Predicting Treatment Repetitions in the Implant Denture Therapy Process

Marzieh Bakhshandeh*, Dennis M.M. Schunselaar†, Henrik Leopold‡ and Hajo A. Reijers§
Department of Computer Science
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
*Email: *m.bakhshandeh@vu.nl, †d.m.m.schunselaar@vu.nl, ‡h.leopold@vu.nl, §h.a.reijers@vu.nl*

***Abstract*—Healthcare can be considerably expensive for both patients and insurance companies. In some cases, high costs in healthcare are an indirect outcome of a low quality of care, for example, when treatments have to be repeated. Unfortunately, identifying the factors that lead to such repetitions is a complex and challenging task. In this paper, we focus on the domain of dental healthcare and develop an approach that can predict treatment repetitions in the context of the implant denture therapy process. The challenges associated with predicting treatment repetitions in this setting are considerable. First, hardly any patient undergoes the exact same series of treatments like another. This results in a high degree of variation in the data. Second, only a few patients experience treatment repetitions. This lead to a highly imbalance in the data. To address these challenges, we develop a prediction technique that particularly exploits the process perspective. What is more, we apply so-called resampling methods to deal with the imbalance in the data. Our resulting model is able to predict treatment repetitions with an AUC value of 0.69.**

***Keywords*-process prediction; treatment repetitions; implant denture therapy**

I. Introduction

Healthcare can be considerably expensive for both patients and insurance companies. Typically, such high costs are caused by a combination of expensive medication, highly customized treatments, and a large number of caregivers that are involved in the delivery of care [1]. Sometimes, however, high costs are also the indirect outcome of a low quality of care [2]. Among others, this is the case when treatments have to be repeated. As an example, consider a patient receiving a dental implant. If certain steps of the implant denture therapy process have to be repeated because of an insufficiently fitting implant or a preventable infection, the costs associated with the implant can be expected to increase considerably.

Unfortunately, identifying the factors that lead to such repetitions is a complex and challenging task. One reason is the large number of parties that may be involved in the delivery of care to a single patient: the general practitioner, medical specialists, medical professionals within a hospital, etc. Another reason is that in many medical domains different patients may receive treatments from varying combinations of these parties for the same health problem. While in theory this variation should not affect the quality of care, data from the healthcare domain suggests that it actually does [3].

In this paper, we focus on the domain of dental healthcare and develop a technique that can predict treatment repetitions in the context of the implant denture therapy process. We chose this particular process because it is one of the most cost-intensive processes in dental therapy. Hence, repetitions in this process have particularly severe financial consequences for both patients and insurance companies. The challenges associated with predicting treatment repetitions in the implant denture therapy process are twofold. First, there are hardly any two patients who receive the exact same series of treatments. Therefore, we need to deal with a high degree of *variation* in the data. Second, the fraction of patients experiencing treatment repetitions is typically smaller than five percent. This means, we also need to deal with a considerable *imbalance* in the data. To address these challenges, we develop a prediction technique that particularly exploits the *process perspective* [4], i.e., the temporal order of treatments. What is more, we apply so-called resampling methods to deal with the imbalance in the data. Our resulting model is able to predict treatment repetitions with an AUC value of 0.69.

The rest of the paper is structured as follows. Section II introduces the implant denture therapy process and elaborates on the problem we address. Section III discusses related work and highlights the innovative aspects of our work. Section IV develops our technique for treatment prediction. Section V presents the results. Finally, Section VI discusses the implications of our work before Section VII concludes the paper.

II. Problem Illustration

The goal of this paper is to develop a technique for predicting treatment repetitions in the implant denture therapy process. Figure 1 shows the main activities of the implant denture therapy process using the Business Process Modeling and Notation (BPMN). It shows that the process starts when a request from a patient is received. Such a request can occur in the context of regular implant consultation or also in the context of an emergency, for instance, when a

