UNIVERSITY OF TARTU
Institute of Computer Science
Software Engineering Curriculum

Cigin Koshy

A Literature Review on Predictive Monitoring of Business Processes

Master's Thesis (30 ECTS)

Supervisor(s): Fabrizio Maria Maggi
Fredrik Milani
Chiara Di Francescomarino

Tartu 2017

Acknowledgement

I would like to thank my supervisors Fabrizio Maria Maggi, Fredrik Milani and Chiara Di Francescomarino for their tremendous support and guidance throughout the Thesis work.

Special thanks to Fredrik Milani, for the constant motivation and being the guiding light, during the entire journey of this Thesis work.

I would also like to express my gratitude to all staff member of our faculty who always helped me and guided me at every step.

I would like to thank my parents back home for sticking by me at the time of need and providing me with much required strength and support throughout.

Special thanks to my sister, for keeping me motivated and focussed on my goals.

Special thanks to my fiancée, for being a constant support throughout this period.

Special thanks to my flatmates Asima & Nazim, for the love and support.

Finally, I would like to thank all my friends and family members back in India, without their prayers and support this would not have been possible.

A Literature Review on Predictive Monitoring of Business Processes

Abstract:

The goal of predictive monitoring is to help the business achieve their goals, help them take the right business path, predict outcomes, estimate delivery time, and make business processes risk aware. In this thesis, we have carefully collected and reviewed in detail all literature which falls in this process mining category. The objective of the thesis is to design a Predictive Monitoring Framework and classify the different predictive monitoring techniques. The framework acts as a guide for researchers and businesses. Researchers who are investigating in this field and businesses who want to apply these techniques in their respective field.

Keywords: Prediction, Process mining, Business Process, Predictive Monitoring

CERCS: P170

Kirjanduse ülevaade äriprotsesside ennetava seire kohta

Lühikokkuvõte:

Oleme läbi vaadanud mitmesuguseid ennetava jälgimise meetodeid äriprotsessides. Prognoositavate seirete eesmärk on aidata ettevõtetel oma eesmärke saavutada, aidata neil võtta õige äri, prognoosida tulemusi ja aega ning muuta äriprotsessid riskantsemaks. Antud väitekirjaga oleme hoolikalt kogunud ja üksikasjalikult läbi vaadanud selle väitekirja teemal oleva kirjanduse. Kirjandusuuringu tulemustest ja tähelepanekutest lähtuvalt oleme hoolikalt kavandanud ennetava jälgimisraamistiku. Raamistik on juhendiks ettevõtetele ja teadlastele, teadustöötajatele, kes uurivad selles valdkonnas, ja ettevõtetele, kes soovivad neid tehnikaid oma valdkonnas rakendada.

Võtmesõnad: Prognoosimine, protsessi kaevandamine, äriprotsess, ennetav seire

CERCS: P170

Table of Contents

1 Introduction 3
1.1 Problem Statement 5
1.2 Contribution 5
2 Research Methodology 6
2.1 Predictive Monitoring Protocol 13
2.2 Predictive Monitoring Elicitation 14
2.2.1 Studies Identification Phase 14
2.2.2 Selection Phase 14
2.2.3 Extraction Phase 15
2.3 Literature Review 16
2.4 Predictive Monitoring Framework 16
3 Conceptual Foundation 17
3.1 Business Process Management 17
3.2 Process Mining 18
3.3 Machine Learning Techniques 21
3.4 Other Important Terminologies 24
4 Literature Review 25
4.1 Time-based Prediction 25
4.1.1 Discussion of the Studies 25
4.1.2 Summary of Time based Prediction 32
4.2 Outcome Prediction 35
4.2.1 Discussion of the Studies 35
4.2.2 Summary of Outcome Prediction 42
4.3 Process Path Prediction 46
4.3.1 Discussion of the Studies 46
4.3.2 Summary of Process Path Prediction 49
4.4 Risk Prediction 51
4.4.1 Discussion of the Studies 51
4.4.2 Summary of Risk Prediction 53
5 Predictive Monitoring Technique Framework Design 54
6 Threats to Validation 57
7 Conclusion 58
8 References 59
I. Appendix 64
II. Predictive Monitoring Framework Implementation 66
III. License 67

Table of Figure

Figure 1 Discovery input and output [1] 3
Figure 2 Conformance input and output [1] 3
Figure 3 3Enhancement input and output [1] 3
Figure 4 Research Methodology 12
Figure 5 Graph showing the distribution of identified study sources 14
Figure 6 Literature selection phase 15
Figure 7 Extraction stage statistics 15
Figure 8 Summarizing accept and reject reasons of papers in extraction phase 16
Figure 9 Business Process Lifecycle [3] 18
Figure 10 Working of a process mining technique [1] 19
Figure 11 An example of an event log [5] 20
Figure 12 Sliding window model [10] 23
Figure 13 An example of a Petri Net 24
Figure 14 An example workflow [16] 26
Figure 15 Implementation using CoCaMa application [17] 27
Figure 16 Sequential time series petri net [23] 29
Figure 17 Depicts a BPMN model of such a process [27] 31
Figure 18 A probabilistic graph of an Insurance Claim scenario [28] 35
Figure 19 Flowchart of operation with the RFID [42] 40
Figure 20 Different path representation [44] 46
Figure 21 i) The YAWL helps participants in choosing the next work item to perform based on risks and (ii) UI to supports participants filling out a form based on risk [52] 52
Figure 22 Predictive Monitoring Framework 54

1 Introduction

Process mining is an important branch of data mining which deals with mining of logs related to workflows, events, businesses processes and control flows. All upcoming and latest trends in technology have one thing in common and that is logs. The art of process mining takes up raw data from these logs and converts them into valuable information. The process mining techniques with these logs as input, can detect, support and predict future business processes.Over the last decade, it is very common to find event logs generated from information systems. These events logs can be considered as a gold mine of data, providing valuable information in the form of detailed historic data. The techniques devised to extract information from these data sources are known as "Process Mining". Process Mining shows ways to discover, monitor and improve processes in a variety of way.

The first kind of process mining is Discovery. A discovery technique takes an event log and generates a business model based on the log without any apriori information. The discovered model is typically a process model (e.g., a Petri net, BPMN, EPC, or UML activity diagram) [1].

Figure 1 Discovery input and output [1]

The second kind of process mining is Conformance. This technique checks by comparing the process model to the event log. The technique helps to understand whether both the model and the log align with each other. The output of this technique is an improvement or an extended model.

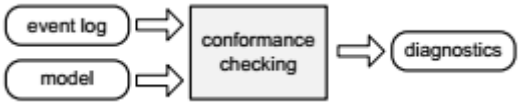

Figure 2 Conformance input and output [1]

The third kind of technique is the Enhancement. This technique aims to improve the business process model by using existing information derived from event logs.

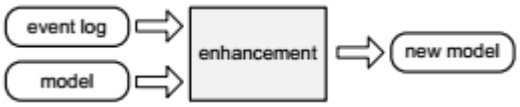

Figure 3 3Enhancement input and output [1]

This Thesis focuses on predictive analytics in business processes which is an enhancement technique in Process mining. Predictive process monitoring uses data mining, machine learning and statistical techniques to predict various business process related information. Predictive process mining can be used to forecast, alert and possibly change the direction of the business process for a positive outcome. A good amount of literature has been published in this field, in a wide domain of businesses ranging from healthcare, logistics to modern IoT based devices. Today business scenarios need to predict Predictive analytics can help predict whether a business scenario or task will complete successfully, what next steps need to be taken to finish a particular task, completion time of a particular business process or the risk estimation associated with a business process. In the logistics industry, the client seeks information pertaining to delivery dates, in hospitals doctors use effective treatment paths of successfully treated patients to treat new patients. In airline industry process mining, could help detect "no show" passengers and avoid over booking related hassles for

customers. In modern IoT based devices, predictive process mining could help devices predict failures, service requirements, and device maintenance in advance.

This Thesis work is dedicated to a literature review of predictive monitoring of business processes, which includes predicting the process outcomes, cycle times, process paths and risk estimations. A systematic literature review was conducted, based on which the categorization and classification of predictive processes were done. Once the literature review and classification were completed, based on the data derived we created a framework. The purpose of the framework is to help create a platform to help future researcher businesses about the current progress in this field and use it to enhance the field further.

The Thesis is structured as follows. In Chapter 1.1 the problem statement is presented. In Chapter 2 the adopted research methodology is introduced and in the following chapter, Chapter 3, a conceptual foundation is prepared, to prepare the readers for the chapters to follow. In Chapter 4, we study in detail the literature collected, and summarize the findings in a table. In Chapter 5 we design the Predictive Monitoring Framework; the framework is introduced and discussed in detail in this chapter. In Chapter 6 we define the threats to validity and in Chapter 7 we conclude.

1.1 Problem Statement

Currently, there is a good amount of literature dedicated to Predictive monitoring of business processes. Sometimes it gets confusing for process mining users to analyze and conclude which technique best suits their requirement. The literature reviewed in this paper show multiple papers which show different categories of prediction and under different contexts and domain. In these proposed predictive monitoring techniques, the kind of prediction appears to be the same (outcome, time, path, risk) but the approaches could be different, i.e. a different statistical or machine learning algorithm maybe in use enhancing the previous version, an extra information found in the logs which had been omitted before may have been used now making the accuracy even better, maybe the approach is entirely the same but the output has been presented using metrics which make more sense.

Hence the main goal of the Thesis is to generate a framework which could easily classify each reviewed literature belonging to the field of predictive monitoring of business processes. Secondly, we need to distinguish each reviewed literature based on a set of attributes. These attributes answer to the question, which technique could be best applied when. These goals will be addressed by the following Research Questions to which we seek answers in this Thesis. The first research question we present is the most important, it is very important for us to review the literature in hand and conclude, under what broad categories can we classify the reviewed literature. As we intend to classify the literature based on kinds of prediction, we need to find the categories of prediction. This leads to our first research question:

Research Question 1 (RQ 1): What aspects of a business process can predictive monitoring predict?

The second research question deals with the application of the proposed techniques. The application can be related to the domain or industry it is being used, whether there is any implementation of the proposed technique

Research Question 2 (RQ 2): How is predictive monitoring currently applied in the industry?

1.2 Contribution

The main contributions of this Thesis can be listed as first, a systematic literature review of predictive monitoring techniques in business processes is conducted. Secondly, based on the literature review, the prediction monitoring approaches are classified and finally, a Predictive monitoring framework is designed to help businesses navigate and use different prediction techniques based on their requirement.

2 Research Methodology

The goal of this Thesis is to define a framework for the Predictive monitoring of business processes thus allowing the systematic comparison of new and existing predictive monitoring approaches under a common guideline.

The real task in this approach is divided into four steps. As described in Figure 4.

The methodology has been divided into four main phases: (i) Predictive Monitoring Technique Protocol (ii) Predictive Monitoring Technique Elicitation (iii) Predictive Monitoring Technique Literature Review (iv) Predictive Monitoring Technique Design

The first phase is the protocol phase, the Systematic literature review is guided by Kitchenham in [2]. Any literature review starts with a planning phase. Planning is done by establishing a protocol which defines the basic structure to carry out the whole systematic literature review. The protocol also defines basic guidelines which need to be maintained throughout the review phase. The main questions to be answered in the protocol are as follows:

1. **Objective**: The research goal or focus is expected to be defined here. What the research expects to be answered at the end.
2. **Keywords and Synonyms**: The main terms that will be used throughout the research. These terms will be used to extract data from electronic databases.
3. **Source Selection Criteria Selection**: Why certain sources of data have been chosen.
4. **Source search methods**: From where the sources have been identified.
5. **Source List**: List of sources
6. **Study Selection Criteria**: Criteria that were used to filter out literature.

The second phase is the Elicitation phase. The elicitation phase consists of (i) study identification. (ii) study selection phase (iii) extraction of the literature. In the second phase, all the literature related to the study is identified from all the sources, this is followed by the first selection phase in which the literature is filtered based on just the keywords and the titles of the literature. This phase is followed by the extraction phase, in this phase further filtering is done based on the abstract described in the literature. If a decision could not be made by reading the abstract, the literature is read in detail.

This is followed by a systematic literature review which is described in Chapter 4. The literature review helps in categorizing the predictive techniques which help in solving RQ1.

Followed up with the PMT Design Phase, in the design phase a Predictive Monitoring Framework is built. The frame work helps in evaluation and comparison of existing technique. The phase has been proposed in response to RQ2.

Once the protocol has been set in phase 1, followed by literature selection in phase 2, the literature review is done in phase 3. Based on the findings and observations in phase 3, a framework is prepared in the fourth phase for evaluating various features of each technique described in reviewed literature (Phase 3). The finding of phase 3 is presented using the framework defined in phase 4, which is the final goal of the Thesis.

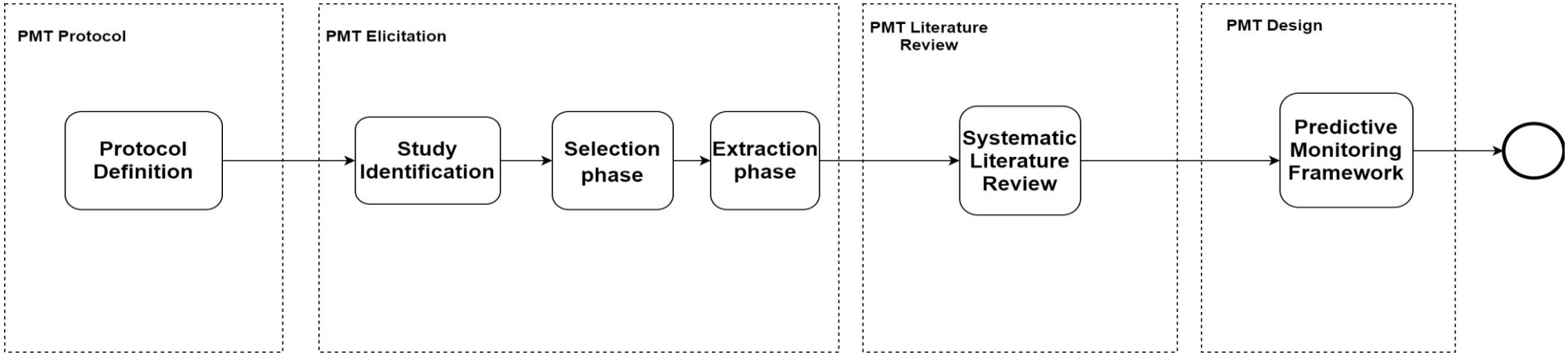

Figure 4 Research Methodology

Now we introduce each section of the Research Methodology defined in Figure 4 in detail and in relation to the Thesis. The subchapter 2.1 and 2.2 discusses in depth about the protocol defined to carry out this Thesis research and how the elicitation phase of the research was carried out. Section 2.3 and 2.4 give a brief introduction to their specific chapters which are discussed in detail as the Thesis progresses.

2.1 Predictive Monitoring Protocol

In the protocol phase, the following details were defined to create an initial platform for literature selection and extraction. The information defined below form the basis for this literature review.:

Objective: The objective of this Thesis is to create a framework for the evaluation of predictive monitoring techniques in business processes using process mining. A framework is established to carry out this evaluation. The framework enables categorical classification of predictive monitoring techniques.

Keywords and Synonyms:

To begin the search for relevant published papers it is necessary to identify the correct keywords. We identified 4 such keywords which are very appropriate for our Thesis topic. Some keywords could have different forms of the same word, it is necessary to include them as well.

- Predictive
- Prediction
- Business process
- Process mining

Source Selection Criteria Selection: Sources which are frequently used and contains a good amount of research papers published in the field of computer science, machine learning, business processes or process mining. Multiple electronic databases were used to cover the maximum available research papers published which are relevant and related to our scope of research

Source search methods: The source for this Thesis is primarily Electronic databases.

Source List:

All the electronic databases which are well known and gave relevant search results based on our keywords have been included in this paper. The databases included are

- Scopus
- Springer
- IEEE
- Science Direct
- ACM

Study Selection Criteria:

The criteria's which decide whether a particular should be included or excluded. The main exclusion criteria's are as follows:

1. Length of literature should not be less than 6 pages
2. Literature should not be a conference paper.
3. Literature should not be short paper or position paper.
4. Literature should propose a prediction technique, model or enhancement in process mining.

2.2 Predictive Monitoring Elicitation

The Elicitation phase of the Predictive Monitoring Technique starts with the Studies Identification phase.

2.2.1 Studies Identification Phase

In the study identification phase, keywords defined in the protocol phase are used to create regular expressions and extract results from the source list identified in the protocol phase. The Boolean expressions developed for extraction of literature in this Thesis is as follows:

("predictive" OR "prediction") AND ("business process" OR "process mining")

Each part in the above Boolean expression surrounded by the ( ) symbol is itself a Boolean expression consisting of synonyms, acronyms, and abbreviations.

