DATABASE AND ML-DRIVEN SOFTWARE TOOL TO PREDICT CNC FOR BPMN MODELS USING TNN AND TNSF

Kopp Andrii
Ph.D., Associate Professor, Head of Software Engineering and Management Intelligent Technologies Department
National Technical University "Kharkiv Polytechnic Institute"

Kudii Dmytro
Ph.D., Associate Professor, Professor of Software Engineering and Management Intelligent Technologies Department
National Technical University "Kharkiv Polytechnic Institute"

Halatova Olha
Assistant of Software Engineering and Management Intelligent Technologies Department
National Technical University "Kharkiv Polytechnic Institute"

Diachenko Bohdan
Master's Student of Software Engineering and Management Intelligent Technologies Department
National Technical University "Kharkiv Polytechnic Institute"

Abstract. The relevance of the study is determined by the growing role of business process modeling using BPMN for analyzing, improving, and automating the activities in organizations, where excessive structural complexity of models negatively affects their comprehensibility and practical value for users. The aim of the work is to develop and experimentally verify an objective approach to assessing the quality of BPMN models from the perspective of structural complexity based on formal graph properties, quantitative metrics, and machine learning methods. In the study, the business process is presented as a directed labeled graph, for which basic size metrics and the CNC network connectivity coefficient are used as key complexity characteristics. A domain-specific model for CNC prediction based on linear regression and a mechanism for classifying BPMN models by complexity levels using threshold values are proposed. Experimental verification was performed on 6137 business process descriptions from the Camunda open repository and showed adequate prediction accuracy with MSE of 0.02 and R-squared of 0.71. The classification results demonstrated high efficiency for BPMN models of moderate and high complexity with an F1-score of up to 0.98. It was concluded that the combination of structural metrics and machine learning provides an objective and scalable assessment of the quality of BPMN models in terms of their understandability to users, therefore, business process models quality.

Keywords: business process model, database, process size, complexity metrics, coefficient of network connectivity, machine learning, software tool.

Introduction. Business process modeling is a key tool for analyzing, improving, and automating organizational activities, and the quality of such models directly affects the effectiveness of management and IT support for business [1–4]. In modern research, the quality of a business process model is considered to be the ability of the model to be understandable and useful to users in various contexts of application, in particular during the redesign and development of information systems [2, 3]. This study focuses on evaluating the quality of business process models described using the BPMN (Business Process Model and Notation) notation from the perspective of their structural complexity, using formal structural properties, quantitative metrics, and machine learning methods, which increases the objectivity and scalability of the assessment process.

Recent studies. Recent research in the field of assessing the quality of business process models in BPMN notation focuses on analyzing structural complexity as a key factor influencing the understandability and modifiability of models in a dynamic organizational environment. It has been proven that these quality attributes can only be objectively assessed through structural metrics, for which correlation, regression, and statistical analysis are used to determine threshold values that allow distinguishing between different levels of process model quality [5, 6]. Further development in this area is linked to the use of intelligent data analysis and machine learning methods for the automated establishment of metric thresholds and the classification of BPMN models by quality levels [7]. At the same time, there is growing interest in intelligent approaches to complexity assessment, in particular with the use of fuzzy logic, which allows for uncertainty and reduces the risks of the negative impact of excessive complexity of business process models on their quality and practical use [8].

Problem statement. The problem of assessing the quality of business process models lies in the lack of a universal and objective approach to determining their complexity and the impact of this complexity on the understandability of BPMN models for users.

A business process model can be formally presented as a directed labeled graph:

$$BPModel = (N, A),$$

where:

– N – is the set of business process elements, such as events E, activities T (tasks and sub-processes), and gateways G;

– A – is the set of sequence flows connecting business process elements.

For $BPModel$ basic size metrics, such as the total number of nodes and the total number of sequence flows, reflect the structural richness of the model:

– total number of nodes – $TNN = |N|$;

– total number of sequence flows – $TNSF = |A|$.

Since the quality of a business process model is defined as “the degree of its comprehensibility to users” [2, 3], the key quality characteristic is complexity, which is assessed using metrics borrowed from graph theory and software engineering.

In particular, the Coefficient of Network Connectivity (CNC) reflects the level of interconnection between the elements of the BPMN model, and an increase in its value increases the likelihood of errors and complicates the perception of the model [5–7]:

$$CNC = \frac{|A|}{|N|},$$

where:

– N – is the set of business process elements, such as events E, activities T (tasks and sub-processes), and gateways G;

– A – is the set of sequence flows connecting business process elements.

