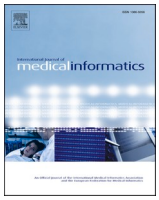
International Journal of Medical Informatics 170 (2023) 104936

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



External validation of prediction models for patient-reported outcome measurements collected using the SELFBACK mobile app☆

Deepika Verma [a,*], Kerstin Bach [a], Paul Jarle Mork [b]

[a] Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway
[b] Department of Public Health and Nursing, Norwegian University of Science and Technology, Trondheim, Norway

ARTICLE INFO

Keywords:
Machine learning
Case-based reasoning
Low-back pain
Neck pain
Patient-reported outcome measurements
Self-reported measures
Outcome prediction

ABSTRACT

Background: External validation is essential in examining the disparities in the training and validation cohorts during the development of prediction models, especially when the application domain is healthcare-oriented. Currently, the use of prediction models in healthcare research aimed at utilising the under-explored potential of patient-reported outcome measurements (PROMs) is limited, and few are validated using external datasets.
Objective: To validate the machine learning prediction models developed in our previous work [29] for predicting four pain-related patient-reported outcomes from the SELFBACK clinical trial datasets.
Methods: We evaluate the validity of three pre-trained prediction models based on three methods— Case-Based Reasoning, Support Vector Regression, and XGBoost Regression—using an external dataset that contains PROMs collected from patients with non-specific neck and or low back pain using the SELFBACK mobile application.
Results: Overall, the predictive power was low, except for prediction of one of the outcomes. The results indicate that while the predictions are far from immaculate in either case, the models show ability to generalise and predict outcomes for a new dataset.
Conclusion: External validation of the prediction models presents modest results and highlights the individual differences and need for external validation of prediction models in clinical settings. There is need for further development in this area of machine learning application and patient-centred care.

1. Introduction

Use of technology to support self-management of musculoskeletal pain is a feasible and promising approach [15,22]. In the SELFBACK project, a mobile app was developed to make weekly tailored self-management plans for users to help them manage back pain and other pain-related symptoms [17]. The self-management plans are tailored to each user based on a set of variables reported by the user in the mobile application. Tools like SELFBACK enable the effective use of technology for bridging the gap between patient-reported outcome measurements (PROMs) and patient-centred care. PROMs serve as a tool to assess and evaluate the health status of a patient from the patient's perspective at any given time point [18]. They may be recorded before, during or after a healthcare intervention and can help in measuring the impact of the intervention given to the patient. From a clinical perspective, the addition of predictive analytics to such healthcare tools could serve to further improve patient-centred care by detecting early signs of deteriorating outcomes, and warning primary caregivers to proactively prevent their occurrence [30,27]. This can therefore help caregivers to optimise the treatment approach for a given patient.

Previous research has shown that integrating technology with healthcare data can support preventive treatment [8,1], hospital re-admissions [23], and prevention of post-surgical complications [11]. To make further advancements in this field, it is important to have a clear understanding of what factors should be considered when deciding the treatment approach for a given patient [10]. From both clinical and machine learning points of view, this translates to deciding features from the available data that may be valuable in predicting a future outcome. Furthermore, external validation is essential to assess the generalisability of the prediction models [16,7].

Most studies that address the prediction of PROMs using machine learning methods have only validated their models internally [28]. Bootstrapping may be an approach suitable for internal validation to compensate for the lack of external validation due to the bias-corrected

☆ This work is the result of the Back-Up research project funded by the European Union Horizon 2020 under grant agreement No 777090.
* Corresponding author.
E-mail addresses: deepika.verma@ntnu.no (D. Verma), kerstin.bach@ntnu.no (K. Bach), paul.mork@ntnu.no (P.J. Mork).

https://doi.org/10.1016/j.ijmedinf.2022.104936
Received 10 August 2022; Received in revised form 18 November 2022; Accepted 22 November 2022
Available online 26 November 2022

estimation of the prediction models [25]. However, the bootstrapping method cannot replace external validation since models often perform better on the dataset they were trained on compared to validation on a different dataset [4]. This effect is often attributed to the overfitting of the model caused by the high variance. Furthermore, since clinical datasets tend to be relatively small, it is unlikely that internal validation would be sufficient as prediction models are prone to overfitting when using small datasets [14].