Electronic databases are the primary sources of relevant material. The devised expression was used in the electronic databases identified in the protocol phase to extract relevant results. The result is as follows:

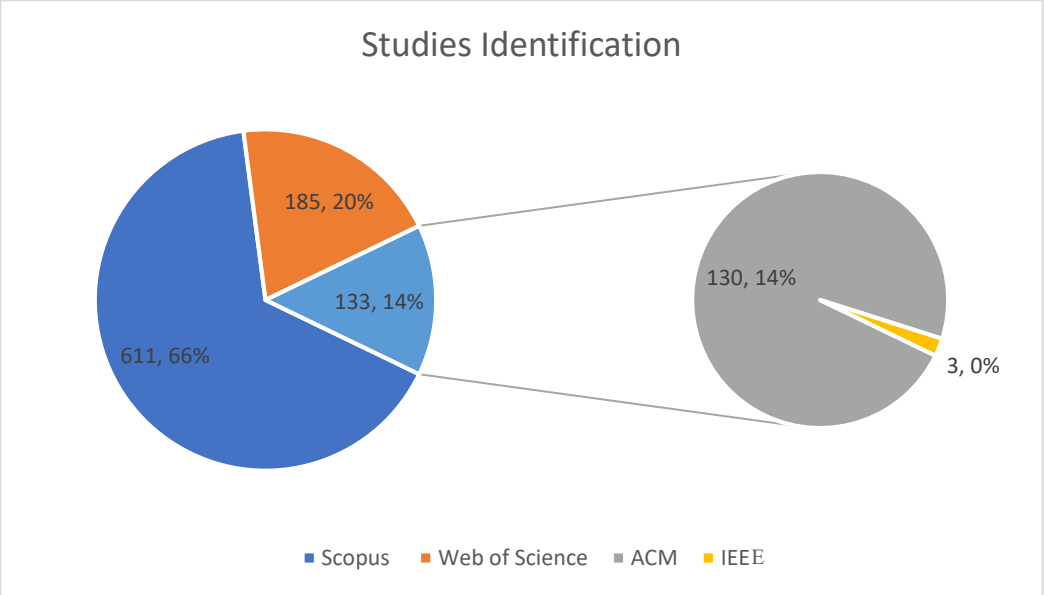

Figure 5 Graph showing the distribution of identified study sources

A total of 929 results were extracted from 5 electronic databases. Scopus contributed to 66% of the results with 611 papers, followed by Web of Science with 185 papers which contributed to 20% to results and finally ACM provided the rest of the 14% papers i.e. 133 papers. 3 papers were identified from IEEE as well.

2.2.2 Selection Phase

In the selection phase, the identified papers are started to be filtered. In this phase, filtering is done based on the title of the paper, the keywords relevant to the paper, and the abstract of the paper. During this phase, if the keyword of the paper contains terms like process mining, business process, predictive modeling, the paper is accepted. The title of the paper also gives much-required information, what the paper is proposing, what it is predicting. If the paper proposes relevant required information, then it is accepted. The main reason for rejection is not following the literature selection criteria which have been stated in the previous chapter.

After the initial round of filtering in this phase using title and keywords, the literature list is further filtered by thoroughly reading the abstract of the paper and the Introduction section. If the information is found relevant in terms of our Thesis guideline, then it is accepted else rejected.

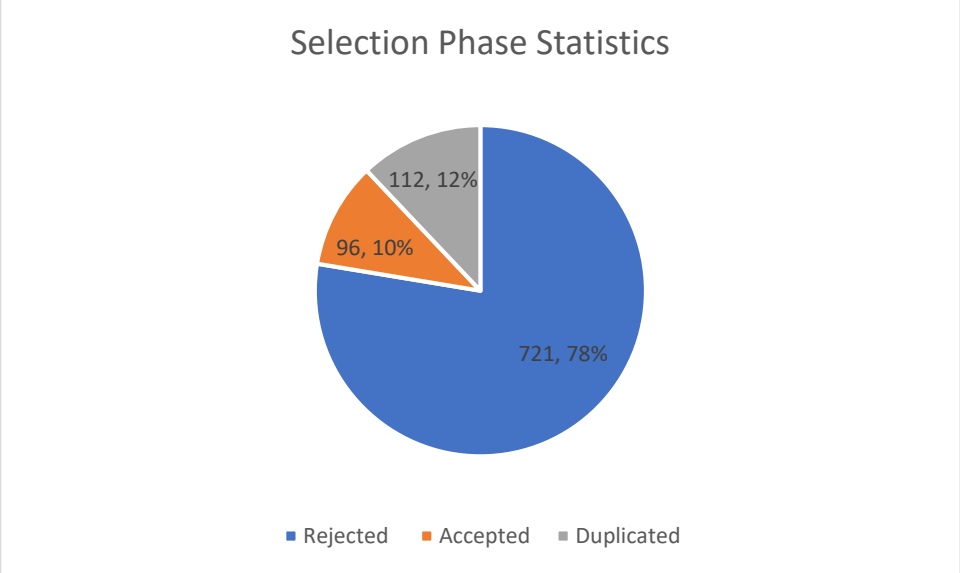

Figure 6 Literature selection phase

According to the chart shown in Figure 6, 78% of the literature gathered from the identification phase were rejected in this phase and only 12 % of the literature was accepted. The

2.2.3 Extraction Phase

In this stage, all the 96 papers, which were selected in the previous stage, the literature list is further filtered by thoroughly reading the abstract of the paper and the Introduction section. If the information is found relevant in terms of our Thesis guideline, then it is accepted else rejected.

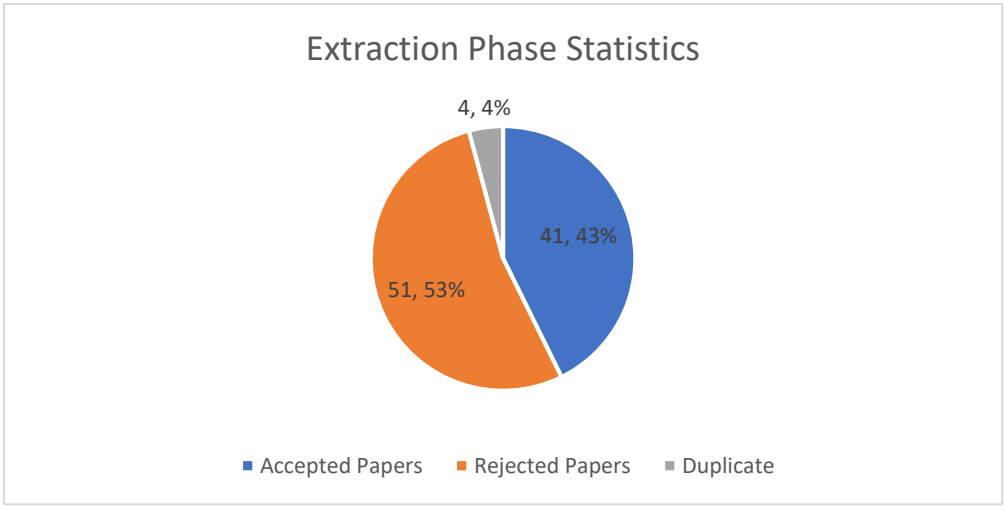

Figure 7 Extraction stage statistics

According to the final statistics, out of a total of 96 papers, 41 papers were accepted for final evaluation and 51 papers were rejected as they were failing to follow the guidelines mentioned in our protocol. Figure 7 shows the final results of this stage and phase.

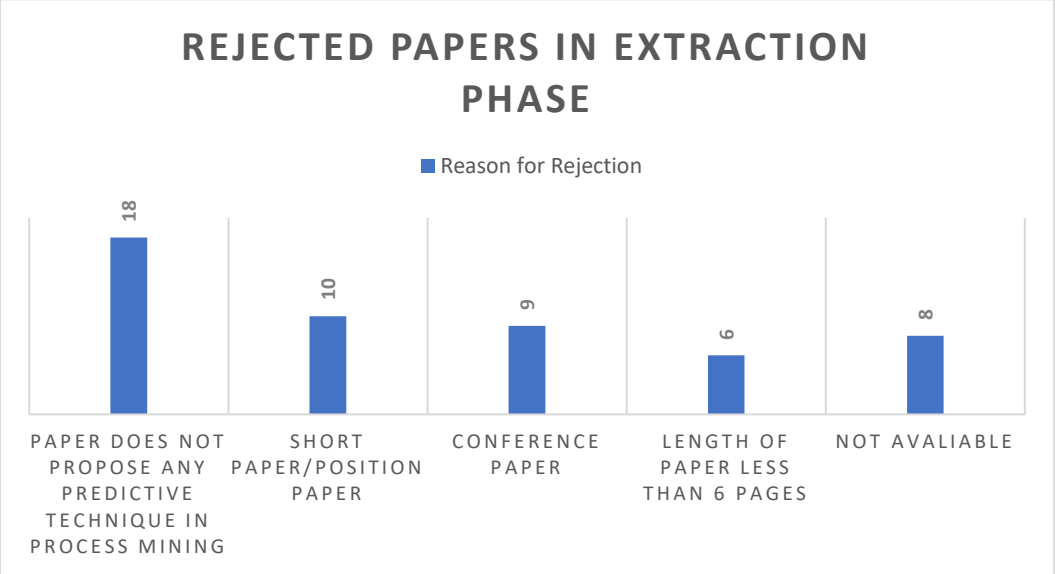

Figure 8 Summarizing accept and reject reasons of papers in extraction phase

In the extraction phase as mentioned earlier, the papers were further filtered. Figure 8 clearly shows the reason for the rejection of literature and the frequency of their occurrence. As the graph indicates, the majority of the rejections were due to the fact that, the paper listed were short paper or position papers. These kinds of paper were considered unfit for this Thesis. The second common reason was that the paper didn't present or discuss any process mining technique. Literature which belonged to the category of conference paper was also rejected. Papers which were shorter than 6 pages were considered ineligible as it was observed they contained limited information. It was also found certain papers which fulfilled all other requirements, couldn't be accessed for reading purposes, hence they had to be rejected as well.

2.3 Literature Review

The literature review is the third phase of the research methodology. In this phase a total of 41 papers which have been filtered in the Elicitation phase are studied in detail and a review of the literature is done. The literature review focusses primarily on the main concept presented in the paper which includes the technique proposed in the paper, the implementation of the proposed technique, its validation and output. In this Thesis, the literature review has been presented in Chapter 4. Every sub chapter is followed up by a summarization table, for easy analysis of the papers being discussed.

2.4 Predictive Monitoring Framework

The Predictive Monitoring Framework is the main contribution of this Thesis. Once the literature has been reviewed, it is carefully analyzed and clustered based on its common properties. Based on the clustering a framework is presented. The framework allows businesses and researchers easy access to analyze and utilize the predictive techniques published till date. The framework is present in Chapter 5.

3 Conceptual Foundation

Before we dive into the literature review section of the Thesis it is very important for us to understand some basic underlying concepts which will help understand the concepts and techniques presented in the chapters following this. In this chapter, primarily we have discussed basic concepts related to Business Process Management, Process Mining, and certain important Machine Learning Techniques.

3.1 Business Process Management

This Thesis is related to helping businesses improve by using predictive business process techniques. Hence it is important for us to understand the underlying concept of Business Process Management and its lifecycle.

Business Process Management (BPM) is the art and science of overseeing how work is performed in an organization to ensure consistent outcomes and to take advantage of improvement opportunities [3]. BPM could help in reducing costs, execution time, error rates and so on. A process in a Business process consists of events, activities, and decision. Improving a business process is not about improving a particular activity or task but the entire process itself.

A business is a collection of sequential logical steps which ultimately aim to serve goods or services to the customer. A business process management approach focussed on managing, analyzing and optimizing all aspects of an organization with the wants and needs of the client. All the tools that support BPM are known as Business Process Management System. BPM lifecycle consist of four phases:

Model: High-level diagram of the business process and the possible flows (BPMN, UML).

Assemble: Model designed is developed, tested and configured.

Deploy: Modelled business process is deployed on BPMS engine.

Manage: Deployed process is analyzed for potential bottle necks.

BPM Lifecycle

The BPM life cycle as the name suggests is the series of events that take place to improve the business process. A BPM process does not start from scratch i.e. initially the problem is identified, followed by identifying the range of issues it could cause.

The BPM lifecycle starts by identifying the processes to the problem and then identifying the relation between these processes. The purpose of engaging in BPM lifecycle is to consistently deliver positive outcomes and deliver maximum value to the organization in servicing its client [3].

Secondly, the performance measure of the business process is measured, which analyzes whether the business process is in good shape or bad shape. Cost related measure is one such measure

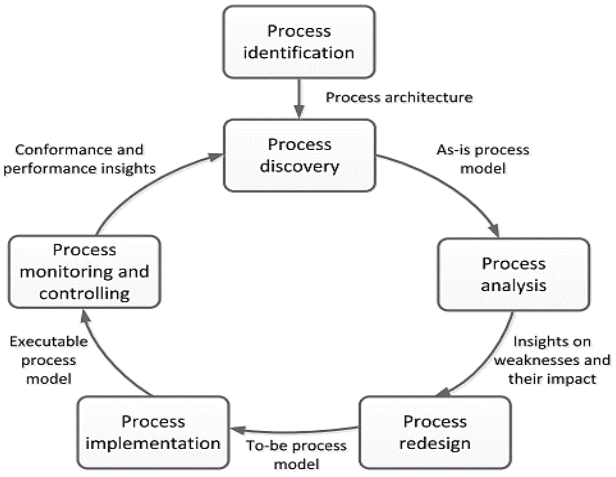

Figure 9 Business Process Lifecycle [3]

which could be used for this kind of analysis. Another measure is the time-based measurement for example cycle time. And finally, the quality measure also adds to the list of recurrent classes which analyses whether a business is in good shape or not.

Process discovery (also called as-is process modeling): As-is models are created to document the current state of the business process

Process analysis: The as-is models created in the above stage are clearly analyzed in this section. The problems related to them are identified and documented. These issues are ordered per their impact on the business and the efforts required to resolve them.

Process redesign (also called process improvement): In this phase changes required for the process are analyzed and their performance measures are compared. Process redesign and process analysis go hand in hand. A to-be model is generated at the end of this phase.

Process implementation. In this phase, the business model is implemented, there by moving from the as-is model to implementing the to-be model.

Process monitoring and controlling: In this phase, the data associated with the above-implemented process are collected and analyzed determining how well the process is performing with respect to its performance measure and objectives.

3.2 Process Mining

As we already learned the main concept involving business processes and the business process management. It is time to look into Process Mining. Process Mining involves extracting valuable information from logs generated by business processes.

Process mining whose goal is to discover, monitor and improve providing techniques and tools to extract knowledge from event logs [4]. The goal of process mining is to extract information from transaction logs [5]. In [5], process mining could be expressed in three different perspectives,

1. The process perspective (How?)
2. The organizational perspective (Who?)
3. The case perspective (What?)

The process perspective looks into the control flow i.e. how a particular task is achieved, the series of tasks involved and so on.

The organizational perspective considers who is involved i.e. it could be the originator of the task. It also looks into people and their role in the organization.

The case perspective takes into consideration the properties associated with a case.

In Figure 10, which is extracted from the Process Mining Manifesto clearly illustrates different tasks, steps, and activities involved in a process mining technique. The process begins with extracting data from the event logs, followed control flow modeling and process modeling and operational support.