This necessitates the use of formal methods and machine learning for a reasonable assessment of the quality of business process models from the perspective of their complexity.

Research results. This section presents results confirming the feasibility of using machine learning methods and intelligent technologies for predicting and classifying the quality of business process models, in particular in the form of binary classification tasks using standard model quality assessment indicators [9, 10]. Modern approaches demonstrate the effectiveness of both classical machine learning algorithms and model parameter optimization methods, in particular based on gradient descent, which allows for increased stability and accuracy of results [11]. Along with this, intelligent information technologies are actively used for automated quality assessment and error detection in BPMN models based on the analysis of their structural properties [12]. This study proposes a novel approach that combines traditional machine learning methods with threshold values of business process model complexity metrics to determine the complexity levels of BPMN models and, accordingly, evaluate their quality in terms of user comprehensibility [2, 3].

Fig. 1 describes an approach in which BPMN models from the database are formalized as graph structures, from which the size structural features TNN and $TNSF$, as well as CNC metric are calculated, after which linear regression is used to predict complexity values and compare them with thresholds of complexity levels for further evaluation of the quality of business process models.

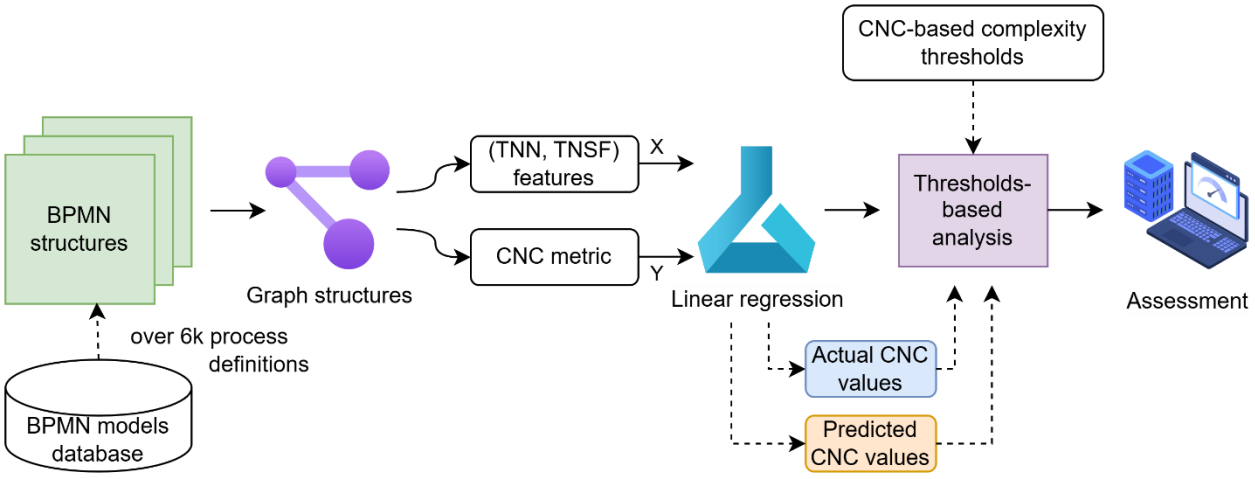

Figure 1. Approach to CNC prediction on business process models database.
Source: created by the authors.

A novel domain-oriented model for forecasting CNC values based on linear regression is proposed, in which parameters that minimize prediction errors are determined based on the available dataset of business process descriptions:

$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} [CNC_i - (\beta_0 + \beta_1 \cdot TNN_i + \beta_2 \cdot TNSF_i)]^2,$$

where:

– n – is the number of processes;
– β_0 – is the intercept;
– β_1 – is the slope for TNN;
– β_2 – is the slope for $TNSF$.

Accordingly, the predicted CNC values can be calculated using the mathematical relationship given below:

$$\widehat{CNC_i} = \beta_0 + \beta_1 \cdot TNN_i + \beta_2 \cdot TNSF_i, i = \overline{1,n}.$$

Complexity assessment based on levels and threshold values for CNC in this case is carried out as follows:

$$C_i(y,\theta) = \begin{cases} 1, & y < 0.4 \wedge \theta = Low, \\ 1, & 0.4 \le y \le 1.1 \wedge \theta = Moderate, \\ 1, & y \ge 1.1 \wedge \theta = High, \\ 0, & else, \end{cases}$$

where:

– y – is the CNC_i for actual or $\widehat{CNC_i}$ for predicted metric values, $i = \overline{1,n}$;
– θ – is the complexity level (Low, Moderate, High) to consider.

