Wrocław University of Science and Technology

Improving Quality and Minimizing Costs of Software Development of 5G Systems at Nokia Using Machine Learning

SOFTWARE ENGINEERING SECTION

of the Committee on Informatics of the Polish Academy of Sciences

Szymon Stradowski

November 2025

Wroclaw University of Science and Technology

Purpose of the presentation

Presenting the:

- dissertation content
- discussion on results
- contributions and thesis summary

Agenda

- Introduction
  - research field
  - research context
  - research approach/methods
- Publication summary
  - for each article [ART1 – 9]
- Implementation and generalization
- Summary
- Q&A
- Backup material
  - references, Nokia test process, 5G specifics, awards and recognitions

Wroclaw University of Science and Technology

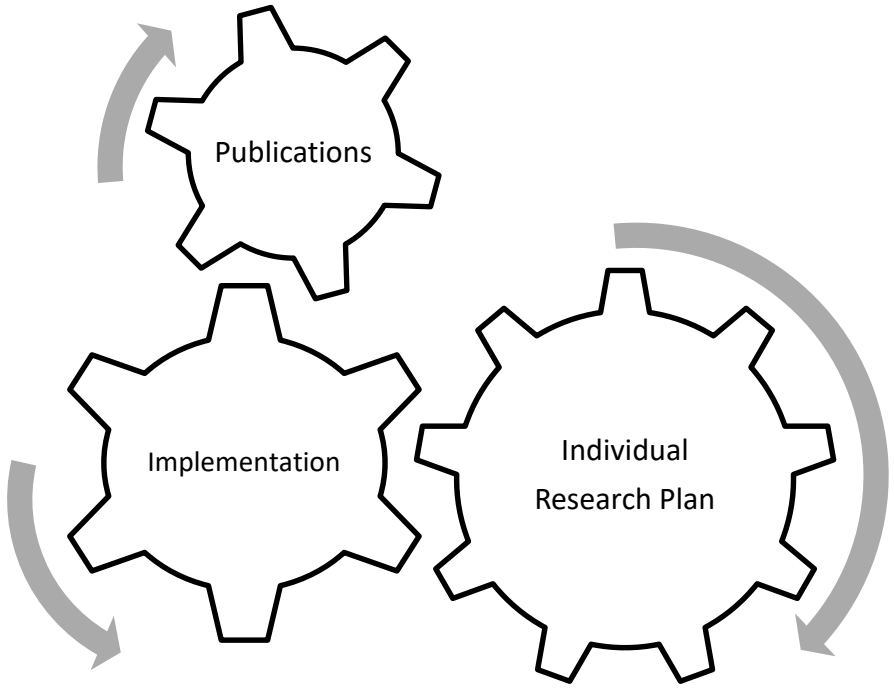
Students profile

- MSc Teleinformatics, PWR (2010)
- MBA, PMP, CBAP, L6S BB etc.
- Since 2011 working in NOKIA
  - Project Office, MN RAN RL1
  - Nokia's focus is mainly 5G
  - 5G is grand scale and complexity [41]

- PhD start: October 2021
- Faculty: Information and Communication Technology (W04)
- Department: Applied Informatics (K45)
- Supervisor:
  - Lech Madeyski, PhD, DSc
- Auxiliary supervisor:
  - Markku Räsänen, NOKIA
- PhD submission: September 2025

Implementation PhD

- Set in a business context (divergence between priorities in practice and science [43])
- Goal is to build synergy between practice and science (win-win-win)
- Achieved by a coherent portfolio of publications (Individual Research Plan -> Implementation -> Publications)



Wroclaw University of Science and Technology

Research field – Machine Learning Software Defect Prediction

Machine Learning (ML) is a branch of artificial intelligence (AI) and computer science which employs algorithms (learners) to imitate the way that humans learn from data, gradually improving predictive performance [38]. Characteristics:

- ability to analyze huge amounts of data,
- enables domain expertise and knowledge discovery,
- vastly popular in research with many solutions and frameworks available,
- industry acceptance is increasing [38],
- hindered by “No free lunch” theorems [39].

Software Defect Prediction (SDP) is one of the supporting activities of the quality assurance (QA) process [14, 15]. The goal is to predict the SW modules that are defect prone and require extensive testing based on various data inputs. Main difficulties:

- low industry uptake [2, 3, 26],
- many open issues in Software Defect Prediction [27, 42],
- robustness, interpretability, costing, scaling, false positives, and more...

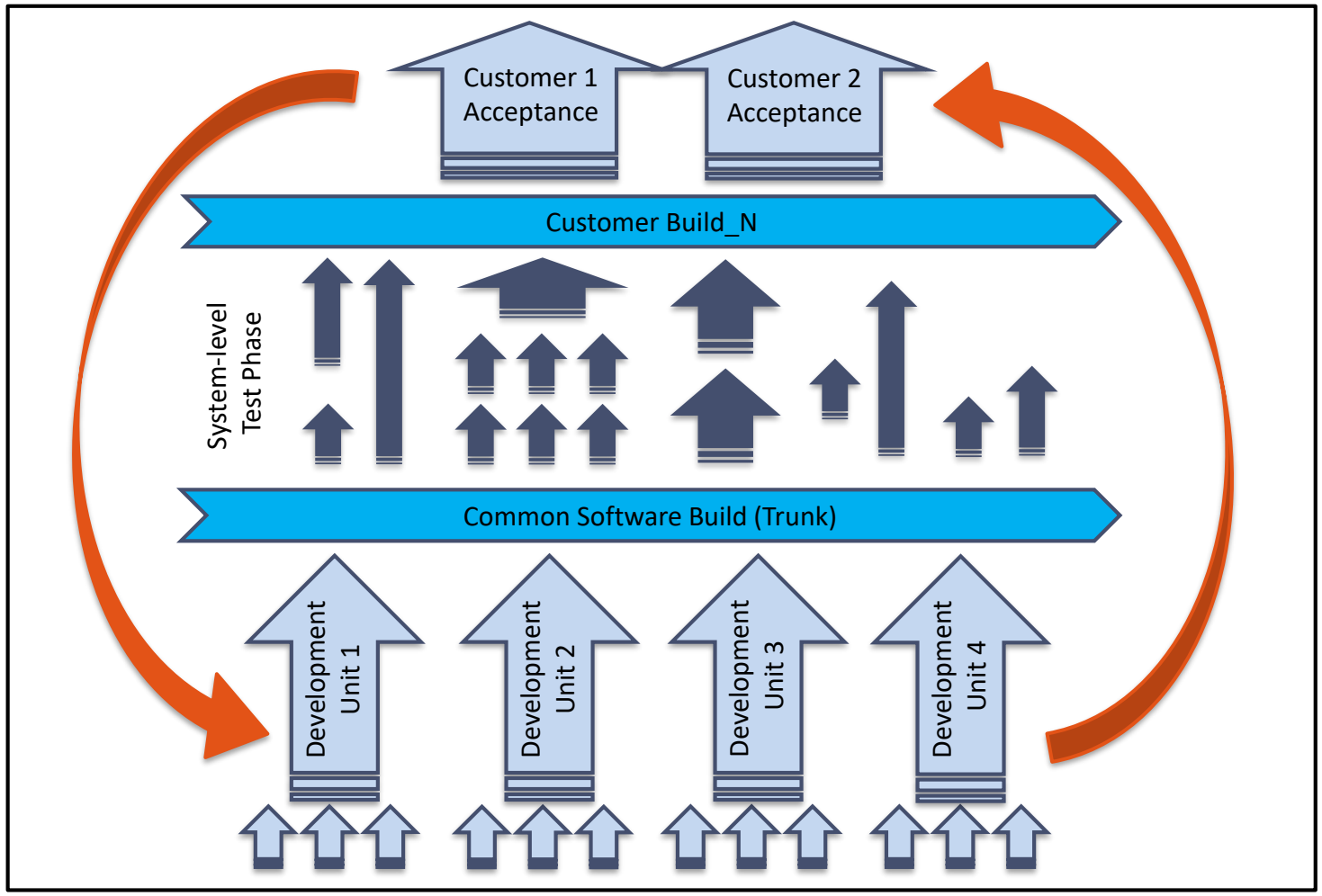
Our approach: high-level test data, not directly code-dependent, novel perspective, lightweight, validated in the real-world and industry-oriented, adhering to the state-of-the-art.