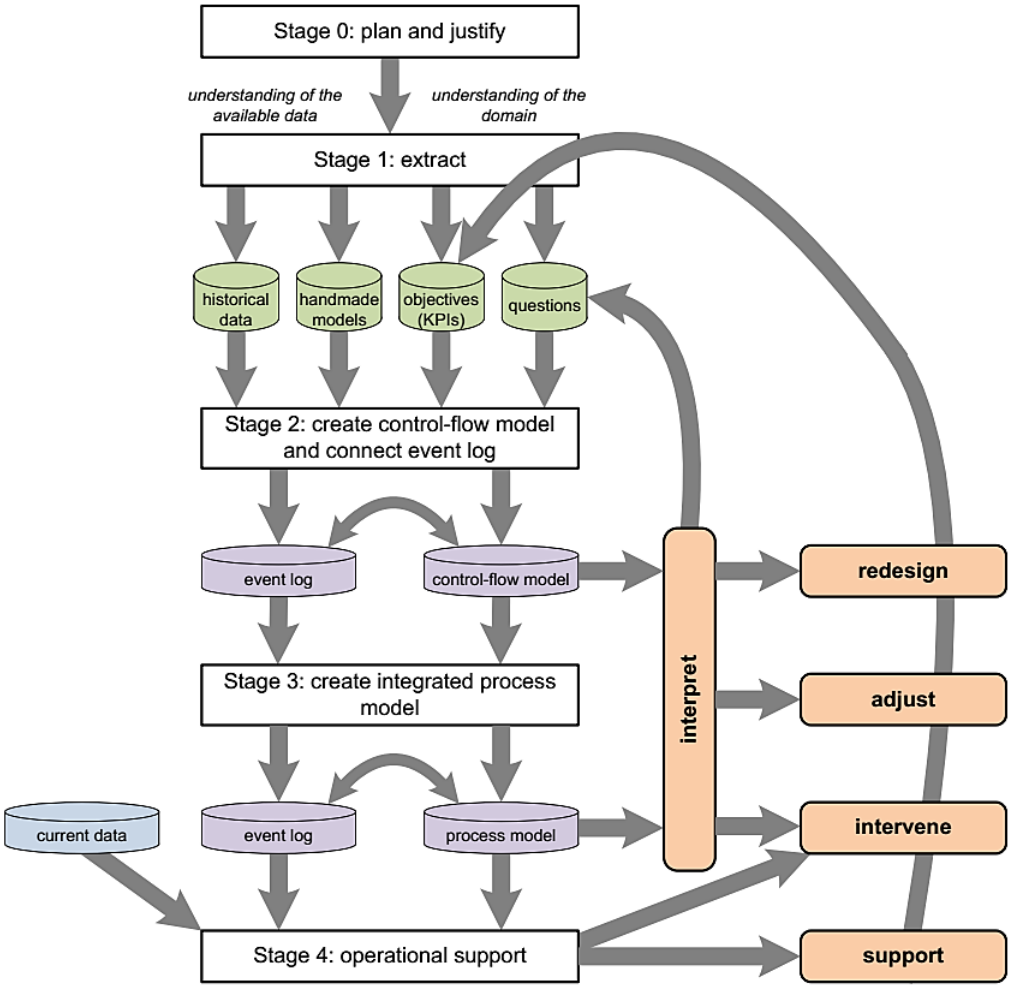

Figure 10 Working of a process mining technique [1]

Event Logs

An event log is a series of data sorted by order of timestamp as shown in Figure 11. The data generally consists of a timestamp, trace id which acts as an identifier of the trace associated with the event, the name of the activity associated with the event. In simpler terms, an event log is a sequence of events sorted in ascending order by their timestamp [6]

case id	activity id	originator	timestamp
case 1	activity A	John	9-3-2004:15.01
case 2	activity A	John	9-3-2004:15.12
case 3	activity A	Sue	9-3-2004:16.03
case 3	activity B	Carol	9-3-2004:16.07
case 1	activity B	Mike	9-3-2004:18.25
case 1	activity C	John	10-3-2004:9.23
case 2	activity C	Mike	10-3-2004:10.34
case 4	activity A	Sue	10-3-2004:10.35
case 2	activity B	John	10-3-2004:12.34
case 2	activity D	Pete	10-3-2004:12.50
case 5	activity A	Sue	10-3-2004:13.05
case 4	activity C	Carol	11-3-2004:10.12
case 1	activity D	Pete	11-3-2004:10.14
case 3	activity C	Sue	11-3-2004:10.44
case 3	activity D	Pete	11-3-2004:11.03
case 4	activity B	Sue	14-3-2004:11.18
case 5	activity E	Clare	17-3-2004:12.22
case 5	activity D	Clare	18-3-2004:14.34
case 4	activity D	Pete	19-3-2004:15.56

Figure 11 An example of an event log [5]

Predictive Process Monitoring

Data extracted from event logs can be practically used for various predictive activities. Prediction based on process mining is an extensively used part of process mining. As our Thesis is about Predictive monitoring, it is important for us to understand what prediction or predictive nature is and understand it with respect to our Thesis.

In business processes, the effectiveness of a managerial decision-making response to variation in the environment depends on the extent to which it can reduce the impact of uncertainty through forecasting. A prediction in process mining is driven by machine learning and data mining techniques. These predictions are fuelled by data collected over a long period and are now being used to predict the future in different scenarios. In this Thesis, we talk about such predictions in the field of process mining.

Process mining generally helps predict the performance of the business operation, the time required to complete a particular task, activity or operation. It helps predict outcomes of a task and in turn, guide the business operation to help it reach its goal. Other predictions that could be met are deadline violation predictions, completion time and next step prediction. Predictions could be done offline and online. Predictions can be done during runtime using the operational support capabilities of process mining. Since in today's date, significant computing power is available, it is possible to compute in real time. There are three operational activities in general, as defined in Process Mining Manifesto [1]:

Detect: As soon as a running process deviates from its actual path an alert is set off.

Predict: Historic data extracted from event logs are sued to build predictive models, which help in making predictions like completion time outcome amongst others.

Recommend: Based on the predictions made, a recommendation model proposes actions to reduce cost or shorten flow time.

All the above activities are performed in offline fashion i.e. the processing does not happen in real-time. On the contrary Operational Support provides a technique to monitor business processes in real time i.e. Online. Assuming a process model generated from an event log and a partial trace, operational support techniques could be utilized for detecting deviations at runtime, predicting remaining processing time (Predict) and recommending the next activity (Recommend) [4].

3.3 Machine Learning Techniques

Now that we have understood basics of Process Mining, it is time to focus on the various machine learning algorithms that are used in Process Mining. It is important to understand the underlying principle of these algorithms inorder to comprehend the techniques better.

Linear Regression

It is the function of the statistical relationships between variables so that one can predict from the other.

$$y_i = b_0 + b_1 x_i + e_i$$

b_0

b_1 = Slope } Parameters x_i, y_i = measured value

e_i = Residue

In a Linear regression, there are only two variables involved. Linear regression helps us to find the relationship between the two variables. In other words, there is one dependent variable and one independent variable. The relationship between the two is established using Linear Regression [7].

Multiple Regression Analysis

Multiple Regression Analysis is a statistical tool helping us to understand the relationship between multiple independent variables and one dependent variable. Once the relation with dependent variables is established, it could be used to make much more accurate predictions relating them.

$$Y' = a + b_1 X_1 + b_2 X_2$$

Y' – this is our dependent variable.

X_i – this is our independent variable.

a - the "Y" intercept.

b_i – the change in Y for each incremental change in X_i

Decision Tree

A decision tree is a set of nodes consisting of root nodes and leaf. The root node is the top most node of the tree and leaf nodes are the nodes at the bottom of the tree.

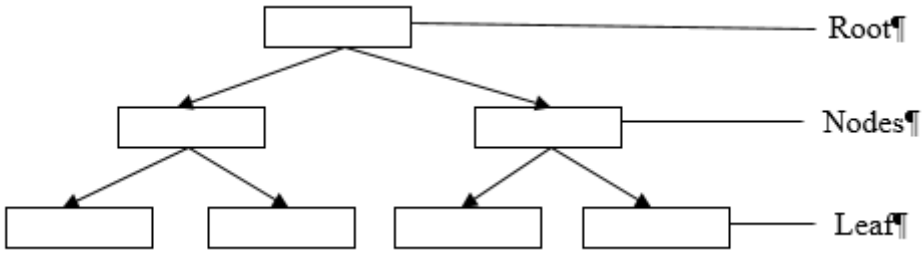

Figure 5: An example of a decision tree

Time series

Time series is the measurement taken over a period of time for a particular event or activity. KPI (Key performance indicators) components can be observed using time series. The series could observe long term, seasonal or periodic pattern.

Sequential Pattern Mining

Sequential pattern mining deals with data represented as sequences (a sequence contains sorted sets of items) [8]. Let us assume we have a database of customers having the following columns: id, booking time, item purchased, amount. Here, each transaction consists of a set of items. And each sequence is a set of transactions ordered according to the timestamp. For decision-making purposes, it is required to obtain sequences with similar behavior.

Trace Clustering

A trace T_n can be defined as, a collection of all the events having the same trace id, where n is the number of events (e).

In trace clustering method, divide and rule approach is applied. It is dependent on a set of profiles, each measuring number of features for each case from a specific perspective. Depending on these feature matrices generated, various distance metrics are applied to compute any two cases in the log. It is followed by the application of a data clustering technique to group closely related cases into subsets [9].

Predictive Clustering Tree

A predictive clustering approach is a clustering approach where in which, once for a particular event its associated cluster is discovered, the prediction for a new instance is dependent on an assignment function which evaluates to which cluster the new instance belongs to. It is believed higher similarity between the instances of the same cluster could help form a more accurate predictor. In a predictive clustering tree, the assignment function is encoded by a decision tree.

C4.5

This algorithm constructs a classifier in the form of a decision tree. C4.5 takes in classified data and predicts the class of new incoming data. Let's suppose a dataset of new born babies with data related to their height, weight, heart beat and various other factors. We need to predict whether the new born is healthy or not healthy. C4.5 has two classes healthy or not healthy. C4.5 is told the class for each new born in the training dataset. Now, using the set of attributes and the new born baby's corresponding class, C4.5 constructs a decision tree which could predict a new born baby's class.

Sliding Window

A sliding window in simple words distinguishes recent data from the past data. It is defined for data stream mining and is similar to First in First out data structure. When an event e_i is acquired and inserted in the window, the latest event e_{i-w} is discarded. w is the size of the window [10].

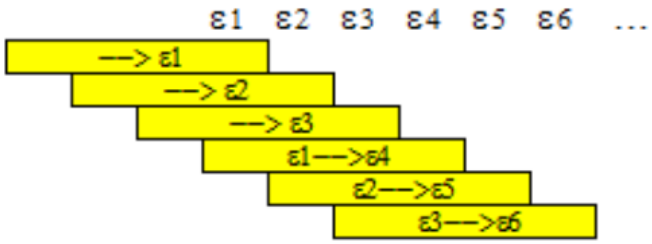

Figure 12 Sliding window model [10]

K means algorithm

It is a clustering algorithm which creates K groups of object, such that object belonging to the same group are similar in nature. It is basically presented in a multidimensional space, where multiple objects are presented using the x and y coordinates. Initially, K points (centroid) are picked to represent the K clusters. Now, every other point picked will be close either of the K initial cluster points (centroids). The centroids are then calculated again based on the new points added. This process continues until all the n points are added. K means generally has issues with outliers and the initial selection of cluster centroid.

Support Vector Machines (SVM)

An SVM is a classifier which classifies data in 2 dimensions using a hyperplane. An easier way to understand hyper plane will be the ball classification example. Imagine a table with red and blue balls spread randomly without much mixing. In such a scenario, we can use a stick and classify the balls into two classes. The balls here represent the data points, black and blue are the two classes and the stick is the hyperplane which is a line in this case. Now imagine if the balls were too mixed, then it is not possible to classify them using a simple line. In this situation, imagine throwing all the balls in mid-air and then classifying them using a sheet of paper. This is SVM in 3 Dimension, the paper acts as the hyperplane.

kNN Algorithm

kNN or K Nearest Neighbour is a supervised classification algorithm. kNN carries out the classification process in a few simple steps. It starts with storing the labeled training data. Now when the unlabelled data is provided for classification, the kNN algorithm first looks for the k closest training data point also known as k nearest neighbor. Secondly, using the neighbor's classes KNN classifies new data. In the case of continuous data Euclidian distance metric is used and for discrete data hamming distance is used for measuring the closest class distance.

ANOVA

ANOVA stands Analysis of Variance. It is a statistical technique which compares differences of mean both among and within a group. ANOVA searches for data variations and where it is found. More specifically ANOVA compares the amount of variation within and between groups. It is used in experimental as well as observational studies.

Ant Colony Optimization

The Ant Colony Optimization (ACO) algorithm as the name suggests is inspired by ants. The ants as they find the shortest path to a food source they leave a substance known as a pheromone, indicating to other worker ants to follow which path. If there is more than one track, the track with more pheromone is chosen. This biological technique was used as an underlying principle for the shortest path algorithm by [11].

3.4 Other Important Terminologies

There are some important terms that have been used regularly in the following chapter. It is important to understand these before we move forward.

Service Level Agreement

A service level agreement is contractually binding agreement between a service provider and a user. It is essential that SLA's are not violated.

Petri Nets

A Petri net is a collection of directed arcs connecting places and transition. Each place can hold a token. Petri nets are used in wide variety of fields like office automation, workflows, manufacturing, programming language, network protocols, real time systems etc.

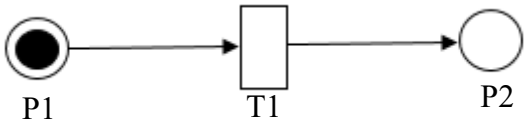

Figure 13 An example of a Petri Net

In Figure 13 an example of a Petri net workflow is clearly shown. In this P1 is a place with a token, T1 is a transition between P1 and P2. When transition T1 is fired token from P1 moves to P2.

YAWL

YAWL is BPM/Workflow system consisting of a powerful modeling language. YAWL helps in data transformation and in full integration with organizational resources and external web services.

KPI (Key Prediction Indicator)

KPI refers to the 4 primary factors which evaluate the performance of a business. These four factors involved cost, quality service, and speed.

Cost is the expenditure involved in carrying out the business operation.

Quality refers to the value of service provided. If the services being offered and giving the expected results.

Service refers to the kinds of facilities provided in a business activity.

Speed, if the business is meeting its requirements at a pace which is relevant to its needs.

These 4 four factors act as the basic prediction indicators.

4 Literature Review

In this chapter, we study in detail the papers listed in the literature list that we generated after rigorous filtering in section 2.2. This chapter has been divided into various categories of prediction based on our study. The literature has been divided into mainly 4 categories (i) Time-based prediction (ii) Process path prediction (iii) Outcome prediction and (iv) Risk Prediction. A detailed introduction regarding each prediction category is provided at the beginning of each prediction inorder to make the reader understand better. Eac prediction category has two subchapters (i) Literature Study and (ii) Literature Summary. The first sub chapter "Literature Study" provides a detailed analysis of the literature related to that particular category. The second subchapter "Literature Summary" provides an excellent overview of the literature study in a tabular manner, easy for the readers to go through. The concepts and techniques mentioned in this chapter have been discussed in detail in chapter 3.

4.1 Time-based Prediction

In this category of prediction, Time is the main attribute. Every client or customer or the user wants to know when a business, activity or a particular task will finish. Accurate time-based prediction is required to fulfill this requirement. Time-based prediction can be classified further into completion time (time taken to complete a business process) [12], remaining time (time left to finish a business process) [6], delay time (amount of delay) [13] and execution time. An example for time-based prediction will be when a client calls an insurance agent, to know the status of his insurance claim. The agent could tell him the expected remaining time of the process [5], when calling a customer care the delay time could be conveyed to the customer to reach the customer representative [13].

In the following sub chapter, we have reviewed a total of 15 papers which fall under this category. We study in detail in detail each literature and try to extract and present information which is valuable for this Thesis topic.

4.1.1 Discussion of the Studies

In [14] the author predicts, the completion time and remaining time of a business process. It is explained using a real-life case study concerning the execution of process instances in an information system for the management of road traffic fines by local police of an Italian municipality. When a driver commits a violation, a policeman opens a new fine. The amount depends on the violation performed. Within 180 days, the fine notification is sent to the offender. The offenders get 60 days to pay his dues or else the fine doubles. If the offender never paid, it is sent to the credit collection department. Two event logs were extracted from this information system and used for the study. The technique proposed in this paper could help in predicting the remaining time associated with such processes. Regression models and Naïve Bayes classifier are used to build the prediction model in this paper. The technique is validated using real life event logs extracted from Information Systems. Logs are of offenders who have not paid the fine in full. These event logs refer to non-overlapping periods in time and contain 1500 log traces and 5000 log traces, respectively. A prom plugin has been implemented for the proposed technique.

In the paper [16], *predicts* the completion time of a process. The author explains historical log data contains a path that was executed in the past. Hidden Markov Model is used to generate and train the predictive model. The model is initialized using the following parameters (i) Observation vector has value 1 or 0 where former means completed while latter means "other state" (ii) Dirichlet distribution is used to initialize the transition and observation matrices (iii) Baum Welch algorithm is used for HMM training.

