INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

SEQUENCE RANDOM FIELDS FOR SUPERVISED EVENT ABSTRACTION EVALUATION

Surendra Kumar Reddy Koduru, Research Scholar, Department of Computer Science and Applications, Monad University, Hapur, U.P.
Dr. Anil Badarla, Professor, Supervisor, Department of Computer Science and Applications, Monad University, Hapur, U.P.

Abstract: Today's much information makes healthy decision-making difficult. Data mining, machine learning, and computational statistics are prominent topics of research that allow empowered individuals to make beneficial judgements to optimise any working domain. Patients rise according to population growth and lifestyle changes, hence healthcare data processing is in high demand. Early illness identification and prognosis prediction are needed to improve human therapy. Health prediction models benefit from data mining.
In recent years, cancer has spread worldwide, hence this study has concentrated on cancer registry data. The thesis's goal is to create a useful cancer prognostic classifier model. Most present systems construct diagnostic prediction models using screening or survey data, which is widely accessible and simple to acquire owing to the insensitive nature of such research. Prognosis prediction involves sensitive patient data. Hospitals and government community registers are the key data sources. Researchers in India cannot access well-maintained computerised hospital records including histopathological data. This study utilised US open access cancer data for all experiments.

This study approach steadily increases prediction accuracy by choosing relevant data mining methods in each step. Prognosis refers to a cancer patient's future illness severity and survival rate. This study seeks to uncover key response characteristics that aid prognosis prediction and enhance prediction models.

Keywords—Process Mining, Rule Augmented, Event Abstraction, Conditional Random Fields

ISSN: 2057-5688

1 INTRODUCTION:

Data mining involves finding hidden patterns in massive datasets. It is a growing discipline that employs statistical, visualisation, machine learning, and other data manipulation and information extraction methods to uncover data linkages and patterns [10, 40]. In this internet and mobile age, analytics is the buzzword in IT and non-IT sectors as data volumes expand exponentially. Data mining and machine learning techniques can solve large data problems, with most applications in healthcare [41]. Patients rise according to population growth and lifestyle changes, hence healthcare data processing is in high demand.

Process mining involves discovery, compliance, and improvement. In [6], automatic process discovery extracts process models from an event log, conformance checking monitors deviations by comparing a model to the event log, and enhancement extends or improves a process model using event log data.

An accurate model of a process enhances the ability to design and implement process requirements in the HIS that support the process, configure any extra needs not included in the system, and assist process analysis. In [6], the author suggests using organisational mining, automated simulation model generation, model extensions, model repair, forecasting process behaviour, and history-based suggestions to expand the study.

Complex healthcare procedures vary throughout time [2]. Patient circumstances and resource (physician, nurse, and other healthcare personnel) activity sequences produce these differences.

CHALLENGES IN MEDICAL DATA MINING

In the medical field, everyone wants great accuracy, and data mining has helped researchers get close. Scientists have few strategies to regulate carcinogenesis. Cancer registries in India and elsewhere gather and categorise all cancer cases to create a framework for analysing and regulating the community's effect and severity of cancer. Since the Indian

ISSN: 2057-5688

cancer registries data collection is private, this study utilised SEER. These registries' standard research solutions don't expose intrinsic data properties. Data mining learning algorithms let us find new relationships and patterns that recognise symptoms and illnesses, assess prognosis, and avert death.

This study identifies and characterises cancer data case studies where process mining has been applied in cancer healthcare, provides an overview of the state of the art, guides researchers on how to apply process mining techniques, methodologies, algorithms, and tools, and highlights some of the benefits of using this discipline.

II RELATED WORK

2.1. Process Discovery Algorithms

Few healthcare data and process mining literature reviews exist. We found various medical data mining reviews [17–23]. Process mining in health care has a brief literature analysis on clinical route case studies [24]. No research gathers, characterises, and contextualises all healthcare process mining case studies.

Process mining extracts useful information from process data. It bridges process science (business process management and operations research) with data science (data mining and predictive analytics) to examine processes using data [8]. Process mining may be used in any industry with processes and data. This study discusses process mining in healthcare.

Healthcare process mining broadly. Process mining involves modelling processes, such as their stages and pathways [22]. Flowcharts or Business Process Modelling Notation (BPMN) may show a process's activity order [23].