To quantitatively assess the error and quality of approximation of the proposed regression model for predicting CNC values, the root mean square error and coefficient of determination indicators are used:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(CNC_i - \widehat{CNC_i}\right)^2,$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(CNC_i - \widehat{CNC_i}\right)^2}{\sum_{i=1}^{n}\left[CNC_i - \left(\frac{1}{n}\sum_{j=1}^{n} CNC_j\right)\right]^2}.$$

The effectiveness of complexity assessment using the CNC metric at a given level is determined using classic elements of the machine learning error matrix, which includes correctly and incorrectly classified positive and negative cases:

$$TP(\theta) = \sum_{i=1}^{n}\left[C_i(CNC_i,\theta) = 1 \wedge C_i\left(\widehat{CNC_i},\theta\right) = 1\right],$$

$$TN(\theta) = \sum_{i=1}^{n}\left[C_i(CNC_i,\theta) = 0 \wedge C_i\left(\widehat{CNC_i},\theta\right) = 0\right],$$

$$FP(\theta) = \sum_{i=1}^{n}\left[C_i(CNC_i,\theta) = 0 \wedge C_i\left(\widehat{CNC_i},\theta\right) = 1\right],$$

$$FN(\theta) = \sum_{i=1}^{n} \left[C_i(CNC_i, \theta) = 1 \wedge C_i\left(\widehat{CNC_i}, \theta\right) = 0 \right].$$

Thus, the main evaluation indicators for classification analysis of complexity according to the CNC metric at the corresponding level are determined as follows:

$$Accuracy(\theta) = \frac{TP(\theta) + TN(\theta)}{TP(\theta) + TN(\theta) + FP(\theta) + FN(\theta)},$$

$$Precision(\theta) = \frac{TP(\theta)}{TP(\theta) + FP(\theta)},$$

$$Recall(\theta) = \frac{TP(\theta)}{TP(\theta) + FN(\theta)},$$

$$F_1(\theta) = 2 \cdot \frac{Precision(\theta) \cdot Recall(\theta)}{[Precision(\theta) + Recall(\theta)]}.$$

The study used an collection of BPMN files from the open GitHub repository Camunda, which contains 3729 business process diagrams created during training sessions based on text scenarios and intended for research use [13]. During the preliminary processing of this collection, it was taken into account that a single BPMN file may contain several process definitions for modeling interaction and collaboration, resulting in 6137 separate business process descriptions for which calculated structural metrics were available [14].

The Python programming language was used for the software implementation of the proposed approach and for conducting experimental research, on the basis of which the corresponding software tool was developed.

The presented below database schema (Fig. 2) reflects the logical structure of the business process model database, which covers the storage of BPMN files, process descriptions, calculated structural metrics, and complexity analysis results, ensuring data integrity and traceability, while the database itself is implemented using the MySQL database management system.