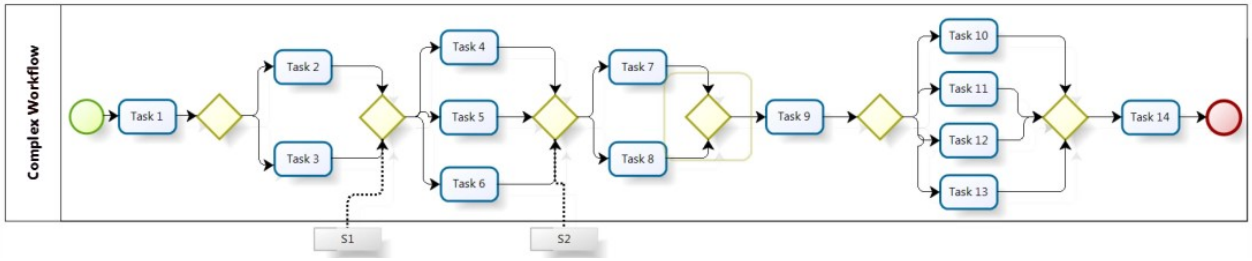

Figure 14 An example workflow [16]

Figure 11 depicts an example workflow pattern. If the tasks that were completed were: T_I, T_2, T_5, T_7, T_9, T_{11}, T_{14}, the observation assigns 1 for each task completed, and 0 otherwise. This can be represented as the observed values: 1,1,0,0,1,0,1,0,1,0,1,0,0,1. Using this information a Hidden Markov Model is generated. The model is used to calculate the probability of occurrence of the path. We use the Forward algorithm (PF) mentioned in the paper, to calculate the probability of occurrence of a given observation sequence. We use this probability value to predict the remaining time for a given sequence of probable tasks. The time to completion can now be obtained from the sample mean function described in the paper. The technique is *evaluated* using the generated event logs, consisting of 1000 concurrent threads recorded for the event occurring from one workflow template. Essentially, each thread is responsible for one instance of the workflow. The technique gives as *output* the average remaining time.

In [17], three kinds of prediction take place, the time based predictive model predicts the remaining time, the next step deadline model predicts the violation of case deadline and the goal based model predicts the next step recommendation of activities. The authors showcase an example using the customer inquiry validation process. The time-based predictor generates all expected predictions. The time-based model estimates the proper remaining time of the partial event log trace and could provide the correct elapsed time and total time. The time-based model also provides the right recommended items that are expected to complete the case in the fastest possible way. The deadline violation predicts the deadline violations when a deadline is set prior to the remaining time. The deadline violation model also helps with recommendations to avoid such violations. The deadline based recommender generates some recommendations that mitigate deadline violation. By offering items that finish the case as quickly as possible. The main *technique* applied is decision mining using decision trees. The proposed technique is *implemented* by creating a plugin for CoCaMa functionality. The technique is *evaluated* using 25 historic cases and five running cases, pertaining to the inquiry process. Those cases are manually populated into the database of CoCaMa. This data set is a starting point for evaluating the quality of predictions and recommendations. The *output* is shown through graphical representations, a progress bar showing remaining time in proposition to total time, the second visualization shows the possible violations, the third visualization shows case goals whether they are supported or unsupported by the current progress. Recommendations are visualized in rows. A recommendation includes the recommended item and reasons why this case item is recommended.

In the paper [18] a technique is proposed to predict the remaining process execution time considering the time passed since last observed event. The paper presents a case study of a

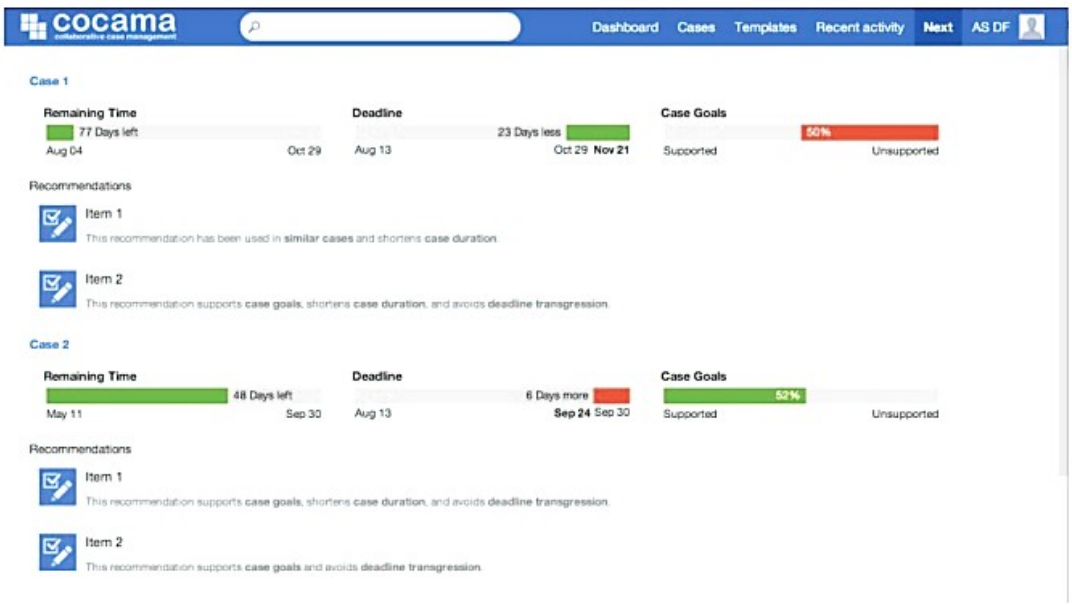

Figure 15 Implementation using CoCaMa application [17]

logistics provider from Netherlands. The logs provided contains entries of the arrival of sea vessels, discharge of container and date of picking up the container by inland transport. The approach presented in the paper tries to predict the remaining process duration for completion of the process. The technique is *evaluated* using a simulated model of real life data mentioned above. From this model 100 traces of execution are generated and stored in the simulated log, besides the simulated model the Petri net model is used as *input*. The technique takes as input Model of the business process, the current trace of the case, current time, number of simulation iteration. The *output* produced shows the remaining process time value. The technique has been *implemented* as a ProM plugin.

In [4], the completion time and next activities are predicted. The proposed prediction model uses sequential pattern mining and is validated using two real life logs ProM and THINK3. ProM concerns repairing telephones of a communication company. The event log contains 11,855 activities and 1,104 cases. THINK3 is an event log presenting 353,490 cases in a company, for a total of 1,035,119 events executed by 103 performers. The completion time is given as output.

Bolt and Sapulveda in their paper [6] used query catalogs. A query catalog is a non-equivalent Partial trace trail occurring in an event log with additional information.

A *Partial trace* PT_m is a sequence that contains first m elements of a trace T_n, where m is the length of the partial trace and m<n and m>0

A *Partial Trace Trail* PTT_n is a sequence of Partial Trace PT_m containing atleast n elements of a particular trace.

In their paper [6], the described technique first compares partial traces. The comparison is done in the following manner

Let e_1 and e_2 be two events, which are considered equivalent if they correspond to the same activity.

***Sequence**:* Two PTT are equivalent by multiset if they have the same set of events regardless of the order.

***Set**:* Two PTT are equivalent if each PTT has atleast one equivalent event in the other PTT

Multiset: Two PTT are equivalent if they have the same set of events regardless of the order.

The *algorithm* does the following, once similar PTT's (Partial Trace Trails) are found, they are stored under each category explained above (set, multiset, sequence). Along with PTT, additional information such as the sum of their remaining time and the frequency of occurrence is also stored. Then based on the variables stored, the average remaining time is calculated. The *technique* is validated using synthetic event logs, generated by a software named Process Log Generator (PLG), which generates Petri Net model from certain pattern probability defined by the user. Event logs are created from these Petri nets generated using simulations. The logs used for *validation* was extracted from a Chilean telecommunication company's systems. The analyzed process in this log is a Technical issue resolution process'', which consists of solving technical issues detected by customers, who call the company and request a solution for their issues. The log consists of 261 traces. Its durations varies from 3.3 h to 58 days. The paper uses *techniques* known as Query catalogs (explained above) to which traces are added, updated and searched (existed)) using basic algorithms mentioned in this paper. The algorithm mentioned in the paper *outputs* the remaining time.

The paper [19] aims to *predict* Runtime prediction of performance measure (remaining process time/steps). The author tries to explain the technique introduced in the paper using a container shipment example. Each container is unloaded from a ship and stored in a dock, and finally carried to a storage area for being stocked. Similarly, while loading the cargo, the container is first placed in a yard and then loaded on a cargo. Various vehicles are used for moving e.g. cranes, straddle carriers and multi trailers. This basic life cycle may be extended with additional transfers, devoted to make the container approach its final embark point or to leave room for other ones. A variety of data attributes are available for each container as context data (of the corresponding process instance), including: the origin and final ports, its previous and next calls, various properties of the ship unloading it, physical features (such as, e.g., size, weight), and some information about its contents. In this paper, additional information like context features for each container: the hour (resp., the day of the week, month) when it arrived, and the total number of containers that were in the port at that moment are also considered. The method is *validated* using logs of a real transshipment system, keeping trace of major logistic activities applied. The *technique* uses Predictive Clustering Tree (PCT) and a discovery algorithm (AA PPM mentioned in the paper) for predicting the average processing time/step.

The paper [20] aims to predict performance measures like remaining time. The solution *algorithm* explained in the paper begins with, (i) Extraction of structural pattern from the logs (ii) Extracting structural pattern that occur frequently in the log. (iii) The patterns are filtered by a function filterPatterns which filter out the top relevant patterns. Once the relevant patterns are detected, a numerical variable is associated with each trace, which are the related targets in the clustering that follows. Finally, each cluster is equipped with a Predictive Performance Model (PPM). The PPM predicts unknown performance values like remaining time based on corresponding trace properties. The abstracted traces are used to predict the remaining time of a process. The author explains using a real-life case study pertaining to transshipment system. The proposed approach was *validated* against the logs of a real transshipment system, more precisely, on a sample of 5336 traces, corresponding to all the containers passed through the system. The *technique* uses clustering based predictive models.

In another paper by Folino [21] the author *predicts* remaining time. The author introduces a new approach where different context related execution scenarios are equipped with separate prediction models. The author *uses* both Predictive Clustering Trees (PCT) and Process Performance Predictors to carry out the proposed technique. The technique is *validated*

using a real-life case study that concerns a transshipment process. A sample of 5336 traces, corresponding to all the containers that passed through the system in the first third of the year 2006 is used.

The paper [22] predicts remaining time. The author explains the paper using a container shipment process. The process begins with unloading the container from a ship and temporary placing it near the dock followed by carrying it to a suitable yard to keep it stored. Different vehicles are used to move the container and basic operations performed on the container are all recorded like MOV, DRG, SHF. Certain SLA's are established that process enactment must not last more than MDT (Maximum Dwell Time), otherwise, penalties are charged on the transshipment company. Another important parameter for this scenario, is the average dwell-time (ADT), i.e., the average halt time for containers in the terminal, which will be also used normalizing time measures. In this paper an ad-hoc predictive clustering approach, capable of detecting different context related execution scenarios (or process variants) is presented, and equips of them with a tailored performance-prediction model. The ultimate goal of the paper is to find a novel kind of predictive model, where performance forecasts for any (unfinished) process instance, are made in two steps: the instance is first assigned to a reference scenario (i.e., cluster), whose performance model is then used to eventually make the forecast. The *technique* uses Predictive Clustering Tree and is *validated* on a real-life scenario, pertaining the handling of containers in a maritime terminal. a series of logistic activities are registered for each container passing through the harbour.

The paper [23] aims to *predict* the duration of an activity. The author explains a scenario of a call center, where a customer calls. The voice receiving unit responds to the call and routes the call to the appropriate station. Depending on the business of the station, the client may have to enter into a queue. If the client is tired of waiting he may end the call. Once the client

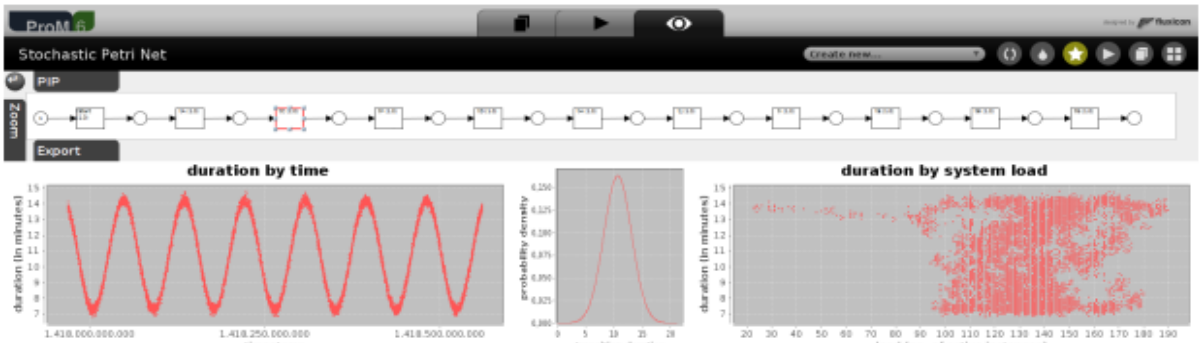

Figure 16 Sequential time series petri net [23]

finishes serving in the queue, he may be connected to a service employee and finish the service and complete the process. To predict the duration of the process, time series effects within the model need to be captured. This paper proposes a Petri net model with time series called Time Series Petri Net (TSPN) as shown in Figure 16. For example, predicting the duration of a single activity which translates to predicting the firing time of the corresponding transition of the Petri net. For this case, the historical data concerning the transition duration are aggregated into a time series i.e. hourly averages. This data is used as training data to fit the model that needs to be compared. Once the models are fit, activity duration is predicted. Therefore, the timestamp is checked at the prediction time and compared with the last observed timestamps. The difference of the timestamps in hours is the forecast. The technique is *evaluated* with synthetic models with their corresponding

processes, a sample of 10; 000 process instances from each model and thereby obtain the simulated event logs. The technique is implemented as a PROM plugin.

The paper [24] proposed technique aimed to *predict* process duration. The author explains the technique using a credit card approval process. The loan approval process starts with an applicant submitting an application. After receiving the application, a clerk could ask for additional information if required. Once the additional required information is received, further checks on applicant's income and credit history are checked. The checks depend on the amount of loan. Once the documents are validated, the application is forwarded to a manager for approval or rejection. In this paper, a workflow simulation model is proposed to tackle such issues *using* event graphs. The simulated logs are analyzed for 3 main problems (i) Suitable resource allocation plan. (ii) performance (process duration) under different case arrival rates and resource allocation plans. (iii) evaluation and prediction from personal performances. The technique is *validated* using a log file of the credit application, part of which is shown. It consists of 17,900 events in total, 1050 cases with 12 different tasks.

In [25]**,** the author aims to predict performance characteristics like waiting time, throughput time and utilization rates. In this paper the author predicts the duration of an activity is observed based on its context. This is done by measuring the time between the events that represent the start and complete transitions for an activity. The measurement is observed from two different angles (i) the resources involved in the activity and (ii) activities preceding the activity (prefix of the trace). Using the resource context function, it is checked whether the resource involved in the execution of an activity influences its duration. The technique is evaluated using the dataset from a loan application process from a Dutch bank, and was originally used in the Business Processing Intelligence Challenge (BPIC) in 2012 The log comprises of 13,087 cases of a loan application process, for which in total 262,200 events have been recorded. There are 36 distinct activities and 69 different resources are involved in this process. The proposed method *uses* ANOVA algorithm and is *implemented* as PROM plugin.

The paper [26] tries to predict deadline violations, a legal deadline such as "Claims need to be handled within three weeks". Various techniques can be used to signal the likelihood of violating such a deadline. For example Taking action X will minimize the risk of violating legal requirement Y. The proposed method is *validated* using a process inside a Dutch municipality based on an event log containing 5187 events related to 796 cases and the technique has been *implemented* as a PROM plugin. The technique gives *uses* decision trees and gives as *output*, the likelihood estimation of deadline violation.

The paper [27] provides a technique for delay prediction. The author explains the process using an example of a Voice Response Unit (VRU) as shown in Figure 17. The customer dials in and is connected to a VRU. The customer can either complete the service via the VRU or transfer it to an agent. Once the service with an agent is over, the customer usually hangs up or in some cases continues the service to VRU or maybe another agent. Customers that seek a service are served by one of the available agents or have to wait in a queue. Hence, activity 'Be Serviced by Agent' comprises a waiting phase and an actual service phase. The customer can abandon the call queue due to impatience. To provide operational analysis for this service process and predict delay of processing, such queues and abandonments must be taken into account explicitly. Queue mining is used to do the delay prediction. The technique outputs average delay time. The *experiments* were conducted on a real-life queue log of a call center of an Israeli bank. The data comes from the Technion laboratory for Service Enterprise Engineering. The log contains an average weekday data on approximately 7000 calls.