HISs help several healthcare procedures nowadays. EHR systems from Epic4, Cerner5, MEDITECH6, Allscripts7, athenahealth8, IBM9, McKesson10, and Siemens11 support healthcare processes. These systems track healthcare procedures [8]. Event logs may be made from process execution data. Process mining methods rely on event logs with process execution data.

III. DATASET & PREPROCESSING METHODS

Process mining procedure and data set are covered here. Health Technology

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

Assessment commonly uses randomised clinical trials. The "gold standard" for Health Technology Assessment, they give the most scientific evidence. Their fundamental drawback is that they are administered to certain groups in targeted situations, which frequently differ from each country's clinical or social reality. Therefore, they may fail to offer sufficient data on the efficacy of provided healthcare technology, especially owing to the profile of patients whose comorbidities and demographics often differ from those utilised in randomised clinical trials.

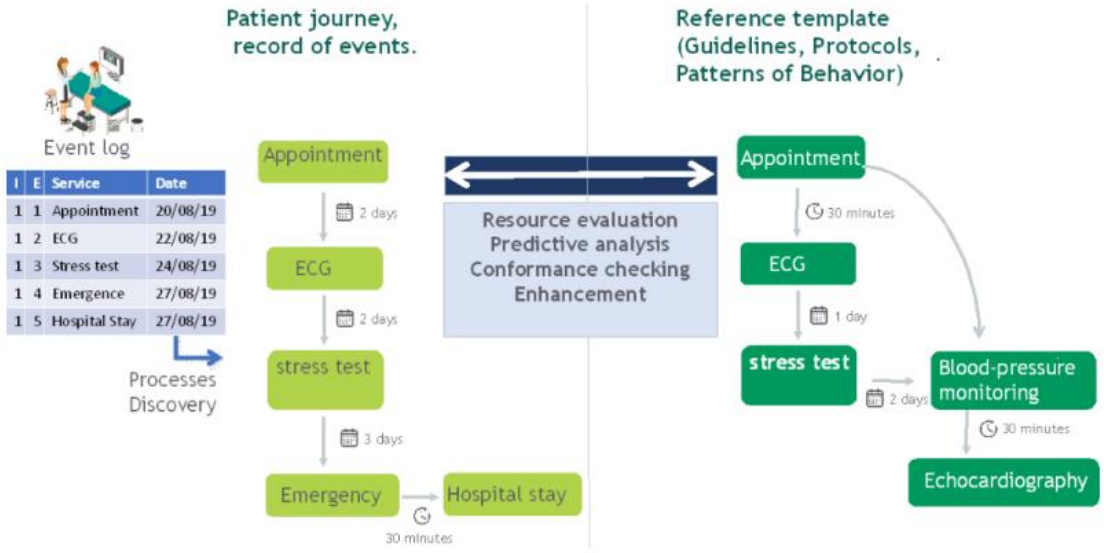

Fig. 1 Example of using Process Mining in Healthcare

3.1 DATA SET DESCRIPTION

Indian sources are collecting cancer data for the study. The American Cancer Society's Cancer Prevention and Early Detection (CPED) statistics illuminated cancer research's present state and needs.

All categorization challenges have been studied using the NCI SEER data set. Before categorization, this data set was extensively pre-processed. SEER proof includes permission letter and postal copy.

The UCI Machine Learning Repository provided the other two dimensionality reduction data sets [95]. This section provides data set details. These classification data were utilised directly to evaluate dimensionality reduction difficulties utilising NN-based PCA algorithms.

ISSN: 2057-5688

SEER Data Dictionary

This thesis utilises SEER_1973_2007_TEXTDATA directory data. The 1973-2021 directory was retrieved because this study began in 2009. This directory includes ASCII population and incident data.

3.2 EXPERIMENTAL SETUP

For preprocessing and basic classifier development, this study employed MATLAB, although for a few advanced classification techniques, Rapidminer operators were used. A MATLAB GUI programme does significant preprocessing and feature extraction. To simplify tool-to-tool transfers, findings were stored in Excel and MATLAB matrix data formats.

