In the context of the practical implications of the research artifact, special attention was paid to the evaluation. We conducted quantitative computer experiments and qualitative interviews and discussions in the experimental and evaluation phase consisting of six main steps (see Fig. 1).

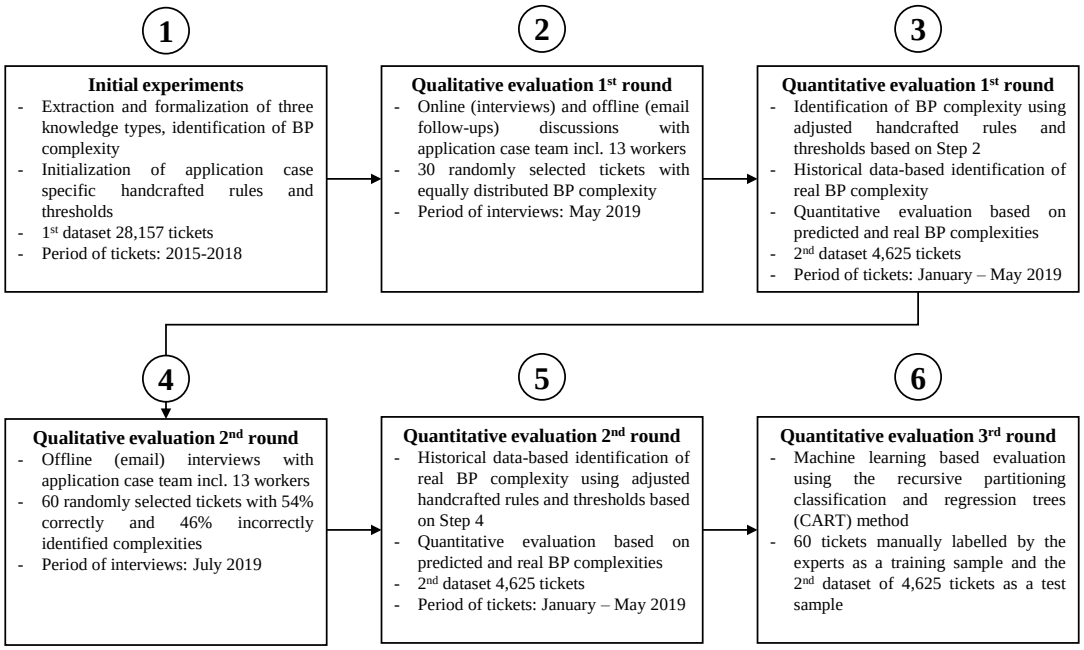

Fig. 1. Evaluation process

In *Step 1*, using the dataset of 28,157 tickets, the initial experiments were carried out to extract and formalize the three knowledge types, set up initial application case specific handcrafted rules and thresholds and, based on these values, identify BP complexity.

In *Step 2*, to evaluate the obtained values, 30 randomly selected tickets with equally distributed predicted *BP complexity* values were presented to the experts, i.e., 13 workers of the application case department. This first qualitative evaluation round, including online discussion (interviews[1]) and offline follow-up sessions, was conducted in May 2019. The interview was divided into three parts. First, we introduced the objectives of the interview, research motivations, theoretical and methodological background. Second, the method of knowledge extraction and BP complexity prediction was illustratively presented using the sample of 30 tickets. Afterward, the experts estimated the IT ticket complexity based on the available historical data regarding IT ticket processing.

The discussion of the discrepancies between *predicted BP complexity* and historical data-based complexity, further referred to as *real BP complexity*, was shifted to offline (email) due to additional information from the IT ticketing system needed to estimate *real BP complexity*. In the scope of the research, *real BP complexity* is exclusively used to evaluate the research artifact. Third, a Q&A session was conducted following a so-called funnel model (Runeson and Höst, 2009), i.e., we started with open questions and moved towards more specific ones regarding possible practical implications of the complexity prediction. Hereby, providing recommendations in the form of templates or historical ticket data, prioritization of an incoming ticket as a dashboard for the correct time and workforce management in the team, and automatic filling in of the ticket complexity field in the IT ticketing system were mentioned as possible use cases of BP complexity concept application. The offline discussion of the BP complexity values yielded the results presented in Table I, point 1. The table gives an overview of qualitative and quantitative evaluation results of BP complexity prediction. Overall precision is the relative number of correctly identified *predicted BP complexity* compared to the whole number of identified *real BP complexity*. Recalls are calculated for each of the three possible *predicted BP complexity* values and represent a fraction of relevant values that have been retrieved over the total amount of relevant values.

Table I. Evaluation results of IT ticket complexity prediction

	low	medium	high
1. Qualitative evaluation of initial experiment results (28,157 tickets) based on 30 tickets – predicted vs. expert complexity			
Recall	69%	55%	67%
Overall precision	63%		
2. Quantitative historical data-based evaluation of follow-up experiments results (4,625 tickets) – handcrafted rules			
Recall	51.2%	28.4%	9.1%
Overall precision	45%		
3. Qualitative evaluation of experiment results (4,625 tickets) based on 60 tickets – real vs. expert complexity			

[1] Please see our evaluation questionnaire used as a preparation for the interviews.

