A -> E -> F -> G -> H	< 1,0,0,0,1,1,1,1,1,0 >

Figure 4.1: Vector representation.

if you comprehended how these various elements are connected so that you can organise comparable cars. PCA can help to find the most critical variables that can differentiate one type of car from another. It accomplishes this by combining the initial variables to create additional "principal components." Next, we can use these principal components to compare the cars because they represent the critical patterns in the data. For instance, the most crucial elements that set one type of car apart are its engine power and speed. Then, with the aid of PCA, we can make a new variable that blends these two variables, allowing you to classify cars according to their speed and power. Table 5.2 shows a PCA-transformed vector representation for one of our evaluations.

Chapter 5

Evaluation

This section outlines the implementation and two evaluations we conducted to confirm the applicability and usefulness of the methods and their implemented tool support.

5.1 Implementation

The proposed methods have been implemented as Python scripts [1] using PM4PY framework [2] and scikit-learn[3] and can be executed by following the instructions in the Readme.md file. In addition, the required dependencies have been mentioned in the Requirements.txt file. The output of the extraction of the alpha attribute (ID-K & ID-R) is two graphs showing the relative importance of each attribute. The instantiating alpha attribute script (IN-V and IN-P) takes an XES log as input and returns an XES log file with cluster labels appended as trace attributes.

5.2 Applicability

We first evaluate the approaches' applicability using them on a real-world event log Sepsis, made available to the public. Then, we illustrate its capability to offer valuable insights into the possible divergences in processing that can arise within a single process. In addition, this evaluation demonstrates how the approach can be leveraged to improve healthcare procedures in intricate settings, underscoring its versatility for deployment in diverse domains.

5.2.1 Data Exploration

Sepsis, a condition characterised by the body's harmful response to infection, is a frequent cause of severe illness and death worldwide [69]. The dataset[4] consists of 1050 patient cases recorded between 2013 and 2015 and includes diagnostic test results, patient demographic information and organisational information. We apply the methods to

[1] `https://github.com/asadTariq666/BPM-Alpha-Attribute-Selection`
[2] https://pm4py.fit.fraunhofer.de/
[3] https://scikit-learn.org/stable/
[4] `https://data.4tu.nl/articles/dataset/Sepsis_Cases_-_Event_Log/12707639`

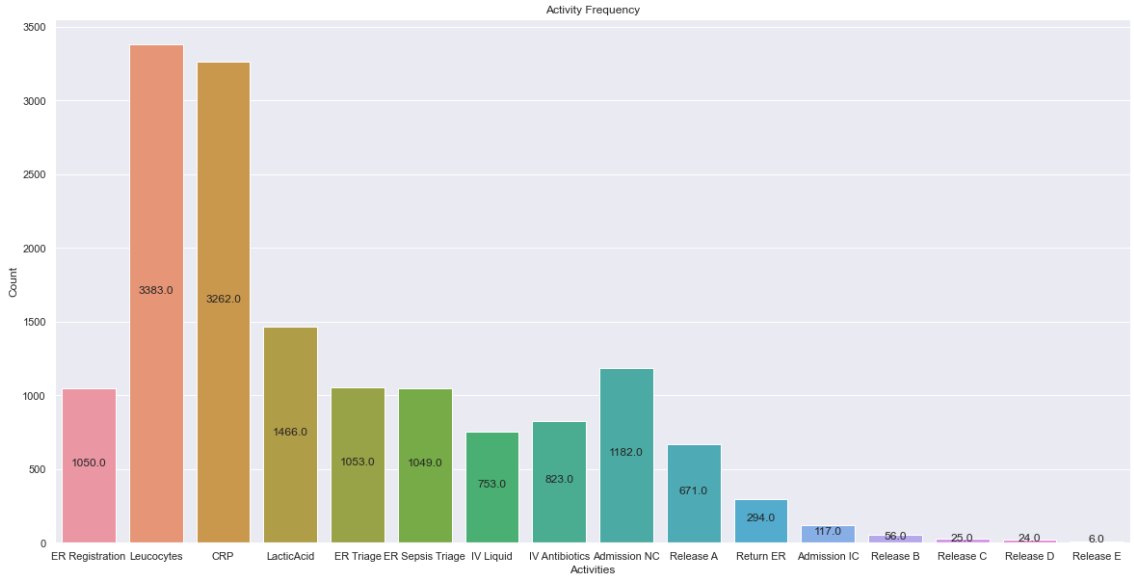


Figure 5.1: Sepsis - Activity count

understand the diagnostic journey of sepsis patients and identify factors that may affect patient outcomes.

The data source format is an XES event log [70] and patient identifiers are kept anonymous using alphabetical case IDs. To read the files, the open-source PM4Py library [71], which supports process mining algorithms in Python, was utilised within a Jupyter Notebook environment. An examination of a selection of records was conducted and determined that the event log file contained 1050 patient cases, with an average of 14 events per patient in their diagnostic journey. A total of 16 unique activities were observed in the process, with the activities of Leucocytes and CRP being the most frequent, occurring almost three times per case. The frequency of each activity is shown in Figure 5.1. Additionally, there were three types of Registration and Triage activities present in the event log, and they were found to be a part of almost every case,

- ER Registration
- ER Triage
- ER Sepsis Triage

There are six possible activities on which the patient's journey ends: either the patient is released with a release code, i.e. `Release A`, `Release B`, `Release C`, `Release D` or `Release E` or returned to the emergency room, i.e. `Return ER`. `Release A` is the most frequent release type, whereas `Release E` is the least frequent release activity. The count of each end activity is shown in Figure 5.3. The event log also contains trace attributes for each case. Following are the trace and event attributes in the event log,

- **Age:** Age of the patients has been anonymised in bins of 5 years. It ranges from 20 to 90 in the data.
- **Case:** Unique Identifier for a patient.

- **Activity:** 16 Activities as discussed above.
- **Transition:** Completion state of Activity/Event.
- **Organisation:** Organisation or Resource carrying out the activity.
- **Diagnose(Type of Diagnose):** 144 unique diagnoses in the event log.
- **Diagnostic Tests:** 22 Boolean Tests with True/False as result values indicate whether the test was conducted for the patient.

The high-level process map of the Sepsis event log is shown in Figure 5.2. A high-level process map is a graphical representation of a business process that presents a high-level picture of the flow of the process, including its activities, decisions and results. A high-level process map helps identify process optimisation opportunities and inform data-driven decisions.

5.2.2 Data pre-Processing

The event log has five distinct release types, i.e. patient discharges. After applying a filter to exclude cases that lacked any release activity, as these cases were deemed incomplete and lacked a definitive end activity, 777 cases remained. We added a trace attribute denoting the release type to the event log. The event log contains trace attributes such as `Case ID`, `Age`, `Transition`, `Organization`, `Activity Count`, `Diagnose`, and 22 `Diagnostic Tests`. After removing the non-contributing trace attributes (case identifier, comments and timestamps), we have 26 trace attributes for our analysis. These features have some missing values, shown in Figure 5.4. Missing data or null values can harm machine learning algorithms by introducing bias, decreasing the sample size and causing errors [72]. Therefore, we removed traces with missing attributes by performing an additional filtering step and the processed event log is left with 729 cases. The values of boolean and categorical features, i.e. diagnostic tests and release type are factorised into enumerated types so that our methods can process them.

5.2.3 Assisted Alpha Attribute Selection

Both methods ID-K and ID-R have been applied to the 26 trace attributes to extract the alpha attributes.

ID-K:

Once the data is in the format our machine learning algorithm can work on, we feed this data to our clustering algorithm. The Elbow method in Figure 5.5 and the Silhouette score in Figure 5.6 helped to determine the optimal number of clusters, i.e. 2 for the clustering algorithm. Based on the clustered data, we analysed the features on which clusters were formed. The resulting graph 5.7 of ID-K shows that `age` and `Diagnose` played an essential role in segregating traces into clusters.

ID-R:

As discussed in Section 4.2, we use the ensemble classification method, the Random Forest classifier, to find the important features in the classification method. In this method, we select `Activity Count` as the target feature, while all other 25 features are selected as

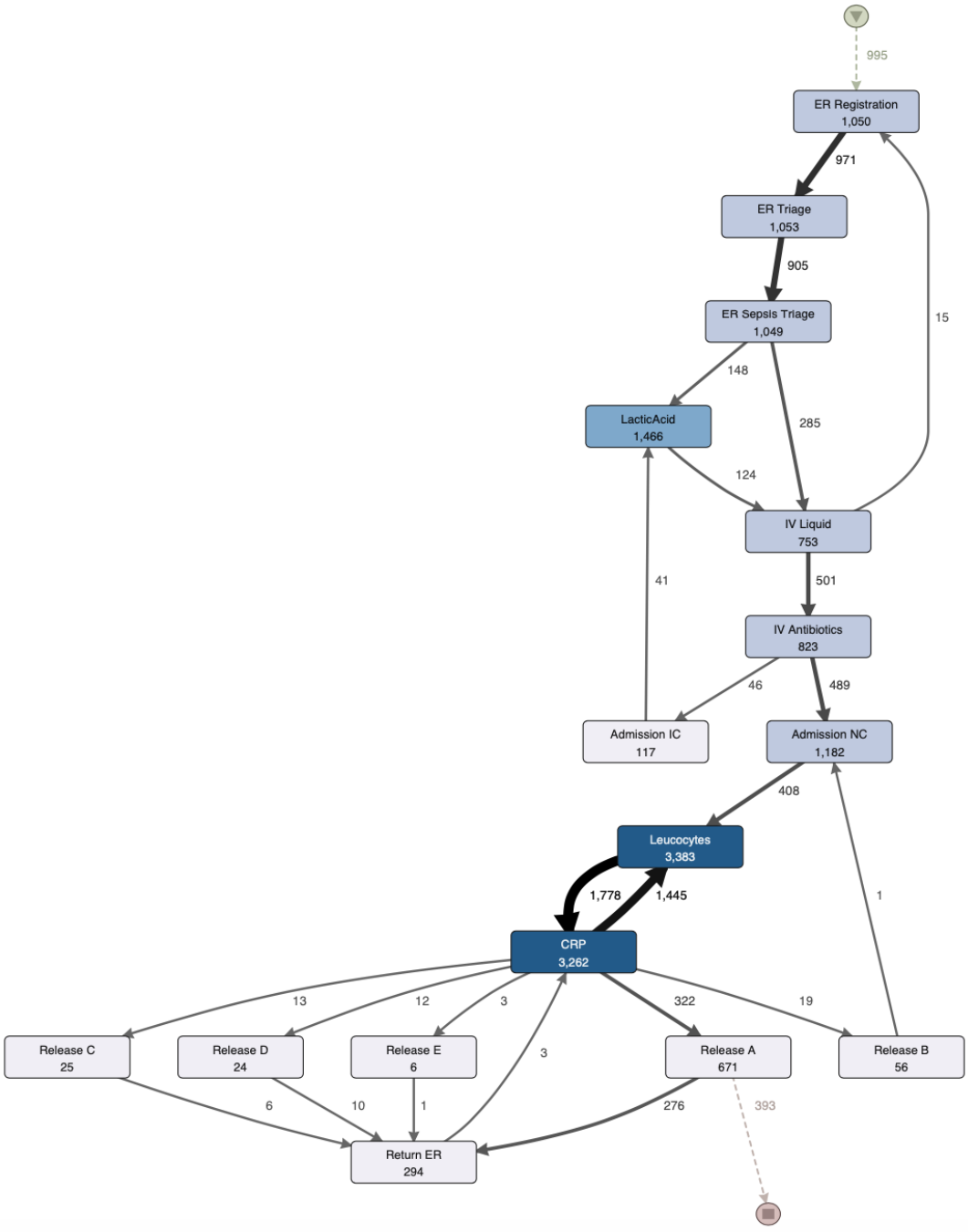


Figure 5.2: Sepsis - Process Map

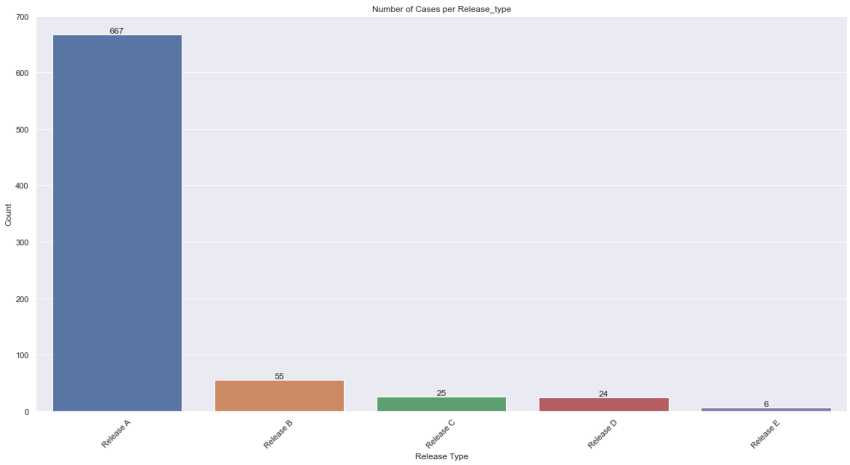


Figure 5.3: Sepsis - Release Activities

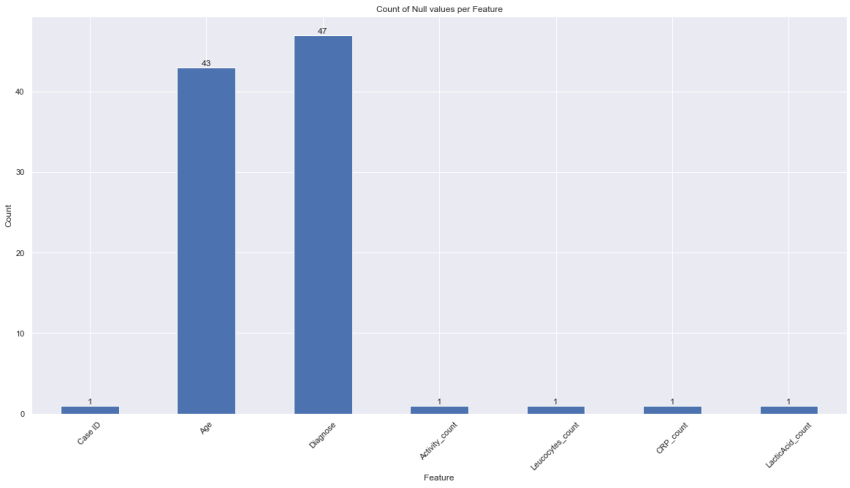


Figure 5.4: Sepsis - Features Null Values

training features. After applying the Random Forest classifier, the resulting graph 5.8 shows that out of the 25 training trace attributes, `Diagnosis` and `Age` are the most impactful in classifying the target feature, Activity Count.