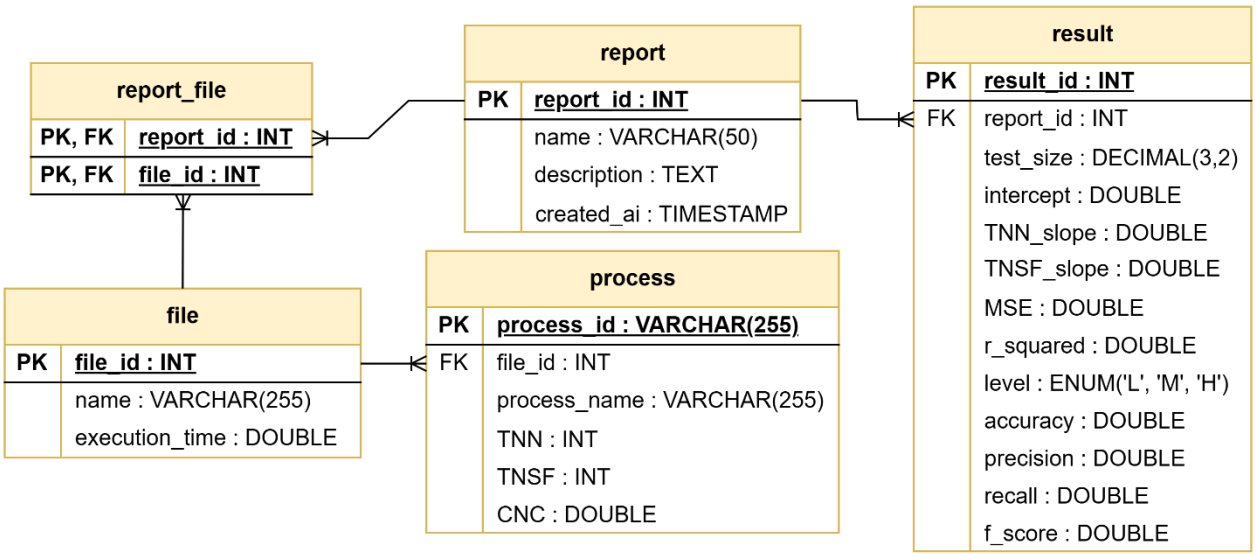

Figure 2. Business process models database structure to handle CNC prediction data.
Source: created by the authors.

To implement the linear regression model, we used the Scikit-Learn machine learning library, which provides tools for building, training, and testing models based on standard algorithms [15]. The model was trained using an 80/20 split of the sample into training and testing parts, resulting in a MSE of 0.02 and a coefficient of determination (R^2) of 0.71, which indicates an adequate level of prediction accuracy and the mode's ability to explain the variation in complexity metric values.

The linear regression model obtained within the study for predicting CNC values is presented below in the form of a corresponding analytical equation:

$$\widehat{CNC} = 0.3901 - 0.0004 \cdot TNN + 0.0332 \cdot TNSF.$$

The results of classifying BPMN models by CNC complexity levels using a linear regression model and threshold values indicate different recognition quality for each class (Fig. 3).

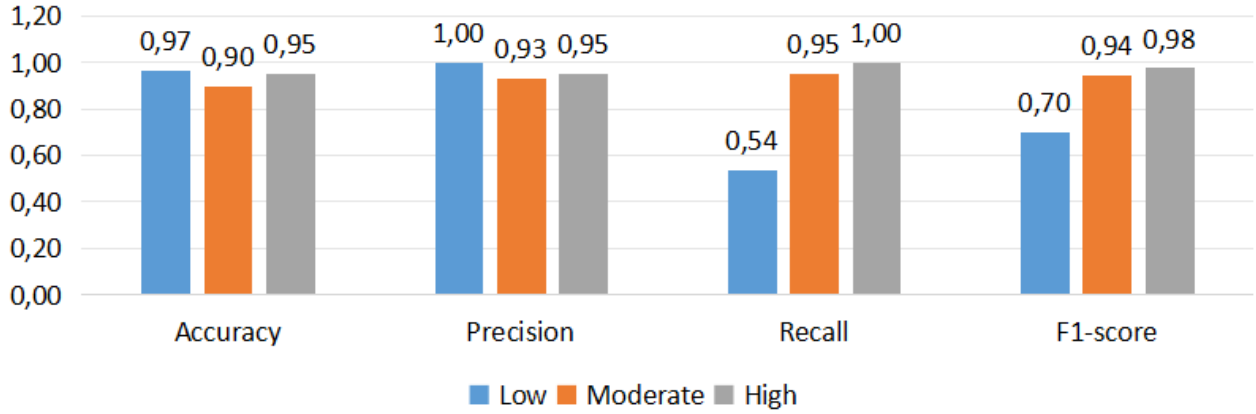

Figure 3. Results of classification efficiency using CNC and threshold values.
Source: created by the authors.

For the low complexity level, a very high Accuracy of 0.97 and ideal Precision of 1.00 were achieved, but Recall of 0.54 indicates that almost half of the models at this level were not detected, resulting in a moderate generalized F1-score of 0.70.

For a moderate level of complexity, all indicators are consistently high, in particular Accuracy of 0.90, Precision of 0.93, Recall of 0.95, and F1-score of 0.94, which indicates a balanced and reliable classification.

For a high level of complexity, Accuracy of 0.95 was obtained with Precision of 0.95 and maximum Recall of 1.00, which means complete detection of complex patterns, and the F1-score of 0.98 confirms the high overall classification efficiency of this level.