Implementation employed 500–30000 data samples with varied scaling and seeding. All models' efficiency was evaluated inside this geographic boundary and interpreted in the relevant chapters. Table 1 shows that the basic and upgraded classifiers for SEER data set were evaluated using prominent labels in conjunction with all characteristics.

Table 1: Implementation features

Sample size	Classifiers	Data Set	Class Labels	Sampling
500		Breast	Survival	
1000		Cancer	Age at	Random
2000	Naïve Bayes	Colorectal	Diagnosis	
5000	Decision Tree	Cancer	Multiple	Stratified
8000 -10000	KNN	Respiratory	Primaries	
13000 - 16000		Cancer	Stage	Balanced
200000		Mixed Type	Grade	
300000				

Operational Flow of Classification Models

The sequence of operation of each model discussed in this thesis has followed the steps given in Figure 2 Models vary in step 4 based on the type of classifiers and the respective architectures.

3.3 DATA PRE-PROCESSING

ISSN: 2057-5688

This section mainly discusses the extensive pre-processing phases of SEER data set.

The incidence patient profile is a single record with 254 characters representing a total of 118 attributes. Remaining region and race based 78 attributes were out of scope with respect to this thesis. The 118 attributes are both nominal and numerical in type. All 118 variables have been utilized as features for the initial phase. The attributes have been reduced from 118 to 37 by performing a thorough preprocessing of data.

IV EXPERIMENT AND RESULTS

The main objective of this Section is to identify the prominent class labels that support the prognosis prediction of cancer in patients who have been identified with positive diagnosis factors.

Impurity Function: A function f is said to be an impurity function if it is defined on the n instances of attributes (a1, a2, …, an) for ai ≥ 0, a = 1, ..., n, sum(aj)=1 such that f is maximum only at (1/n, 1/n,…, 1/n); f is minimum at (1, 0, 0, ..., 0), (0, 1, 0, ..., 0), ..., (0, 0, ..., 0, 1) and f is symmetric to (a1, a2, …, an).

Posterior Probability: It is directly proportional to the product of likelihood function and the prior probability. The posterior probability is calculated based on the Bayes theorem as given in equation

$$P\left(C_j/_X\right) = P(C_j)P\left(X/_{C_j}\right)/P(X)$$

Distance Metric: To find the distance between the parameters of any two samples many distance metrics has been used in the literature. For example the equation 2 defines the Euclidean distance measure for n attributes. It finds the difference from one sample with that of the other. Another form of metric named, similarity measure finds the closeness of the two samples and has been used for the categorical variables as defined in equation 3

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

$$sim(a,b) = \sum_{i=0}^{n} w_i S(a_i, b_i)$$

Where w_i is the weight factor and S is set as 1 if a_i=b_i and 0 otherwise

Algorithm:

Algorithm : Decision Tree

Classification-DT(D,A,t)

ISSN: 2057-5688

Input : Training data set = {(x1,c1),(x2,c2),.....(x3,c3)} and attribute list A= (a1, a2,..., ak)
Output : Decision Tree
Step 1 : Let t={}
Step 2 : Calculate A' the attribute with maximum info gain with D
Step 3 : Add A' to t
Step 4 : For all ai
in A repeat step 5-8
Step 5 : Branch at A' in t for all values of A'
Step 6 : Create subsets of Ds with values of a in A'
Step 7 : If A=D , then create leaf node with label c in C found in A'
Step 8 : Else call DT(Ds, A\ {A'}, t).

The variable t in step1 is an attribute list that keeps adding the attributes in all iteration. A' is the attribute list with maximum info gain and Ds is the split set passed to the next iteration of the algorithm.

The model uses 60:40 ratios for test and training samples. The top labels with good fitness that are selected from this model will be considered as the prominent labels. The classification phase of this model has been implemented using the Rapidminer for all combination of cancer data sets from the processed data.

Model proposed for selecting the prominent label for cancer prognosis prediction. All ten labels have been selected initially as input to the model and kept a threshold of average accuracy as 50% for prominent labels. As a result only the top five labels have been considered as prominent labels. The label grade has been included in this section but the same has been rejected for further research as the performance of this label has been fluctuating based on the sample size and the choice of the classifier. Considering the inconsistent nature, label grade has been discarded later.