5.2.4 Instantiating a new alpha attribute

In our study, we developed a methodology for instantiating new alpha attributes based on the original and dimensionally reduced vector representation. This was accomplished by transforming the vectors through techniques such as dimensional reduction via principal component analysis (PCA). Subsequently, the newly transformed data was subsequently utilised to instantiate new alpha attributes in cluster labels, creating a modified event log with these new alpha attributes.

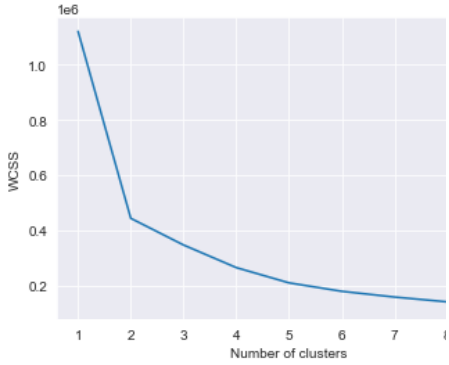


Figure 5.5: Elbow Method

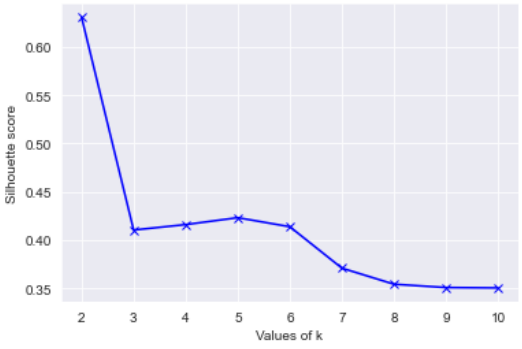


Figure 5.6: Silhouette Score

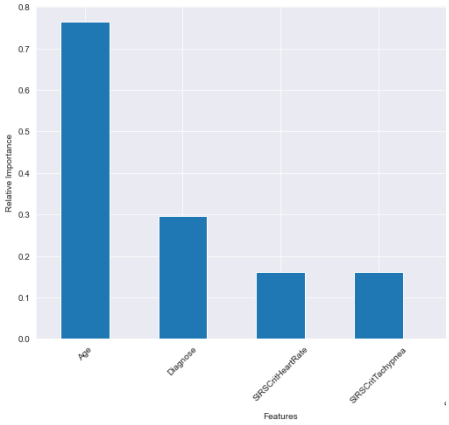


Figure 5.7: ID-K: Important Features

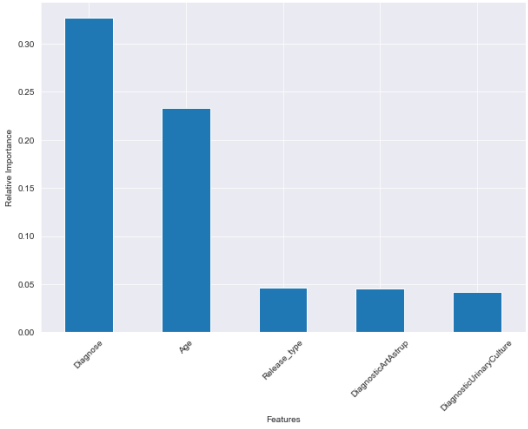


Figure 5.8: ID-R: Important Features

IN-V:

As described in Section 4.3, we formed a vector based on the set of activities and the count of repeating activities. Among the 16 total activities, five occurred more than once in some cases. These activities are `Leucocytes`, `CRP`, `LacticAcid`, `Admission NC` and `Admission IC`. We created our input vector by merging vectors `Set of Activities` and `Frequent Activity Count`. The vector representation of the first 5 cases in the event log is shown in the table 5.1. The Silhouette score method graph shows that the score increases with the increase in the number of clusters 5.9. The Silhouette score method returns an adequate cluster number with the maximum score. Next, we apply k-means clustering on the transformed dataset. The k-means clustering algorithm analyses this vector representation. Each trace in the event log is assigned a cluster label and a new event log is returned with a cluster label instantiated as the alpha attribute.

IN-P:

We apply Principal Component Analysis on the vector representation from the IN-V method and reduce the dimensions of the data set while retaining as much of the variation in the data as possible. The principal components = 7, are calculated by the Explained

Table 5.1: Sepsis Merged Vector (Set + Frequent Activity Count)

Sepsis:

Case ID	Vector
A	<0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 7, 7, 1, 1, 1, 0>
B	<0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0>
C	<0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0>
D	<0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0>
E	<0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 5, 4, 1, 1, 1, 0>

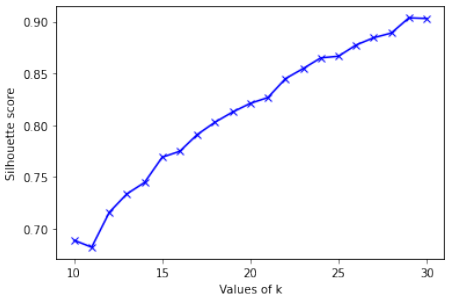


Figure 5.9: IN-V: Silhouette Score

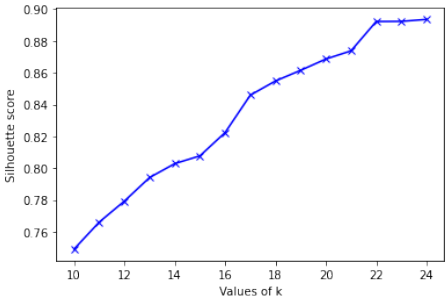


Figure 5.10: IN-P: Silhouette Score

Variance method [73], shown in Figure 5.11. Based on the principal components number, PCA is applied and the dimensions of the dataset are reduced. The top 5 cases of the transformed vector after applying PCA are shown in Table 5.2. The Silhouette score method graph shows that the score increases with the increase in the number of clusters 5.10. Therefore, the Silhouette score method returns an adequate cluster number with the maximum score. Next, we apply K-means clustering on the transformed dataset. Each trace in the event log is assigned a cluster label and a new event log is returned with a cluster label instantiated as the alpha attribute.

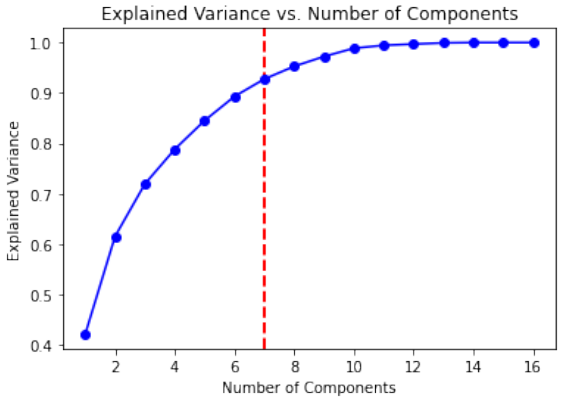


Figure 5.11: IN-P: Principal Components vs Explained Variance

5.2.5 Scoping Analysis

In the selected candidate alpha attributes, `Diagnose` has more than 100 distinct values, and we take the top ten most frequent ones `C`,`B`,`E`,`H`,`G`,`D`,`K`,`R`,`Q` and `S`. As a result, we have ten sublogs, one for each selected diagnosis. For example, for `Age`, we partition the values

Table 5.2: SEPSIS: PCA Transformed Vector

SEPSIS - PCA:

Case ID	PCA Transformed Vector
A:	<-0.47, -0.12, -0.42, -0.14, -0.05, -0.03, 0.03>
B:	<0.40, -0.95, 0.28, -0.17, -0.11, -0.23, -0.04>
C:	<-0.75, 0.30, 0.43, -0.20, 0.05, 0.03, 0.01>
D:	<0.40, -0.95, 0.28, -0.17, -0.11, -0.23, -0.04>
E:	<-0.47, -0.12, -0.42, -0.14, -0.05, -0.03, 0.03>

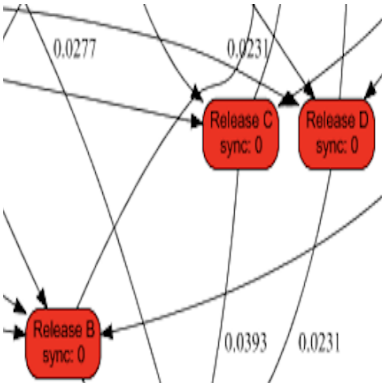


Figure 5.12: Sepsis Age Comparison

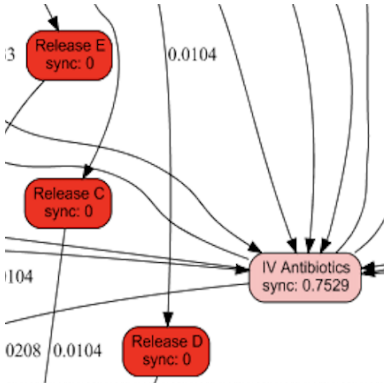


Figure 5.13: Sepsis Diagnosis Comparison

into 10-year periods, resulting in eight sublogs and they are 0-20, 20-30, 30-40, 40-50, 50-60, 60-70, 70-80 and 80-90.

5.2.6 Process Comparison and Results

To evaluate the approach, we use PROM plugins Process comparator plugin [74] and Earth Movers' Stochastic Conformance Checking plugin [75]. The methods `ID-K` and `ID-R` suggested `Age` and `Diagnose` as the alpha attributes, we compared the journeys of youngest and eldest age groups in our analysis, i.e. Age 20-30 vs Age 80-90. Figure 5.12 (see also A.1) shows the differences between the cohorts mentioned above. The sync = 0 shows that the activity is non-existent in one of the logs. It is clear that the patient journey for elderly patients is quite different than that of young patients, where no patient from the younger group has gone through the activities `Release B`, `Release C`, `Release D` and `Release E`.

Similarly, we compared two cohorts of diagnosis, i.e. `B` and `I`. Figure 5.13 (see also A.2) shows the differences between these cohorts, patients diagnosed with diagnosis `I` have not gone through the activities `Release C`, `Release D` and `Release E`.

Also, we presented the concept and results to a healthcare expert. She showed interest in the idea and recommended that grouping patients by diagnosis or age would be helpful. With this strategy, patient outcomes can be improved overall and treatments and care plans can be tailored to particular patient needs. The endorsement letter with her comments is in the appendix section A.6.

Furthermore, we also evaluated the performance of our proposed clustering methods, IN-V and IN-P, on the sepsis data set. To compare the sublogs generated by these clustering methods, we analysed the differences between the sublogs generated by different cluster

labels. For instance, we compared the sublogs generated by IN-V with cluster labels 12 and 15; `sublog12` has 59 traces and `sublog15` has 51. The analysis shows that the activities `IV liquid`, `IV Antibiotics` and `Lactic Acid` are absent in sublog12. Figure 5.14 shows the difference between sublog12 and 15, focusing on the activity IV liquid. Similarly, we compared the sublogs produced by the IN-P clustering technique. According to our analysis, the sublog4 and sublog12 that the IN-P technique produced contained 42 and 52 traces, respectively. Notably, our research showed that the two sublogs shared only three activities. In addition, eight activities in sublog12 were not in sublog4, and five activities were absent from both sub logs as shown in Figure 5.15 (see also A.3). Overall, our evaluation results demonstrate the effectiveness of our proposed approach in analysing patient journey data and generating sublogs that capture the process flow of different patient groups. These sublogs can be further used for process analysis and improvement and can assist healthcare professionals in making informed decisions for personalised patient care. In addition, these findings provide insight into the many processes and flow connected to various patient groups, which can impact interventions in personalised healthcare.

By evaluating our approach using PROM plugins and receiving expert feedback, we have demonstrated that patient journeys for older patients and those diagnosed with certain illnesses may differ significantly from younger patients or those with different diagnoses. Our proposed clustering techniques, IN-V and IN-P, have also produced promising results in creating sublogs that capture the process flow of distinct patient groups. These sublogs can offer valuable insights for healthcare professionals, aiding in making informed decisions for personalised patient care and process enhancement. Our research has improved comprehension of the various processes and connections associated with different patient groups. We emphasise the significance of categorising patients based on age and diagnosis for personalised healthcare interventions.

5.3 Usefulness

The second evaluation reveals how the approach can offer significant insights into the diversities arising during a single process, thereby improving healthcare procedures in intricate settings. The results also demonstrate the potential of this approach to be utilised in various other domains by applying our methods in a case study performed at the emergency department (ED) admissions of Beth Israel Deaconess Medical Center.

5.3.1 Data Exploration

MIMIC-IV-ED is an emergency department (ED) admissions database at the Beth Israel Deaconess Medical Center. It includes vital signs, triage data, medication reconciliation, administration, and discharge diagnoses of roughly 425 000 ED stays. Human care is rationed in the emergency department (ED) to provide the best patient care feasible, as emergency departments are usually environments with limited resources [76]. The data is recorded from 2011 to 2019 and is derived from MIMIC [76]. The data-set includes structured data such as demographics, vital signs, laboratory values, medications and diagnoses and unstructured free-text clinical notes.