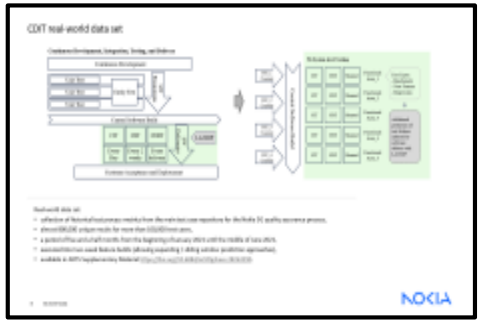
Wroclaw University of Science and Technology

Research context – NOKIA 5G

NOKIA 5G Software Development Life Cycle:

- environment: Continuous Development, Integration, and Testing (CDIT),
- goal: add ML SDP to black-box system-level testing,
- outcome: improve quality and decrease costs with ML SDP.





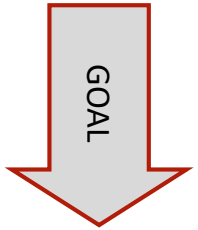
Wroclaw University of Science and Technology

Research approach

Thesis:

ML SDP can be adapted to complement the existing quality assurance processes in system-level testing in Nokia 5G to improve quality and decrease costs, with modeling predictions that enable human understanding.

- Streamlining software defect finding and improving current processes by increasing quality and minimizing the cost.
- Being understandable to humans or provide opportunities to explain proposed decisions using the eXplainable Artificial Intelligence approach.

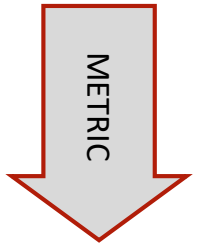


Adapt an ML SDP solution to complement the existing quality assurance processes in system-level testing in Nokia 5G to improve quality and decrease costs, with modeling predictions that enable human understanding.

Can an ML-based solution complement the system-level testing of the Nokia 5G product to streamline software defect finding?

Can XAI be used to meaningfully interpret ML SDP models for Nokia 5G system-level testing?

Can ML SDP be cost-effective when used as an additional quality assurance process within Nokia 5G system-level testing?



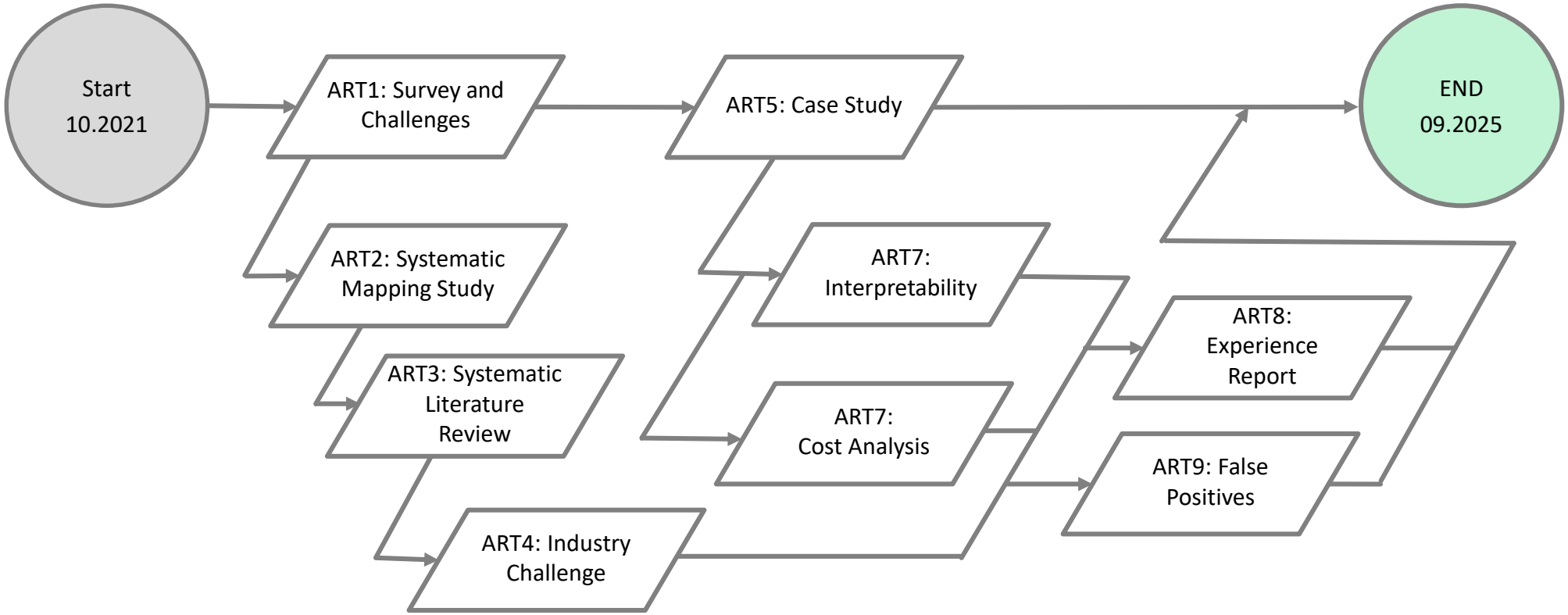
Quantitative and qualitative analysis of the implemented ML SDP solution in terms of:

- predictive performance (*Mathews Correlation Coefficient and Precision*),
- interpretability (*new domain knowledge discovery reflected by the number of designed improvement actions based on the Feature Importance analysis*),
- profitability (*Return On Investment and Benefit-Cost Ratio*).

Wroclaw University of Science and Technology

Research plan

- Nine separate but coherent publications [ART1-9].
- Many characteristics and practices of action [23] and case study research [24].
- Timeline and chronology followed a natural flow of events.
- Rigorous peer review according to the respective publishers' requirements for top scientific journals and conferences in the field of software engineering.
- Each article offers a unique set of contributions to science and practice, as well as follows the reproductible research principles [25].



Wroclaw University of Science and Technology

ART 1 - Exploring the Challenges in Software Testing of the 5G System at Nokia: A Survey

Journal: Information and Software Technology (Elsevier, IF 3.9, 140 pts)
Submission: February 2022, major revision: May 2022, acceptance: September 2022
DOI: 10.1016/j.infsof.2022.107128
Supplementary material: https://doi.org/10.5281/zenodo.6945430

Contributions

- Description of MoW of Nokia 5G and definition of the main challenges in system-level testing.
- 17 predetermined challenges under evaluation criteria
  importance / urgency / difficulty.
- 1 open question (what is missing?).
- 2 demographic questions: role & experience.

Methodology

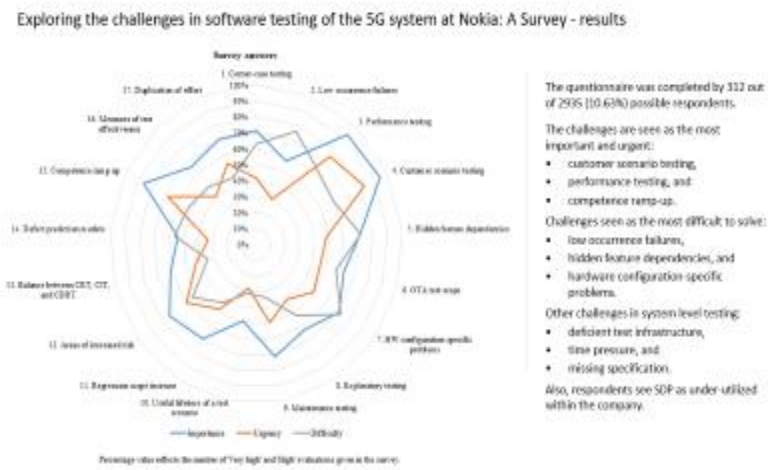
- Guidelines for empirical study of software engineering challenges set in a real business context [28-30].
- Goal Question Metric approach [22].
- MS Forms tool using five-point Likert scale + “IDK”.
- Pre-survey results analysis.
- Post-survey results analysis.