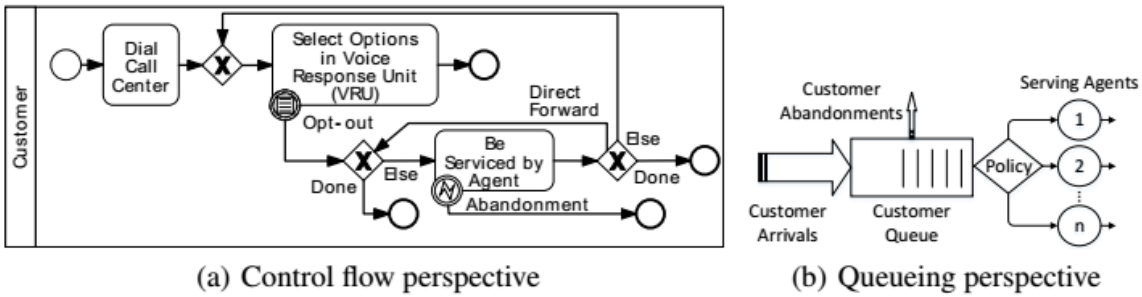

Figure 17 Depicts a BPMN model of such a process [27]

4.1.2 Summary of Time based Prediction

Sub prediction category	*What is predicted*	*Domain*	*Algorithms Used*	*Implementation*	*Output*	*Reference*
Completion time	Completion time and next activity prediction	Issue to resolution	Sequential pattern mining (FAST algorithm)	Not mentioned	Completion time and probability of next activity	Completion Time and Next Activity Prediction of Processes Using Sequential Pattern Mining [4]
Completion time	Completion time and remaining time	Multi Domain	Regressor models, Naıve Bayes classifier	ProM plugin	Remaining time estimation	Data-aware remaining time prediction of business process instances [14]
Completion time	time for completion	Multi Domain	Hidden Markov Model	no implementation	Average remaining time is shown	A Test-bed for the Evaluation of Business Process Prediction Techniques [16]
Remaining time	Remaining time, next step recommendation, case deadline violation	Issue to resolution	Decision mining, decision trees	Plugin extending CoCaMa functionality.	Graphical representation of remaining time, next step recommendations.	Next Step Recommendation and Prediction based on Process Mining in Adaptive Case Management [17]
Remaining time	Remaining process execution time	Logistics	Stochastic Petri Net Model	ProM plugin	Remaining process time value	Prediction of Remaining Service Execution Time

						Using Stochastic Petri Nets with Arbitrary Firing Delays [18]
Remaining time	Remaining process time	Multi Domain	Query Catalogs	No implementation	Remaining time	Process Remaining Time Prediction Using Query Catalogs [6]
Remaining time	Remaining process time/steps	Logistics	Predictive Clustering Trees	Prom plugin	Remaining time/steps using metric average processing time/step	A Data-Driven Prediction Framework for Analyzing and Monitoring Business Process Performances [19]
Remaining time	remaining time	Logistics	Clustering based predictive model	no implementation	Average remaining time is shown	Adaptive Trace Abstraction Approach for Predicting Business Process Performances [20]
Remaining time	Remaining process time	Logistics	Predictive Clustering Tree	No implementation	Average remaining time	Context-Aware Predictions on Business Processes: An Ensemble-Based Solution [21]
Remaining time	Remaining time prediction	Logistics	Predictive Clustering Tree	Prototype of ProM plugin developed	Maximum dwell time, average dwell time is shown	Discovering Context-Aware Models for Predicting Business Process Performances [22]

Process duration	Duration of an activity	Issue to resolution	Time series	Implemented in PROM plugin	Time-based visualization output graph, average time	Time Series Petri Net Models [23]
Process duration	Duration of a process	Financial institute	Simulation	No implementation	Average time	Workflow simulation for operational decision support using event graph through process mining [24]
Waiting time	Waiting time, throughput time and utilization rates	Financial institute	ANOVA	Prom plugin		A Generic Framework for Context-Aware Process Performance Analysis [25]
Deadline violation	Deadline violation	Government	Decision trees	Prom plugin	Likelihood estimation of deadline violation	Auditing 2.0: Using Process Mining to Support Tomorrow's Auditor [26]
Delay	Delay prediction	Multi Domain	Queue mining	No implementation	Average delay time	Queue Mining – Predicting Delays in Service Processes [27]

4.2 Outcome Prediction

Our everyday routines come along with a lot of prediction challenges right from birth: “Is the baby going to be healthy?”, “when will the baby come out?”. When we are in school “Am I going to pass the exam?”, when we appear for an interview, “Will I clear this interview?”. As humans, we always look forward to our future and the outcomes it holds for us.

The same set of questions holds true for business processes and their associated outcomes. In this chapter, we discuss Outcome prediction, as the name suggests the techniques described in this chapter explains whether a current business process will complete successfully or not, whether the business goal will be achieved or not. This category of predictions also includes papers linking exception, violations and failures in business processes. Failure or exceptions occur in business processes when a process deviates from its said path or exceeds the time frame allotted or due to failure in compliance.

In this section, a total of 16 papers are presented. The chapter has been divided into two subchapters, the first subchapter 4.2.1 discusses in detail each literature belonging to this prediction category. The second subchapter section 4.2.2 gives a summary of the papers discussed.

4.2.1 Discussion of the Studies

The technique described in the paper [28] helps in *predicting* the outcome at certain decision points and how these predictions at this decision impact further path. The author explains the technique using an automobile insurance company scenario. The scenario has the following characters CSR (Customer Service Representative), a claim handler (CH), an automobile repair shop (ARS) and the police department (PD). The role of the CSR and PD are limited to one action. In Figure 18, we examine the decision point carShouldBeTotaled, this particular decision point has three potential outcomes are possible. The contents of the document are examined at this decision point and analyzed under what conditions they are further sent to the point sendRepairRequest and further to the point approveAdditionalRepairs. In order to formulate the decision problem, the values of the document content variables (six attributes in this scenario) that are accessible to carShouldBeTotaled are examined. To *validate* this approach an automobile insurance claims scenario was designed and implemented on a simulator. The *technique uses* Ant Colony Optimization technique.

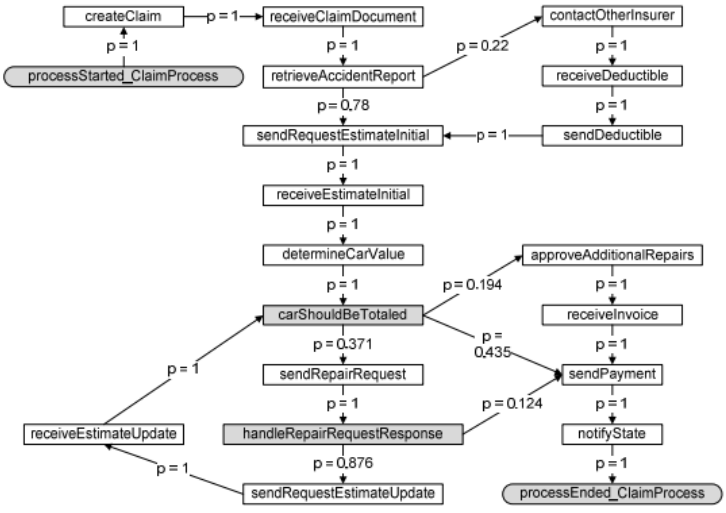

Figure 18 A probabilistic graph of an Insurance Claim scenario [28]

The paper [29] predicts the outcome of each ongoing case. The technique has been explained using the debt recovery process. The process is initiated when the creditor transfers the debt to a company. This showcases the debtor has already defaulted the payment. If the collection specialist considers the case to be irreparable, a phone call is made to the debtor followed by an inquiry/reminder letter. The debtor is expected to provide an expected payment date, in which case no additional action is taken during the present week. Alternatively, the collection specialist and the debtor propose to the creditor about forwarding the debt to an outside debt collection agency (send to encashment) or about sending a warning letter to the debtor on the same matter. The final decision about issuing an encashment warning to the debtor and/or sending the debt to encashment is made by the creditor. If there is no advancement in collecting the debt after 7 days (e.g., the payment was not received on the provided date or the debtor has neither answered the phone nor the reminder letter), the procedure is repeated. The goal of our prediction is to determine outcome cases of cases which have not received payment 8 weeks after the initiation of the debt recovery process. The technique uses text mining algorithms and sequence classification to carry out the process. The proposed method has been implemented in Python. The technique has been validated using two real-life datasets pertaining to (i) the debt recovery (DR) process of an Estonian company that provides credit management service for its customers (creditors), and (ii) the lead-to-contract (LtC) process of a due diligence service provider in Estonia. The technique output a positive prediction if the threshold value is lower than the prediction probability then a positive prediction is made else no prediction is made.

In [15], the probability of a particular outcome is *predicted*. The paper contains three case studies, we will be explaining one of them here. One of the case studies mentioned in the paper describes a telephone repair process. The process starts with device registration after which the phone is sent to the problem detection department where the defect is analyzed. Once analyzed the phone is sent to the repair department, which consists of two teams, simple defects repairing team and complex defects repairing team. Once the defect effect is resolved, the case is archived and the device is sent to the customer. The company tries to fix the defect in a limited number of times. The optional properties are defect type (ten categories), phone type (three categories), defect fixed (true or false) and number of repairs (0, 1, 2, 3). The model presented in the paper tries to predict the properties of an event. Given a running case, the technique helps in answering questions like “what is the activity of the next event?", “who is the resource triggering the next event?" and so on. The validation log of telephone repair contains 1104 cases and 11855 events. Predictive clustering tree technique is used in this approach. The output is a PCT (Predictive Clustering Tree) with a yes/no of all the associated properties value.

In [30] the outcome of an ongoing case is predicted along with remaining time. It is similar to paper [19] by Bevacqua which is a data driven approach. The author tries to explain the technique introduced in the paper using a container shipment example. Each container is unloaded from a ship and stored in a dock, and finally carried to a storage area for being stocked. Similarly, while loading the cargo, the container is first placed in a yard and then loaded on a cargo. Various vehicles are used for moving e.g. cranes, straddle carriers and multi trailers. This basic life cycle may be extended with additional transfers, devoted to make the container approach its final embark point or to leave room for other ones. A variety of data attributes are available for each container as context data (of the corresponding process instance), including: the origin and final ports, its previous and next calls, various properties of the ship unloading it, physical features (such as, e.g., size, weight), and some information about its contents. In this paper, additional information like context features for each container: the hour (resp., the day of the week, month) when it arrived, and the total number of containers that were in the port at that moment are also considered. The model uses regression models and

EM algorithm. The technique is *validated* using a collection of 5336 log traces generated by a real transshipment system. Each trace stores a sequence of major logistic activities. The model is *implemented* for testing purposes as a standalone application. The output of a probabilistic model is a probability value of belonging to the particular cluster.

The paper [31] aims to *predict* future behavior of business processes. The technique uses a table look up approach, known as History. It uses as input an entire sequence of event, selects from the training set all process instances that begin with the sequence and estimate the probability of the next event will be of a given event type by counting how many of the selected sequences continue with an event of the given type and dividing this count by the number of process instances that begin with the selected sequences; that is, the prediction is based on the relative frequency with which the given event type occurs after a particular sequence. The event type with the highest probability is predicted to be the next one. The technique is validated using real-world event logs from the 2012 and 2013 BPI challenges (van Dongen 2012, 2013). From the 2012 challenge, the author gathered anonymized data from a Dutch financial institute's loan or overdraft application process— consisting of 262,000 events in 13,087 instances. From the 2013 challenges, two logs from Volvo IT Belgium that describe incident- and problem-management processes were used. The incident-management log contains 3,777 instances with 36,730 events of 11 types, and the problem management log contains 744 instances with 4,045 events of 5 types. The *technique uses* EM Algorithm and a probabilistic model. The technique *outputs* probability estimation of next event.

The paper [32] *predicts* violation of an aggregate performance constraint. The author proposes a technique which addresses the problem of predicting whether the process instance in each time window will infringe or violate an aggregate performance constraint at a series of checkpoints. At each checkpoint t, a two-level estimation must be carried out, which are: (i) what performance outcome each ongoing (i.e. not finished by t) process instance will yield, and (ii) how many process instances will start in the rest of the window (i.e. after t), and what their aggregate performance outcome will be. The *algorithm* used is based on time series and clustering algorithms. The *technique* is validated using logs pertaining to the handling of containers in a maritime terminal. A log of 5336 traces, registering the history of all the containers that transited through the harbour in the first third of 2006. The *output* generates a binary classification of violation prediction.

In [33], the case predicts the likelihood of a case yielding in the event of an intervention such as audit. The Irish Tax and Customs authority has been developing the use of data mining techniques and putting up analytics at the core of its business process. The department has successfully built models predicting possible non-compliance/ tax evasion, and liquidation. The models aim to *predict* the likelihood of a case yielding in the event of an interventionion, such as an audit. In addition, all audits completed by Revenue in the year after the models had been created were assessed *using* the model probability to yield a score, and a significant correlation exists between the expected and actual outcome of the audits. This process has been *developed by* SAS and it forms a solid framework for analysis. An *evaluation* of the Predictive model using the latest REAP run was conducted in mid-January 2011. The method used was 'Back Validation' of the 2010 yield prediction model, with closed audits (i.e. known results) from 2010. These audits were not used in the training of the model; hence the model can be assessed on the basis of how well it predicted events unknown to it at the time of its creation approximately 475 cases (total of 9500 cases).

The paper [34] *predicts* the outcome of a loan application process. In this technique, the author observes the following, the loan approval process starts with a customer submitting an application and sending the application with status Approval, Cancellation or Rejection. Each

case contains AMOUNT_REQ which states the amount requested by the applicant. For each event, the extract shows the type of event, life cycle stage (Schedule, Start, Complete), a resource indicator and the time of event completion. The author *uses* Classification and Regression Tree (CART) to predict the outcome of a loan approval process. The technique is *validated* using a real life event log of the loan and overdraft approvals process from a bank in Netherlands and is analyzed using process mining and other analytical techniques.. The event log is comprised of a total of 262,200 events within these 13,087 cases.

In [35] the KPI (Key Prediction Indicator) of a business is predicted. The author uses classification and regression techniques to implement the prediction model. The presented architecture has been implemented on top of an internal release of a Mayor software vendor's BPMS which also integrates CEP functionality. The technique has been validated using synthetic logs of a simple repair process.

The literature [36] aims at *predicting* costs associated with a process by combining cost data with historic information to help in the decision-making process. The technique proposed in literature could be explained by using an example of a simple telephone repair process. The process starts with the registration of the faulty phone sent by the customer. Which is followed by the test department. The test categorizes the defect and sends it to the solve department and a letter to the customer explaining the problem. The solve department has 2 distinct teams to solve simple and complex defects. Once the defect is resolved it is sent back to the test department for testing. If the test passes, it is finally sent back to the customer. In this process cost annotation is done, there by assigning a cost to an individual employee or a role (e.g. Tester). Cost rate for each process, simple/static costs, cost of an activity is calculated based on the time it takes. All types of cost like fixed cost, labour cost, overheads are included. By combining additional cost information, it becomes easier to assess cost impacts during decision making processes. Simulation logs were used for *validating* this technique. A dataset with 1000 completed cases of the telephone repair process was used for validation. The log contains timestamps (task start date and completion date), by whom, phone type, defect type, and the repair outcome were also used. A Prom Plugin for cost reporting and cost prediction is *implemented*, the plugin can be found in the "CostBasedAnalysis" package in ProM repository. The *output* of the technique is a cost annotated transition system model (activates and resources) with the cost associated with each.

The paper [37] predicts the probability of achievement of a goal. The proposed technique is evaluated on an event log which is taken from a phone repair process and is publicly available and used in some research for evaluations. The log contains 11855 events from 12 different events in 1104 cases, each case representing a phone terminal repair process (register, analyze defect, repair, test repair, archive and etc.) (http://www.processmining.org/logs/start). A decision tree based CART is used is used to implement this technique. The output of the technique is a table showcasing false positive and false negative outcomes.

The paper [38] proposes a technique to predict whether a business goal is achieved or not. The *example* mentioned in the paper is a cement manufacturing company and the scenario explained is of concrete generation. The goal of the business process is to get an acceptable strength factor of the concrete generated. In the manufacturing process, seven activities are executed sequentially, and the age of the concrete is checked in the quality test. Finally, a compressive strength of the concrete is evaluated based on eight measurements as described in the paper. In real-time monitoring, a process instance is observed through attributes along with real-time progress so that it is updated additionally as the monitoring time elapses. The final performance is evaluated based on whether the strength generated was acceptable or not. If the concrete generated is of acceptable strength, then the performance is determined to be a success or else

a failure. The author *uses* Support Vector Machines (SVM) Classifiers to implement the proposed technique. The *evaluation* of the technique is done on real life logs which were modified bit. The dataset used for *validation* was of the concrete compressive strength dataset presented by Yeh (1998), in the UCI machine learning repository (Frank and Asuncion, 2010). The dataset is composed of 1,030 instances and each instance is a set of eight quantitative input variables and one quantitative output variable, which correspond to kinds of measurements from the course of making a concrete and the corresponding compressive strength of the finished concrete. The *output* is presented as the probability of achieving a business goal.