This paper presents an evaluation of the prediction models developed in our previous work [29] using an external dataset. In our previous work, a twofold feature selection approach that combines correlation and data-driven similarities in Case-Based Reasoning (CBR) was used to identify relevant features for predicting a set of PROMs in the SELFBACK dataset. The features selected were used to build prediction models using three methods—CBR, Support Vector Regression (SVR), and XGBoost Regression (XGB).

2. Methods

2.1. Dataset

The dataset used for external validation consists of PROMs collected from patients with non-specific neck and/or low back pain in a randomised controlled trial (RCT II[1]) with the help of questionnaires to evaluate the effectiveness of the SELFBACK decision support system (DSS) in a secondary care setting [15]. The dataset used for training the models consisted of PROMs collected from patients with non-specific low back pain during an earlier RCT (I[2]) with the help of questionnaires to evaluate the SELFBACK DSS in a primary care setting [22].

Fig. 1 shows how the data collection was carried out in the two RCTs. The collected data is categorised into Baseline, Tailoring, and Follow-Up (FU). Only data from baseline and the 3-month follow-up 2 (FU-2 data) is used to train and evaluate the prediction models. In total, the training dataset includes 218 patients while the external validation dataset includes 75 patients that completed at least the FU-2 questionnaire. The external validation dataset is a subset of the data collected in RCT II. A detailed account of the data collection in the two RCTs can be found in Sandal et al. [22] and Marcuzzi et al. [15].

During data collection (for the external validation dataset), eligible patients who accepted to join the study answered questionnaires at different time points: (1) at the time of intake: Baseline questionnaire (*Baseline Data*), (2) at the end of every week: Tailoring questionnaire (*Tailoring Data*), (3) at the end of 6-weeks, 3-months, 6-months: Follow-Up questionnaire ((*FU Data*)). The questionnaires include validated clinical measures of pain level, pain self-efficacy, work-ability, mood, physical activity, sleep quality, functional ability, and fear avoidance. In addition to the clinical measures, the baseline questionnaire also includes questions regarding patient demographics such as age, height, weight, education, employment type, and family. Based on the patients' responses at baseline, the SELFBACK mobile application recommends an exercise plan and educational elements along with tracking their number of steps every day from a wearable device (Xiaomi Mi Band 3). Exercise completion and education readings are self-reported in the app [21].

Target Outcomes

The training dataset originally comprised 47 features. In our previous work, we focused on six target outcomes. However, due to exclusion of one outcome in RCT II, two outcomes—Roland Morris Disability Questionnaire and Numeric Pain Rating Scale—had to be removed from this experimental evaluation due to feature dependency. Instead, we focus on the four secondary outcomes that were chosen to represent a diversity of domains; Workability index (WAI, range: [0,10]), Pain Self Efficacy Questionnaire (PSEQ, range: [0,60]), Fear Avoidance Belief Questionnaire (FABQ, range: [0,30]) and Global Perceived Effect Scale (GPE, range: [-5,+5]). We use the features previously selected for each target outcome based on the training dataset [29] and evaluate the generalisability of the models using the external dataset. Table 1 gives a brief summary of the various features used in this work. Marcuzzi et al. [15] give a more comprehensive summary of all the features collected at various time points in the RCT II.

2.2. Prediction models

Prediction models using three machine learning methods were trained on the completed PROMs collected in RCT I—CBR, SVR, and XGB—to predict the four target outcomes reported by patients in RCT II.

2.3. Feature selection

Two feature selection approaches were applied in the previous work to select features from the baseline dataset for each of the chosen outcomes to be predicted: i) a twofold hybrid approach that uses statistical correlation [20] to filter out the most correlated features, followed by a final selection of features based on CBR model built using data-driven local similarity modelling approach [26] (similarity modelling carried out in *myCBR workbench* [2]) and ii) an ensemble approach that uses permutation feature importance to select features with XGBoost as the base regressor [9].

2.4. Hyperparameter optimization

Before the two machine learning algorithms—SVR and XGB—were trained on the training dataset, their hyperparameters were tuned using grid search to optimize their performance on the dataset. Grid search was used to perform an exhaustive search through a pre-defined set of hyperparameter space for each learning algorithm to identify their optimal hyperparameters [12]. Regarding the CBR models, as there are no hyperparameters involved, this step was not required.