Recall	62%	10%	0%
Overall precision	54%		
4. Quantitative historical data-based evaluation of follow-up experiments results (4,625 tickets) – handcrafted rules			
Recall	73.9%	71.9%	40.7%
Overall precision	61.75%		
5. Quantitative historical data-based evaluation of follow-up experiments results (4,625 tickets) – CART based rules			
Recall	75.6%	61.6%	50.2%
Overall precision	62.27%		

In the discussions, we obtained the following findings for qualitative improvements: (1) enrichment of the DML taxonomy and BS lexicon with the one- and bi-grams indicating simple vs. complex problem solving, (2) development of the handcrafted rules and thresholds, and (3) identification of necessary historical ticket data allowing to calculate the *real BP complexity*. Hence, we amended the mentioned vocabularies with such one- and bigrams as "(no, not) affected", "(no) PSO" (Projected Service Outage), "(no) impact", "(no, short, zero) downtime", "test", "(no, not) production", "(no, not) prod". We also added German equivalents of such adverbs as "no", "not", i.e. "kein(e)", "nicht", for the case of English-German ticket texts. Next, the following handcrafted rules and historical data were selected to identify the real complexity: (i) the presence of the mentioned one- and bi-grams in the IT ticketing system fields "Impact description" and "Brief description" of the ticket (RegEx (Prasse *et al.*, 2015) based free text search), (ii) number of tasks per ticket (count of tasks, integer data type), (iii) number of configuration items, specifically applications involved in the ticket (count of applications, integer data type), (iv) risk type of ticket (enumeration, ordinal scale of "low", "medium", "high").

In *Step 3*, we obtained the second dataset of 4,625 tickets with the historical data necessary to identify *real BP complexity*, as discussed with the experts in Step 2. The obtained historical ticket data were used to identify *real BP complexity*. These first quantitative evaluation results were not satisfying, revealing the overall precision of approximately 45% with the following recalls of *predicted BP complexity* – low 51.2%, medium 28.4%, and high only 9.1% (see Table I, point 2).

Therefore, in *Step 4*, we conducted the second qualitative evaluation round in the form of an interview in an offline (email) mode in July 2019. For this purpose, 60 randomly selected tickets with *predicted and real BP complexity values* were presented to the experts. The sample contained 54% correctly and 46% incorrectly identified complexities with the random structure of low, medium, and high values. The goal was to adjust the rules and thresholds for the identification of *real BP complexity* based on the historical data. In the offline discussions, the cases of discrepancies between identified *real BP complexity* and the one assigned by the experts were reviewed in detail (for the evaluation results, see Table I, point 3).

Finally, in *Step 5*, we conducted the second quantitative evaluation round. Using adjusted rules regarding the keywords and thresholds for the historical ticket data, such as number of applications and tasks, we achieved an improvement resulting in a better prediction (see Table I, point 4).

Additionally, in *Step 6*, to compare the evaluation results, we applied a machine learning (ML) based approach, i.e., the recursive partitioning classification and regression trees (CART) method (Podgorelec *et al.*, 2002) with complexity parameter cp=0.056 and measures of the error in classification xerror=0.39. For this purpose, we used the mentioned set of 60 tickets manually evaluated by the experts as a training sample and a dataset of 4,625 tickets as a test sample. The results can be seen in Table I, point 5. Comparing the evaluation results of points 4 and 5 in Table I, we observe relatively consistent results and can conclude that the performance of our method is acceptable. Looking into ML based ticket classification approaches in the literature, sophisticated ML classification pipelines report accuracy in a rather broad range from 30% to 90% (Banerjee *et al.*, 2012; Mandal *et al.*, 2019).

The dataset structure obtained at the end of the experiments and evaluation is presented in Table II. Considering both datasets, we could identify some clear trends. Hence, in the DML distribution, the predominant values are *routine* and *semi-cognitive*, with only a few *cognitive* values. This trend follows a general understanding and expectation of the distribution of daily tasks. In the BS distribution, there is an evident discrepancy between the two datasets. In the first case, the prevalent BS is *medium* (68.5%). Generally, CHM workers tended to use the BS intensifiers (capitalizations, special characters, punctuation) to highlight certain text parts since the IT ticket processing software did not support standard text highlighting functions like bold or cursive letters, underlining, colours. Thus, we observe most tickets of *medium* BS in the first dataset. In the second dataset, the majority of tickets evidence *low* BS (63.2%). Such a discrepancy can be explained by the different sizes of the datasets and their imbalance. The *high* BS is distributed almost equally in both datasets.

The distribution values of Readability demonstrate a trend similar to that of DML. The most common values are *effortless* and *involving effort*, with relatively few *telegraphic* values. The most frequent value in the first dataset is *effortless*, and in the second – *involving effort*. The IT ticket complexity values of both datasets demonstrate comparable distributions, i.e., prevailing *low* complexity tickets followed by *medium* and, afterward, *high*.

Table II. Distribution statistics of DML, BS, Readability, and IT ticket complexity

DML (objective knowledge)	BS (subjective knowledge)	Readability (meta-knowledge)	IT ticket complexity
1) Dataset of 28,157 tickets			

routine – 60% semi-cognitive – 39% cognitive – 1%	low – 8.7% medium – 68.5% high – 22.8%	effortless – 53.1% involving effort – 43.6% telegraphic – 3.3%	low – 56.3% medium – 26.8% high – 16.9%
2) Dataset of 4,625 tickets			
routine – 48.6% semi-cognitive – 49.5% cognitive – 1.9%	low – 63.2% medium – 11.3% high – 25.5%	effortless – 35.5% involving effort – 52.8% telegraphic – 11.7%	low – 52.4% medium – 31.7% high – 15.9%