The paper [39] *predicts* the likelihood of achieving a business goal. The author explains the technique described in his paper using the *example* of a patient diagnosis. During the process execution, the doctor takes decisions on therapies and on the dosage of medicines to be administered to the patient. The process starts with lab test results provided to the doctor by the patient. Based on the tests, the doctor formulates a diagnosis. Then, the doctor prescribes a therapy. The goal for the doctor could be, every diagnosis is eventually followed by the patient recovery. By exploiting data related to the clinical history of other patients with similar characteristics, this technique aims at providing the process participants with predictions about whether the patient will recover or not. Whenever the doctor has to make a decision (e.g., prescribe the type of therapy or choose the dose of a medicine), recommendations are provided about the options for which it is more likely that the patient will recover. The technique is *validated* using BPI challenge 2011, event log. This log pertains to a healthcare process and, in particular, contains the executions of a process related to the treatment of patients diagnosed with cancer in a large Dutch academic hospital. The whole event log contains 1, 143 cases and 150, 291 events distributed across 623 event classes (activities). Each case refers to the treatment of a different patient. The technique is *implemented* using decision trees and a ProM plugin has been developed as part of the implementation. The *output* is shown as a probability value, the value estimates the chances of achieving the business goal.

The paper [40] *predicts* possible process disruption ahead of time. The author explains, for *example*, an airplane takes goods from the JFK International Airport (USA) to Amsterdam Airport Schiphol (NL), where they are transferred to a truck sent by a Logistics Service Provider (LSP) and transported to a destination in Utrecht (NL). The main aim of LSP is to deliver the goods on time, for which the connection point in Amsterdam is very critical. If the airplane has to divert and land at a different airport (e.g., due to a thunderstorm near Amsterdam), the LSP has to cancel (or re-route) the truck that was sent to Schiphol, and in parallel reserve another vehicle to pick up the cargo at the new location. In order for these corrective actions to be effective, it is crucial that the LSP is aware of the airplane diversion as soon as possible, which implies constantly monitoring the task in charge of air transport. The technique described in this paper helps in predicting such disruptions up front. The technique was *evaluated* using a case study on AirFreight transportation consisting of 119 logs of events reporting flight data in the U.S. during May 2013 (98 regular flights, 21 diverted). Data were gathered from Flightstats, a data provider for air traffic information. The *technique uses* Support Vector Machines for its implementation.

The paper [42] predicts the process parameters for quality assurance. Figure 19 describes the complete working of the RFID based technique in the garment industry.

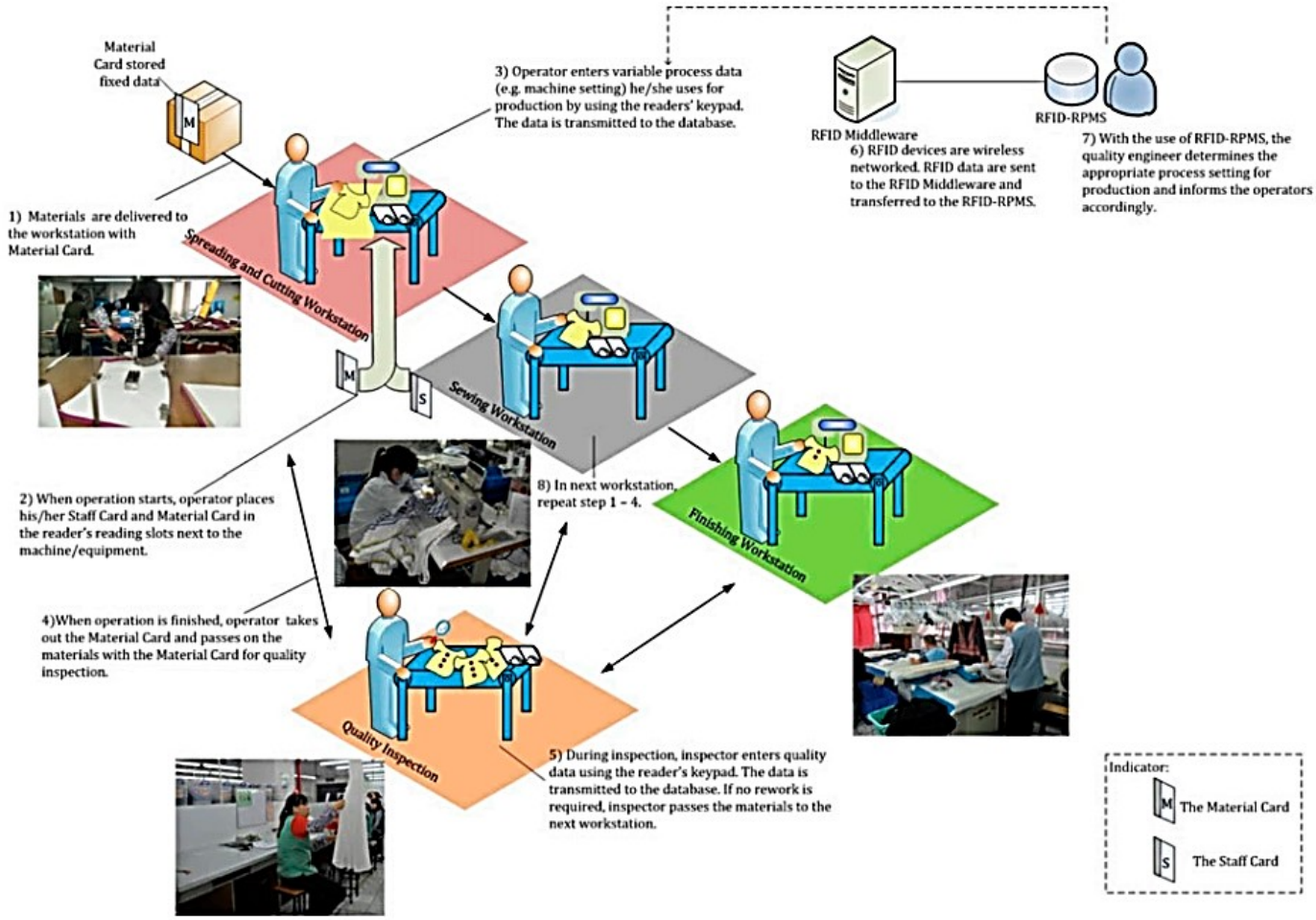

Figure 19 Flowchart of operation with the RFID [42]

Each workstation is provided with an RFID based card. The card contains machine setting values which need to be optimized for optimum performance. The RFID based method uses fuzzy mining algorithm and is *validated* using a case study, conducted in a Hong Kong-based garment manufacturing company located in Shenzhen, China. The technique is implemented in a garment industry using RFID based technology. The initials input taken are the initial parameters.

The paper [43] presents a technique to minimize overprocessing by predicting cases of rejection or knockout. The author explains the technique using an example of a loan application process. When a customer applies for a loan, a loan application is filled up, which is followed up by 3 check process the identity check, the creditworthiness check and finally the document verification process. A negative outcome of any of these checks leads to a negative result. The checks are performed by bank clerks. The order in which the checks are conducted has changed. A knockout session starts as soon as a case is lodged. Thus, the only features can be used to build this predictive model is the case. attributes, i.e. the information provided by the borrower at the time of lodging the application. These features can be grouped into three categories. Demographical features which include age of the loan borrower, their gender, country of residence, language, educational background, employment and marital status. Financial

features describe the borrower's financial well-being and include information about their income, liabilities, debts, credit history, home ownership, etc. Finally, the third group includes loan features, such as the amount of the applied loan, and its duration, maximum acceptable interest rate, the purpose of the loan and the application type (timed funding or urgent funding). The model aims to predict reject probabilities and knockout cases right at the start avoiding overprocessing. The technique is *validated* using the loan origination process of Bondora, an Estonian peer to-peer lending marketplace; the second one originates from an environmental permit request process carried out by a Dutch municipality, available as part of the CoSeLoG project. The proposed framework *uses* classification and regression models. It is *implemented* using as a set of scripts for the statistical software R. A package containing the R scripts, the datasets, and the evaluation results are available at http://apromore.org/platform/tools.The *output* is shown as the maximum likelihood decision value.

4.2.2 Summary of Outcome Prediction

Sub prediction category	*What is predicted*	*Domain*	*Algorithms Used*	*Implementation*	*Output*	*Reference*
Outcome prediction	Predict outcome at certain points in the business process	Insurance	Ant colony optimization, decision tree	No implementation	Probability estimation of which path to take	Predictive Analytics for Semi-structured Case Oriented Business Processes [28]
Outcome prediction	Outcome of each ongoing case	Financial Institute	Text mining algorithms and Sequence classification	Implementation in Python (https://github.com/irhete/PredictiveMonitoringWithText)	Probability estimation	Predictive Business Process Monitoring with Structured and Unstructured Data [29]
Outcome prediction	Remaining processing time and the probability of a particular outcome for running cases	Multi Domain	Predictive Clustering Tree	No implementation	Yes/No and all the associated properties value	Process Mining to Forecast the Future of Running Cases [15]
Outcome prediction	Outcome of any ongoing process case, or the remaining time)		Regression models, EM Algorithm	No Implementation	Probability value of belonging to a particular cluster	A Cloud-Based Prediction Framework for Analyzing Business Process [30]

Outcome prediction	Predicts future behavior of business processes	Financial Institute	EM Algorithm, probabilistic model	No implementation	Probability estimation of next event	Comprehensible Predictive Models for Business Processes [31]
Outcome prediction	Outcome each ongoing process instance and violation	Logistics	Time series algorithm and clustering algorithms	No implementation	Binary classification of violation	A prediction framework for proactively monitoring aggregate process-performance indicators [32]
Outcome prediction	Predict the likelihood of a case yielding in the event of an intervention such as audit	Government Organization		This process has been developed by SAS and it forms a solid framework for analysis	Predict the likelihood of a case yielding, probability score in case of an audit	Predictive Analytics in the Public Sector: Using Data Mining to Assist Better Target Selection for Audit [33]
Outcome prediction	Outcome of a loan application process	Financial Institute	Classification and Regression Tree (CART)	No implementation	Not mentioned	Process Mining-Driven Optimization of a Consumer Loan Approvals Process [34]
Outcome prediction	KPI values	Issue to Resolution	Classification and Regression algorithms	Implemented on Mayor software vendor's BPM Suite	Not mentioned	preCEP: Facilitating Predictive Event-Driven Process Analytics [35]

Cost	Prediction of cost associated with a business process	Issue to resolution	Not mentioned	ProM Plugin	A transition system model with cost annotation	A Framework for Cost-Aware Process Management Cost Reporting and CostPrediction [36]
Business goal	Goal achievement	Issue to resolution	Decision tree based on CART algorithm	no implementation	False positive and false negative	Process Mining Approach Based on Partial Structures of Event Logs and Decision Tree Learning [37]
Business goal	Whether a business goal is achieved or not	Manufacturing Industry	SVM Classifier	Not mentioned	Probability of achieving the final goal	Periodic performance prediction for real-time business process monitoring [38]
Business goal	Estimation or likelihood of achieving a business goal	Healthcare	Decision tree	Implemented as a ProM plugin	Probability of achieving a business goal per case and recommendations (confidence value)	Predictive Monitoring of Business Processes [39]
Process disruption	Predicting possible process disruption ahead of time	Logistics	SVM	no implementation		Predictive Task Monitoring of Business Processes [40]

Quality assurance	Process parameters for quality assurance	Manufacturing Industry	Fuzzy association rule mining algorithm	Implemented in a garment industry using RFID based technology	Process parameters	An RFID-based recursive process mining system for quality assurance in the garment industry [42]
Predict rejection cases or knockout	Minimizing over processing by predicting cases of rejection or knockout and required processing time	Multi domain	classification and regression models	R scripts http://apromore.org/platform/tools.	Maximum likelihood to decision value	Minimizing Overprocessing Waste in Business Processes via Predictive Activity Ordering [43]

4.3 Process Path Prediction

Process Path prediction is an important category of prediction, which as the name suggest helps in predicting the process path of a business process.. For example, in a hospital scenario, a doctor could predict possible treatment sequences or paths based on successful patient records and carry out his treatment accordingly. This category of prediction is also helpful in avoiding failures by predicting process misbehaviour or deviation from actual path. An important part of this prediction is to determine the future tasks This helps businesses to decide, the best possible next step to complete a task successfully.

For example, in an insurance case, opening a claim, getting an accident report and confirming the accident with a witness could lead to payment for the damage caused to a car more likely than opening a claim, running an audit and confirming the invalidity of the policy. Selecting the right process path will lead to the right desired outcome.

A total of 7 papers falls under this category. These papers have been studied in detail and presented in the following section. A summary of the literature has been presented in the subchapter 4.3.2.

4.3.1 Discussion of the Studies

The paper [44] proposes a *technique* for process path prediction. In an insurance case, opening a claim, getting an accident report and confirming the accident with a witness could lead to getting paid for the damage to the car more likely than opening a claim, running an audit and confirming the invalidity of the policy. Predicting the result of the case will reveal the effectiveness of audit vs claim department to the business owner, which helps the business owner decide how to invest in business units for improving customer satisfaction. The example presents two parallel paths as shown in Figure 20. The author *implements* the proposed techniques using certain algorithms described in the paper to select correct path representational model and decision tree. The model is *validated* by simulating traces of sample business process that contains a company's marketing strategy for a business process. The technique takes as *input* a business model and the execution traces and gives out the prediction regarding which is the best model.

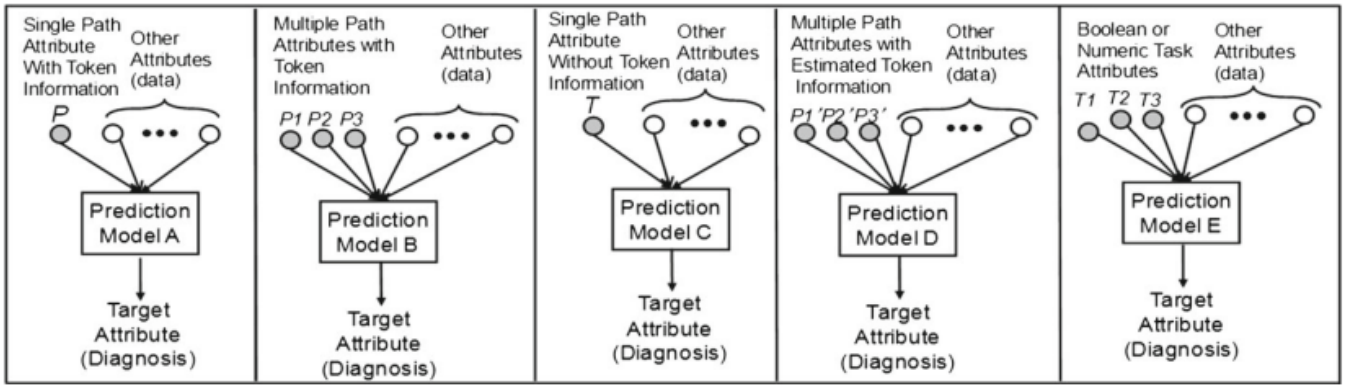

Figure 20 Different path representation [44]

The paper [47] targets to *predict* a future sequence of events. The author describes a scenario where the model is given a partial sequence of events, the model tries to predict the likelihood of achieving a future sequence of events. To *validate* this approach, a real-life event log from the business process intelligence challenge in 2012 is used. The log is a Dutch financial institute and contains about 262,000 events in 13,084 process instances. The described technique *uses* Probabilistic finite automata (PFA) and Maximum a posteriori (MAP) estimation. The *output* of the technique is the likelihood estimation.

The technique presented in [46] describes what event is likely to follow the current event. The framework requires historical event data and fits it to a probabilistic model. Using a real-time data feed, it determines probabilistically what event is likely to follow. The functionality is implemented using machine learning research and grammatical inference, which is a discipline concerned with learning formal languages from data. The final goal is to provide a basis for implementing process intelligence approaches capable of anticipating opportunities and threats before they occur, which gives managers a window of opportunity to take suitable actions. The technique *uses* a regularized Predictive Finite Automata (PFA) with EM-based parameter estimation combined with grid search. The framework is *implemented* in Java is publicly available (https://github.com/DominicBreuker/RegPFA). The *output* of the event is given as the probability of the next event occurrence.

The paper [10] predicts next process step. The author explains, in large IT companies, the maintenance of a process or processes is implemented across a large number of IT systems. Over time it so happens that the company loses track and the process evolves in an uncontrollable fashion. There is different execution sequence for different processes. The solution is to 'divide' the process into smaller groups of tasks/steps and at each group, build a model accordingly. another strategy addressing the complexity and diversity of process data which partitioning process data into groups of sequences of similar characteristics. Mathematical models are chosen according to the properties of the sequences in the resulting groups. The strength of this method is that it keeps the sequential form of the process, discriminates and adequately deals with different representative prototypes. Logs from two real processes (DS1, DS2) of a multinational telecommunications company were used for validation purposes. DS1 is a legacy process, consisting of 2367 process instances of different lengths. DS2 is also a real process with 11839 entries, 576 process instances with different lengths. The *output* of the technique is the likelihood estimation of the next step.