Results

- 312 responses out of 2935 invited (10.63%) with respondents from 8 countries.
- 127 open answers.
- Challenges that are most significant in terms of:
  importance / urgency / difficulty.
- Analysis of the implications for industry.
- Analysis of the implications for science.
- Demographics overview.
- Generalizability discussion.

Discussion

- Spider chart visualizing perceived importance and urgency (industry), or difficulty (academia).



Wroclaw University of Science and Technology

ART 2 - Machine Learning in Software Defect Prediction: A Business-Driven Systematic Mapping Study

Journal: Information and Software Technology (Elsevier, IF 3.9, 140 pts)
Submission: May 2022, major revision: June 2022, acceptance: November 2022
DOI: 10.1016/j.infsof.2022.107128
Supplementary material: https://doi.org/10.5281/zenodo.7375768

Contributions

- ML SDP in a big-picture overview.
- Keyword analysis on Scopus database.
- 1222 papers found -> 742 papers analyzed.

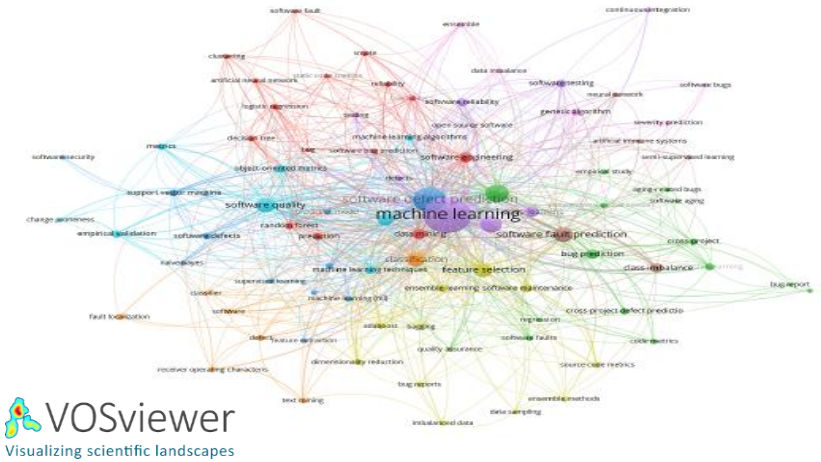
Results

We build several maps of keywords for analysis, researcher cooperation, trends in time analysis.

- Number of publications is increasing YoY.
- Number of keywords is increasing YoY.
- Learners: Decision Trees prevail in published research.
- Datasets: NASA, PROMISE, open-source vs. industry.
- Only 32 publications in vivo (!)
- Emerging trends: just-in-time, cross-project, deep learning, XAI.
  - Pre-2000 → Foundations.
  - 2000–2010 → First ML defect prediction & empirical methods
  - 2010–2015 → Methodological maturity & DL emergence
  - 2016–2020 → Optimization, XAI, JIT & industry readiness
  - 2021–2025 → Industrial deployments & building trust

Methodology

- PRISMA 2000 standard for secondary studies [31].
- VOSviewer tool for visualization.



Implications

- Industry papers are scarce despite growing interest in ML SDP and require much more effort before full-scale industry deplyment.

Wroclaw University of Science and Technology

ART 3 - Industrial Applications of Machine Learning Software Defect Prediction: Literature Review

Journal: Information and Software Technology (Elsevier, IF 3.9, 140 pts)
Submission: June 2022, major revision: January 2023, acceptance: March 2023
DOI: 10.1016/j.infsof.2023.107192
Supplementary material: https://doi.org/10.5281/zenodo.7476403

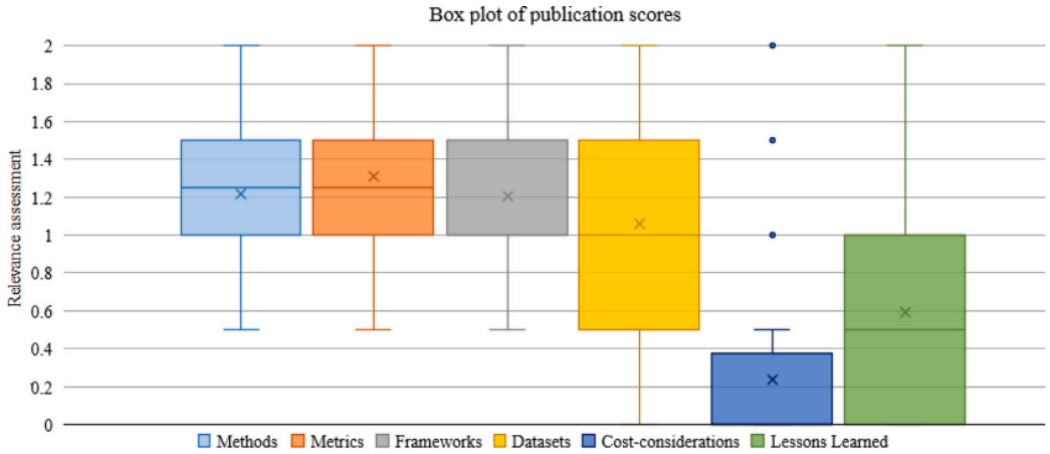
Contributions

- Follow-up to the Systematic Mapping Study [ART2].
- 6 online databases in scope.
- 397 papers identified -> 32 papers analyzed.
- A list of selected papers focused on the industry application of ML SDP.
- A synthesis of the current state-of-the-art, describing the details of successful industry applications.

Methodology

- SEGRESS standard for secondary studies [32].
- Quasi-Gold Standard (QGS).
- Sensitivity = 68%, Precision = 4%.
- Quality of evidence evaluation for:
  methods, metrics, frameworks, data sets, cost considerations, and lessons learned.

Results



Discussion

- Number of real-world publications is low.
- “No free lunch” theorem strongly visible.
- Only 2 papers on cost considerations and a handful of experience reports.
- Further effort on bridging the gap is needed (critical in lessons learned and cost-benefit analyses).

Wroclaw University of Science and Technology

ART 4 - Can we Knapsack Software Defect Prediction? Nokia 5G Case

Conference: 45th International Conference on Software Engineering (IEEE/ACM, CORE A*, 200 pts)
Submission: February 2023, acceptance: April 2023, presentation: May 2023
Location: Melbourne, Australia
DOI: 10.1109/ICSE-Companion58688.2023.00104

Contributions

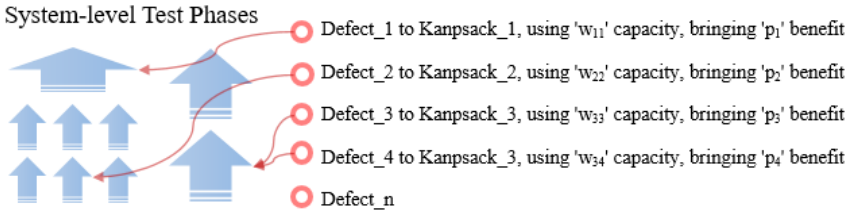
- NOKIA CDIT context description.
- Challenge of scaling ML SDP to multi-level process.
- Definition of ML SDP as Multidimensional Knapsack Problem (MKP) [44].

Problem definition

Industry challenge: As software products become larger and more complex, the test infrastructure costs needed for quality assurance grow similarly. However, most ML SDP solutions address only singular test phases rather than the overall agile Software Development Life Cycle (SDLC) [1, 40].

Hypothesis:
Can ML-based SDP successfully complement test case assignment to different test phases and provide sufficient explanation on made decisions?

Solution proposal



(MKP) maximize:

$$z = \sum_{j=1}^{n} p_j x_j \qquad (1)$$

Subject to:

$$\sum_{j=1}^{n} w_{ij} x_j \leqslant c_i, \quad i = 1, 2, \ldots\ldots, m \qquad (2)$$

$$x_j \in \{0, 1\}, \quad j = 1, 2, \ldots\ldots, n \qquad (3)$$

Discussion

- Preconditions for the solution.
- Benefits and challenges.
- Potential user stories.
- Next steps: How to create data sets that allow further research?