MIMIC-IV information is divided into four major categories:

- **Patient data** contains essential demographic information like age, gender, ethnicity,

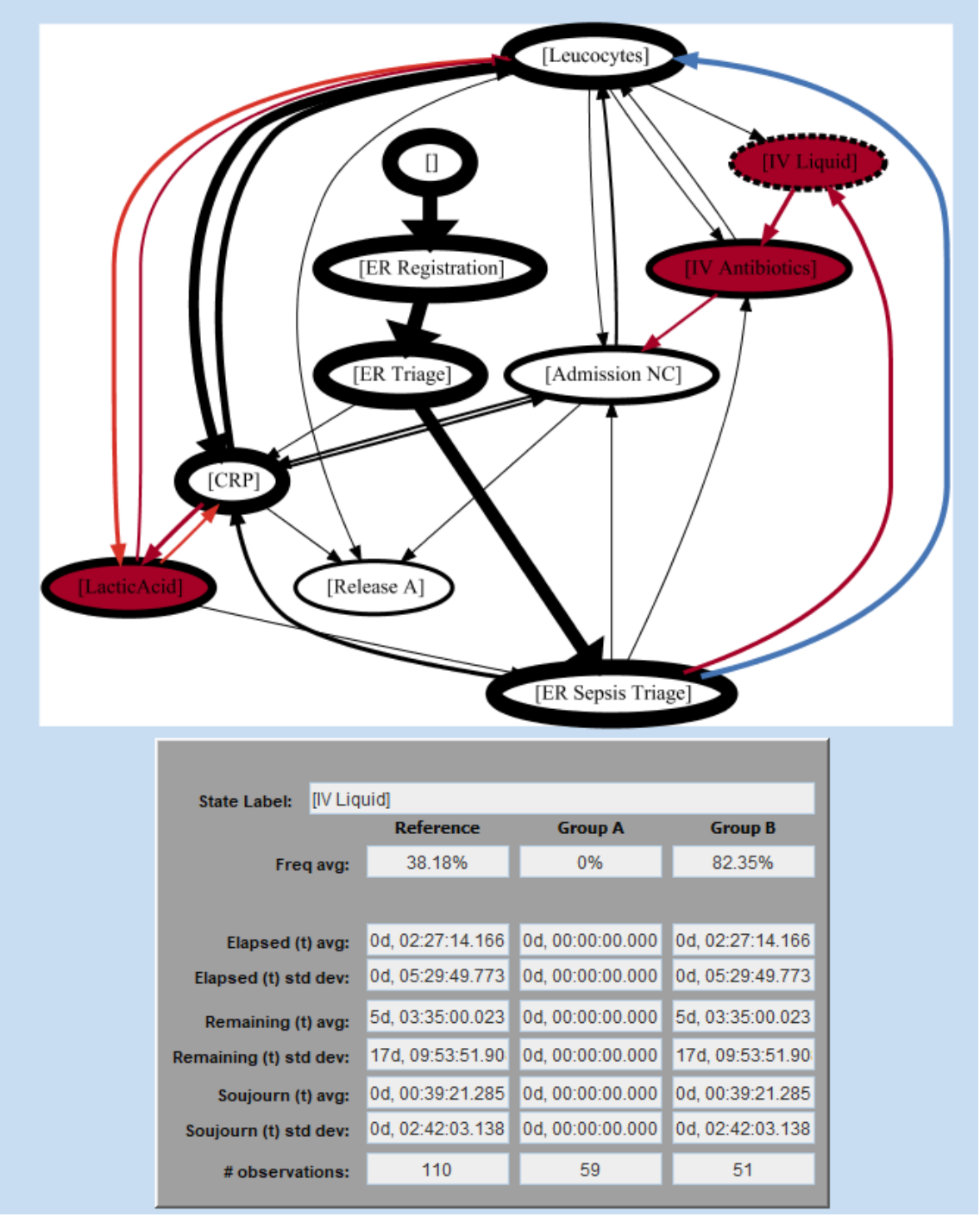
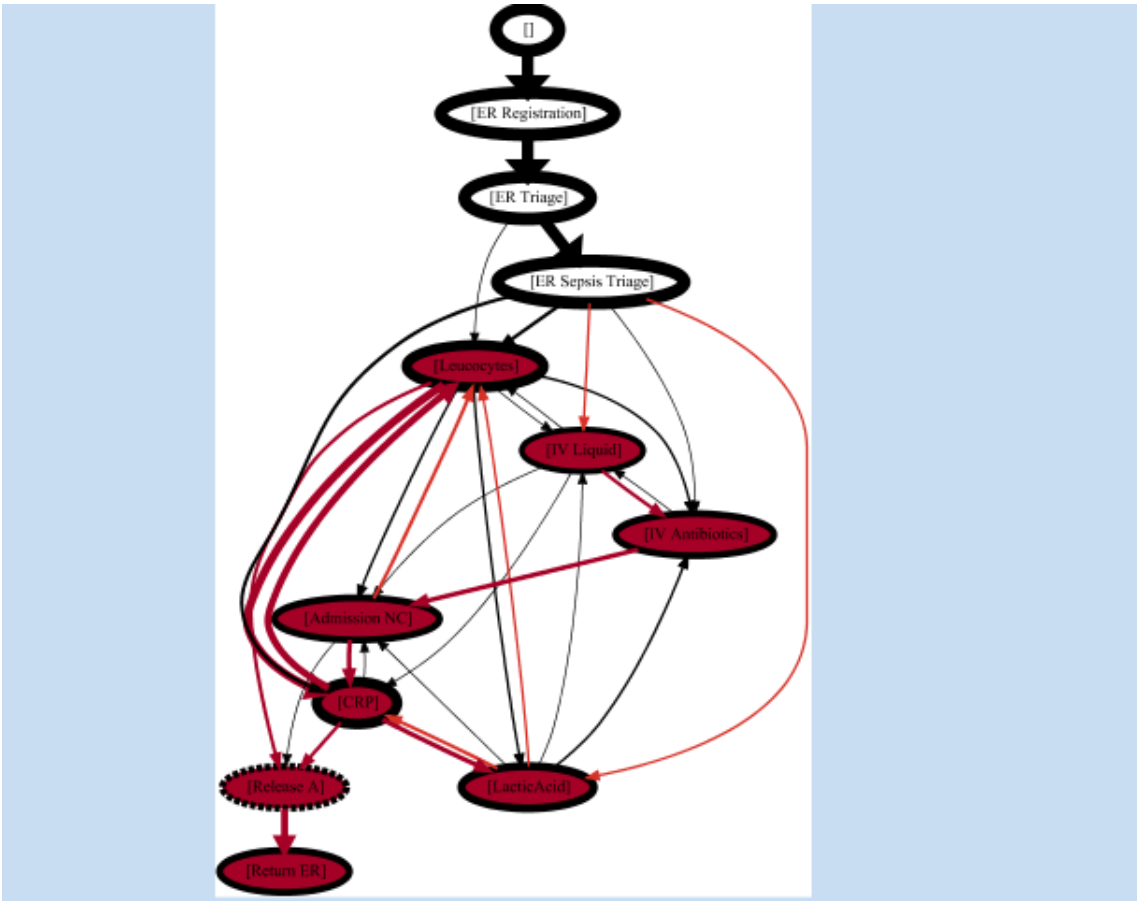


Figure 5.14: Sepsis-IN-V: Sublogs 12 and 15



State Label: [Release A]

	Reference	Group A	Group B
Freq avg:	53.19%	0%	96.15%
Elapsed (t) avg:	7d, 11:27:11.360	0d, 00:00:00.000	7d, 11:27:11.360
Elapsed (t) std dev:	5d, 19:14:12.041	0d, 00:00:00.000	5d, 19:14:12.041
Remaining (t) avg:	78d, 12:40:42.32	0d, 00:00:00.000	78d, 12:40:42.32
Remaining (t) std dev:	101d, 17:02:57.0	0d, 00:00:00.000	101d, 17:02:57.0
Soujourn (t) avg:	78d, 12:39:51.74	0d, 00:00:00.000	78d, 12:39:51.74
Soujourn (t) std dev:	101d, 17:03:36.7	0d, 00:00:00.000	101d, 17:03:36.7
# observations:	94	42	52

Figure 5.15: Sepsis-IN-P: Sublogs 4 and 12

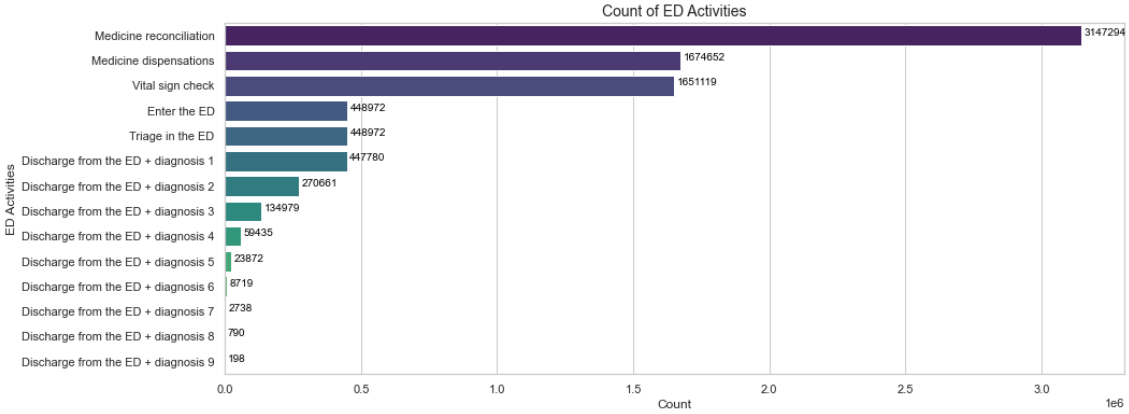


Figure 5.16: Mimic - Activity Count

admission and discharge dates, and clinical information like diagnoses, procedures and comorbidities.

- **Clinical data** contains structured and unstructured data, such as vital signs, laboratory values, medications, fluid balance, nursing notes, physician progress notes and discharge reports.
- **Administrative data** include hospital stay information such as ICU admission and discharge dates, hospital duration of stay, and insurance information.
- **Researchers Data** contains information about how the MIMIC-IV database was used, such as the researcher's name, affiliation, contact information, and details about the research project and any publications that resulted from the use of the data.

MIMIC-IV stores metadata in a hierarchical structure that represents the data's organisation. Each data category is subdivided into tables containing information about particular data elements. Tables for vital signs, laboratory measurements and medications, for example, are included in the clinical data group. There are 14 activities performed in the process and the breakdown of them are shown in Figure 5.16.

5.3.2 Data pre-Processing

We selected the event attributes `temperature`, `heartrate`, `resprate`, `o2sat`, `sbp`, `dbp`, `pain`, `acuity`, `chiefcomplaint` and `icd_title` and lifted them to the trace level. Other attributes with a high percentage of null values or identifiers, timestamps or comments were dropped; 10 trace attributes were used further. Out of the initial 448 972 cases, 436 737 cases remained after filtering traces with missing values. Table 5.3 shows the count of the top ten `icd_title` and `chiefcomplaint`, respectively, in the data-set.

5.3.3 Assisted alpha attribute selection

To extract the alpha attributes, ID-K and ID-R have been applied to the ten trace attributes.

Table 5.3: Count of Top 10 Diagnosis and Chief Complaints

Diagnosis:

Diagnosis	Count
HYPERTENSION NOS	28354
Essential (primary) hypertension	22398
Chest pain, unspecified	13749
CHEST PAIN NOS	13120
DIABETES UNCOMPL ADULT	12711
Unspecified abdominal pain	11096
Type 2 diabetes mellitus without complications	9241
ABDOMINAL PAIN OTHER SPECIED	9226
Fall on same level, unspecified, initial encounter	8218
UNSPECIFIED FALL	7447

Chief Complaints:

Complaint	Count
Chest pain	11930
Abd pain	11348
Dyspnea	6286
ABD PAIN	5277
s/p Fall	5227
SI	5219
ETOH	4991
Wound eval	4776
Headache	4052
Back pain	3350

ID-K:

Once the data is in the format our machine learning algorithm can work on, we feed this data to our clustering algorithm. The Elbow method in Figure 5.17 helped to determine the optimal number of clusters, i.e. 3 for the clustering algorithm. Based on the clustered data, we analysed the features on which clusters were formed. The resultant graph5.18 of ID-K shows that `icd_title` and `Acuity` played an important role in segregating traces into differnt clusters.

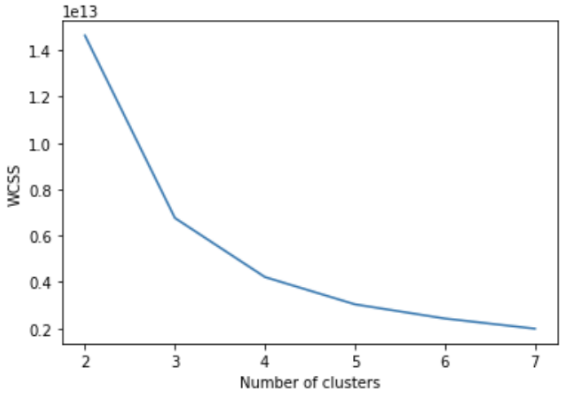


Figure 5.17: Mimic Elbow Method

5.3.3.1 ID-R:

As discussed in Section 4.2, we use the ensemble classification method, the Random Forest classifier, to find the critical features in the classification method. In this method, we select `Activity Count` as the target feature, while all other nine features are selected as training features. After applying the Random Forest classifier, the resulting graph 5.19 shows that out of the nine training trace attributes, `chiefcomplaint`, `icd_title` and `heartrate` are the most impactful in classifying the target feature, Activity Count.