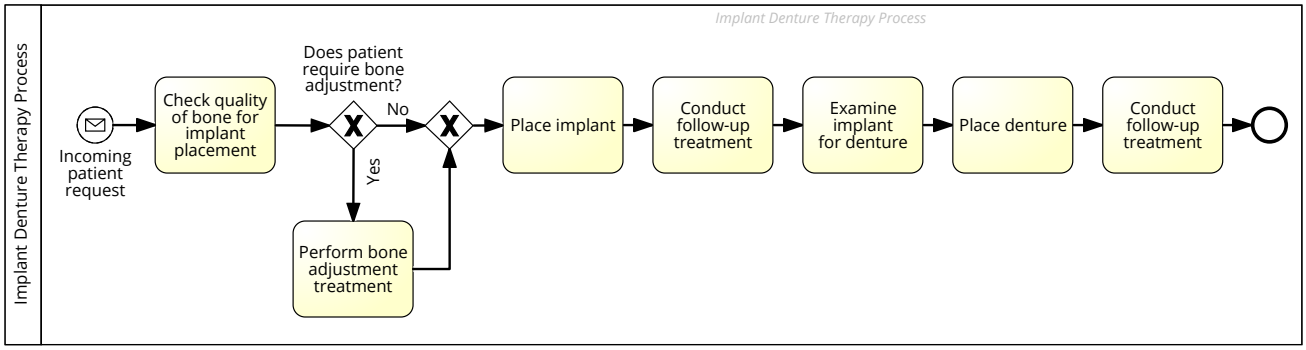


Figure 1: Implant denture therapy process (simplified)

tooth is broken in an accident. Afterwards, several steps are conducted before placing the actual implant. Most importantly, the caregiver will check whether a bone adjustment is required (the jaw bone degenerates after losing a tooth and might not be sufficient for "holding" the implant). If this is the case, a respective bone adjustment treatment is performed. Then the implant is placed. Once the implant is suitable for the denture, the denture is placed in the jaw. If everything went satisfactory, i.e., the implant fits well and the patient did not experience any infections, the process ends as illustrated in Figure 1.

While this is the case for the majority of patients, some patients also encounter problems. In some cases, the denture needs be replaced or the implant placement has to be repeated. In a very few cases, even both is necessary. Since such repetitions are very costly for the patient and the insurance company, they should be avoided to the largest possible extent. Given this setting, we operationalize the goal of this paper as a multi-class classification problem with four distinct classes:

- Class 0: No repetition
- Class 1: Repetition of the implant placement
- Class 2: Repetition of the denture placement
- Class 3: Repetition of implant and denture placement

The challenges associated with developing an appropriate prediction model for this problem are twofold. First, there are hardly any two patients receiving the exact same series of treatments, resulting in a high degree of *variation* in the data. Second, the fraction of patients experiencing treatment repetitions is typically smaller than five percent, which means the data is also highly imbalanced. In the next section, we review related work in order to show to what extent prior work has addressed these challenges.

III. Related Work

In this paper, we adopt a process perspective, i.e., we exploit information about the temporal order of the treatments a patient undergoes. Thus, this paper most closely relates to work in the area of *process mining* [5]. Within the field of process mining, various approaches exist for predicting relevant aspects of a particular case. These include predicting the time remaining until completion [6], [7], [8], the risks involved with possible future events [9], predicting the case outcome [10], and predicting which future events will probably be executed [8].

Within this paper, we are particularly interested in predicting the case outcome, i.e., we want to predict whether a patient will undergo additional treatments. Existing strategies to predict the case outcome include the application of decision and regressions trees [10], the combination of the control-flow perspective with the data perspective [11], and the analysis of how specific data attributes change over time [12]. What all these approaches have in common is that they do not take a potential imbalance in the data into account. Hence, they are not directly applicable to our setting. At the same time, the technique presented in this paper can provide a valuable addition for their applicability to other (medical) domains.

IV. Predicting Treatment Repetitions

In this section, we develop our technique for predicting treatment repetitions in the implant denture therapy process. Section IV-A presents the data set we use. Section IV-B introduces the features we selected for our prediction model. Section IV-C discusses the classifier selection. Section IV-D explains how we address the data imbalance challenge. Finally, Section IV-E describes how we selected the best prediction model.

A. Data

For this paper we used a data set we obtained from an health insurance company in the Netherlands. The data set consists of 1,048,576 dental treatments that have been conducted for 36,527 patients between 2009 to 2015. For each treatment, we had information about the executer and the day the treatment has been performed. Based on this information, we were able to reconstruct the individual process each patient experienced in terms of the order of treatments. In the following, we will refer to this individual

Table I: Data set characteristics

Characteristic	Total	Avg. per case
Number of patients	36,527	1
Number of executers	7,249	2
Number of treatments	1,048,576	91
Number of (distinct) treatments	1,255	31
Variability	97 %	not applicable
Duration	2009-2015	24 months

Table II: Feature scores based on UVFS and RFFI

Feature	UVFS	RFFI
Patient sex	1.194	0.033
Process compliance	8.807	0.030
Bone adjustment before the implant	117.411	0.188
No. of treatments before the implant	17.330	0.644
Implant executor	7.874	0.042
Type of denture	16.735	0.064

process as a *case*. Table I summarizes the key characteristics of the data set. Note that we differentiate between the total number of treatments and the distinct number of treatments since several treatments, such as check-ups, are conducted multiple times with in a single case.