Conclusions. The following main results were obtained in this study:

1. A formalized approach to evaluating the quality of BPMN models based on their structural complexity is proposed, combining graph representation of models, TNN and TNSF size metrics, and the CNC network connectivity coefficient with linear regression and threshold complexity levels.
2. The developed regression model for CNC forecasting, trained on 6137 business process descriptions, showed adequate accuracy with a MSE of 0.02 and a

coefficient of determination of 0.71, confirming its ability to explain the variation in model complexity.

3. The experimental classification of BPMN models by complexity levels demonstrated high efficiency for moderate and high levels with F1-score of 0.94 and F1-score of 0.98, respectively, with Recall of 0.95 for moderate and Recall of 1.00 for high respectively.

4. For a low level of complexity, high Accuracy (of 0.97) and Precision (of 1.00) were achieved, but the lower Recall value (of 0.54) revealed the limitations of the formal approach and justified the need for further expansion of the set of evaluation features.

References:

1. Moreno-Montes de Oca, I., Snoeck, M., Reijers, H. A., & Rodríguez-Morffi, A. (2015). A systematic literature review of studies on business process modeling quality. Information and Software Technology, 58, 187–205. https://doi.org/10.1016/j.infsof.2014.07.011

2. Krogstie, J. (2016). Quality of business process models. In Quality in business process modeling (pp. 53–102). Springer. https://doi.org/10.1007/978-3-319-42512-2_2

3. Boomsma, R. D. (2017). An evaluation of thresholds for business process model metrics [Master's thesis, Eindhoven University of Technology]. Research Portal Eindhoven University of Technology. https://research.tue.nl/en/studentTheses/an-evaluation-of-thresholds-for-business-process-model-metrics/

4. Kopp, A., & Orlovskyi, D. (2021). Towards the method and information technology for evaluation of business process model quality. In Communications in Computer and Information Science (Vol. 1487, pp. 93–118). Springer. https://doi.org/10.1007/978-3-030-77592-6_5

5. Sánchez-González, L., García, F., Mendling, J., & Ruiz, F. (2010). Quality assessment of business process models based on thresholds. In Lecture Notes in Computer Science (pp. 78–95). Springer. https://doi.org/10.1007/978-3-642-16934-2_9

6. Sánchez-González, L., García, F., Ruiz, F., & Mendling, J. (2012). Quality indicators for business process models from a gateway complexity perspective. Information and Software Technology, 54(11), 1159–1174. https://doi.org/10.1016/j.infsof.2012.05.001

7. Kbaier, W., & Ghannouchi, S. A. (2019). Determining the threshold values of quality metrics in BPMN process models using data mining techniques. Procedia Computer Science, 164, 113–119. https://doi.org/10.1016/j.procs.2019.12.161

8. Kopp, A., Orlovskyi, D., & Litvinova, U. (2023). Fuzzy logic-based software tool for business process model complexity assessment. In 2023 IEEE 4th KhPI Week on Advanced Technology (KhPIWeek) (pp. 1–6). IEEE. https://doi.org/10.1109/khpiweek61412.2023.10312963

9. El aachab, Y., Kaicer, M., & Jouilil, Y. (2023). Binary Classification with Supervised Machine Learning: A Comparative Analysis. Applied Mathematics & Information Sciences, 17(4), 589–598. https://doi.org/10.18576/amis/170407

10. Araveeporn. A., & Wanitjirattikal, P. (2024). Comparison of Machine Learning Methods for Binary Classification of Multicollinearity Data. 44–49. https://doi.org/10.1145/3686592.3686600

11. Arora, U. (2022). A First Order Iterative Method for Numerical Optimization of Machine Learning Models. Advances and Applications in Mathematical Sciences, 21(6), 3493–3502. https://www.mililink.com/upload/article/314959707aams_vol_216_april_2022_a42_p3493-3502_urvashi_arora.pdf

12. Yanholenko, O., Kopp, A., Godlevskyi, M., & Orlovskyi, D. (2025). Intelligent information technology for business process model quality and error volume assessment. CEUR Workshop Proceedings, 3983. https://ceur-ws.org/Vol-3983/paper14.pdf

13. BPMN for research. https://github.com/camunda/bpmn-for-research

14. Camunda's BPMN 2.0 Exploratory Dataset. https://www.kaggle.com/datasets/andreykopp/camundas-bpmn-2-0-exploratory-dataset

15. Scikit-Learn. https://scikit-learn.org/