Wroclaw University of Science and Technology

ART 5 - Predicting Test Failures Induced by Software Defects: A Lightweight Alternative to ML SDP (1)

Journal: Journal of Systems and Software (Elsevier, IF 3.3, 100 pts)
Submission: April 2024, major revision: December 2024, acceptance: February 2025
DOI: https://doi.org/10.1016/j.jss.2025.112360
Supplementary material: https://doi.org/10.6084/m9.figshare.28263290

Contributions

- We proposed and developed a Lightweight Alternative to SDP (LA2SDP) that predicts test failures induced by software defects to allow pinpointing defective software modules.
- Evaluation of the proposal in a real-world Nokia 5G scenario.
- Four different iterations of research with growing/better content, and research effort lasting over two years.

Methodology

- R → MLR3 framework [18] + DALEX for interpretability [19].
- Five supervised machine learning algorithms with their tuned versions.
- The Matthews Correlation Coefficient (MCC) for performance evaluation [35], with precision as secondary metric.
- Expanding and sliding window time-based approach.

Results

The main implications for our case study:

- CatBoost with consistently high MCC and precision across multiple tasks.

Honorable mentions:

- Random Forest with exceptional precision and with acceptable MCC, but quite unstable,
- Tuned Naïve Bayes with highest MCC performance on the last task, but low precision.

In conclusion, even relatively simple learners and existing data base offer satisfactory results (MCC>0.3).

Discussion

- LA2DP is feasible in vivo using limited data readily available within the Nokia 5G system-level test process CDIT.
- Widely available learners and existing metrics offer satisfactory results with imposed expectations (lightweight (to build initial inroads), using existing data, enabling interpretability).
- The most important features are related to the week of execution, test instance, and responsible organization.
- Data sets, code, and results published for reproducible research.

Wroclaw University of Science and Technology

ART 5 - Predicting Test Failures Induced by Software Defects: A Lightweight Alternative to ML SDP (2)

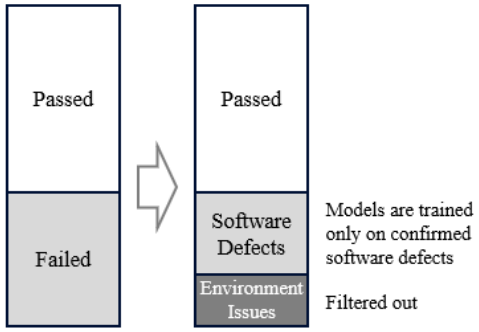
General approach

Final version of the piloted solution:

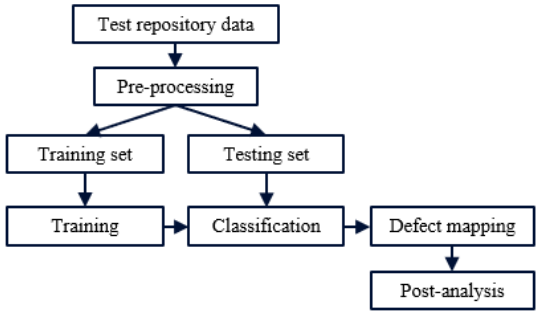
GOAL: A Software defect prediction solution that can work with system-level test process data and complement other defect prediction and test case selection and prioritization mechanisms in the company.

INPUT: Readily available system-level test repository data for NOKIA 5G.

ML SDP framework to predict test failures only related to software defects (without test and environmental issues). Using existing historical test failure data to software defect mapping.



1. We use the existing test repository data base.
2. Filter out irrelevant test failures.
3. Use the proposed ML SDP framework to predict test failures only related to software defects (without test and environment issues).
4. Analyze the builds between test executions for software changes and mapping predicted test failures to past defects and fixed modules.



OUTPUT: Prediction of TEST RUNs that are likely to fail (due to software defects) that allows pinpointing faulty software modules without expensive test re-execution using the using the mapping of predicted test failures to past defects and corrected modules.

Supplement material on Figshare:

- R/LA2SDP.R - code in R using mlr3 R package,
- sessionInfo.txt - version information about R, used packages, and OS,
- renv.lock - the lockfile that records enough metadata about every package so that the computational environment can be re-installed on a new machine.

Results overview

Five classic learners and tuned versions (using Hyperband with 10-fold CV and MCC optimization, under a time-bounded tuning budget):

- Classification Tree (ct), with tuned complexity parameter for a CART classifier.
- Light Gradient-Boosting Machine (lgbm), with tuned key hyperparameters of a LightGBM dart classifier (iterations, learning rate, min data in leaf, num leaves).
- CatBoost Gradient Boosting (cb), with tuned number of iterations and tree depth.
- Random Forest (rf), with tuned number of trees.
- Naïve Bayes (nb), tuned with sample imputation and mode imputation for missing data and Laplace smoothing using random search.

Task	Model	MCC	ACC	Recall	Prec.	Fbeta	AUC	TP	TN	FP	FN
2_5	ct	0.179	0.996	0.126	0.260	0.169	0.569	60	139212	171	417
2_5	ct.tuned	0.158	0.995	0.136	0.190	0.159	0.600	65	139105	278	412
2_5	lgbm	0.092	0.996	0.029	0.292	0.053	0.942	14	139349	34	463
2_5	lgbm.tuned	**0.339**	0.997	0.130	0.886	0.227	0.930	62	139375	8	415
2_5	cb	**0.342**	0.997	0.145	0.812	0.246	0.955	69	139367	16	408
2_5	cb.tuned	**0.328**	0.997	0.117	0.918	0.208	0.928	56	139378	5	421
2_5	rf	**0.330**	0.997	0.111	**0.981**	0.200	0.949	53	139382	1	424
2_5	rf.tuned	**0.336**	0.997	0.117	0.966	0.209	0.792	56	139381	2	421
2_5	nb	0.127	0.942	0.568	0.033	0.063	0.834	271	131510	7873	206
2_5	nb.tuned	0.189	0.958	0.700	0.055	0.103	0.934	334	133696	5687	143
3_6	ct	0.087	0.956	0.021	0.429	0.040	0.510	60	61596	80	2787
3_6	ct.tuned	0.086	0.955	0.021	0.415	0.041	0.511	61	61590	86	2786
3_6	lgbm	0.116	0.956	0.029	0.529	0.055	0.734	82	61603	73	2765
3_6	lgbm.tuned	0.115	0.956	0.028	0.529	0.054	0.701	81	61604	72	2766
3_6	cb	0.105	0.956	0.016	**0.742**	0.032	0.800	46	61660	16	2801
3_6	cb.tuned	0.055	0.956	0.006	0.593	0.011	0.810	16	61665	11	2831
3_6	rf	-0.001	0.956	0.000	0.000	0.000	0.762	0	61675	1	2847
3_6	rf.tuned	0.053	0.956	0.005	0.636	0.010	0.701	14	61668	8	2833
3_6	nb	0.259	0.916	0.389	0.233	0.291	0.728	1108	58025	3651	1739
3_6	nb.tuned	**0.276**	0.896	0.497	0.210	0.296	0.832	1415	56367	5309	1432
2_6	ct	0.088	0.956	0.021	0.440	0.040	0.510	59	61602	75	2788
2_6	ct.tuned	0.087	0.956	0.021	0.434	0.040	0.510	59	61600	77	2788
2_6	lgbm	0.108	0.955	0.037	0.387	0.067	0.719	104	61512	165	2743
2_6	lgbm.tuned	0.165	0.957	0.041	0.715	0.078	0.683	118	61630	47	2729
2_6	cb	0.150	0.957	0.040	0.620	0.075	0.812	114	61607	70	2733
2_6	cb.tuned	0.139	0.957	0.026	0.802	0.050	0.802	73	61659	18	2774
2_6	rf	0.105	0.956	0.012	**0.971**	0.024	0.761	34	61676	1	2813
2_6	rf.tuned	0.102	0.956	0.012	**0.919**	0.024	0.708	34	61674	3	2813
2_6	nb	0.249	0.913	0.386	0.222	0.282	0.712	1100	57812	3865	1747
2_6	nb.tuned	**0.295**	0.904	0.499	0.230	0.315	0.831	1421	56924	4753	1426