Besides the data set, we also collected information about the implant denture therapy process. To this end, we had several meetings with implant knowledge experts such as dentist, dental implantologist, oral surgeons, and denturists. The outcome was a reference process that defined the boundaries of how the implant denture therapy process should be implemented in practice.

B. Feature Selection

To identify appropriate prediction features, we anlayzed our data set with respect to potentially discriminating features. As a result, we selected a set of six features:

- Patient sex (female / male)
- Process compliance (true / false)
- Bone adjustment before the implant (yes / no)
- Number of treatments taken before the implant (integer)
- Implant executor (dentist / surgeon)
- Type of denture (lower denture / upper denture / full denture)

The most notable and also innovative feature is the feature *process compliance*. It is only *true* if the executor (i.e. the dentist or the surgeon) followed the reference process we defined together with the knowledge experts. In case additional treatments were conducted or certain treatments were skipped, the feature *process compliance* is *false*. It is worth highlighting that, with exception of the number of treatments taken before the implant, all features have binary values. In order to check whether the selected features were adequate, we used two algorithms: univariate feature selection (UVFS) and random forest feature importance (RFFI). Table II shows the scores obtained for each feature using the two algorithms.

The results from Table II illustrate that the univariate feature selection considers *bone adjustment*, the *number of treatments taken before the implant*, and the *type of denture* as most informative. The random forest feature importance confirms this assessment particularly for the features *bone adjustment* and the *number of treatments taken before the implant*. However, given the scores and the potentially interesting domain insights based on the full set of features, we decided to include all of them in the prediction model.

C. Classifier Selection

Selecting appropriate classification algorithms is a key task when developing a prediction technique. However, unfortunately, there is no silver bullet. Hence, we analzyed related work to identify classification algorithms that have been applied in similar healtchare contexts [13], [14]. As a result, we selected three promising candidates: (1) decision trees (DTs), (2) random forests (RFs), and (3) and Support Vector Machines (SVMs). The advantage of decision trees is that they are easy to interpret and that their visualization is often considered useful for identifying interesting patterns [15]. Random forests avoid overfitting and can deal well with unstructured data. SVMs have been found to deal well with high variation, also when the variation only occurs in a small fraction of the data.

D. Dealing with the Data Imbalance

We identified two strategies to deal with the imbalance in our data set, i.e. the fact that only a few patients experienced repetitions: (1) We defined weights for each of the four classes when training the predictive model. With this approach the classification algorithm is configured in such a way that all classes receive weights that are inversely proportional to the class frequencies in the input data. In this way, we ensured that the predictive model was constructed in a balanced way with respect to each class. (2) We used techniques for the generation of synthetic samples. More specifically, we used advanced resampling methods to generate new samples from the minority classes to obtain a balanced data set. Unlike other oversampling techniques, the employed techniques SMOTE [16] and ADASYN [17] do not replicate existing samples. The core idea of these methods is to resample to training set but to keep the validation set with the original distribution.

In order to select the most appropriate data balancing strategy, we compared the normalized confusion matrices we obtained for each of the classifiers when using the different techniques to balance the data set. Our goal was to identify the model with the highest proportion of true positives in the

Table III: Results for different classifiers in combination with SMOTE

		Precision	Recall	F1 Score
Class 0	Decision tree depth 5	0.904	0.086	0.154
	Random forest depth 5	0.934	0.288	0.434
	SVM OvO	0.936	0.310	0.464
	SVM OvA	0.938	0.248	0.375
Class 1	Decision tree depth 5	0.084	0.321	0.133
	Random forest depth 5	0.097	0.318	0.148
	SVM OvO	0.107	0.385	0.167
	SVM OvA	0.085	0.164	0.110
Class 2	Decision tree depth 5	0.021	0.621	0.041
	Random forest depth 5	0.020	0.408	0.039
	SVM OvO	0.020	0.329	0.037
	SVM OvA	0.021	0.521	0.041
Class 3	Decision tree depth 5	0.011	0.233	0.021
	Random forest depth 5	0.008	0.200	0.015
	SVM OvO	0.004	0.167	0.007
	SVM OvA	0.017	0.442	0.033

minority classes. Preferring a model with a high proportion of true positives in the minority classes over others can be justified by the fact that patients generally do not require implant repetitions or denture replacements. Therefore, we focused on the cases in which it was necessary to repeat the implant, replace the denture or both.