5.3.4 Instantiating a new alpha attribute

In our study, we developed a methodology for instantiating new alpha attributes based on the original and dimensionally reduced vector representation. This was accomplished by transforming the vectors through techniques such as dimensional reduction via principal component analysis (PCA). Subsequently, the newly transformed data was utilised to instantiate new alpha attributes in cluster labels, creating a modified event log with these

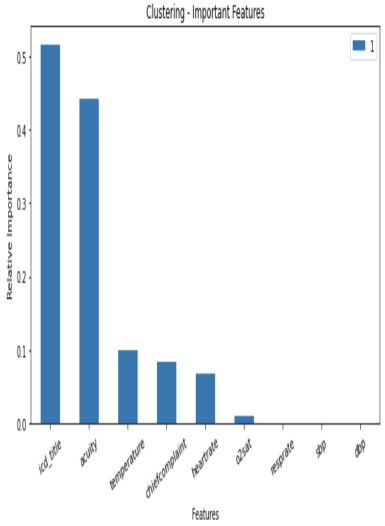


Figure 5.18: ID-K: Important Features

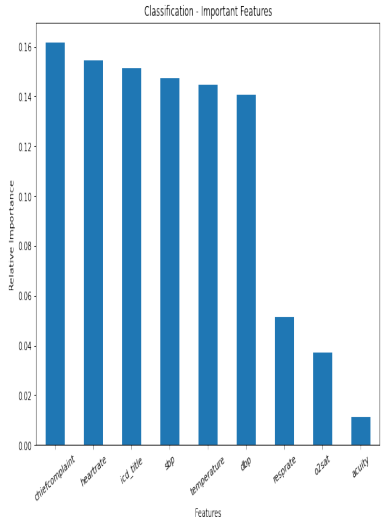


Figure 5.19: ID-R: Important Features

Table 5.4: MIMIC Merged Vector (Set + Frequent Activity Count)

	Case ID	Vector
MIMIC:	30000177:	<1 , 1 , 1 , 1 , 1 , 1 , 1 , 1 , 1 , 1 , 1 , 0 , 0 , 0 , 15 , 6 , 1>
	30000252:	<1 , 1 , 1 , 1 , 1 , 1 , 1 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 7 , 5 , 6>
	30000443:	<1 , 1 , 1 , 0 , 1 , 1 , 1 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 12 , 1 , 2>
	30000573:	<1 , 1 , 1 , 1 , 1 , 1 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 11 , 14 , 2>
	30000679:	<1 , 1 , 1 , 1 , 1 , 1 , 1 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 10 , 13 , 4>

new alpha attributes. We were dealing with a huge data set with vast data in it. As a result, we decided to sample a data set to make it easier to handle. We made sure, though, that the sample we selected reflected the entire data collection and contained the same amount of activities as in the original event log. This allowed us to analyse the data effectively without compromising the accuracy of the data we were using. The filtered event log has 21 154 traces and total 401 457 events. The high-level process map of the Mimic event log is shown in Figure 5.20.

IN-V:

As described in Section 4.3, we formed a vector based on the set of activities and the count of repeating activities. Among the 14 total activities, three occurred more than once in some cases. These activities are `Vital sign check`, `Medicine dispensations` and `Medicine reconciliation`. We created our input vector by merging vectors `Set of Activities` and `Frequent Activity Count`. The vector representation of the first 5 cases in the event log is shown in the table 5.4. The Silhouette score method graph shows that the score increases with the increase in the number of clusters 5.21.

IN-P:

We apply Principal Component Analysis on the vector representation from the IN-V method and reduce the dimensions of the data set while retaining as much of the variation in the data as possible. First, the principal components = 6, are calculated by the Ex-

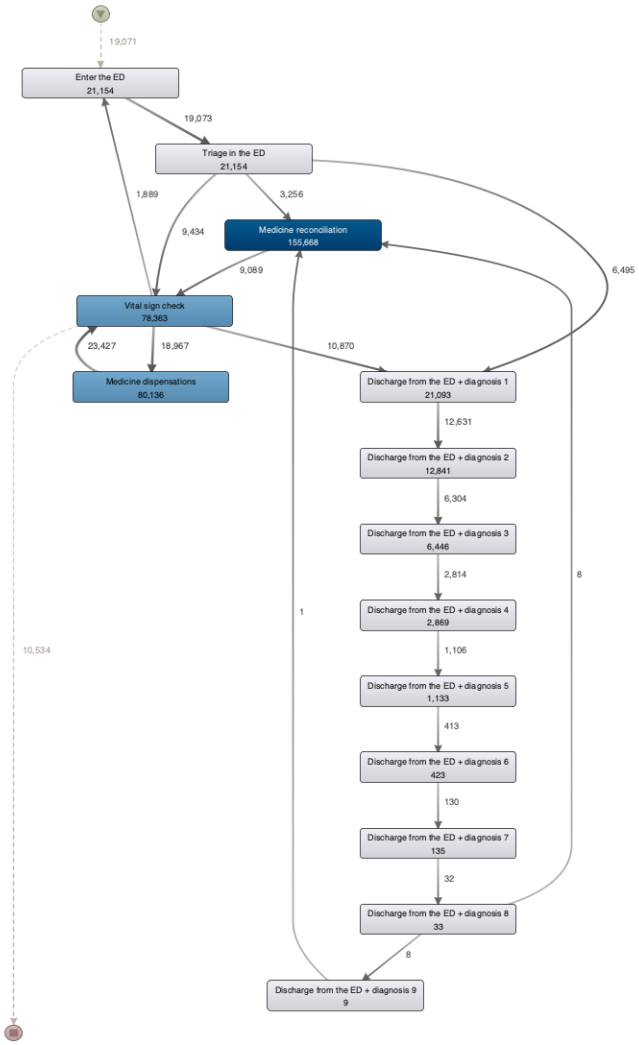


Figure 5.20: Mimic - Process Map

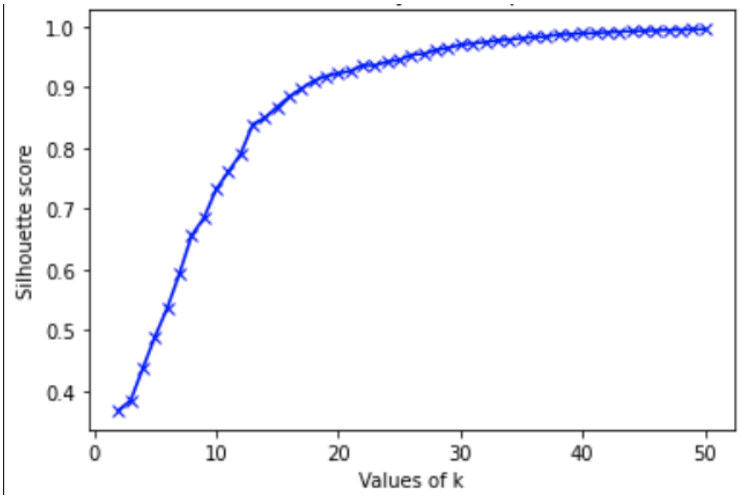


Figure 5.21: IN-P: The Silhouette Score

plained Variance method [73], shown in Figure 5.22. Based on the principal components number, PCA is again then applied and the dimensions of the dataset are reduced. The top 5 cases of the transformed vector after applying PCA are shown in Table 5.5. Next, we apply K-means clustering on the transformed data set. Each trace in the event log is assigned a cluster label and a new event log is returned with a cluster label instantiated as the alpha attribute.

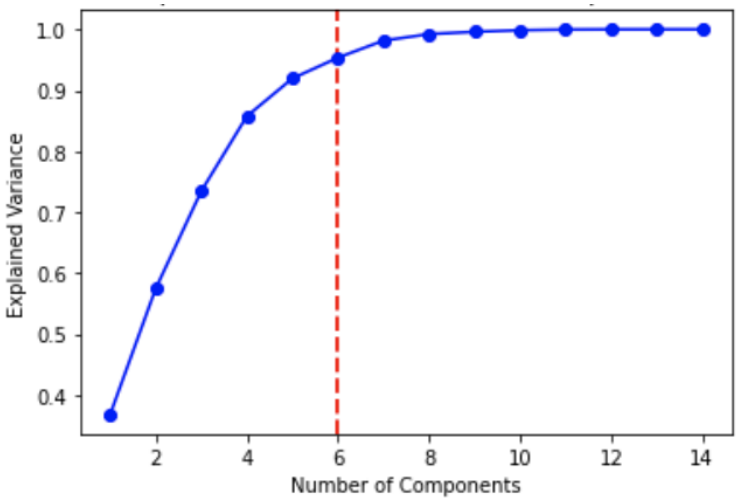


Figure 5.22: IN-P: Principal Components vs Explained Variance

5.3.5 Scoping analysis.

Here, we find comparable sublogs based on the `icd_title` attribute, which has more than a thousand values, it is impossible to compare these all one-on-one. So we select the top ten most frequent values and generate ten event logs by filtering the event log based on these values.

Table 5.5: MIMIC: PCA Transformed Vector

	Case ID	PCA Transformed Vector
MIMIC - PCA:	30000177:	<1.29 , -0.27 , -0.02 , 0.75 , 0.7 , 0.04>
	30000252:	<0.05 , -0.35 , 0.02 , -0.51 , 0.2 , 0>
	30000443:	<-0.02 , 0.42 , -0.61 , -0.51 , 0.2 , -0.05>
	30000573:	<-0.58 , -0.5 , -0.09 , 0.2 , -0.04 , -0.01>
	30000679:	<-0.58 , -0.5 , -0.09 , 0.2 , -0.04 , -0.01>

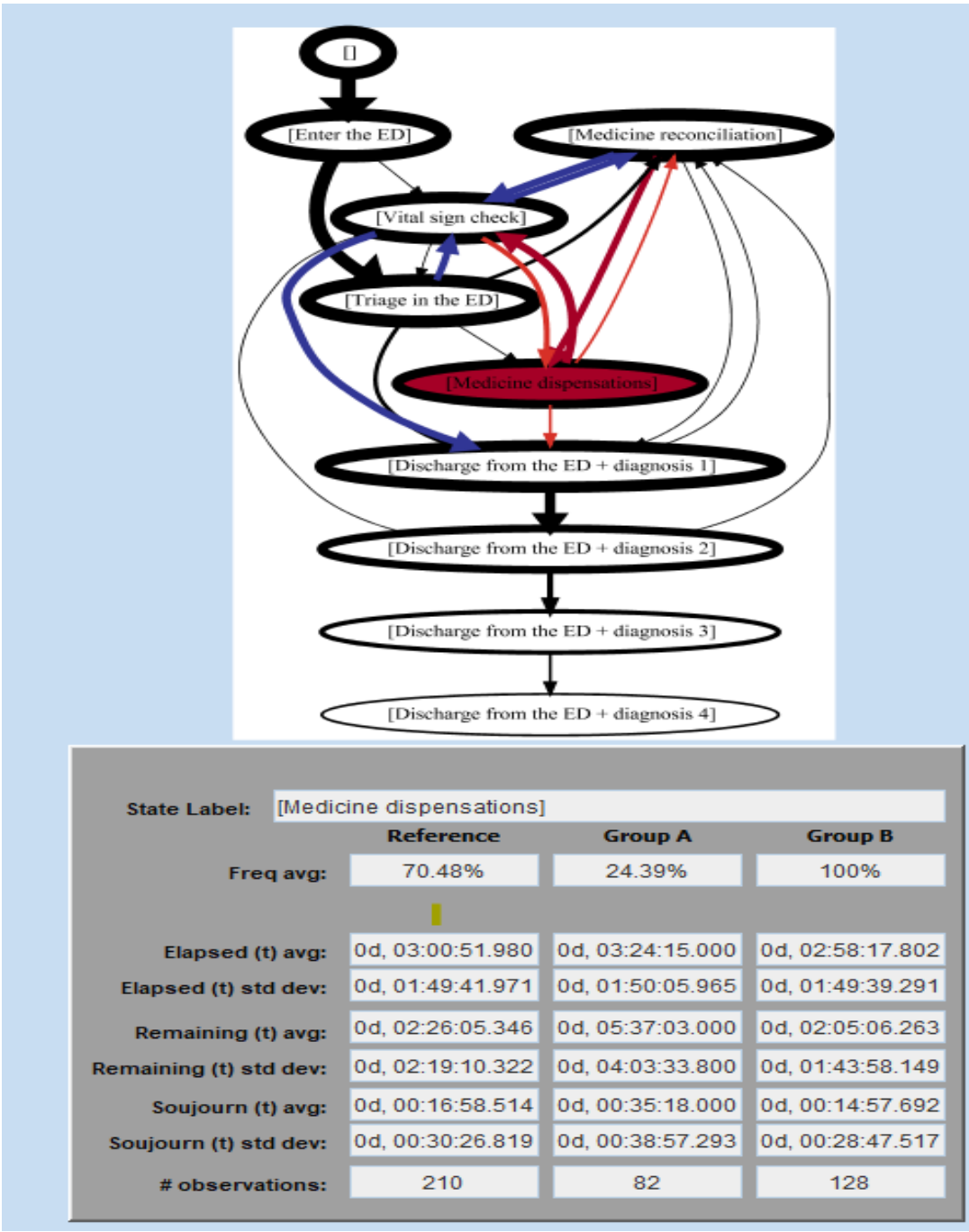
5.3.6 Process Comparison and Results

For our evaluation of the MIMIC event log, we evaluated the performance of our proposed clustering methods, IN-V and IN-P. To compare the sublogs generated by these clustering methods, we analysed the differences between the sublogs generated by different cluster labels. For instance, we compared the sublogs generated by IN-V with cluster labels 10 and 17; `sublog10` has 82 traces and `sublog17` has 128. The analysis in Figure 5.23 (see also A.4) shows the activity `Medicine Dispensations` is present in all 128 traces of sublog17 whereas its frequency in sublog10 is not even 25% as compared to that of sublog17. Also, the most frequent activity path followed by `Vital Sign Check` activity in the sublog10 is `Medicine Dispensations` and for sublog17 is `Medicine reconciliation`.

Similarly, we compared the sublogs, sublog7 and sublog13 produced by the IN-P clustering technique. The analysis in Figure 5.24 (see also A.5) shows `Medicince reconciliation` is present in all traces of sublog13 whereas its frequency in sublog7 is just over 25% as compared to sublog13.

Overall, our evaluation results demonstrate the effectiveness of our proposed approach in analysing patient journey data and generating sublogs that capture the process flow of different patient groups. These sublogs can be further used for process analysis and improvement and assist healthcare professionals in making informed decisions for personalised patient care. In addition, these findings provide insight into the many processes and flow connected to various patient groups, which can impact interventions in personalised healthcare.