2.5. Evaluation metrics

The metrics used to evaluate the results in the experiments are Mean Absolute Error (MAE) and Normalized Mean Absolute Error (NMAE). MAE is the average of the absolute errors, i.e., the difference between the observed and predicted value. While there are several ways to normalize error, we normalized the MAE using the max–min method (see Eq. 2) for each outcome to get NMAE in the range [0,1]. This brings the results on the same scale and simplifies comparison across different models and outcomes.

$$MAE = \left(\frac{1}{n}\right)\sum_{i=1}^{n}|\widehat{y_i} - y_i| \tag{1}$$

$$NMAE = \left(\frac{MAE}{y_{max} - y_{min}}\right) \tag{2}$$

3. Experiments & results

The methods were implemented in Python [19] in jupyterLab notebook[3] using Scikit-learn [6] and *myCBR* Rest API[4] [3] was used for querying the CBR models developed in *myCBR workbench*. SVR and XGB models were 10-fold cross-validated during the training phase.

Table 2 summarises the results of the experiments. The SVR and XGB models gave the lowest prediction error for *WAI* and *FABQ* at *1.68* and

[1] https://clinicaltrials.gov/ct2/show/NCT04463043
[2] https://clinicaltrials.gov/ct2/show/NCT03798288
[3] https://jupyter.org/
[4] https://github.com/ntnu-ai-lab/mycbr-sdk

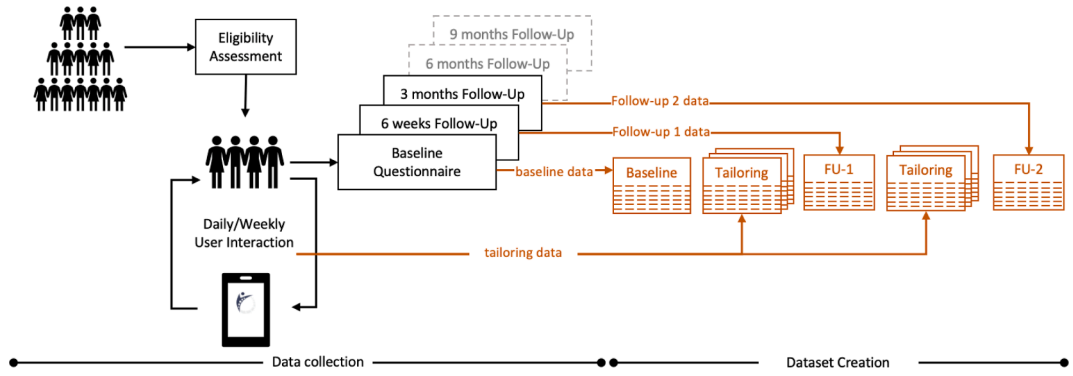


Fig. 1. Overview of data collection in two RCTs that evaluated the SELFBACK DSS. The different data components are indicated by the orange boxes. Only data from baseline and the 3-month follow-up (FU-2 data) is used to train and evaluate the prediction models.

Table 1
Summary of the SELFBACK RCT I & II dataset features used in this work. Abbreviated feature names in the bracket include the specific sub-scale scores used in this work either as a predictor or as a target outcome. Features predicted at FU-2 are marked with an asterisk (*).

Feature	Description
Age	Age of the participant in years
Body Mass Index (BMI)	Calculated using reported weight and height
Workability Index (WAI*)	Used to assess work-ability of an individual using an 11-point numeric rating scale
Pain Self-Efficacy Questionnaire (PSEQ*, PSEQ_2)	Used to assess the participants' level of confidence in carrying out specific activities despite their pain using ten items, each measured on a 6-point scale.
EuroQoL 5-dimension (EQ5D, EQ5D_mobility)	Used to assess health-related quality of life using five items, each scored 0–5
Brief Illness Perception Questionnaire (BIPQ_life, BIPQ_pain_continuation, BIPQ_concern, BIPQ_symptoms)	Used to evaluate participants' illness perception using eight items on an 11-point numeric rating scale
Pain Intensity (Pain_1year, Pain_worst)	Perceived intensity of low back and/or neck pain measured by a 11-point numerical rating scale
Sleep (Sleep_wakeup)	Sleep problems assessed by four self-report items which provide information needed to diagnose insomnia according to the DSM-V criteria
Fear-Avoidance Belief Questionnaire (FABQ*)	Physical activity sub-scale used to measure participant's beliefs about how physical activity affects their low back and/or neck pain using five items, each scored 0–6
Patient-Specific Functional Scale (PSFS)	Used to evaluate changes in participant's ability to perform up to two self-selected activities regarded as important by them using an 11-point score
Global Perceived Effect (GPE*)	Used to investigate the effect of the intervention as perceived by the participant using one item scored −5 to 5