ART 5 - Predicting Test Failures Induced by Software Defects: A Lightweight Alternative to ML SDP (3)

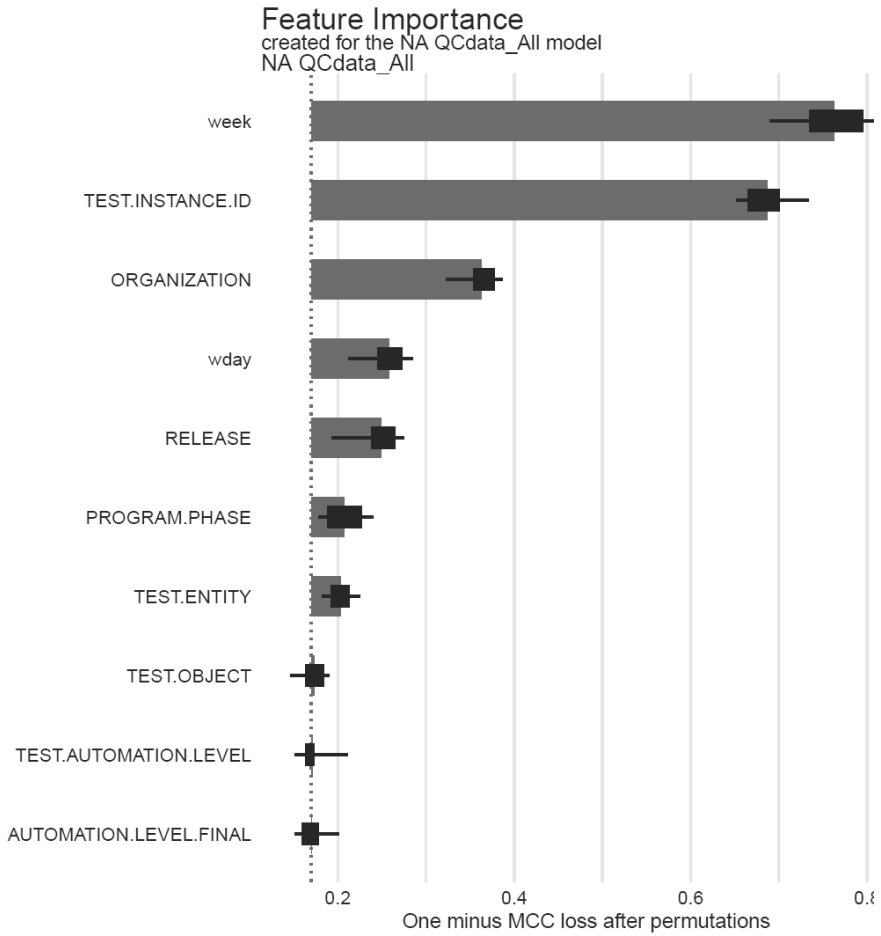
General results

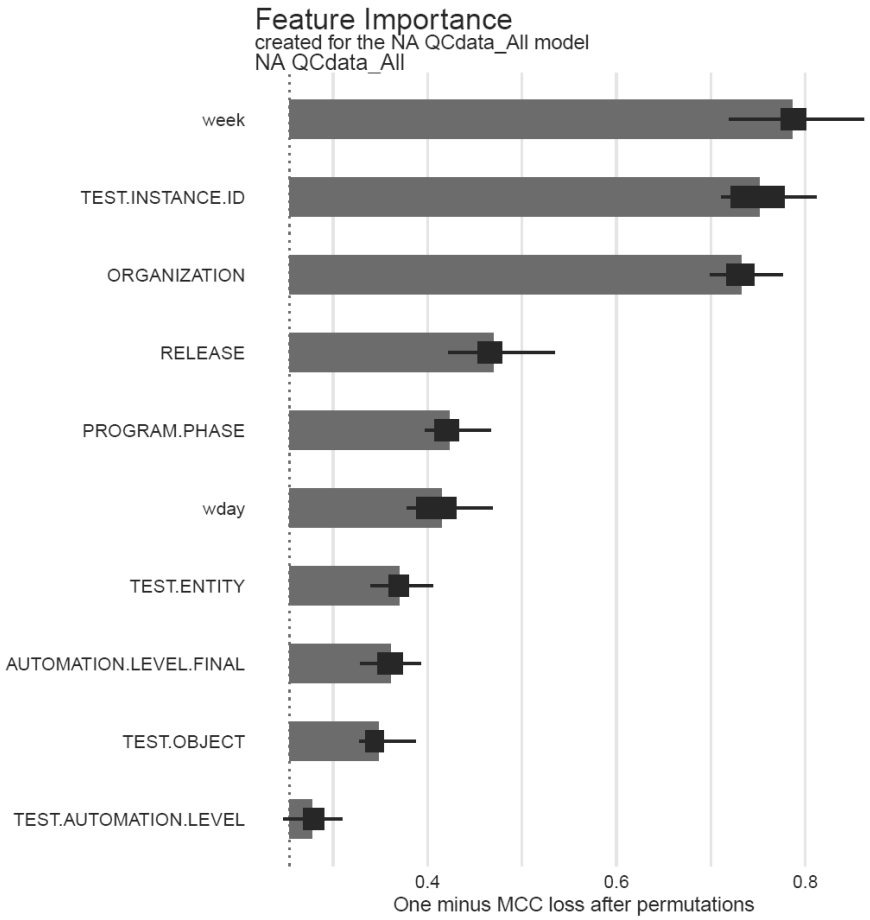
We illustrate the use of the permutation-based variable-importance evaluation by applying it to the CatBoost Gradient Boosting and the Random Forest models on the data set that is composed of the longest window — Task1_6.

The most important features are related to the: WEEK, TEST.INSTANCE.ID, and ORGANISATION for both best models.

Furthermore, the discoveries brought new domain knowledge and process improvement opportunities to the organization described in a dedicated paper [ART6].

Graphical representation of Feature Importance





Wroclaw University of Science and Technology

ART 6 - Interpretability/Explainability applied to Software Defect Prediction: Industrial Perspective

Journal: IEEE Software (IEEE, IF 3.3, 100 pts)
Submission: May 2024, major revision: August 2024, accepted: November 2024
DOI: 10.1109/MS.2024.3505544

Contributions

- Based on the results of our underlying study [ART5].
- Expectations:
  - post-hoc and model agnostic, not impacting performance,
  - facilitating knowledge discovery and domain expertise,
  - actionable results,
  - supporting the stakeholder management process,
  - having positive business impact.

Practitioners Perspectives

- Focus Groups on practitioners' expectations.
- New technology readiness can be viewed as a summarized impact of four personality dimensions: optimism, innovativeness, discomfort, and insecurity [37].
- Each group has different expectations and needs.
- Stakeholder management enables establishing and maintaining effective working relationships (BAbok [13], PMbok [14]).

Stakeholder Management

- Five groups of stakeholders for ML SDP XAI in Nokia and provided an excerpt from our management strategy matrix.