By evaluating our approach using PROM plugins and receiving expert feedback, we have demonstrated that patient journeys for older patients and those diagnosed with certain illnesses may differ significantly for patients with different diagnoses. Our proposed clustering techniques, IN-V and IN-P, have also produced promising results in creating sublogs that capture the process flow of distinct patient groups. These sublogs can offer valuable insights for healthcare professionals, aiding in making informed decisions for personalised patient care and process enhancement. Our research has improved comprehension of the various processes and connections associated with different patient groups. We emphasise the significance of categorising patients based on age and diagnosis for personalised healthcare interventions.



State Label:	[Medicine dispensations]		
	Reference	**Group A**	**Group B**
Freq avg:	70.48%	24.39%	100%
Elapsed (t) avg:	0d, 03:00:51.980	0d, 03:24:15.000	0d, 02:58:17.802
Elapsed (t) std dev:	0d, 01:49:41.971	0d, 01:50:05.965	0d, 01:49:39.291
Remaining (t) avg:	0d, 02:26:05.346	0d, 05:37:03.000	0d, 02:05:06.263
Remaining (t) std dev:	0d, 02:19:10.322	0d, 04:03:33.800	0d, 01:43:58.149
Soujourn (t) avg:	0d, 00:16:58.514	0d, 00:35:18.000	0d, 00:14:57.692
Soujourn (t) std dev:	0d, 00:30:26.819	0d, 00:38:57.293	0d, 00:28:47.517
# observations:	210	82	128

Figure 5.23: Sepsis-IN-V: MIMIC 10 and 17

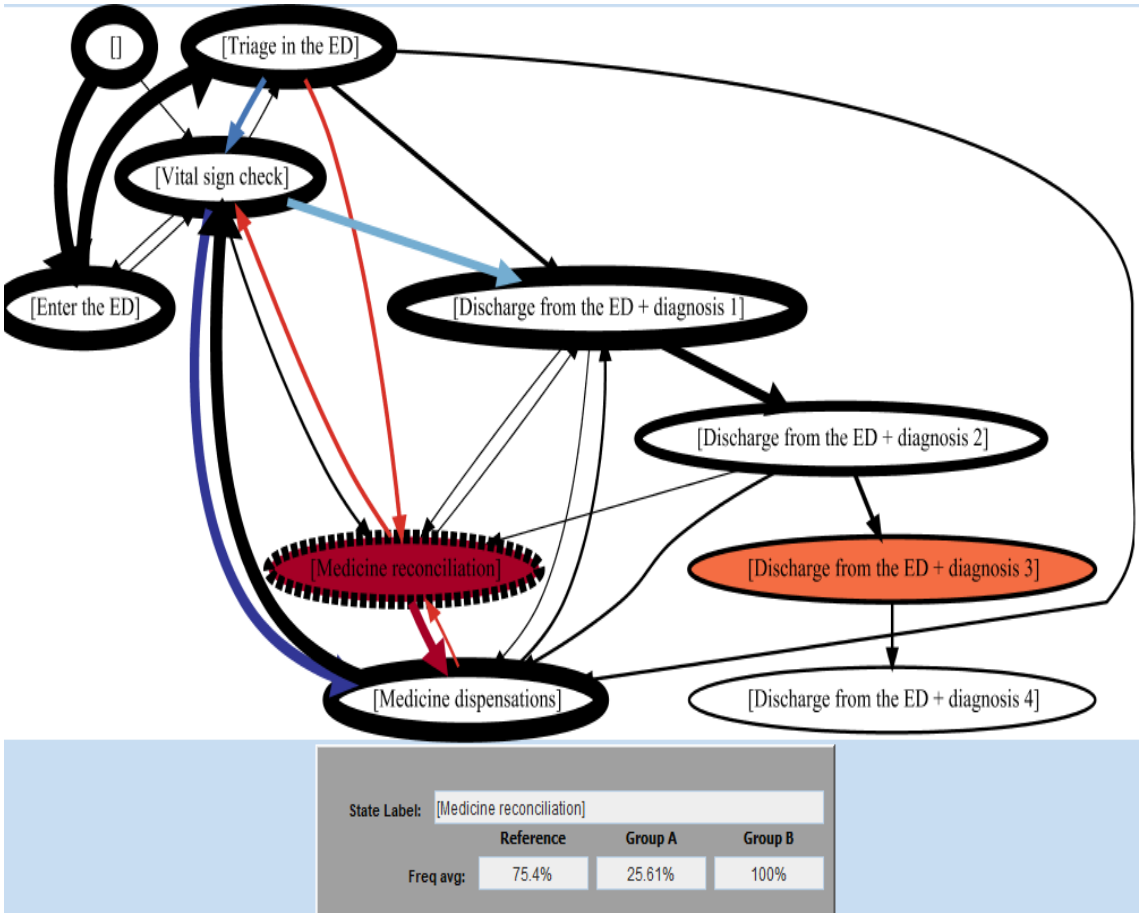


Figure 5.24: MIMIC-IN-P: Sublogs 7 and 13

Chapter 6

Discussion

An approach for analysing patient journey data and creating sub-logs for various patients was described in the initial assessment. The assessment used the PROM plugins Process Comparator and Stochastic Conformance Checking to evaluate the efficacy of the method. Two patient cohorts were compared in the evaluation based on age and diagnosis, and the results were discussed with a healthcare professional who positively responded to the approach. The examination also examined the sub-logs produced after using the IN-V and IN-P clustering algorithms to the Sepsis data set. The results showed that the suggested method is effective for looking at patient journey data and creating sub-logs that show the process flow of various patient groups. This can assist healthcare professionals in making defensible choices about individualised care and process improvement. Additionally, constructing sub-logs that capture the process flow of various patient groups using the proposed clustering methods IN-V and IN-P showed promising results.

The effectiveness of the suggested clustering approaches, IN-V and IN-P, was examined in more detail in the second evaluation, which focused on the MIMIC event log. The evaluation contrasted and examined the differences among the sub-logs produced by various cluster labels. The results demonstrated that the suggested clustering techniques successfully generate sub-logs representing various patient groups' process flow.

Overall, both evaluations showed how well the proposed method looked at patient journey data and created sub-logs that reflect the process flow of different patient groups. For healthcare professionals, these sub-logs can offer insightful information to help them make defensible choices concerning individualised patient treatment and process improvement. The evaluations also showed how well the suggested clustering techniques produce sub-logs that depict the process flow of various patient groups.

Although the study's findings are encouraging, some limitations must be considered. The method of evaluation employed is one of the key drawbacks. Although the PROM plugins Process Comparator and Earth Movers' Stochastic Conformance Checking help assess the suggested method, they might only be appropriate for some data-set types. For instance, using these plugins to evaluate the technique's efficacy could be challenging when the data is highly complicated.

The use of clustering as a method of analysis also has limitations. Although the results from the IN-V and IN-P clustering algorithms were encouraging, they might only be

appropriate for some data-set types. Furthermore, these clustering techniques might not produce reliable sub-logs, for example, when the data is heavily contaminated with noise. In light of the peculiarities of the data set being examined, it is crucial to thoroughly assess the applicability of the clustering approach.

Processing large data-sets for the study can also take longer. Large amounts of data must be processed as part of the suggested approach, which could take some time. Therefore, it might be required to use powerful processing equipment to speed up the process in circumstances where the data is extensive. This drawback emphasises the need to consider the necessary computational resources when putting the suggested technique into effect.

While the study's findings are encouraging, it's vital to consider the limitations mentioned above. Nevertheless, healthcare practitioners can effectively apply the suggested strategy and clustering methods and make judgments for individualised patient care and process improvement by realising these limits. Additionally, future studies might address these constraints to increase the effectiveness and adaptability of the suggested technique in other healthcare settings

Chapter 7

Conclusion

The study aimed to develop a framework for selecting alpha attributes or creating new ones using machine learning methods in the healthcare domain. Therefore, the model's performance has been assessed using the Mimic and Sepsis datasets. In addition, the model's effectiveness on various datasets was compared as part of the evaluation procedure. The following discussion summarises the study's contributions, limitations and future research directions.

Firstly, it created a new technique by applying machine learning techniques for choosing alpha attributes that blend feature engineering with feature selection. Secondly, The IN-P and IN-V methods are proposed for alpha attribute instantiation using the trace clustering technique. Homogeneous traces are grouped in a similar cluster and a cluster label is assigned to the group, which is treated as the new alpha attribute. Both methods aim to improve the performance of process mining models by providing additional information about the traces. Compared to manual feature selection methods, the suggested method has a substantial benefit: it is automated, scalable and reproducible.

To evaluate the proposed methods, they were tested on two datasets, Sepsis and Mimic. The results showed that alpha attributes suggested by ID-K and ID-R and new alpha attributes instantiated by IN-V and IN-P could divide event logs into smaller groups to make the processes more transparent and devise group-specific treatment plans. The approach was also discussed with a healthcare expert and she also endorsed the approach.

The study also found several risks to the validity of the suggested methodologies, including user reliance, generalisability to other datasets, unpredictability in data quality and ease of use. Therefore, these restrictions should be considered when utilising the approaches on different datasets.

The proposed methods offer a practical strategy for picking and instantiating alpha attributes from event logs. Still, more investigation is required to evaluate their applicability to other datasets and to resolve the issues raised in this work.

Future Work: Future research might build on the suggested approaches and assess them in broader datasets to examine how well they function in various settings and domains. To further improve the precision and effectiveness of predictive modelling, hybrid approaches that combine the above methodologies with other feature selection and feature

engineering techniques can be investigated. Additionally, analysing the interpretability and explainability of the chosen or instantiated alpha characteristics can offer insight into the underlying mechanisms and raise the models' openness and credibility. The suggested techniques for extracting alpha attributes from event logs are semi-automated and involve several phases that call for user participation. Although this method has produced encouraging results, its degree of user dependence may prevent it from being widely used and scaled. Therefore, a potential direction for future work is to look into eliminating this dependency and fully automating the suggested procedures. One option is to incorporate the suggested techniques as a plug-in for the Process Mining framework (ProM), offering a standardised and approachable application interface for the techniques. By automating parameter tweaking, the plug-in could decrease the human input requirement while enhancing the approaches' effectiveness and dependability. The plug-in may also make combining the suggested methodologies with other process mining methods and tools easier, enabling a more thorough and efficient study of event logs.

In conclusion, the proposed methods have shown the potential to improve the performance of process mining and predictive modelling performance by identifying relevant attributes in event logs. To improve the accuracy and application of the approaches, further research should address the constraints and potential risks to validity. The suggested techniques can be expanded upon, hybrid approaches explored and the interpretability and explainability of the selected or instantiated alpha attributes investigated. These initiatives can assist numerous areas and applications while advancing the fields of process mining and predictive modelling.

Bibliography

[1] Wil M. P. van der Aalst. *Process Mining - Data Science in Action, Second Edition*. Springer, 2016. ISBN 978-3-662-49850-7. doi: 10.1007/978-3-662-49851-4. URL https://doi.org/10.1007/978-3-662-49851-4.

[2] Suriadi Suriadi, Ronny S Mans, Moe T Wynn, Andrew Partington, and Jonathan Karnon. Measuring patient flow variations: A cross-organisational process mining approach. In *APBPM*, pages 43–58. Springer, 2014.

[3] Jorge Munoz-Gama, Niels Martin, et al. Process mining for healthcare: Characteristics and challenges. *JBI*, 127:103994, 2022.

[4] Robert Andrews, Kanika Goel, Paul Corry, Robert Burdett, Moe Thandar Wynn, and Donna Callow. Process data analytics for hospital case-mix planning. *JBI*, 129: 104056, 2022.

[5] Wil Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, volume 136. 01 2011. ISBN 978-3-642-19344-6. doi: 10.1007/978-3-642-19345-3.

[6] Sander J. J. Leemans, Shiva Shabaninejad, Kanika Goel, Hassan Khosravi, Shazia W. Sadiq, and Moe Thandar Wynn. Identifying cohorts: Recommending drill-downs based on differences in behaviour for process mining. In *ER*, volume 12400 of *LNCS*, pages 92–102. Springer, 2020. doi: 10.1007/978-3-030-62522-1_7. URL https://doi.org/10.1007/978-3-030-62522-1_7.

[7] Alifah Syamsiyah, Alfredo Bolt, Long Cheng, Bart FA Hompes, RP Jagadeesh Chandra Bose, Boudewijn F van Dongen, and Wil MP van der Aalst. Business process comparison: A methodology and case study. In *BIS*, pages 253–267, 2017.

[8] Sander J. J. Leemans, Wil M. P. van der Aalst, Tobias Brockhoff, and Artem Polyvyanyy. Stochastic process mining: Earth movers' stochastic conformance. *Inf. Syst.*, 102:101724, 2021. doi: 10.1016/j.is.2021.101724. URL https://doi.org/10.1016/j.is.2021.101724.

[9] Mieke Jans, Michael Alles, and Miklos Vasarhelyi. Process mining of event logs in internal auditing: a case study. 10 2012. URL https://www.researchgate.net/publication/268344444_Process_mining_of_event_logs_in_internal_auditing_a_case_study.

[10] Mieke Jans, Michael Alles, and Miklos Vasarhelyi. Process mining of event logs in auditing: Opportunities and challenges. *SSRN Electronic Journal*, 02 2010. doi: 10.2139/ssrn.1578912. URL https://www.researchgate.net/publication/

228302047_Process_Mining_of_Event_Logs_in_Auditing_Opportunities_and_Challenges.

[11] Yu-Chien Ko and Hamido Fujita. *Gaussian Representations of K-Means Clusters: Case Study of Educational Process Mining of UCI*. 09 2020. ISBN 9781643681146. doi: 10.3233/FAIA200584. URL https://www.researchgate.net/publication/345099728_Gaussian_Representations_of_K-Means_Clusters_Case_Study_of_Educational_Process_Mining_of_UCI.