4.04, respectively, using the features selected by the hybrid method. XGB and CBR gave the lowest prediction error for *PSEQ* and *GPE* at *8.04* and *1.30*, respectively, using the features selected by the permutation feature importance method. For the MAE, the error for *PSEQ* and *FABQ* is higher compared to the other two outcomes, however, considering the NMAE, these errors are comparable to that of the other two outcomes.

4. Discussion

The external validation sample in this work included PROMs from 75 patients, while the training and internal validation included 218 participants. In the internal validation of the models in our previous work (see Supplementary file), CBR and SVR gave the lowest MAE for *WAI* and *FABQ* at *1.14* and *3.60*, respectively, using the features selected by the hybrid method, while CBR gave the lowest MAE for *PSEQ* and *GPE* at *5.95* and *1.49*, respectively, using the features selected by the feature importance method. Comparing these figures to the results in Table 2, we can see that the models show slightly worse performance for the external dataset, which is usually expected. While the results for *PSEQ* appear worse in the external validation, when considering the performance of the same best-performing model in the internal validation for the outcome (XGB), the model in fact fared better on the external dataset (MAE 8.04) that on the training dataset (MAE 17.1). Although the predictive power was low, the evaluation suggest that the prediction models can be applied to a new dataset. The approach for selecting features seems to have negligible influence on the performance of the prediction models.

Training and testing a predictive model on the same dataset is by and large not considered optimal, especially when the predictions should be used to support clinical decision-making [24]. At a minimum, our evaluation substantiates the potential of both the PROMs and utility of machine learning methods for PROMs, while also highlighting the need for external validation and further development of prediction models. Future work should compare the predictions made by clinicians versus machine learning methods to fully assess the usefulness of machine learning methods in this field.

4.1. Study limitation

The fact that this work is based on patient-reported data may be considered a limitation owing to the limited reliability of subjective datasets [5]. Further, it is difficult to fully assess the extent of adequacy of the features selected for clinical judgement since clinicians themselves have a hard time selecting the most valuable or informative features [13].

5. Conclusion

To conclude, the external validation of prediction models presents modest results and highlights the need for further development in this area of machine learning application. While the results are still far from being applicable in a clinical setting, they nevertheless show potential in the methods as well as PROMs data. More research is prudent to further this field of machine learning application.

Table 2
Results of prediction of target outcomes at FU-2 using different feature selection methodologies and regression methods for the intervention group (size of the dataset: 75 patients). Values in each cell are MAE/NMAE pairs. Numbers in bold highlight the lowest MAE/NMAE pair. Abbreviations: **WAI**-Workability Index, **PSEQ**-Pain Self Efficacy Questionnaire, **FABQ**- Fear Avoidance Belief Questionnaire, **GPE**-Global Perceived Effect Scale, **n**-number of features, **CBR**-Case-based Reasoning, **SVR**-Support Vector Regression, **XGB**-XGBoost, **PFI**-Permutation Feature Importance.

	Feature Selection Methodology							
	Correlation + CBR				PFI + XGB			
Outcome [range]	n	CBR	SVR	XGB	n	CBR	SVR	XGB
WAI [0,10]	4	2.04/0.204	**1.68/0.168**	1.94/0.194	1	1.90/0.190	1.91/0.191	1.92/0.192
PSEQ [0,60]	3	9.97/0.166	9.49/0.158	8.66/0.144	2	10.28/0.171	9.8/0.163	**8.04/0.134**
FABQ [0,30]	1	5.54/0.184	4.09/0.133	**4.04/0.134**	6	5.24/0.174	4.34/0.144	4.74/0.158
GPE [-5,5]	2	1.32/0.132	1.92/0.192	1.55/0.155	3	**1.30/0.130**	1.60/0.160	1.43/0.143

Summary Table
What was already known:

- PROMs are a valuable source of information, but few studies have explored the application of machine learning methods to facilitate clinical decision support [10].
- Promising development in the application of machine learning methods on PROMs for identifying predictors of and predicting outcomes [27].
- External validation is an additional, but desired step in the development of machine learning prediction models [4].