For each of the selected classification algorithms (DTs, RFs, and SVMs), we trained six different models by applying different variations of the previously introduced data balancing strategies. This resulted in a total of 24 predictive models. For each predictive model, we computed precision, recall, and the F1 score. The results revealed that the data balancing method SMOTE produced the best results for all classifiers. More specifically, the models that were trained with synthetic samples obtained by SMOTE delivered the highest number of true positives in the minority classes. As a result, we selected SMOTE to deal with the data imbalance challenge.

E. Selection of best prediction model

To select the best prediction model, we analyzed which model yielded the best results in terms of precision, recall, and F1 score for the minority classes. Table III gives an overview of the results obtained for each model and each class.

Focusing on the overall F1 score, the results emphasize the challenge associated with accurately predicting the minority classes 1, 2, and 3. Analyzing the detailed results for each class reveals that the SVM One-versus-One (OvO) classifier delivered the best results for classes 0 and 1. That is, it was most adequate for correctly labelling patients who will not need any repetitions and patients who will need to repeat the implant placement. For class 2 (patients who will need a replacement of the denture) the results are quite similar for all classifiers. However, the F1 scores produced by the DT classifier and the SVM One-versus-All (OvA) classifier are slightly higher. With respect to class 3 (patients who will need to repeat both the implant and the denture placement), the DT classifier and the SVM OvA classifier yielded the highest F1 scores. Nevertheless, it is necessary to emphasize that the prediction accuracy for this group of patients is poor for all classifiers.

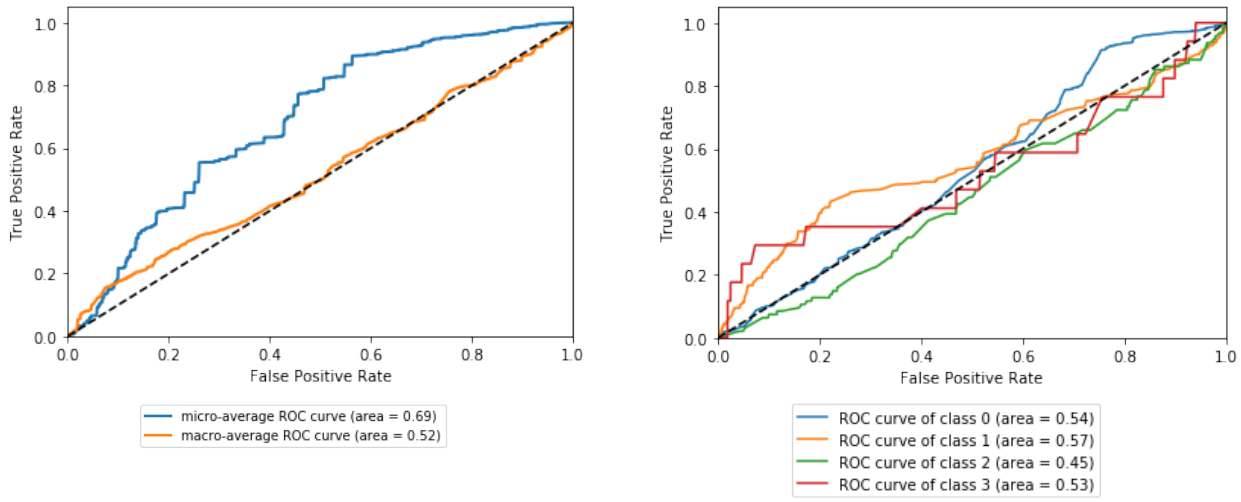
Based on these results, we chose the SVM OvO classifier as the best classifier because it produced the best F1 scores for the minority classes 1, 2, and 3. This type of SVM implementation trains a separate classifier for each pair of labels and, therefore, is less sensitive to problems associated with imbalanced data sets.