[12] Kingsley Okoye, Abdel-Rahman Tawil, Usman Naeem, Syed Islam, and Elyes Lamine. *Semantic-Based Model Analysis Towards Enhancing Information Values of Process Mining: Case Study of Learning Process Domain*, pages 622–633. 01 2018. ISBN eBook: ISBN 978-3-319-60618-7 Softcover ISBN: 978-3-319-60617-0. doi: 10.1007/978-3-319-60618-7_61. URL https://link.springer.com/chapter/10.1007/978-3-319-60618-7_61.

[13] Rita Marques, Miguel Mira da Silva, and Diogo R. Ferreira. Assessing agile software development processes with process mining: A case study. In *20th IEEE Conference on Business Informatics, CBI 2018, Vienna, Austria, July 11-14, 2018, Volume 1 - Research Papers*, pages 109–118, 2018. doi: 10.1109/CBI.2018.00021. URL https://doi.org/10.1109/CBI.2018.00021.

[14] Saulius Astromskis, Andrea Janes, and Michael Mairegger. A process mining approach to measure how users interact with software: an industrial case study. In *Proceedings of the 2015 International Conference on Software and System Process, ICSSP 2015, Tallinn, Estonia, August 24 - 26, 2015*, pages 137–141, 2015. doi: 10.1145/2785592.2785612. URL https://doi.org/10.1145/2785592.2785612.

[15] Arjel Bautista, Syed Akbar, Anthony Alvarez, Tom Metzger, and Marshall Reaves. Process mining in information technology incident management: A case study at volvo belgium. In *Proceedings of the 3rd Business Process Intelligence Challenge co-located with 9th International Business Process Intelligence Workshop (BPI 2013), Beijing, China, August 26, 2013*, 2013. URL http://ceur-ws.org/Vol-1052/paper2.pdf.

[16] Christian Fleig, Dominik Augenstein, and Alexander Maedche. Process mining for business process standardization in ERP implementation projects - an SAP S/4 HANA case study from manufacturing. In *Proceedings of the Dissertation Award, Demonstration, and Industrial Track at BPM 2018 co-located with 16th International Conference on Business Process Management (BPM 2018), Sydney, Australia, September 9-14, 2018*, pages 149–155, 2018. URL http://ceur-ws.org/Vol-2196/BPM_2018_paper_31.pdf.

[17] Mahendrawathi ER, Hanim Maria Astuti, and Ika Rakhma Kusuma Wardhani. Material movement analysis for warehouse business process improvement with process mining: A case study. In *Asia Pacific Business Process Management - Third Asia Pacific Conference, AP-BPM 2015, Busan, South Korea, June 24-26, 2015, Proceedings*, pages 115–127, 2015. doi: 10.1007/978-3-319-19509-4_9. URL https://doi.org/10.1007/978-3-319-19509-4_9.

[18] Alessandro Bettacchi, Alberto Polzonetti, and Barbara Re. Understanding production chain business process using process mining: A case study in the manufacturing scenario. In *Advanced Information Systems Engineering Workshops - CAiSE 2016*

International Workshops, Ljubljana, Slovenia, June 13-17, 2016, Proceedings, pages 193–203, 2016. doi: 10.1007/978-3-319-39564-7_19. URL https://doi.org/10.1007/978-3-319-39564-7_19.

[19] Stefan Tönnissen and Frank Teuteberg. Using blockchain technology for cross-organizational process mining - concept and case study. In *Business Information Systems - 22nd International Conference, BIS 2019, Seville, Spain, June 26-28, 2019, Proceedings, Part II*, pages 121–131, 2019. doi: 10.1007/978-3-030-20482-2_11. URL https://doi.org/10.1007/978-3-030-20482-2_11.

[20] R. S. Mans, Helen Schonenberg, Minseok Song, Wil M. P. van der Aalst, and Piet J. M. Bakker. 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, January 28-31, 2008, Revised Selected Papers*, pages 425–438, 2008. doi: 10.1007/978-3-540-92219-3_32. URL https://doi.org/10.1007/978-3-540-92219-3_32.

[21] Sjoerd van der Spoel, Maurice van Keulen, and Chintan Amrit. Process prediction in noisy data sets: A case study in a dutch hospital. In *Data-Driven Process Discovery and Analysis - Second IFIP WG 2.6, 2.12 International Symposium, SIMPDA 2012, Campione d'Italia, Italy, June 18-20, 2012, Revised Selected Papers*, pages 60–83, 2012. doi: 10.1007/978-3-642-40919-6_4. URL https://doi.org/10.1007/978-3-642-40919-6_4.

[22] Angelo Corallo, Mariangela Lazoi, Roberto Paiano, and Fabrizio Striani. Application of process mining in teleconsultation healthcare: Case study of puglia hospital. In *ICIST '20: 10th International Conference on Information Systems and Technologies, Lecce, Italy, 4-5June, 2020*, pages 32:1–32:13, 2020. doi: 10.1145/3447568.3448540. URL https://doi.org/10.1145/3447568.3448540.

[23] Marco Pegoraro, Madhavi Bangalore Shankara Narayana, Elisabetta Benevento, Wil M. P. van der Aalst, Lukas Martin, and Gernot Marx. Analyzing medical data with process mining: a COVID-19 case study. *CoRR*, abs/2202.04625, 2022. URL https://arxiv.org/abs/2202.04625.

[24] Zhichao Zhou, Yong Wang, and Lin Li. Process mining based modeling and analysis of workflows in clinical care - A case study in a chicago outpatient clinic. In *Proceedings of 11th IEEE International Conference on Networking, Sensing and Control, ICNSC 2014, Miami, FL, USA, April 7-9, 2014*, pages 590–595, 2014. doi: 10.1109/ICNSC.2014.6819692. URL https://doi.org/10.1109/ICNSC.2014.6819692.

[25] Gert-Jan de Vries, Ricardo Alfredo Quintano Neira, Gijs Geleijnse, Prabhakar M. Dixit, and Bruno Franco Mazza. Towards process mining of EMR data - case study for sepsis management. In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017) - Volume 5: HEALTHINF, Porto, Portugal, February 21-23, 2017*, pages 585–593, 2017. doi: 10.5220/0006274405850593. URL https://doi.org/10.5220/0006274405850593.

[26] Gustavo Bernardi Pereira, Eduardo Alves Portela Santos, and Marcell Mariano Corrêa Maceno. Correction to: Process mining project methodology in healthcare: a case study in a tertiary hospital. *Netw. Model. Anal. Health Informatics Bioinform.*,

9(1):44, 2020. doi: 10.1007/s13721-020-00247-6. URL https://doi.org/10.1007/s13721-020-00247-6.

[27] Gijs Geleijnse, Himalini Aklecha, Mark Vroling, Rob Verhoeven, Felice N. van Erning, Pauline A. Vissers, Joos C. A. M. Buijs, and Xander A. Verbeek. Using process mining to evaluate colon cancer guideline adherence with cancer registry data: a case study. In *AMIA 2018, American Medical Informatics Association Annual Symposium, San Francisco, CA, November 3-7, 2018*, 2018. URL https://knowledge.amia.org/67852-amia-1.4259402/t007-1.4262189/t007-1.4262190/2977349-1.4262947/2971507-1.4262944.

[28] Melike Bozkaya, Joost Gabriels, and Jan Martijn E. M. van der Werf. Process diagnostics: A method based on process mining. In *International Conference on Information, Process, and Knowledge Management, eKNOW 2009, Cancun, Mexico, February 1-7, 2009*, pages 22–27, 2009. doi: 10.1109/eKNOW.2009.29. URL https://doi.org/10.1109/eKNOW.2009.29.

[29] Álvaro Rebuge and Diogo R. Ferreira. Business process analysis in healthcare environments: A methodology based on process mining. *Inf. Syst.*, 37(2):99–116, 2012. doi: 10.1016/j.is.2011.01.003. URL https://doi.org/10.1016/j.is.2011.01.003.

[30] Wil M. P. van der Aalst. *Process Mining: Data Science in Action*. Springer, Heidelberg, 2 edition, 2016. ISBN 978-3-662-49850-7. doi: 10.1007/978-3-662-49851-4.

[31] Maikel L. van Eck, Xixi Lu, Sander J. J. Leemans, and Wil M. P. van der Aalst. PM ^2 : A process mining project methodology. In *Advanced Information Systems Engineering - 27th International Conference, CAiSE 2015, Stockholm, Sweden, June 8-12, 2015, Proceedings*, pages 297–313, 2015. doi: 10.1007/978-3-319-19069-3_19. URL https://doi.org/10.1007/978-3-319-19069-3_19.

[32] Alfredo Bolt, Massimiliano de Leoni, and Wil M. P. van der Aalst. A visual approach to spot statistically-significant differences in event logs based on process metrics. In *Advanced Information Systems Engineering - 28th International Conference, CAiSE 2016, Ljubljana, Slovenia, June 13-17, 2016. Proceedings*, pages 151–166, 2016. doi: 10.1007/978-3-319-39696-5_10. URL https://doi.org/10.1007/978-3-319-39696-5_10.

[33] Joos C. A. M. Buijs and Hajo A. Reijers. Comparing business process variants using models and event logs. In *Enterprise, Business-Process and Information Systems Modeling - 15th International Conference, BPMDS 2014, 19th International Conference, EMMSAD 2014, Held at CAiSE 2014, Thessaloniki, Greece, June 16-17, 2014. Proceedings*, pages 154–168, 2014. doi: 10.1007/978-3-662-43745-2_11. URL https://doi.org/10.1007/978-3-662-43745-2_11.

[34] Carsten Cordes, Thomas Vogelgesang, and Hans-Jürgen Appelrath. A generic approach for calculating and visualizing differences between process models in multidimensional process mining. In *Business Process Management Workshops - BPM 2014 International Workshops, Eindhoven, The Netherlands, September 7-8, 2014, Revised Papers*, pages 383–394, 2014. doi: 10.1007/978-3-319-15895-2_32. URL https://doi.org/10.1007/978-3-319-15895-2_32.

[35] Sander J. J. Leemans, Shiva Shabaninejad, Kanika Goel, Hassan Khosravi, Shazia W.

Sadiq, and Moe Thandar Wynn. Identifying cohorts: Recommending drill-downs based on differences in behaviour for process mining. In *Conceptual Modeling - 39th International Conference, ER 2020, Vienna, Austria, November 3-6, 2020, Proceedings*, pages 92–102, 2020. doi: 10.1007/978-3-030-62522-1_7. URL https://doi.org/10.1007/978-3-030-62522-1_7.

[36] Sander J. J. Leemans, Dirk Fahland, and Wil M. P. van der Aalst. Scalable process discovery and conformance checking. *Softw. Syst. Model.*, 17(2):599–631, 2018. doi: 10.1007/s10270-016-0545-x. URL https://doi.org/10.1007/s10270-016-0545-x.

[37] Angelina Prima Kurniati, Ciarán McInerney, Kieran Zucker, Geoff Hall, David C. Hogg, and Owen A. Johnson. A multi-level approach for identifying process change in cancer pathways. In *Business Process Management Workshops - BPM 2019 International Workshops, Vienna, Austria, September 1-6, 2019, Revised Selected Papers*, pages 595–607, 2019. doi: 10.1007/978-3-030-37453-2_48. URL https://doi.org/10.1007/978-3-030-37453-2_48.

[38] Stefania Montani, Giorgio Leonardi, Silvana Quaglini, Anna Cavallini, and Giuseppe Micieli. Improving structural medical process comparison by exploiting domain knowledge and mined information. *Artif. Intell. Medicine*, 62(1):33–45, 2014. doi: 10.1016/j.artmed.2014.07.001. URL https://doi.org/10.1016/j.artmed.2014.07.001.

[39] Giorgio Leonardi, Manuel Striani, Silvana Quaglini, Anna Cavallini, and Stefania Montani. Leveraging semantic labels for multi-level abstraction in medical process mining and trace comparison. *J. Biomed. Informatics*, 83:10–24, 2018. doi: 10.1016/j.jbi.2018.05.012. URL https://doi.org/10.1016/j.jbi.2018.05.012.

[40] Suriadi Suriadi, Ronny Mans, Moe Thandar Wynn, Andrew Partington, and Jonathan Karnon. Measuring patient flow variations: A cross-organisational process mining approach. In *Asia Pacific Business Process Management - Second Asia Pacific Conference, AP-BPM 2014, Brisbane, QLD, Australia, July 3-4, 2014. Proceedings*, pages 43–58, 2014. doi: 10.1007/978-3-319-08222-6_4. URL https://doi.org/10.1007/978-3-319-08222-6_4.

[41] Joos C. A. M. Buijs, Boudewijn F. van Dongen, and Wil M. P. van der Aalst. Towards cross-organizational process mining in collections of process models and their executions. In *Business Process Management Workshops - BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part II*, pages 2–13, 2011. doi: 10.1007/978-3-642-28115-0_2. URL https://doi.org/10.1007/978-3-642-28115-0_2.

[42] Brian Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Cluster Analysis*, volume 5th. 01 2011. ISBN 9780470978443. doi: 10.1002/9780470977811. URL https://www.researchgate.net/publication/318392855_Cluster_Analysis.

[43] Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping Multidimensional Data - Recent Advances in Clustering*, pages 25–71. 2006. doi: 10.1007/3-540-28349-8_2. URL https://doi.org/10.1007/3-540-28349-8_2.

[44] Fareed Zandkarimi, Jana-Rebecca Rehse, Pouya Soudmand, and Hartmut Hoehle. A generic framework for trace clustering in process mining. In *2nd International Conference on Process Mining, ICPM 2020, Padua, Italy, October 4-9, 2020*, pages

177–184, 2020. doi: 10.1109/ICPM49681.2020.00034. URL https://doi.org/10.1109/ICPM49681.2020.00034.

[45] Gianluigi Greco, Antonella Guzzo, Luigi Pontieri, and Domenico Saccà. Mining expressive process models by clustering workflow traces. In *Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004, Proceedings*, pages 52–62, 2004. doi: 10.1007/978-3-540-24775-3_8. URL https://doi.org/10.1007/978-3-540-24775-3_8.