PROFILE	DESCRIPTION	STRATEGY	ACTIONS	XAI IMPACT
Group 1	Sponsors and decision-makers, with high-influence on the project	Manage closely, regular reporting	Gather expectations, measure engagement, report regularly, communicate the timelines and achievements	Provide very precise information on gained knowledge, high-level evidence of XAI business value, impact on business metrics, potential regulatory and societal requirements.
Group 2a	Technical staff (believers & agnostics), using the solution	Engage, frequent communication	Gather feedback and improvement proposals, measure engagement, provide frequent updates, enable room for innovation, offer training, consult	Offer as much output information as possible on all XAI aspects and results.
Group 2b	Technical staff (skeptics), using the solution	Advocate and convince, moderate communication	Gather feedback, measure engagement, provide training, provide evidence of good predictive performance, use precise informative messaging	Offer feature importance analysis, limiting false positives, domain expertise increase
Group 3a	Management staff (believers)	Satisfy expectations,less frequent communication	Gather input, consult on business cases, measure engagement, focus on business impact and effectiveness increase	Provide evidence of predictive performance and efficiency improvements due to XAI, showcase future possibilities, develop quality improvement plans
Group 3b	Management staff (skeptics & agnostics)	Satisfy expectations, less frequent communication	Gather feedback, measure engagement, focus on business impact and effectiveness increase	Provide evidence of added business value, demonstrate initial inroads, use XAI for knowledge discovery, show positive business impact examples and cost-benefit considerations
Group 4	Underlying process owners	Consult, occasional updates	Gather process input, provide project updates and results for inroads	Discover relevant/interesting information, provide evidence of process efficiency improvements
Group 5	Technical and management staff outside of the project	Monitor and keep informed, occasional updates	Open chat for questions, provide infrequent and mass updates on purpose and concept (newsletter, all hands, etc.)	Use XAI as a source of interesting facts and talking points to raise interest in the project

Discussion

- Not-yet-explored stance on the subject of interpretability and expand the understanding of the field from a real-world perspective.
- XAI used to interpret and explain ML SDP models for Nokia 5G system-level testing, and the results obtained are actionable, help achieve a positive business impact and support the stakeholder management process.

Wroclaw University of Science and Technology

ART 7 - Costs and Benefits of Machine Learning Software Defect Prediction: Industrial Case Study

Conference: 32nd ACM Symposium on the Foundations of Software Engineering (ACM, CORE A*, 200 pts)
Submission: February 2024, major revision: April 2024, presentation: July 2024
Location: Porto de Galinhas, Brazil
DOI: 10.1145/3663529.3663831

Contributions

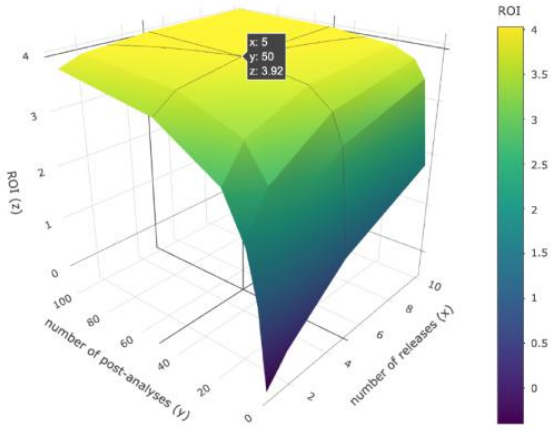
- Case Study on evaluating monetary costs and benefits of ML SDP in vivo (value-based software engineering [20]).
- Real scenario calculations, based on assumptions and estimates provided by Nokia practitioners (lightweight and heavyweight approach).
- Framework for reproduction and building custom scenarios.

Methodology

- Return on investment (ROI) and Benefit-cost ratio (BCR) for lightweight and advanced use cases.
- Profitability calculations conducted based on the general cost model [33].
- Discussion and recommendations on good practices to evaluate the cost-effectiveness of ML SDP in vivo.

Results

- The project life-span, number of releases, and number of post-analyses, and cost of escaped defect affected ROI and BCR more significantly than the predictive performance of ML.



Discussion

- The calculated ROI was between ,0.53 and 3.73 for the lightweight and between -0.71 and 3.51 for the advanced approach.
- Consequently, lightweight software defect prediction is commercially feasible (positive business-case) and can offer a higher return on investment than heavier but more prediction-effective solutions.

Wroclaw University of Science and Technology

ART 8 - Bridging the Gap between Academia and Industry in Machine Learning Defect Prediction

Conference: 38th International Conference on Automated Software Engineering (IEEE/ACM, CORE A*, 200 pts)

Submission: May 2023, acceptance: August 2023, presentation: September 2023

Location: Kirchberg, Luxembourg

DOI: 10.1109/ASE56229.2023.00026

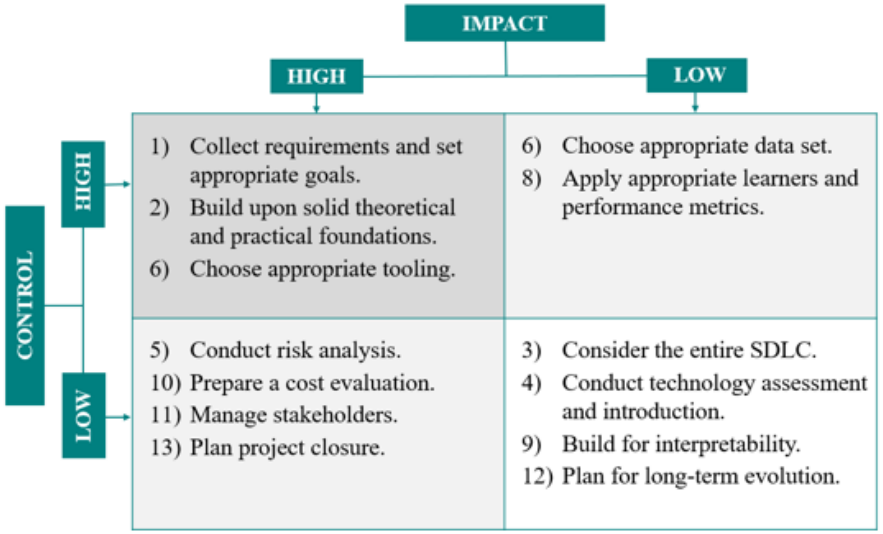
Contributions

- Experience paper → based on most important observations and lessons learned gathered during a large-scale research effort and introduction of ML SDP to the system-level testing in Nokia 5G.
- Thirteen considerations for bridging the gap between industry and academia.

Methodology

- Guided by the global standard of the business analysis body of knowledge (BABOK Guide [13]).
- Feedback from a selected group of Nokia experts and reflect the discussions observed during the planning, execution, and conclusion of the project.
- Control-impact matrix for prioritization.

Considerations



Discussion

- Analysis provides which considerations influence the chances of final success at the lowest amount of time and effort spent.
- High Impact & High Control:
  1) Collect correct requirements and goals,
  2) Build upon solid theoretical and practical foundations,
  6) Choose appropriate data set.
- Note: results are context-specific while the considerations are generic.

Wroclaw University of Science and Technology

ART 9 - "Your AI is impressive, but my code does not have any bugs" Managing false positives

Journal: Science of Computer Programming (Elsevier, IF 1.5, 40 pts)
Submission: January 2025, major revision: March 2025, accepted: May 2025
DOI: 10.1016/j.scico.2025.103320

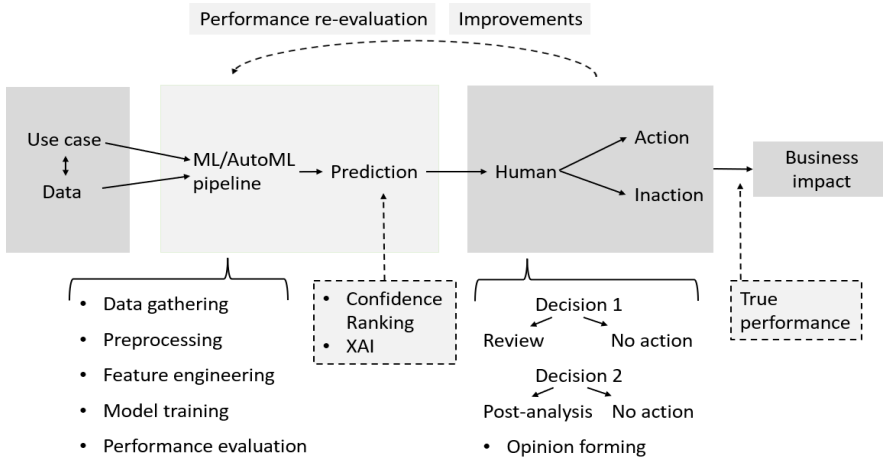
Contributions

- Exploration of the challenges of integrating ML SDP into a larger Quality Assurance process, explicitly from the practitioner's perspective.
- Analysis of false positives' impact and related challenges, as well as the existing mitigation possibilities.
- A real-life motivating example and generalizability discussion.