V. Results

To illustrate the results for the selected SVM OvO model with SMOTE, we employ ROC curves and the corresponding AUC (area under the curve) values. ROC curves are a graphical representation of the proportion of true positives (TPR = True Positive Rate) versus the proportion of false positives (FPR = False Positive Rate) and often used to illustrate the diagnostic capabilities of a binary classifier. The AUC value represents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The AUC value varies between 0 and 1. An uninformative classifier would yield an AUC value of 0.5, a perfect classifier would respectively yield an AUC value of 1.0. Figure 2 shows the ROC curves with the selected SVM OvO classifier. Since we address a multi-class classification problem, we binarized the output. Figure 2a shows the micro-average and the macro-average values for all classes. Figure 2b shows the results for each class separately.

Looking into the details, Figure 2a shows that the AUC value for the macro-average is considerably lower than the AUC value for the micro-average (0.52 versus 0.69). This is the case because the micro-average takes into account the TPR and FPR from a global point of view, while the macro-average takes into account the TPR and FPR values of each class first and then calculates the average. This again emphasizes that the results for the majority class (class 0) are significantly higher than the results obtained for the minority classes (1, 2, and 3). If the results for class 0 are not weighted based on the size of the class (as done in the micro-average), the resulting AUC value is considerably lower.

Figure 2b also shows the ROC curves and AUC values for each class separately. We calculated the ROC curves for each class by interpreting the output of our technique from the perspective of a binary classifier. That means the ROC curve for class 0 only differentiates whether a patient belonged to class 0 (positive) or not (negative). The disaggregated results show that the highest AUC value was obtained for class 1



(a) Overall results

(b) Results for each class

Figure 2: ROC curves for the selected SVM Classifier (OvA) using SMOTE

(0.57). However, the AUC values obtained for class 0 and class 3 are quite close (0.54 and 0.53 respectively). The AUC value for class 2, by contrast, was much lower (0.45), showing that the probability of obtaining false positives for this class is higher than the probability of obtaining true positives.

From a general perspective, these results show that our multi-class classifier is able to successfully predict treatment repetitions for a patient with a probability of 69%. However, the ability of our classifier to correctly do so depends on the specific class a patient belongs to. Patients who will not require an implant repetition or denture replacement (class 0), patients who will only need an implant repetition (class 1), and patients who will need both an implant repetition and a denture replacement (class 3) were correctly predicted in more than half of the cases. Patients who will only need a denture replacement are only correctly predicted with a probability of 45%. This highlights the trade-off in the context of building a multi-class classifier based on data with high variability. One the one hand, it is valuable to be able to predict different types of problems that can occur. On the other hand, the overall accuracy can be expected to turn out lower than for binary prediction models.

VI. Discussion

The work presented in this paper has implications for research as well as for practice.

From a *research* perspective, we showed that leveraging features resulting from a process view is a promising strategy to deal with the variability in a data set. Features such as *the number of treatments taken before the implant* or *process compliance* have the advantage of abstracting from the large variability in the process flow and having discriminating power even though hardly any patient undergoes the exact same series of treatments like another patient. What is more, we showed that the data resampling method SMOTE is a viable choice to deal with imbalanced data. As pointed out earlier, more than 90% of all cases progressed satisfactorily over time, without the need for additional interventions after the implant or denture placement. Despite this considerable imbalance, SMOTE enabled us to develop a model yielding an AUC value of 0.69. Comparing this value to results obtained in similar settings, this represents a competitive outcome.

From a *practice* perspective, our work has especially implications for the medical domain. Analyzing the role of the individual features in our model revealed a number of interesting findings with potentially considerable impact. First, we observed that the gender of a patient influences the likelihood of a successful treatment without repetitions. We found that male patients are much more likely to experience implant repetitions or denture replacements than female patients. Second, we found that the executer plays an important role. Under comparable circumstances, an implant placed by a surgeon was much more likely to result in treatment repetitions than an implant placed by a dentist. While these findings cannot be generalized at this stage, they provide interesting input for both investigating current medical practices as well as for further studies.

VII. Conclusion

In this paper, we developed a technique for predicting whether a patient will experience treatment repetitions in the context of the implant denture therapy process. To this end, we had to overcome two main challenges. First, we had to deal with the high degree of variation in the data set,

resulting from the fact that there are hardly any cases where two patients follow the exact same series of treatments. Second, we had to deal with the imbalance in the data, resulting from the fact that only a fraction of the patients experience treatment repetitions.