What this study adds:

- Emphasizes the need for external validation of predictive models for clinical datasets.
- Machine learning models can generalise and predict PROMs, provided that predictors are generalisable and hold predictive power.
- Corroborates the utility of machine learning methods in predictive modelling for clinical datasets of subjective nature.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is funded by the Back-UP and selfBACK EU project. The Back-UP project is funded by the European Union's H2020 research and innovation programme under grant agreement No. 777090. The self-BACK project is funded by the European Union's H2020 research and innovation programme under grant agreement No. 689043.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ijmedinf.2022.104936.

References

[1] J. Andrews, R. Harrison, L. Brown, L. MacLean, F. Hwang, T. Smith, E.A. Williams, C. Timon, T. Adlam, H. Khadra, et al., Using the nana toolkit at home to predict older adults' future depression, J. Affect. Disorders 213 (2017) 187–190.
[2] K. Bach, K.D. Althoff, Developing Case-Based Reasoning Applications Using myCBR 3, in: I. Watson, B.D. Agudo (Eds.), Case-based Reasoning in Research and Development, Proceedings of the 20th International Conference on Case-Based Reasoning (ICCBR-12), LNAI 6880, Springer, 2012, pp. 17–31.
[3] K. Bach, B.M. Mathisen, A. Jaiswal, Demonstrating the mycbr rest api., in: ICCBR Workshops, 2019, pp. 144–155.
[4] S. Bleeker, H. Moll, E.a. Steyerberg, A. Donders, G. Derksen-Lubsen, D. Grobbee, K. Moons, External validation is necessary in prediction research:: A clinical example, J. Clin. Epidemiol. 56 (2003) 826–832.
[5] A. Bookstein, A. Lindsay, Questionnaire ambiguity: A rasch scaling model analysis. Graduate School of Library and Information Science. University of Illinois...., 1989.
[6] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, ECML PKDD Workshop: Languages for Data Mining and Machine Learning (2013) 108–122.
[7] F. Cabitza, A. Campagner, F. Soares, L.G. de Guadiana-Romualdo, F. Challa, A. Sulejmani, M. Seghezzi, A. Carobene, The importance of being external. methodological insights for the external validation of machine learning models in medicine, Computer Methods and Programs in Biomedicine 208 (2021) 106288.
[8] A.M. Chekroud, R.J. Zotti, Z. Shehzad, R. Gueorguieva, M.K. Johnson, M. H. Trivedi, T.D. Cannon, J.H. Krystal, P.R. Corlett, Cross-trial prediction of treatment outcome in depression: a machine learning approach, Lancet Psych. 3 (2016) 243–250.
[9] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res. 20 (2019) 1–81.
[10] A. Giga, How health leaders can benefit from predictive analytics, in: Healthcare management forum, SAGE Publications Sage CA, Los Angeles, CA, 2017, pp. 274–277.
[11] A.H. Harris, A.C. Kuo, Y. Weng, A.W. Trickey, T. Bowe, N.J. Giori, Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? Clin. Orthopaed. Related Res. 477 (2019) 452.
[12] F. Hutter, L. Kotthoff, J. Vanschoren, Automated Machine Learning, Springer, 2019.
[13] A.F. Leuchter, I.A. Cook, L.B. Marangell, W.S. Gilmer, K.S. Burgoyne, R. H. Howland, M.H. Trivedi, S. Zisook, R. Jain, J.T. McCracken, M. Fava, D. Iosifescu, S. Greenwald, Comparative effectiveness of biomarkers and clinical indicators for predicting outcomes of ssri treatment in major depressive disorder: results of the brite-md study, Psychiatry research 169 (2009), https://doi.org/10.1016/j.psychres.2009.06.004.
[14] A. Luedtke, E. Sadikova, R.C. Kessler, Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder, Clin. Psychol. Sci. 7 (2019) 445–461.
[15] A. Marcuzzi, K. Bach, A.L. Nordstoga, G.F. Bertheussen, I. Ashikhmin, N. A. Boldermo, E.N. Kvarner, T.I.L. Nilsen, G.H. Marchand, S.O. Ose, et al., Individually tailored self-management app-based intervention (selfback) versus a self-management web-based intervention (e-help) or usual care in people with low back and neck pain referred to secondary care: protocol for a multiarm randomised clinical trial, BMJ Open 11 (2021), https://doi.org/10.1136/bmjopen-2020-047921.
[16] K.G. Moons, R.A. Donders, T. Stijnen, F.E. Harrell Jr, Using the outcome for imputation of missing predictor values was preferred, Journal of clinical epidemiology 59 (2006) 1092–1101.
[17] P.J. Mork, K. Bach, A decision support system to enhance self-management of low back pain: protocol for the selfback project, JMIR research protocols 7 (2018) e167.
[18] E.C. Nelson, E. Eftimovska, C. Lind, A. Hager, J.H. Wasson, S. Lindblad, Patient reported outcome measures in practice, Bmj 350 (2015).
[19] T.E. Oliphant, Python for scientific computing, Computing in Science & Engineering 9 (2007) 10–20.
[20] K. Pearson, F. Galton, Vii. note on regression and inheritance in the case of two parents, Proceedings of the Royal Society of London 58 (1895) 240–242.
[21] L.F. Sandal, K. Bach, C.K. Øverås, M.J. Svendsen, T. Dalager, J.S.D. Jensen, A. Kongsvold, A.L. Nordstoga, E.M. Bardal, I. Ashikhmin, et al., Effectiveness of app-delivered, tailored self-management support for adults with lower back pain–related disability: a selfback randomized clinical trial, JAMA internal medicine 181 (2021) 1288–1296.
[22] L.F. Sandal, M.J. Stochkendahl, M.J. Svendsen, K. Wood, C.K. Øverås, A. L. Nordstoga, M. Villumsen, C.D.N. Rasmussen, B. Nicholl, K. Cooper, P. Kjaer, F. S. Mair, G. Sjøgaard, T.I.L. Nilsen, J. Hartvigsen, K. Bach, P.J. Mork, K. Søgaard, An app-delivered self-management program for people with low back pain: Protocol for the selfback randomized controlled trial, JMIR Res Protoc 8 (2019) e14720, https://doi.org/10.2196/14720.
[23] N. Schiltz, M. Dolansky, D. Warner, K. Stange, S. Gravenstein, S. Koroukian, Impact of instrumental activities of daily living limitations on hospital readmission: an observational study using machine learning, J. Gen. Intern. Med. (2020), https://doi.org/10.1007/s11606-020-05982-0.