The paper [48] aims to predict the next step in a business process. To *predict* an outcome of a business process, the main idea is to allocate a number of similar sequences from historical data, expecting the process would have a similar outcome. kNN Algorithm (K Nearest Neighbour) chosen as the predictive model here. A number of sequences are considered in which the search is for K sequences which are similar to the sequence provided. The similarity of the sequences is determined using a distance measure. The distance measure used in this paper is the edit distance. In order to exploit the temporal characteristics of process data, the technique combines KNN with sequence alignment from biology to form a sequential KNN. The technique is *validated* using two datasets (DS1) consisting of a 9 month telecommunication line fault repair records. The second log (DS2) deals with broadband fault repair records of 1 month. The technique gives as *output* the likelihood estimation of the next step.

The paper [49] uses process mining technique to *predict* traffic overload while changing a network topology. The main idea presented in this paper is to look for network topology changes and redistribute traffic in case of overload nodes prediction. NS-3 technique is used to create the dynamic mesh network and process mining is used to extract information regarding network topology and network traffic. A Petri net model is generated as the *output*. The generated model is used to *detect* traffic overload in networks.

The paper [41] aims at *predicting* abnormal termination. Pre-processing aims at defining an upper control limit of a LOF (Local Outlier Factor) value by analyzing historical process instances as training data. After calculating their LOF values, an upper control limit (UCL) can be derived by applying kernel density estimation (KDE) to LOF values. This preprocessing should be performed periodically to reflect currently recorded instances in updating the UCL. In the real-time monitoring phase, an ongoing instance is monitored continuously, and attributes are generated gradually and recorded in each monitoring period. From observed attributes, k plausible instances are generated by modified KNNI (k nearest neighbour

imputation). After calculating their LOF values, the probability of abnormal termination is predicted by comparing the distribution of LOF values with the UCL. If the probability exceeds the predefined threshold, the chances for an abnormal termination rises. The method is *validated* using randomly generating 10,000 instances composed of 6 attributes from these distributions. The *technique uses* KNNI (k nearest neighbour imputation) based local outlier factor algorithm. The *output* is presented as a probability of abnormal termination.

4.3.2 Summary of Process Path Prediction

Sub prediction category	*What is predicted*	*Domain*	*Algorithms Used*	*Implementation*	*Output*	*Reference*
Path prediction	Process path prediction	Multi Domain	Algorithm to select the correct path representation model, decision trees	No implementation		Leveraging path information to generate predictions for parallel business processes [44]
Path prediction	Predict future sequence of events	Financial institute	Probabilistic finite automata (PFA), Maximum a posteriori (MAP) estimation		Likelihood estimation of next event	Designing and Evaluating an Interpretable Predictive Modeling Technique for Business Processes [45]
Path prediction	Predictive likely type of event to follow	Multi Domain	Probabilistic finite automata (PFA), Maximum a posteriori (MAP) estimation combined with grid search	The framework is implemented entirely in Java and is publicly available at https://github.com/DominicBreuker/RegPFA	the probability of next event occurrence	Designing and Implementing a Framework for Event-based Predictive Modelling of Business Processes [46]

Path prediction	Next step prediction	Telecommunication	K means, Markov models, Sequence alignment	No implementation	Likelihood estimation of the next step	Sequential Clustering for Event Sequences and Its Impact on Next Process Step Prediction [10]
Path prediction	Process next step	Issue to resolution	KNN (K Nearest Neighbour), Markov model, Sequence alignment	No implementation	Likelihood estimation of next step	A hybrid model for business process event and outcome prediction [48]
Network traffic prediction	Predict traffic overload	Network	Not mentioned	no implementation	Petri Net model	Traffic Prediction in Wireless Mesh Networks Using Process Mining Algorithms [49]
Abnormal termination	Prediction of abnormal termination		KNNI (k nearest neighbor imputation) based local outlier factor (LOF) algorithm	no implementation	Probability of abnormal termination	Real-time business process monitoring method for prediction of abnormal termination using KNNI-based LOF prediction [41]

4.4 Risk Prediction

Many business goals come with predefined risks, like issuing a new credit card to a client or approving a loan. These processes involve risk and it is important to predict the amount of risk involved, before approving a client. In [50] the author talks about the installation of an ATM machine, he describes the risks involved if the machine malfunction. The assessment of such risk factors and providing recommendations to reduce such risks is the goal of this chapter.

4.4.1 Discussion of the Studies

In [50], the author proposes an approach to *predict* the risk involved in the business process and the process outcome. Paper talks about Risks that threaten the achievement of overall process goals like completing the majority of cases within a time period or within a given budget. The *example* given in the paper describes the transactional processing services to banks which include issuing new credit cards, installation of new ATM and POS (Point of Sales), card transaction processing etc. The process starts with client bank lodges a request for adding a new terminal on the host (task Lodge Request). Then an engineer looks into the information provided by the bank and adds more information by the terminal on the host. Then the information is double checked by another engineer. Then finally security keys are loaded to enable safe communication. In this scenario, risky process behaviours are explained that could cause cost overrun for example a wrong currency could be assigned to an ATM terminal. Consequently, incorrect amount of money being withdrawn causing financial loss for a bank or a certain client. Another case could be when bank launches an urgent case and the case is not started immediately breaking the SLA between bank and processing center, resulting in a monetary penalty. In another scenario, process behavior is considered which could have an impact on the outcome of a process. When a high number of Urgent Cases arrive in a short period, employees will focus on urgent cases, neglecting other kinds of cases. Such a scenario may not sound risky, but multiple occurrences of such scenarios could turn out to have higher risk levels. The technique described in the paper involves replaying the event log on a Petri net with and creating non-parametric regression models. The model described in this paper helps in predicting the risk involved before its arrival. A *plugin* Process Risk Evaluation" for ProM framework has been developed. The plugin takes event logs in XES formats and process models in PNML formats as input.

In the paper published by Conforti in the following year [51] aims to monitor the risks. The method described in this paper is explained using an approval process for a personal loan or overdraft. The process starts with the submission of an application. The financial department can already reject the application ending the process or pre accept it for further processing. The application is handled by a staff who first adds missing information to the application until the application is completed. Once the application is finished, an offer is sent to the customer. Following which periodic calls are made to the customer. After the customer made her decision, the application will be finally assessed while adding still missing information. The corresponding log contains traces with events that cover a period of six months, i.e., from October 2011 to March 2012. The log is pre-processed using two phase approach, first infrequent labels are removed and secondly infrequent behavior in the log is removed. Having pre-processed the log, a process model is extracted. In case an instance already executed the task more often than the defined, the left value of the risk condition must not exceed 1, as it shows the probability that the instance exceeds the specified maximum amount of executions. Whenever the probability exceeds the defined threshold, the PING (Process Instance Graph) is created and the risk propagation algorithm is triggered. The technique is

validated using a process for a personal loan or overdraft application in the context of a real-world scenario from a Dutch financial institute. The corresponding log data was released as part of the BPI Challenge held in conjunction with the 8th International Workshop on Business Process Intelligence 2012. The PRISM technique is *implemented* in Workflow Management System Camunda. The task uses similarity-weighted process instance graph (PING) and a risk propagation algorithm. The method *outputs* risk propagation as TRUE/FALSE.

In paper [52], a recommendation system that supports process participants in taking risk-informed decisions, with the goal of reducing risks that may arise during process execution is proposed. The author explains in his paper a claim handling scenario. The scenario starts when a new risk is received from a customer. Upon receiving an assessment is done and the customer is contacted to inform regarding the assessment. The customer may provide additional documents such as ("Receive Incoming Correspondence"), which need to be processed ("Process Additional Information") and the claim may need to be reassessed. After contacting the customer, a payment order is generated and authorized to process the payment. During the entire process, the customer needs to be kept updated about the status of the project as follow ups. Three kinds of faults that could be encountered in the claim handling process are the Over-time fault, Customer-dissatisfaction fault., Cost overrun fault. The *technique* proposed in this paper measures the severity of these models and provides recommendations keep them in mind. The technique is *validated* using the claim handling process and related event data, of a large insurance company kept under the condition of anonymity. The event data recording about one year of completed instances (total: 1065 traces). The proposed technique has been implemented on top of YAWL BPM system shown in. The UI to support participants in choosing the next work item to perform based on risks (Figure 7).

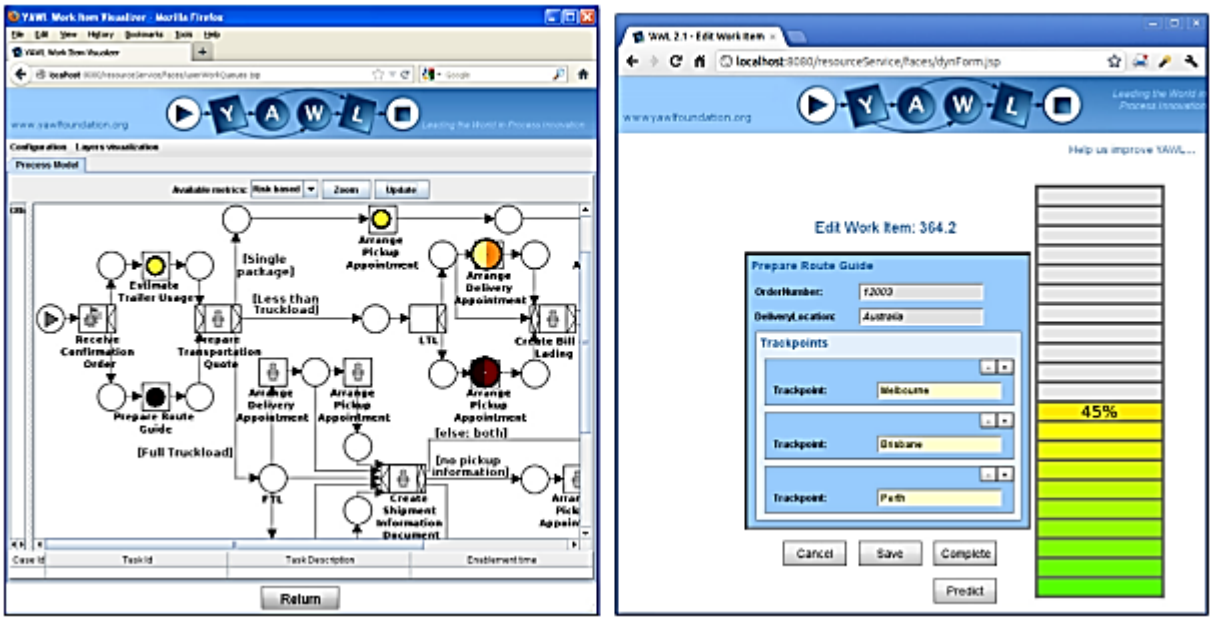

Figure 21 i) The YAWL helps participants in choosing the next work item to perform based on risks and (ii) UI to supports participants filling out a form based on risk [52]

4.4.2 Summary of Risk Prediction

Sub prediction category	*What is predicted*	*Domain*	*Algorithms Used*	*Implementation*	*Output*	*Reference*
Process Risk Prediction	Process risk and process outcome	Multi Domain	Non-Paramtric regression model	ProM Plugin	Mean process risk at time t, Total process risk at time t, aggregate process outcome at time t.	Evaluating and predicting overall process risk using event logs [50]
Process Risk Prediction	Risk monitoring	Financial institute	Process instance graph (PING) and a risk propagation algorithm	Implemented as an extension of the workflow management system Camunda	Risk propagation TRUE/FALSE	PRISM – A Predictive Risk Monitoring Approach for Business Processes [51]
Process Risk Prediction and Recommendation	Recommendation system helping take risk informed decisions	Insurance	Decision trees	YAWL BPM system	UI informing next work item to perform based on risks	A recommendation system for predicting risks across multiple business process instances [52]

5 Predictive Monitoring Technique Framework Design

In this section, we define the predictive monitoring framework. The framework has been designed to categorically classify the techniques and approaches identified during the systematic literature review phase. The main aim of this framework is to address the needs of businesses with a guide, so as to make them aware regarding what can be predicted in a business process, so that, based on their needs, business people could search for the best possible match of existing methods and techniques suitable to their requirement. The match is based on what is predicted, the algorithms used, how the technique has been validated, how the outputs are presented and whether the described technique has a working implementation or not. Other than businesses even researchers could use the framework. Researchers could get an overview of the overall research progress that has been made so far in this research category of predictive monitoring in business processes and generate new ideas or dive into an unexplored segment for further research.

Prediction Category	*Sub Prediction Category*	*What is Predicted*	*Example*	*Validation*	*Domain*	*Family of Algorithm*	*Implementation*	*Input*	*Output*	*Reference*
Time based										
Outcome										
Process Path										
Risk										

Figure 22 Predictive Monitoring Framework

The field "What to predict" serves as the basis for classifying each prediction approach. Figure 7 shows the outline of the predictive monitoring framework.

***What to Predict**:* This predictive monitoring framework field categorically distinguishes each predictive monitoring approach into 4 distinct categories. Each of which has been explained in the previous section. Each prediction category clearly defines what to predict. The categories of prediction are *Time based prediction, Outcome based prediction, Process path prediction and Risk prediction*. We concluded to generalize our prediction categories into these 4 categories after an intense systematic literature review and several rounds of discussion.

This particular column in the framework will help businesses focus precisely on the kind of prediction they are looking for filtering out the rest. This is helpful for business people using the framework providing them with a concise and precise view.

Example: The example field in the PMF showcases examples which are extracted from the corresponding literature. The example field helps understand the problem, the author of the paper is trying to solve. It gives framework users a better understanding about the paper and

the problem it is trying to resolve. In some scenarios, the example could also show the possible solution for the problem.

This particular column in the PMF (Predictive Monitoring Framework) helps business people understand the issues the particular technique is trying to solve. This column could be used by businesses to understand if the issue or problem explained is similar to there's or is something they could face. Hence it is an important column which could be matched to find the right suitable technique.

Validation: It is important to know whether a predictive monitoring technique or approach has been tested appropriately. Validation could be done on real-life logs, simulated logs or case studies. In this field, it is clearly mentioned whether testing has been done and if yes, further details are mentioned here.

This particular column in the PMF, helps business people understand whether the proposed technique is validated and if so how. It is important for businesses to know whether the technique they are planning to adopt has been tested and validated,

Domain: Domain specifies the industry, business or field the technique or approach specifically belongs to or shows expertise in. If no approach or field is specified, in particular, the technique is described as for all businesses (Multi domain). In some scenarios where no specific domain is mentioned, if the approach is tested for example on a healthcare log, we consider the domain of the approach to be specialized in healthcare as well.

This particular column in the PMF is a very important feature. Business people from a specific domain can straight away explore the prediction techniques that could be applied in their respective domain. It could also be used as a filter to look into a particular domain.

Family of Algorithm: The algorithm used in the described predictive mining technique is mentioned here. If some customized algorithm or techniques are being used then it is made sure that the family or root to which the algorithm belongs to is mentioned. This field gives a clear picture about the data mining or machine learning algorithms that are being used.

This particular column in the PMF helps businesses understand the underlying algorithms used in the defined prediction model. It is important for businesses to know this information and will be helpful for them in understanding, how to integrate the model with their current business setup.

Implementation: In this column, it is specified whether the technique or approach being described has any kind of implementation. The implementation is considered as mainly standalone or plugins. Plugins could be on top on ProM or other business suites. It is clearly mentioned in this column.

This particular column in the PMF helps businesses with information about any existing implementation of the stated prediction technique.

Input: In this section, the kind of input which is required by the predictive monitoring technique or approach is mentioned. Inputs are mostly event logs.

This particular column in the PMF helps in understanding the kind of input required for a prediction model. It is important so as to know if businesses could provide the required input.

Output: In this column, it is clearly mentioned what kind of output is presented as output to the end user. Output could be of different types like in [6] average remaining time is considered as the output type, in [53] completion time is considered as the output type. Readers get a clear overview of the kind of output that is expected from the literature.

This particular column in the PMF is a very important feature. It is important for business people to know how the results of a particular prediction model will be presented. It is important for business people to make sure they understand the presented output and it is what they expect as an end result of the prediction model.

Reference: In this column, the literature to which the technique or approach belongs to is mentioned.

PMF Usage Example

Let us assume an insurance company ABC corporation is looking for a claims handling solution which could predict the outcome of a business process. We will explain how the company could use our framework to find a matching solution.

As the company is looking to predict an outcome, it has two options, it could either look for domain Insurance and filter all the available solutions or it could look for the prediction category "Outcome prediction". Either way the user will be able to gather valuable information.

If the user filters the list using the domain as "Insurance" and kind of prediction as "Outcome prediction" they will be matched with a technique, if it exists, that is proposed or implemented.

However, sometimes, it is wise to use Multi domain also while filtering the domain column, as these could also be useful.