Table 2: Top five prominent class labels

Legend	Class Label	No of Classes
1	Survival	3
2	Age at Diagnosis	5
3	Multiple Primaries	5

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

4	Stage	5
5	Grade3	5

The selection of the above three results from many of the experimental outcome has been made as a representation from each collection. That is the first result represent the best classifier for all combination, the second result for frequent individual data set used in all classifiers and the last one represents the top prominent label in all classifiers.

Three combinations of results have been presented using different forms of visualization. Line graph depicts the gradation of the results, the bar scale depicts the intergroup comparison of classifiers and the tabular view depicts the actual data points.

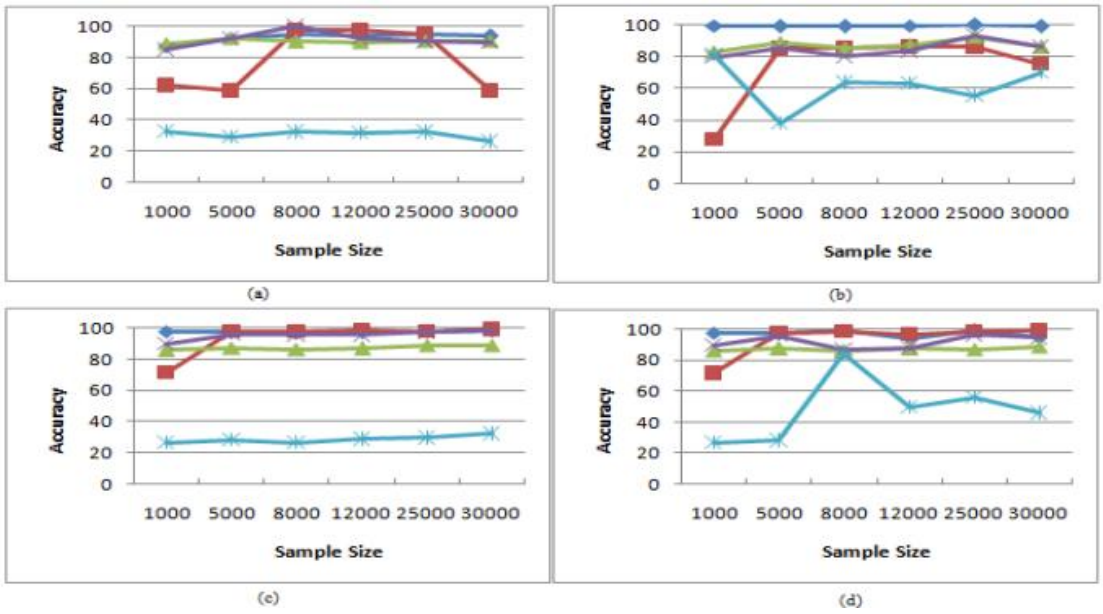

Fig 2 : Comparisons of all prominent labels in the following order of data set (a) Breast Cancer (b) Colorectal Cancer (c) Respiratory Cancer (d) Mixed Cancer (e) Legend numbers as given in Table 4.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

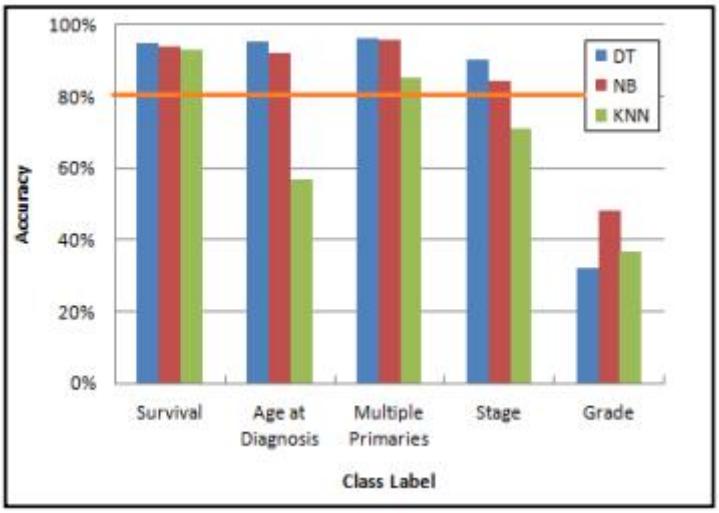

Fig 3 : Comparative performance of prominent labels using all classifiers for breast cancer data set with sample size 25000