[46] Ana Karla Alves de Medeiros, Antonella Guzzo, Gianluigi Greco, Wil M. P. van der Aalst, A. J. M. M. Weijters, Boudewijn F. van Dongen, and Domenico Saccà. Process mining based on clustering: A quest for precision. In *Business Process Management Workshops, BPM 2007 International Workshops, BPI, BPD, CBP, ProHealth, RefMod, semantics4ws, Brisbane, Australia, September 24, 2007, Revised Selected Papers*, pages 17–29, 2007. doi: 10.1007/978-3-540-78238-4_4. URL https://doi.org/10.1007/978-3-540-78238-4_4.

[47] R. P. Jagadeesh Chandra Bose and Wil M. P. van der Aalst. Context aware trace clustering: Towards improving process mining results. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA*, pages 401–412, 2009. doi: 10.1137/1.9781611972795.35. URL https://doi.org/10.1137/1.9781611972795.35.

[48] Alex Meincheim, Cleiton dos Santos Garcia, Júlio César Nievola, and Edson Emílio Scalabrin. Combining process mining with trace clustering: Manufacturing shop floor process - an applied case. In *29th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2017, Boston, MA, USA, November 6-8, 2017*, pages 498–505, 2017. doi: 10.1109/ICTAI.2017.00082. URL https://doi.org/10.1109/ICTAI.2017.00082.

[49] Michelle Taub, Allison Banzon, Tom Zhang, and Zhongzhou Chen. Tracking changes in students' online self-regulated learning behaviors and achievement goals using trace clustering and process mining. *Frontiers in Psychology*, 13:813514, 03 2022. doi: 10.3389/fpsyg.2022.813514.

[50] Alifah Syamsiyah, Alfredo Bolt, Long Cheng, Bart F. A. Hompes, R. P. Jagadeesh Chandra Bose, Boudewijn F. van Dongen, and Wil M. P. van der Aalst. Business process comparison: A methodology and case study. In *Business Information Systems - 20th International Conference, BIS 2017, Poznan, Poland, June 28-30, 2017, Proceedings*, pages 253–267, 2017. doi: 10.1007/978-3-319-59336-4_18. URL https://doi.org/10.1007/978-3-319-59336-4_18.

[51] Fabian Veit, Jerome Geyer-Klingeberg, Julian Madrzak, Manuel Haug, and Jan Thomson. The proactive insights engine: Process mining meets machine learning and artificial intelligence. In *BPM demos*, volume 1920. CEUR-WS.org, 2017. URL https://ceur-ws.org/Vol-1920/BPM_2017_paper_192.pdf.

[52] Ghalia Tello, Gabriele Gianini, Rabeb Mizouni, and Ernesto Damiani. Machine learning-based framework for log-lifting in business process mining applications. In *BPM*, page 232–249. Springer-Verlag, 2019. ISBN 978-3-030-26618-9. doi: 10.1007/978-3-030-26619-6_16. URL https://doi.org/10.1007/978-3-030-26619-6_16.

[53] Till Becker and Wacharawan Intayoad. Context aware process mining in logistics. *Procedia CIRP*, 63:557–562, 12 2017. doi: 10.1016/j.procir.2017.03.149.

[54] Wil MP van der Aalst, AJP Martin Weijters, Laura Maruster, and Ming Song. Trace clustering in process mining. *IJITKM*, 5(2):123–144, 2011.

[55] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. `https://www.researchgate.net/publication/222451107_Rousseeuw_PJ_Silhouettes_A_Graphical_Aid_to_the_Interpretation_and_Validation_of_Cluster_Analysis_Comput_Appl_Math_20_53-65`.

[56] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley, 1990. ISBN 978-0-47031680-1.

[57] Dhendra Marutho, Sunarna Hendra Handaka, Ekaprana Wijaya, and Muljono. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 International Seminar on Application for Technology of Information and Communication*, pages 533–538, 2018. doi: 10.1109/ISEMANTIC.2018.8549751.

[58] David J Ketchen and Christopher L Shook. The application of cluster analysis in strategic management research: An analysis and critique. *SMJ*, 17(6):441–458, 1996.

[59] Chaoqun Ma and Jianhong Wu. *Data Clustering: Theory, Algorithms, and Applications*, volume 20. 01 2007. doi: 10.1137/1.9780898718348.

[60] Peter Rousseeuw. Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65, 11 1987. doi: 10.1016/0377-0427(87)90125-7.

[61] Anil Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31:651–666, 06 2010. doi: 10.1016/j.patrec.2009.09.011.

[62] Turki Alghamdi and Nadeem Javaid. A survey of preprocessing methods used for analysis of big data originated from smart grids. *IEEE Access*, 10, 02 2022. doi: 10.1109/ACCESS.2022.3157941.

[63] Fokrul Mazarbhuiya. A novel approach for clustering periodic patterns. *International Journal of Intelligence Science*, 07:1–8, 01 2017. doi: 10.4236/ijis.2017.71001.

[64] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16:645 – 678, 06 2005. doi: 10.1109/TNN.2005.845141.

[65] Adele Cutler, David Cutler, and John Stevens. *Random Forests*, volume 45, pages 157–176. 01 2011. ISBN 978-1-4419-9325-0. doi: 10.1007/978-1-4419-9326-7_5.

[66] Wil Aalst. *Extracting Event Data from Databases to Unleash Process Mining*, pages 105–128. 02 2015. ISBN 978-3-319-14429-0. doi: 10.1007/978-3-319-14430-6_8.

[67] L Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001. doi: 10.1023/A:1010950718922.

[68] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A%3A1010933404324.

[69] Konrad Reinhart, Thomas Goolsby, Mark Crowther, Karel Cvachovec, Robert Bell, Christopher Duggan, Alison Fox-Robichaud, John Gennari, Stephan Jakob, Kristof Jochmans, et al. Recognizing sepsis as a global health priority—a who resolution. *NEJM*, 377(5):414–417, 2017.

[70] Giovanni Acampora, Autilia Vitiello, Bruno N. Di Stefano, Wil M. P. van der Aalst, Christian W. Günther, and Eric Verbeek. IEEE 1849: The XES standard. *IEEE Comput. Intell. Mag.*, 12(2):4–8, 2017. doi: 10.1109/MCI.2017.2670420. URL https://doi.org/10.1109/MCI.2017.2670420.

[71] Fraunhofer Institute for Open Communication Systems. pm4py, 2021. URL https://pm4py.fit.fraunhofer.de/. https://pm4py.fit.fraunhofer.de/.

[72] Sotiris Kotsiantis, I. Zaharakis, and P. Pintelas. Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, 26:159–190, 11 2006. doi: 10.1007/s10462-007-9052-3.

[73] Ian Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374:20150202, 04 2016. doi: 10.1098/rsta.2015.0202.

[74] Alfredo Bolt, Massimiliano de Leoni, and Wil Aalst. A visual approach to spot statistically-significant differences in event logs based on process metrics. pages 151–166, 06 2016. ISBN 978-3-319-39695-8. doi: 10.1007/978-3-319-39696-5_10.

[75] Sander Leemans, Wil Aalst, Tobias Brockhoff, and Artem Polyvyanyy. Stochastic process mining: Earth movers' stochastic conformance. *Information Systems*, 102: 101724, 02 2021. doi: 10.1016/j.is.2021.101724.

[76] Jia Wei, Zhipeng He, Chun Ouyang, and Catarina Moreira. Mimicel: Mimic-iv event log for emergency department. 2022.