There are multiple ways one could traverse the framework and find suitable results. It is upto the user how he or she wishes to use the Predictive Monitoring Framework.

6 Threats to Validation

Construct Validity

Our research questions (RQ's) may not be able to cover all the literature listed in the paper.

Internal Validity

In order to identify all relevant literature which, fall under this Thesis topic without any bias, 5 databases were searched for. The databases were searched with Boolean expressions consisting of primary most important keywords. The literature filtering was carried out extensively in three phases which have been explained in section 2.2 of the Thesis. The filtering was carried out under the supervision of the supervisors.

External Validity

The papers included following the exclusion criteria which has been defined in the paper in section 2.1, due to which papers had to be rejected from the review process due to these criteria's.

Conclusion Validity

We defined a framework which could be used by researchers and business people to apply predictive monitoring techniques in their respective fields. We have shown an example in section 0 how businesses could use our framework to find the right technique.

7 Conclusion

This thesis presents a Predictive Monitoring Framework for the systematic comparison of predictive techniques in process mining. The framework classifies prediction techniques into mainly four categories (i) Time-based prediction (ii) Outcome prediction (iii) Process path prediction (iv) Risk prediction. Under each category, the framework has identified 9 features which could be used by businesses to filter or match the prediction technique they require. The features are (i)What is predicts (ii)An example to explain its usage (iii) Domain (iv)Validation (v) Implementation (vi) Input and (vii) Output. The framework is expected to be used by businesses and researchers. Business people who want to apply these techniques in their respective domain and Researchers who want to study and develop new prediction models.

In this Thesis, we have managed to find answers to the Research Questions (RQ) we raised at the beginning of this thesis.

Research Question 1 (RQ 1): What aspects of a business process can predictive monitoring predict?

This research question is answered by the Predictive Monitoring Framework (2.4), classifying the predictive monitoring techniques into 4 sub categories and further into subcategories.

Research Question 2 (RQ 2): How is predictive monitoring currently applied in the industry??

The second research question has been answered in the literature review section by citing examples of prediction techniques applied in different domains

In future, this research could be extended further by benchmarking the techniques which have been classified so far. The implementation of these techniques could be tested against a common log and performance of each technique can be reviewed.

8 References

[1] W. V. D. Alast, “Process Mining Manifesto”.

[2] B. Kitchenham and S. Charters, “Guidelines for performing systematic literature reviews in software engineering,” *Technical report, EBSE,* 2007.

[3] M. Dumas, M. L. Rosa, J. Mendling and H. A. Reijers, Fundamentals of Business Process Managemen, 2013.

[4] M. Ceci, P. Lanotte, F. Fumarola, D. Cavallo and D. Malerba, “Completion Time and Next Activity Prediction of Processes Using Sequential Pattern Mining”.

[5] W. v. d. Aalst, H. Reijers, A. Weijters, B. v. D. A. A. d. Medeiros, M. Song and H. Verbeek, “Business Process Mining: An Industrial Application”.

[6] A. Bolt and M. Sepúlveda, “Process remaining time prediction using query catalogs,” 2014.

[7] A. Solomon and M. Litoiu, “Business process performance prediction on a tracked simulation model,” *Proceedings - International Conference on Software Engineering,* 2011.

[8] F. Masseglia, M. Teisseire and P. Poncelet, “Sequential Pattern Mining”.

[9] M. Song, C. G. ther and W.M.P. van der Aalst, Tree CLustering in Process Mining.

[10] M. Le, D. Nauck, B. Gabrys and T. Martin, “Sequential Clustering for Event Sequences and Its Impact on Next Process Step Prediction,” *15th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, IPMU 2014,* 2014.

[11] M. Dorigo, M. Birattari and T. Stutzle, “Ant Colony Optimization”.

[12] A. Rogge-Solti and M. Weske, “Prediction of remaining service execution time using stochastic petri nets with arbitrary firing delays”.

[13] W. M. G. A. M. A. Senderovich A., “Queue mining for delay prediction in multi-class service processes,” 2015.

[14] M. Polato, A. Sperduti, A. Burattin and M. De Leoni, “Data-aware remaining time prediction of business process instances,” *Proceedings of the International Joint Conference on Neural Networks,* 2014.

[15] S. Pravilovic, A. Appice and D. Malerba, “Process Mining to Forecast the Future of Running Cases,” *NEW FRONTIERS IN MINING COMPLEX PATTERNS, NFMCP 2013,* 2014.

[16] S. Pey, S. Nepal and S. Chen, “A Test-bed for the Evaluation of Business Process Prediction Techniques,” *ColiaborateCom 2011 - Proceedings of the 7th International Conference on Collaborative Computing: Networking, Applications and Worksharing,* 2011.

[17] S. Huber, M. Fietta and S. Hof, “Next Step Recommendation and Prediction Based on Process Mining in Adaptive Case Management,” *Proceedings of the 7th International Conference on Subject-Oriented Business Process Management,* 2015.

[18] Rogge-Solti and M. reas Weske, “Prediction of Remaining Service Execution Time Using Stochastic PetriNets with Arbitrary Firing Delays,” *SERVICE-ORIENTED COMPUTING, ICSOC 2013,* 2013.

[19] A. C. Bevacqua, M. Folino, F. Guarascio and L. Massimo Pontieri, “A Data-Driven Prediction Framework for Analyzing and Monitoring BusinessProcess Performances,” *ENTERPRISE INFORMATION SYSTEMS, ICEIS 2013,* 2014.

[20] A. Bevacqua, M. Carnuccio, F. Folino, M. Guarascio and L. Pontieri, “Adaptive trace abstraction approach for predicting business process performances,” *21st Italian Symposium on Advanced Database Systems, SEBD 2013,* 2013.

[21] F. Folino, M. Guarascio and L. Pontieri, “Context-aware predictions on business processes: An ensemble-based solution,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2013.

[22] F. Folino, M. Guarascio and L. Pontieri, “Discovering Context-Aware Models for Predicting Business Process Performances,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2012.

[23] A. Solti, L. Vana and J. Mendling, “Time series petri net models enrichment and prediction,” *Lecture Notes in Business Information Processing,* 2017.

[24] Y. Z. H. L. C. J. R. J. Liu, “Workflow simulation for operational decision support using event graph through process mining,” *DECISION SUPPORT SYSTEMS,* 2012.

[25] B. F. A. Hompes, J. C. A. M. Buijs and W. M. P. van der Aalst, “A Generic Framework for Context-Aware Process Performance Analysis,” *ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS: OTM 2016 CONFERENCES,* 2016.

[26] W. Van Der Aalst, K. Van Hee, J. Van Der Werf and M. Verdonk, “Auditing 2.0: Using Process Mining to Support Tomorrow’s Auditor,” 2010.

[27] A. Senderovich, M. Weidlich and A. M. A. Gal, “Queue Mining - Predicting Delays in Service Processes,” *ADVANCED INFORMATION SYSTEMS ENGINEERING (CAISE 2014),* 2014.

[28] G. Lakshmanan, S. Duan, P. Keyser, T, F. Curbera and R. Khalaf, “Predictive Analytics for Semi-structured Case Oriented Business Processes,” *Lecture Notes in Business Information Processing,* 2011.

[29] I. Teinemaa, M. Dumas, F. Maria Maggi and C. DiFrancescomarino, “Predictive Business Process Monitoring with Structured and UnstructuredData,” *BUSINESS PROCESS MANAGEMENT, BPM 2016,* 2016.

[30] E. Cesario, F. Folino, M. Guarascio and L. Pontieri, “A Cloud-Based Prediction Framework for Analyzing Business ProcessPerformances,” *AVAILABILITY, RELIABILITY, AND SECURITY IN INFORMATION SYSTEMS, CD-ARES2016, PAML 2016,* 2016.

[31] D. Breuker, M. Matzner, P. Delfmann and J. Becker, “Comprehensible Predictive Models for Business Processes,” *MIS QUARTERLY,* 2016.

[32] F. Folino, M. Guarascio and L. Pontieri, “A Prediction Framework for Proactively Monitoring Aggregate Process-Performance Indicators,” *Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOCW,* 2015.

[33] D. Cleary, “Predictive Analytics in the Public Sector: Using Data Mining to Assist Better Target Selection for Audit,” *Proceedings of the European Conference on e-Government, ECEG,* 2011.

[34] A. Bautista, L. Wangikar and S. Akbar, “Process Mining-Driven Optimization of a Consumer Loan Approvals Process,” *Lecture Notes in Business Information Processing,* 2013.

[35] B. Schwegmann, M. Matzner and C. Janiesch, “preCEP: Facilitating Predictive Event-Driven Process,” 2013.

[36] M. T. a. L. W. Z. Wynn, A. H. M. ter Hofstede and W. Nauta, “A Framework for Cost-Aware Process Management: Cost Reporting and CostPrediction,” *JOURNAL OF UNIVERSAL COMPUTER SCIENCE,* 2014.

[37] H. Horita, H. Hirayama, T. Hayase, Y. Tahara and A. Ohsuga, “Process Mining Approach Based on Partial Structures of Event Logs andDecision Tree Learning,” *PROCEEDINGS 2016 5TH IIAI INTERNATIONAL CONGRESS ON ADVANCED APPLIEDINFORMATICS IIAI-AAI 2016,* 2016.

[38] B. Kang, D. Kim, S. Kang and -H, “Periodic performance prediction for real-time business process monitoring,” *Industrial Management & Data Systems,* 2012.

[39] F. M. Maggi, C. Di Francescomarino, M. Dumas and C. Ghidini, “Predictive Monitoring of Business Processes,” *ADVANCED INFORMATION SYSTEMS ENGINEERING (CAISE 2014),* 2014.

[40] C. Cabanillas, C. Di Ciccio and J. B. A. Mendling, “Predictive Task Monitoring for Business Processes,” *BUSINESS PROCESS MANAGEMENT, BPM 2014,* 2014.

[41] B. K. Kang and S.-H. Dongsoo Kang, “Real-time business process monitoring method for prediction of abnormal termination using KNNI-based LOF prediction,” *EXPERT SYSTEMS WITH APPLICATIONS,* 2012.

[42] C. K. H. Lee, G. T. S. Ho, K. L. Choy and G. K. H. Pang, “A RFID-based recursive process mining system for quality assurance inthe garment industry,” *INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH,* 2014.

[43] I. Verenich, M. Dumas, L. Rosa, Marcello, F. M. Maggi and C. Di Francescomarino, “Minimizing Overprocessing Waste in Business Processes via PredictiveActivity Ordering,” *ADVANCED INFORMATION SYSTEMS ENGINEERING (CAISE 2016),* 2016.

[44] M. Unuvar, G. T. Lakshmanan and Y. N. Doganata, “Leveraging path information to generate predictions for parallel business processes,” *Knowledge and Information Systems,* 2016.

[45] D. Breuker, P. Delfmann, M. Matzner and J. Becker, “Designing and Evaluating an Interpretable Predictive Modeling Technique for Business Processes,” *BUSINESS PROCESS MANAGEMENT WORKSHOPS(BPM 2014),* 2015.

[46] J. Becker, D. Breuker, P. Delfmann and M. Matzner, “Designing and Implementing a Framework for Event-based Predictive Modelling of Business Processes,” *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI),* 2014.

[47] K. Krinkin and E. Kalishenko, “Traffic Prediction in Wireless Mesh Networks Using Process Mining Algorithms,” *PROCEEDINGS OF THE 11TH CONFERENCE OF OPEN INNOVATIONS ASSOCIATION FRUCT,* 2012.

[48] M. Le, B. Gabrys and D. Nauck, “A hybrid model for business process event prediction,” *Res. and Dev. in Intelligent Syst. XXIX: Incorporating Applications and Innovations in Intel. Sys. XX - AI 2012, 32nd SGAI Int. Conf. on Innovative Techniques and Applications of Artificial Intel.,* 2012.

[49] B. Kang, D. Kim and S.-H. Kang, “Traffic Prediction in Wireless Mesh Networks Using Process Mining Algorithms,” *EXPERT SYSTEMS WITH APPLICATIONS,* 2012.

[50] A. Pika, W. M. P. Van Der Aalst, M. Wynn, C. Fidge and A. H. M. Ter Hofstede, “Evaluating and predicting overall process risk using event logs,” *Information Sciences,* 2016.

[51] R. F. Conforti, J. Sven Merscheid and M. Roeglinger, “PRISM – A Predictive Risk Monitoring Approach for Business Processes,” *BUSINESS PROCESS MANAGEMENT, BPM 2016,* 2016.

[52] R. Conforti, M. De Leoni, M. La Rosa, W. Van Der Aalst and A. Ter Hofstede, “A recommendation system for predicting risks across multiple business process instances,” *Decision Support Systems,* 2015.

[53] S. M. S. M. Van Der Aalst W.M.P., “Time prediction based on process mining,” 2011.

[54] S. Jalali and I. Bhatnagar, “Leveraging internet of things technologies and equipment data for an integrated approach to service planning and execution,” *Proceedings - 2015 IEEE Region 10 Symposium, TENSYMP 2015,* 2015.

I. Appendix

List of Literature Surveyed

1. Completion Time and Next Activity Prediction of Processes Using Sequential Pattern Mining [4]
2. Data-aware remaining time prediction of business process instances [14]
3. Next Step Recommendation and Prediction based on Process Mining in Adaptive Case Management [17]
4. Prediction of Remaining Service Execution Time Using Stochastic Petri Nets with Arbitrary Firing Delays [18]
5. Process Mining to Forecast the Future of Running Cases [15]
6. Process Remaining Time Prediction Using Query Catalogs [6]
7. A Data-Driven Prediction Framework for Analyzing and Monitoring Business Process Performances [19]
8. A Generic Framework for Context-Aware Process Performance Analysis [25]
9. A Test-bed for the Evaluation of Business Process Prediction Techniques [16]
10. Adaptive Trace Abstraction Approach for Predicting Business Process Performances [20]
11. Auditing 2.0: Using Process Mining to Support Tomorrow's Auditor [26]
12. Context-Aware Predictions on Business Processes: An Ensemble-Based Solution [21]
13. Queue Mining – Predicting Delays in Service Processes [27]
14. Time Series Petri Net Models [23]
15. Workflow simulation for operational decision support using event graph through process mining [24]
16. Discovering Context-Aware Models for Predicting Business Process Performances [22]
17. A Framework for Cost-Aware Process Management Cost Reporting and CostPrediction [36]
18. Periodic performance prediction for real-time business process monitoring [38]
19. Predictive Analytics for Semi-structured Case Oriented Business Processes [28]
20. Predictive Business Process Monitoring with Structured and Unstructured Data [29]
21. Predictive Monitoring of Business Processes [39]
22. Predictive Task Monitoring of Business Processes [40]
23. Process Mining Approach Based on Partial Structures of Event Logs and Decision Tree Learning [37]

24. Real-time business process monitoring method for prediction of abnormal termination using KNNI-based LOF prediction [41]

25. A Cloud-Based Prediction Framework for Analyzing Business Process [30]

26. A prediction framework for proactively monitoring aggregate process-performance indicators [32]

27. A RFID-based recursive process mining system for quality assurance in the garment industry [42]

28. Comprehensible Predictive Models for Business Processes [31]

29. Minimizing Overprocessing Waste in Business Processes via Predictive Activity Ordering [43]

30. Predictive Analytics in the Public Sector: Using Data Using Data Mining to Assist Better Target Selection for Audit [33]

31. Process Mining-Driven Optimization of a Consumer Loan Approvals Process [34]

32. Leveraging path information to generate predictions for parallel business processes [54]

33. Traffic Prediction in Wireless Mesh Networks Using Process Mining Algorithms [49]

34. Designing and Evaluating an Interpretable Predictive Modeling Technique for Business Processes [45]

35. Designing and Implementing a Framework for Event-based Predictive Modelling of Business Processes [46]

36. Sequential Clustering for Event Sequences and Its Impact on Next Process Step Prediction [10]

37. A hybrid model for business process event and outcome prediction [48]

38. Evaluating and predicting overall process risk using event logs [50]

39. A recommendation system for predicting risks across multiple business process instances [52]

40. PRISM – A Predictive Risk Monitoring Approach for Business Processes [51]

41. preCEP: Facilitating Predictive Event-Driven Process Analytics [35]

II. Predictive Monitoring Framework Implementation

As the framework is too large a file to be included the Thesis, we have provided a link below to have a look at the framework with all the relevant predictive monitoring data filled in.

Predictive Monitoring Framework - Google Doc

III. License

Non-exclusive licence to reproduce Thesis and make Thesis public

I, **Cigin Koshy**,

(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:
   1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and
   1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

of my thesis

A Literature Review on Predictive Monitoring of Business Processes

(title of Thesis)

supervised by Fabrizio Maria Maggi, Fredrik Milani, Chiara Di Francescomarino

(supervisor's name)

2. I am aware of the fact that the author retains these rights.
3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **15.08.2017**