[24] G.C. Siontis, I. Tzoulaki, P.J. Castaldi, J.P. Ioannidis, External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination, Journal of clinical epidemiology 68 (2015) 25–34.
[25] E.W. Steyerberg, F.E. Harrell, Prediction models need appropriate internal, internal–external, and external validation, Journal of clinical epidemiology 69 (2016) 245–247.
[26] D. Verma, K. Bach, P.J. Mork, Modelling similarity for comparing physical activity profiles - a data-driven approach, in: M.T. Cox, P. Funk, S. Begum (Eds.), CBR Research and Development, Springer, Cham, 2018, pp. 415–430.
[27] D. Verma, K. Bach, P.J. Mork, Application of machine learning methods on patient reported outcome measurements for predicting outcomes: A literature review, Informatics 8 (2021) 56, https://doi.org/10.3390/informatics8030056.
[28] D. Verma, K. Bach, P.J. Mork, Application of machine learning methods on patient reported outcome measurements for predicting outcomes: A literature review, Informatics 8 (2021), https://doi.org/10.3390/informatics8030056.
[29] D. Verma, K. Bach, P.J. Mork, Using automated feature selection for building case-based reasoning systems: An example from patient-reported outcome measurements, Int. Conf. Innovative Tech. Appl. Artif. Intell., Springer. (2021) 282–295.
[30] H.J. White, J. Bradley, N. Hadgis, E. Wittke, B. Piland, B. Tuttle, M. Erickson, M. E. Horn, Predicting patient-centered outcomes from spine surgery using risk assessment tools: a systematic review, Curr. Rev. Musculosk. Med. 13 (2020) 247.