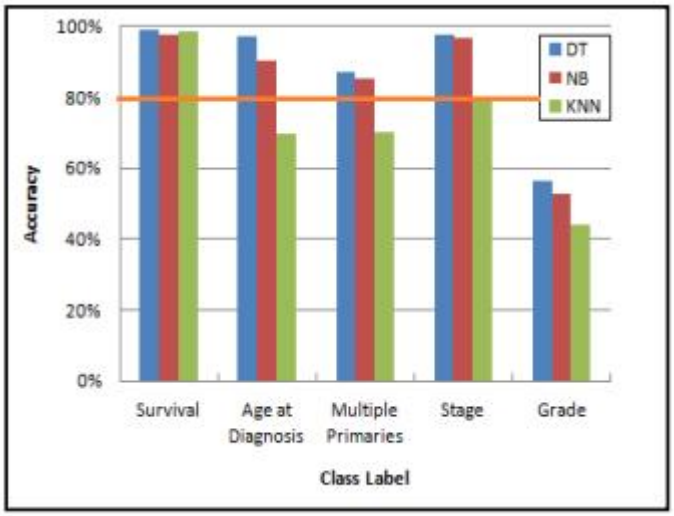

Fig 4 : Comparative performance of prominent labels using all classifiers for mixed cancer data set with sample size 25000

The performance of the decision tree and Naïve Bayes classifiers are relatively better than the k-NN classifier in Figure 8 and Figure 9. The horizontal line at 80% accuracy marks a threshold limit that decides on the class labels as prominent labels for further study. From Figure 8 and Figure 9 it is evident that the label grade is not an appropriate member to predict the prognosis both in the

ISSN: 2057-5688

individual and in the mixed cancer types. Thus the grade has been discarded from top five and only the top four labels have been decided to be prominent.

Table 3: Survival prediction accuracy of all cancer types

Classifier	Breast	Colorectal	Respiratory	Mixed	Sample Size
DT	95.14	99.34	97.21	**99.78**	10000
NB	93.23	96.43	95.12	**99.34**	
Knn	92.63	99.08	95.21	**99.65**	
DT	93.45	**99.89**	97.43	94.23	20000
NB	93.45	96.98	97.67	95.34	
Knn	93	**99.76**	**99.99**	92.34	

The analysis result given in Figure 7 reveals that the label survival scores the top rank of all four labels and so the performance insights of this label for all data sets and all three classifiers have been presented in Table 4. The above insight has been presented due to the high variations of results found in all data sets in samples between the range 10000 and 20000. The bold text of Table 4 highlights the top cancer type in each classifier and the underlined text highlights the best classifier for each cancer type.

V CONCLUSION

The objective of identifying the prominent class labels to predict the prognosis of disease in patients has been achieved using the well known classifier in data mining. The four prominent labels filtered from ten labels marks the success of initial phase of research that provides possibilities to explore further in the SEER data set, as only survival has been used by many researchers in the literature. Based on the average accuracy performance of the overall experiment, survival, multiple primaries, stage and age at diagnosis in the given order have been identified as the prominent response variable, where as grade performed very low and the same has been discarded. The efficiency of the performance of all four prominent labels has been presented based on the simple random selection of samples ranging from 1000 to 30000. Thus the hypothesis “More than one

prominent label exists in prognosis prediction problems" has been proved. This constructive outcome of this research leads the practitioners and software tools to choose response variable from the pool of prominent labels for prognosis prediction. The limitation of the model is to standardize the entire process for generic factors and for many diseases.

VIII REFERENCES

[1] W. M. P. van der Aalst, Process mining: Discovery, conformance and enhancement of business processes. Springer Science & Business Media, 2011.

[2] W. M. P. van der Aalst, A. J. M. M. Weijters, and L. Maruster, "Workflow mining: Discovering process models from event logs," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 9, pp. 1128–1142, 2004.

[3] C. W. Gunther and W. M. P. van der Aalst, "Fuzzy mining–adaptive ¨ process simplification based on multi-perspective metrics," in Business Process Managem-ent. Springer, 2007, pp. 328–343.