Discussion

1. The iterative nature of QA necessitates the expansion of the entire ML SDP workflow.
2. Confidence of the particular prediction/classification result with a probability ranking is a viable tool to help reduce false positives, complementary to increasing the predictive performance of the created solutions.

1. Holistic ML SDP workflow



2. Confidence ranking

The designed confidence ranking can lower the number of executed post-analyses based on given risk tolerance.

Algorithm 1. Probability evaluation for ML SDP.

Input: training data: $(x_1, y_1), ..., (x_m, y_m)$
Input: new data $(x'_1, y'_1), ..., (x'_n, y'_n)$
Input (optional): risk tolerance level: Z
1) Learners train on 'training data' & make predictions R for *newdata*
2) Performance is evaluated with MCC metric
3) Local explainability provided with XAI techniques
4) Probability C is evaluated for predictions R, where $c_n = C(r_n)$
Output: prediction set R for *newdata*
Output: probability set C for R; where $c_n, if\ c_n > Z$
Output: prediction set R' sorted by C

Summary

THESIS:
ML SDP can be adapted to complement the existing quality assurance processes in system-level testing in Nokia 5G to improve quality and decrease costs, with modeling predictions that enable human understanding.

SUMMARY:
We have designed, implemented, and validated a lightweight approach to software defect prediction (LA2SDP) to build ML SDP models for Nokia 5G system-level testing. Specifically, our results demonstrate that the solution complements the existing quality assurance processes in a way that can decrease costs and increase product quality. The evidence resulting from the implemented ML SDP adaptation in terms of predictive performance, as well as interpretability and profitability, has led to a recommendation to the company to commercialize a similar solution in the future.

MRQ1:
Can an ML-based solution complement the system-level testing of the Nokia 5G product to streamline software defect finding?

Answer:
Machine learning software defect prediction can be successfully applied to the system-level testing of Nokia 5G, as our LA2SDP solution achieved the target performance of MCC>0.3. Furthermore, we have implemented time-based splits with expanding and sliding window approaches to enable analyzing sustainability over time.

MRQ2:
Can XAI be used to meaningfully interpret ML SDP models for Nokia 5G system-level testing?

Answer:
We successfully used XAI to interpret and explain ML SDP models for Nokia 5G system-level testing, and the obtained results are actionable as well as help achieve a positive business impact. Furthermore, the discoveries brought new domain knowledge and process improvement opportunities, as well as supported stakeholder management efforts.

MRQ3:
Can ML SDP be cost-effective when used as an additional quality assurance process within Nokia 5G system-level testing?

Answer:
ML SDP can be cost-effective in complementing Nokia's existing 5G system-level test process. The calculated ROI values were between 0.53 and 3.73 for the lightweight and -0.71 and 3.51 for the advanced approach.

Wroclaw University of Science and Technology

Summary

The presented work has provided the following contributions:

1

A design of an original solution to a scientific problem and application of the results of the conducted research in the industrial context of Nokia 5G system-level testing.

2

A novel approach to machine learning software defect prediction (LA2SDP) by adaptation based on data from a test repository (test failures induced by confirmed software defects) as an alternative or addition to more established approaches based on software or change metrics.

3

Insight into an industry perspective on adopting a new technology and bridging the gap between industry and academia, with the commercial value has been confirmed as the approval for a large-scale implementation in Nokia is progressing.

4

Thesis confirmation and answers to three main research questions (MRQs) validating the solution in vivo, as well as several supporting research questions (RQs) expanding the studied subject. Furthermore, nine papers describing our efforts were published in top Software Engineering journals and conferences.

Generalizability

Software engineering is a field where many problems arise and are solved in context and only after researchers identify commonalities and differences, adapt solutions to different situations, and generalize over time by building a body of knowledge from gained experience [36].

Moreover, throughout the project, we used a systematic and sequenced approach to maximize the applicability of our findings to other commercial large-scale software projects.

Preconditions for ML2SDP solution implementation:

- Availability of historical test data in sufficient quantity and quality to enable meaningful predictions.
- Tracking of failed test results to reasons including confirmed defect reports, which in turn are tracked to software modules.
- Possibility to analyze, interpret, and act upon the predictions to execute additional intervention and post-analysis.
- Technological and organizational readiness to implement AI-based solutions on a wide scale (for industrial contexts).

Additionally:

- We applied formal methods to support external validity, which concerns the extent to which it is possible to generalize our research findings to other contexts.
- A detailed discussion of the threats to validity is provided for each research article [ART1-9].

Finally, we used industry-accepted standards for enabling wider commercial adoption:

- Project Management – PMBOK [13],
- Software Testing – ISTQB [14, 15],
- Business Analysis – BABOK [49],
- Process Improvement – Lean6Sigma [50].

Wroclaw
University
of Science
and Technology

Next Steps

Further enhancement driving predictive performance further:

- new ensembles and trending algorithms,
- better sampling techniques, further hyperparameter optimization, and feature selection/extraction.

Longitudinal predictive performance study on how the model's behavior changes when learning on iterative new data in a recurring cadence (in our case, a two-week feature build cycle.

There is also significant business potential to explore the non-final state of test results and transform our approach into a multi-class classification problem.

As the company continues to grow and improve the introduced process, further insights into the operational characteristics of the newly introduced processes will lead to new findings, challenges, and conclusions.

Finally, each of the studies included in the dissertation provides its own suggestions on specific future research directions

Wroclaw University of Science and Technology

References (1/4)

1. S. Stradowski and L. Madeyski, “Exploring the Challenges in Software Testing of the 5G System at Nokia: A Survey,” Information and Software Technology, 2022, 153:107067, https://doi.org/10.1016/j.infsof.2022.107128
2. S. Stradowski and L. Madeyski, “Machine learning in software defect prediction: A Business-Driven Systematic Mapping Study,” Information and Software Technology, 2022, 155:107128, https://doi.org/10.1016/j.infsof.2022.107067
3. S. Stradowski and L. Madeyski, “Industrial applications of software defect prediction using machine learning: A business-driven systematic literature review,” Information and Software Technology, 2023, 159:107192, https://doi.org/10.1016/j.infsof.2023.107192
4. S. Stradowski and L. Madeyski, “Can we Knapsack Software Defect Prediction? Nokia 5G Case,” IEEE/ACM 45th International Conference on Software Engineering (ISCE), 2023, https://doi.org/10.1109/ICSE-Companion58688.2023.00104
5. L. Madeyski and S. Stradowski, “A Lightweight Approach to Software Defect Prediction and its Industrial Application in Nokia 5G System-Level Testing,” , Journal of Systems and Software, 2025, 223:112360, doi: https://doi.org/10.1016/j.jss.2025.112360
6. S. Stradowski and L. Madeyski, “Interpretability/Explainability applied to Machine Learning Software Defect Prediction: An Industrial Perspective,” IEEE Software, 2024, doi: https://doi.org/10.1109/MS.2024.3505544
7. S. Stradowski and L. Madeyski, “Costs and Benefits of Machine Learning Software Defect Prediction: Industrial Case Study,” 32nd ACM International Conference on the Foundations of Software Engineering (FSE), p. 92–103, doi: https://doi.org/10.1145/3663529.3663831
8. S. Stradowski and L. Madeyski, “Bridging the Gap between Academia and Industry in Machine Learning Software Defect Prediction: Thirteen Considerations,” IEEE/ACM 38th International Conference on Automated Software Engineering (ASE), 2023, https://doi.org/10.1109/ASE56229.2023.00026
9. S. Stradowski and L. Madeyski, "Your AI is impressive, but my code does not have any bugs" Managing false positives in industrial contexts, Science of Computer Programming, 2025, 246:103320, https://doi.org/10.1016/j.scico.2025.103320
10. Nokia Corporation, “Nokia Annual Report 2022,” 2023, https://www.nokia.com/about-us/investors/results-reports/
11. Nokia Corporation, “Nokia Annual Report 2024,” 2025, https://www.nokia.com/about-us/investors/results-reports/
12. The 3rd Generation Partnership Project, “3GPP REL15,” 2021, https://www.3gpp.org/release-15
13. International Institute of Business Analysis, “Babok: A Guide to the Business Analysis Body of Knowledge,” 2015, no. t.3, https://www.iiba.org/career-resources/a-business-analysis-professionals-foundation-for-success/babok/
14. International Software Testing Qualifications Board, |Foundation Level Syllabus v3.1.1,| 2021. URL: https://www.istqb.org/certifications/certified-tester-foundation-level
15. International Organization for Standardization, “Software and systems engineering — software testing,” 2013. https://www.iso.org/standard/45142.html