In order to overcome these challenges, we exploited the process perspective and, among others, defined process-related prediction features such as process compliance. What is more, we employed the resampling method SMOTE to deal with the imbalance in the data. Our final multi-class prediction model yielded an AUC value of 0.69, which means that we can successfully predict treatment repetitions in 69% of the cases. Given the multi-class setting and the high degree of variation in the data, this can be considered a satisfying result. Besides overcoming technical challenges, our work also revealed a number of domain-specific insights. Among others, we found that men are more likely to undergo treatment repetitions and that the risk of treatment repetitions is higher when an implant is placed by a surgeon than by a dentist.

In future work, we plan to extend our technique and predict additional aspects. For instance, we would like to be able to differentiate between necessary and unnecessary repetitions (from a medical perspective). What is more, we want to test additional prediction methods. While our analysis of related work indicated that particularly decision trees, random forests, and SVMs are promising techniques, we also believe that deep learning technique might represent a viable choice in this setting. Finally, we would like to test our technique on data sets from other (medical) domains. In this way, we can learn about the generalizability of our technique.

REFERENCES

[1] P. Homayounfar, "Process mining challenges in hospital information systems," in *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on*. IEEE, 2012, pp. 1135–1140.

[2] B. Chaudhry, J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S. C. Morton, and P. G. Shekelle, "Systematic review: impact of health information technology on quality, efficiency, and costs of medical care," *Annals of internal medicine*, vol. 144, no. 10, pp. 742–752, 2006.

[3] E. G. L. de Murillas, W. M. P. van der Aalst, and H. A. Reijers, "Process mining on databases: Unearthing historical data from redo logs," in *International Conference on Business Process Management*. Springer, 2015, pp. 367–385.

[4] M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst, "PM ^2 : A process mining project methodology," in *Advanced Information Systems Engineering - 27th International Conference, CAiSE 2015, Stockholm, Sweden, June 8-12, 2015, Proceedings*, 2015, pp. 297–313.

[5] W. M. P. Van der Aalst, "Data mining," in *Process Mining*. Springer, 2011, pp. 59–91.

[6] B. F. van Dongen, R. A. Crooy, and W. M. P. van der Aalst, "Cycle time prediction: When will this case finally be finished?" in *On the Move to Meaningful Internet Systems: OTM 2008, OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Monterrey, Mexico, November 9-14, 2008, Proceedings, Part I*, 2008, pp. 319–336.

[7] W. M. P. van der Aalst, M. H. Schonenberg, and M. Song, "Time prediction based on process mining," *Inf. Syst.*, vol. 36, no. 2, pp. 450–475, 2011.

[8] N. Tax, I. Verenich, M. La Rosa, and M. Dumas, "Predictive business process monitoring with LSTM neural networks," *CoRR*, vol. abs/1612.02130, 2016. [Online]. Available: http://arxiv.org/abs/1612.02130

[9] R. Conforti, M. de Leoni, M. La Rosa, W. M. P. van der Aalst, and A. H. M. ter Hofstede, "A recommendation system for predicting risks across multiple business process instances," *Decision Support Systems*, vol. 69, pp. 1–19, 2015.

[10] M. de Leoni, W. M. P. van der Aalst, and M. Dees, "A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs," *Inf. Syst.*, vol. 56, pp. 235–257, 2016.

[11] C. Di Francescomarino, M. Dumas, F. M. Maggi, and I. Teinemaa, "Clustering-based predictive process monitoring," *CoRR*, 2015.

[12] A. Leontjeva, R. Conforti, C. Di Francescomarino, M. Dumas, and F. M. Maggi, "Complex symbolic sequence encodings for predictive monitoring of business processes," in *Business Process Management - 13th International Conference, BPM 2015, Innsbruck, Austria, August 31 - September 3, 2015, Proceedings*, 2015, pp. 297–313.

[13] J. Futoma, J. Morris, and J. Lucas, "A comparison of models for predicting early hospital readmissions," *Journal of biomedical informatics*, vol. 56, pp. 229–238, 2015.

[14] A. Sharafoddini, J. A. Dubin, and J. Lee, "Patient similarity in prediction models based on health data: a scoping review," *JMIR medical informatics*, vol. 5, no. 1, 2017.

[15] Y. M. Chae, S. H. Ho, K. W. Cho, D. H. Lee, and S. H. Ji, "Data mining approach to policy analysis in a health insurance domain," *International journal of medical informatics*, vol. 62, no. 2, pp. 103–111, 2001.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[17] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. IEEE, 2008, pp. 1322–1328.