[4] J. M. E. M. Van der Werf, B. F. van Dongen, C. A. J. Hurkens, and A. Serebrenik, "Process discovery using integer linear programming," in Applications and Theory of Petri Nets. Springer, 2008, pp. 368–387.

[5] A. J. M. M. Weijters and J. T. S. Ribeiro, "Flexible heuristics miner (fhm)," in Proceedings of the 2011 IEEE Symposium on Computational Intelligence and Data Mining. IEEE, 2011, pp. 310–317.

[6] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, "Discovering block-structured process models from event logs - a constructive approach," in Application and Theory of Petri Nets and Concurrency, ser. LNCS. Springer, 2013, pp. 311–329.

[7] R. P. J. C. Bose and W. M. P. van der Aalst, "Abstractions in process mining: A taxonomy of patterns," in Business Process Management, ser. LNCS. Springer, 2009, pp. 159–175.

[8] C. W. Gunther, A. Rozinat, and W. M. P. van der Aalst, "Activity ¨ mining by global trace segmentation," in Business Process Management Workshops, ser. LNBIP. Springer, 2010, pp. 128–139.

[9] B. F. van Dongen and A. Adriansyah, "Process mining: Fuzzy clustering and performance visualization," in Business Process Management Workshops, ser. LNBIP. Springer, 2010, pp. 158–169.

ISSN: 2057-5688

[10] C. W. Gunther and H. M. W. Verbeek, "XES-standard definition," ¨ BPMcenter.org, 2014.

[11] T. van Kasteren, A. Noulas, G. Englebienne, and B. Krose, "Accurate ¨ activity recognition in a home setting," in Proceedings of the 10th International Conference on Ubiquitous Computing. ACM, 2008, pp. 1–9.

[12] E. M. Tapia, S. S. Intille, and K. Larson, "Activity recognition in the home using simple and ubiquitous sensors," in Pervasive Computing, ser. LNCS, A. Ferscha and F. Mattern, Eds. Springer, 2004, pp. 158– 175.

[13] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in Pervasive Computing, ser. LNCS, A. Ferscha and F. Mattern, Eds. Springer, 2004, pp. 1–17.

[14] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," ACM SIGKDD Explorations Newsletter, vol. 12, no. 2, pp. 74–82, 2011.

[15] R. Poppe, "A survey on vision-based human action recognition," Image and Vision Computing, vol. 28, no. 6, pp. 976–990, 2010.

[16] L. Chen and C. Nugent, "Ontology-based activity recognition in intelligent pervasive environments," International Journal of Web Information Systems, vol. 5, no. 4, pp. 410–430, 2009.

[17] D. Riboni and C. Bettini, "OWL 2 modeling and reasoning with complex human activities," Pervasive and Mobile Computing, vol. 7, no. 3, pp. 379–395, 2011.

[18] T. van Kasteren and B. Krose, "Bayesian activity recognition in ¨ residence for elders," in Proceedings of the 3rd IET International Conference on Intelligent Environments. IEEE, 2007, pp. 209–212.

[19] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in Proceedings of the 18th International Conference on Machine Learning. Morgan Kaufmann, 2001.

[20] L. R. Rabiner and B.-H. Juang, "An introduction to hidden Markov models," ASSP Magazine, vol. 3, no. 1, pp. 4–16, 1986.

[21] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Machine Learning, vol. 29, no. 2-3, pp. 131–163, 1997.

[22] E. Kim, S. Helal, and D. Cook, "Human activity recognition and pattern

ISSN: 2057-5688

discovery," Pervasive Computing, vol. 9, no. 1, pp. 48–53, 2010.

[23] W. Reisig, Petri nets: an introduction. Springer Science & Business Media, 2012, vol. 4.

[24] T. Murata, "Petri nets: Properties, analysis and applications," Proceedings of the IEEE, vol. 77, no. 4, pp. 541–580, 1989.

[25] H. M. W. Verbeek, J. C. A. M. Buijs, B. F. Van Dongen, and W. M. P. van der Aalst, "ProM 6: The process mining toolkit," in Proceedings of the Business Process Management Demonstration Track, ser. CEURWS.org, 2010, pp. 34–39.