Appendices

Appendix A

Sepsis Sublogs Comparisons

A.1 Sepsis: Age Comparison

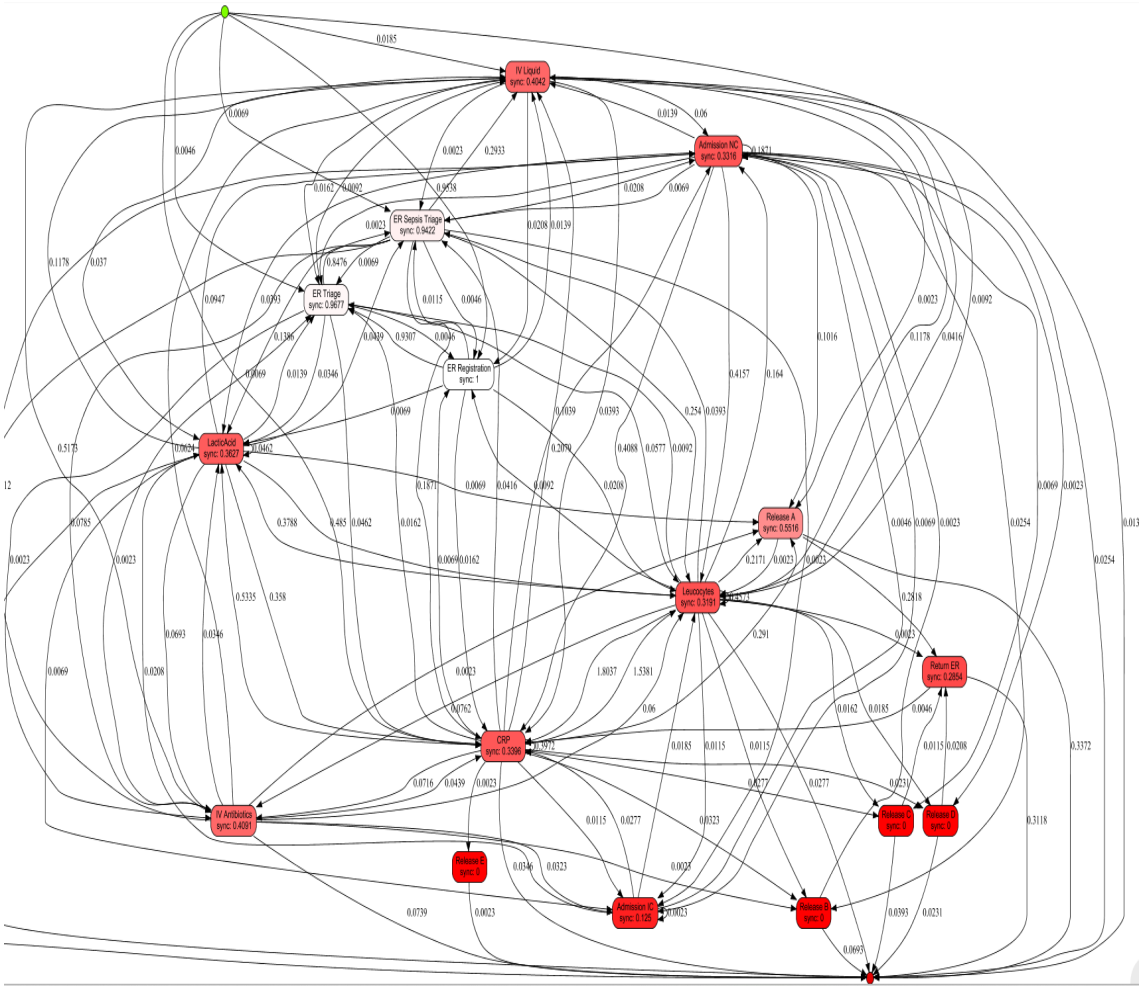


Figure A.1: Sepsis Age Comparison

A.2 Sepsis: Diagnosis Comparison

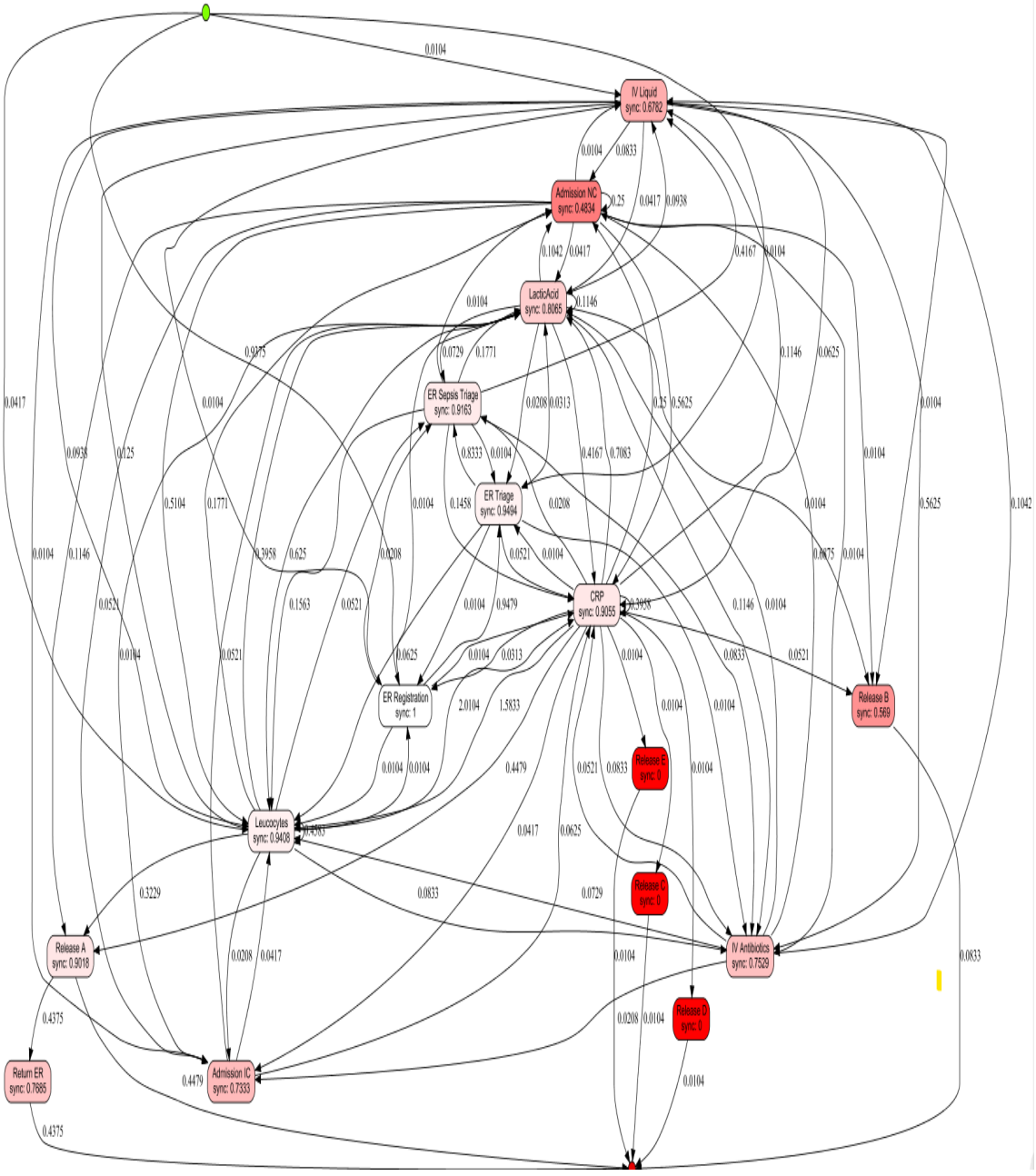


Figure A.2: Sepsis Diagnosis Comparison

A.3 Sepsis: IN-P

A.4 MIMIC: IN-V

A.5 MIMIC: IN-P

A.6 Evaluation Letter

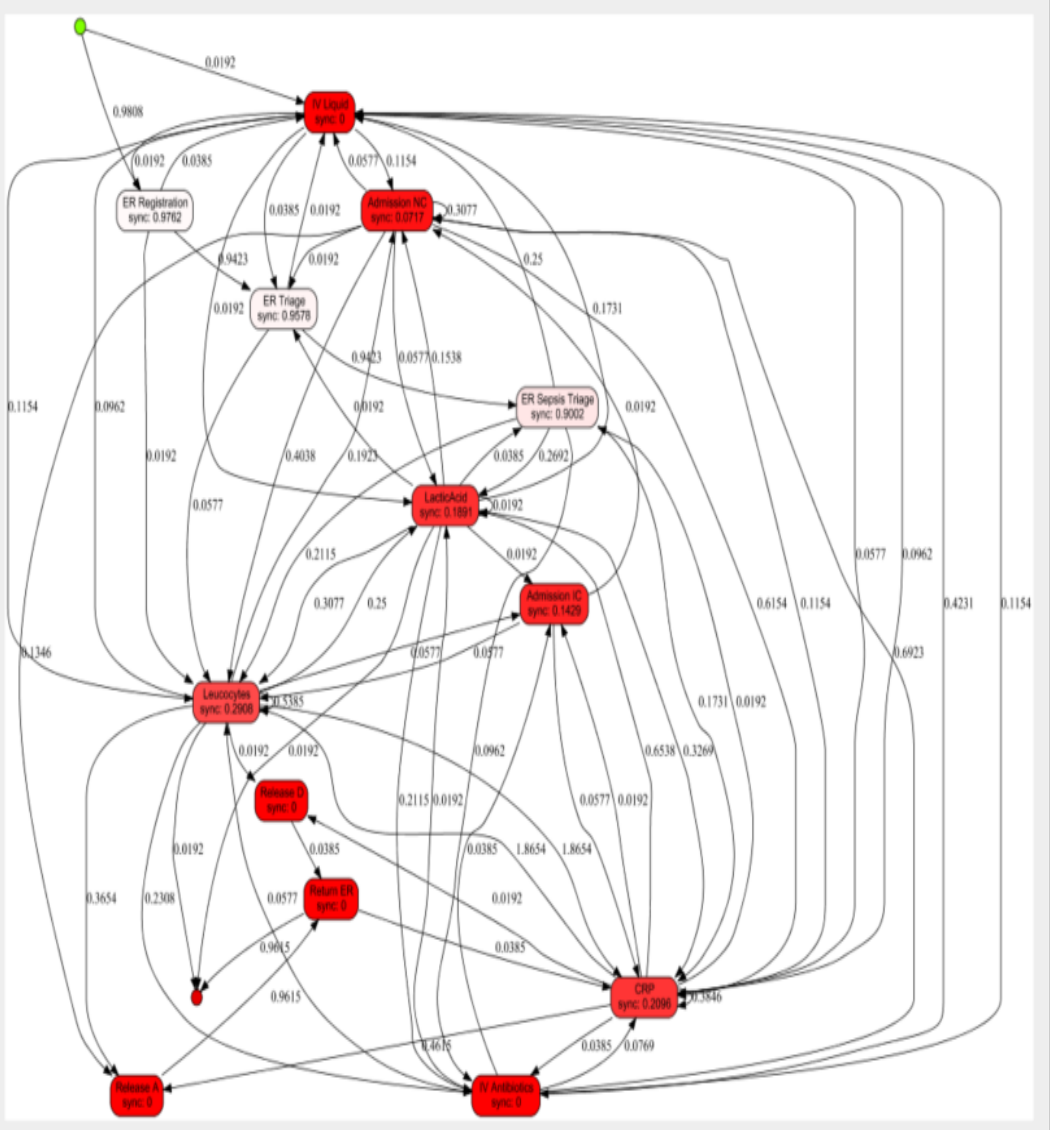


Figure A.3: Sepsis: IN-P - Sublogs 4 and 12

Figure A.4: MIMIC: IN-V - Sublogs 4 and 12

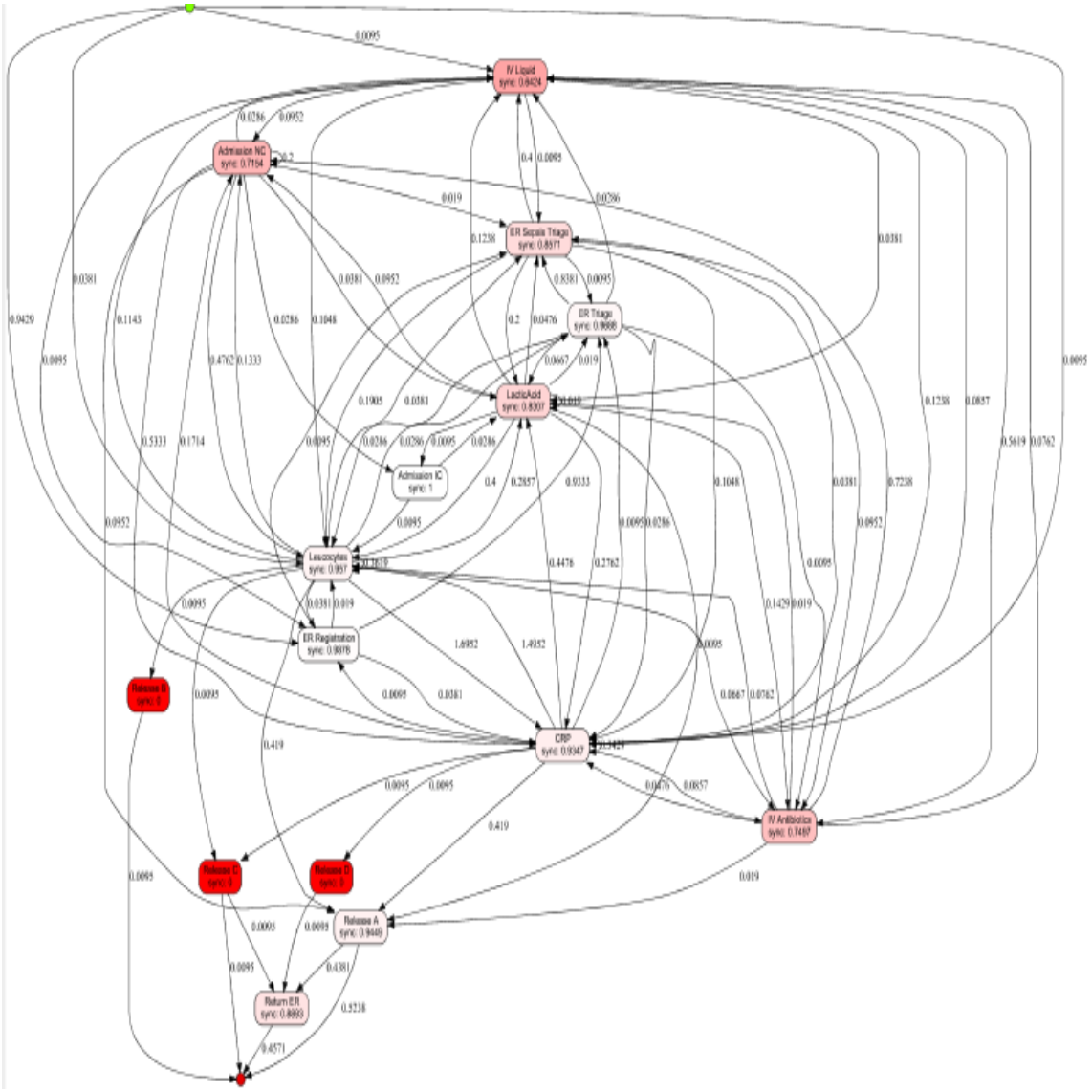


Figure A.5: MIMIC: IN-P - Sublogs 7 and 13

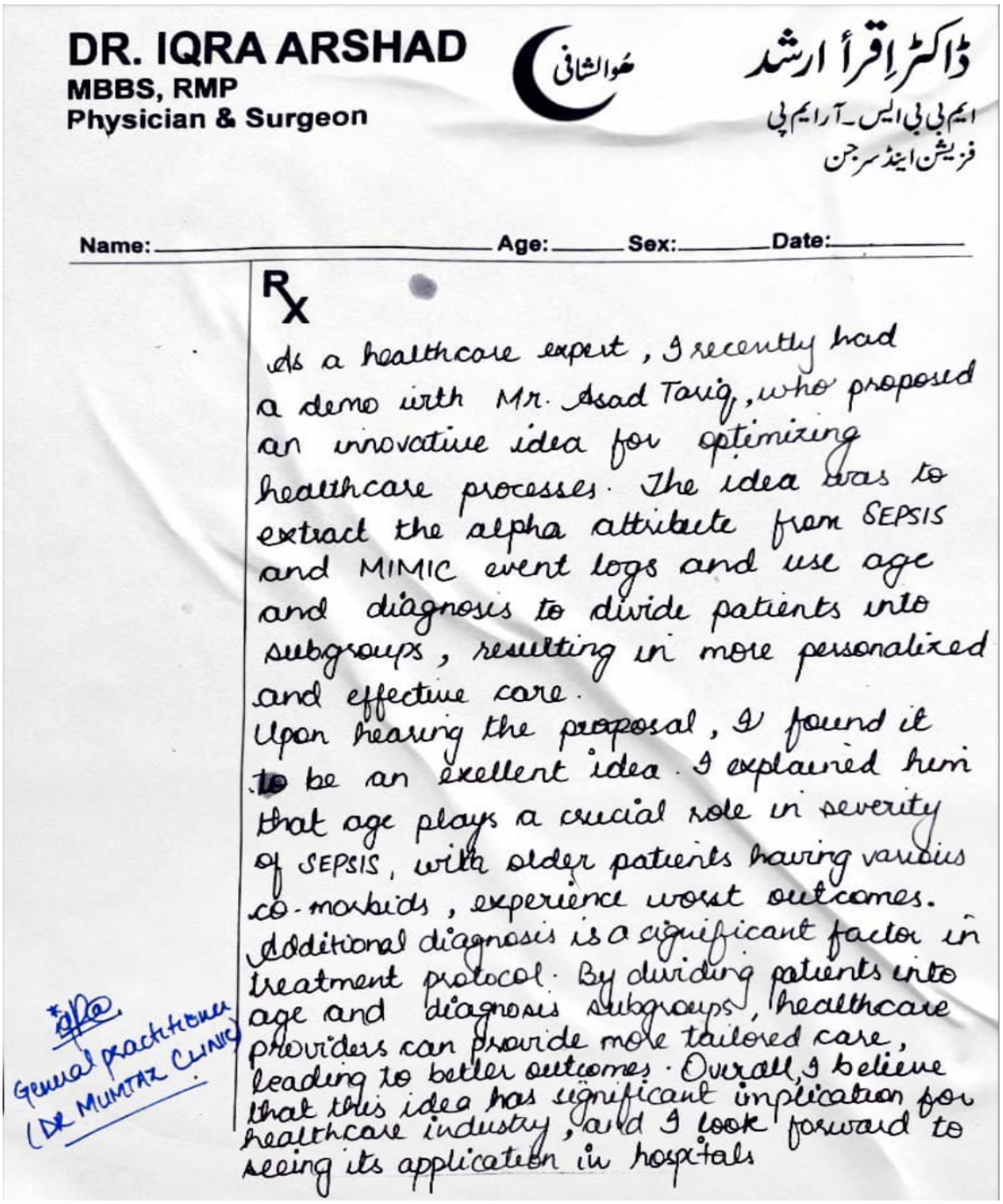


Figure A.6: Evaluation letter

Acknowledgments

In the name of Allah, the most merciful and the most beneficent. First and foremost, I express my sincere gratitude to my wife, Shaista Anjum, for her unwavering support, guidance, and encouragement throughout this academic journey. In addition, I am genuinely grateful for her patience and faith, which have been instrumental in my success. I also extend my most profound appreciation to my supervisor, Univ-Prof. Dr. Ir. Sander J.J. Leemans, for his invaluable guidance, expertise, and unwavering support. Without his mentorship, this research would not have been possible. I am immensely grateful for the opportunity to work on such an exciting project and for the constructive feedback that has significantly improved the quality of this work. I must also acknowledge the contribution of my dear friends, Faizan Hassan and Muhammad Abdullah, for their invaluable assistance and support throughout this journey. Their editing help, feedback sessions, and discussions were instrumental in shaping the ideas and arguments presented in this research. Lastly, I express my heartfelt gratitude to my parents for allowing me to pursue my academic aspirations at a prestigious university. Despite the challenges, their unwavering support and encouragement have been an immense strength and motivation. Their unwavering commitment to my education is a debt I can never fully repay.