Wroclaw University of Science and Technology

References (2/4)

16. M. Shafi, A.F. Molisch, P.J. Smith, T. Haustein, P. Zhu, P.D. Silva, F. Tufvesson, A. Benjebbour, G. Wunder, “5G: a tutorial overview of standards, trials, challenges, deployment, and practice,” IEEE J. Sel. Areas Commun. 35(6):1201–1221, 2017, http://dx.doi.org/10.1109/JSAC.2017.2692307
17. Y. Qi, G. Yang, L. Liu, J. Fan, A. Orlandi, H. Kong, W. Yu, Z. Yang, 5G overtheair measurement challenges: overview, IEEE Trans. Electromagn. Compat. 59 (6) (2017) 1661–1670, http://dx.doi.org/10.1109/TEMC.2017.2707471
18. M. Lang, M. Binder, J. Richter, P. Schratz, F. Pfisterer, S. Coors, Q. Au, G. Casalicchio, L. Kotthoff, B. Bischl, mlr3: A modern object-oriented machine learning framework in R, Journal of Open-Source Software (2019). doi:10.21105/joss.01903
19. P. Biecek, DALEX: Explainers for complex predictive models in R, Journal of Machine Learning Research 19 (2018) 1–5.
20. B. Boehm, “Value-based software engineering: Reinventing,” SIGSOFT Software Engineering Notes 28:2, 2003, https://doi.org/10.1145/638750.638775
22. V. R. Basili, G. Caldiera, H. D. Rombach, “The Goal Question Metric Approach,” volume I, John Wiley & Sons, 1994
23. M. Staron, “Action Research in Software Engineering - Theory and Applications,” Springer Cham, 2020, https://doi.org/10.1007/978-3-030-32610-4
24. P. Runeson, M. Höst, A. Rainer, B. Regnell, “Case Study Research in Software Engineering. Guidelines and Examples,” Wiley, Hoboken, 2012.
25. L. Madeyski, B. Kitchenham, “Would wider adoption of reproducible research be beneficial for empirical software engineering research?,” Journal of Intelligent & Fuzzy Systems, 32:509–1521, 2017, https://doi.org/10.3233/JIFS-169146
26. R. Rana, M. Staron, J. Hansson, M. Nilsson, and W. Meding, “A framework for adoption of machine learning in industry for software defect prediction,” Proceedings of the 9th International Conference on Software Engineering and Applications (ICSOFT-EA’14), 2014, https://doi.org/10.5220/0005099303830392
27. I. Arora, V. Tetarwal, A. Saha, “Open issues in software defect prediction,” Procedia Computer Science, 46 906–912, 2015, https://doi.org/10.1016/j.procs.2015.02.161
28. B. Kitchenham, S. Pfleeger, “Personal opinion surveys, in: Guide to Advanced Empirical Software Engineering,” Springer London, 63–92, 2008, https://doi.org/10.1007/978-1-84800-044-5_3
29. M. Kasunic, “Designing an Effective Survey,” Software Engineering Institute, 2005, https://doi.org/10.1184/R1/6573062.v1
30. J. Linåker, S. Sulaman, R. Maiani de Mello, M. Höst, “Guidelines for Conducting Surveys in Software Engineering,”, 2017, https://doi.org/10.1109/MS.2017.265100233
31. M. J. Page et al., “The prisma 2020 statement: an updated guideline for reporting systematic reviews,” BMJ 372, 2021, https://doi.org/10.1136/bmj.n71
32. B. A. Kitchenham, L. Madeyski, D. Budgen, “Segress: Software engineering guidelines for reporting secondary studies,” IEEE Transactions on Software Engineering, 2023, https://doi.org/10.1109/TSE.2022.3174092
33. S. Herbold, “On the Costs and Profit of Software Defect Prediction,” IEEE Transactions on Software Engineering, 47(11):2617-2631, 2019, https://doi.org/10.1109/TSE.2019.2957794

References (3/4)

34. V. Garousi, M. Felderer, “Worlds Apart - Industrial and Academic Focus Areas in Software Testing,” IEEE Software, 34:38-45, 2017, https://doi.org/10.1109/MS.2017.265100233
35. D. Chicco, G. Jurman, The matthews correlation coefficient (mcc) should replace the roc auc as the standard metric for assessing binary classification, BioData Mining 16 (2023) 4. https://doi.org/10.1186/s13040-023-00322-4.D.
36. L. Briand, D. Bianculli, S. Nejati, F. Pastore, M. Sabetzadeh., “The Case for Context-Driven Software Engineering Research: Generalizability Is Overrated,” IEEE Software 34, 5 2017, https://doi.org/10.1109/MS.2017.3571562
37. P. Godoe and T. Johansen, “Understanding adoption of new technologies: Technology readiness and technology acceptance as an integrated concept,” Journal of European psychology students 3, 2012, https://doi.org/10.5334/jeps.aq
38. K. Meinke and A. Bennaceur, “Machine Learning for Software Engineering Models, Methods, and Applications,” Proceedings of the IEEE/ACM 40th International Conference on Software Engineering (ICSE’18), 2018, https://doi.org/10.1145/3183440.3183461
39. D. H. Wolpert and W. Macready, “No free lunch theorems for optimization,” IEEE Transactions on Evolutionary Computation, 1:67–82, 1997, https://doi.org/10.1109/4235.585893
40. S. Masuda, Y. Nishi, and K. Suzuki, “Complex software testing analysis using international standards,” Proceedings of the IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW’20), 2020, https://doi.org/10.1109/ICSTW50294.2020.00049
41. P. Zhang, X. Yang, J. Chen, and Y. Huang, “A survey of testing for 5g: Solutions, opportunities, and challenges,” China Communications, 16(1):69-85, 2019, https://doi.org/10.12676/j.cc.2019.01.007
42. M. Lanza, A. Mocci, and L. Ponzanelli, “The tragedy of defect prediction, prince of empirical software engineering research,” IEEE Software, 33(6):102-105, 2016, https://doi.org/10.1109/MS.2016.156
43. S. Stradowski, "ML SDP in RAN System-level testing - Project Documentation.docx," version 2.2, pp. 1-24, NOKIA, February 2025
44. J. Puchinger, G. Raidl, and U. Pferschy, “The multidimensional knapsack problem: Structure and algorithms,” INFORMS Journal on Computing, 22(2):250-265, 2010, https://doi.org/10.1287/ijoc.1090.0344
45. D. V. Carvalho, E. M. Pereira, J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” Electronics 8, 2019, https://doi.org/10.3390/electronics8080832
46. J. Jiarpakdee, C. Tantithamthavorn, J. Grundy, “Practitioners’ Perceptions of the Goals and Visual Explanations of Defect Prediction Models,” IEEE/ACM 18th International Conference on Mining Software Repositories (MSR), Madrid, Spain, 2021, https://doi.org/10.48550/arXiv.2102.12007
47. V. H. S. Durelli, R. S. Durelli, S. S. Borges, A. T. Endo, M. M. Eler, D. R. C. Dias, and M. P. Guimaraes, “Machine learning applied to software testing: A systematic mapping study,” IEEE Transactions on Reliability, 68(3):1189-1212, 2019, https://doi.org/10.1109/TR.2019.2892517
48. S. Hosseini, B. Turhan, and D. Gunarathna, “A systematic literature review and meta-analysis on cross project defect prediction,” IEEE Transactions on Software Engineering, 45(2):111-147, 2019, https://doi.org/10.1109/TSE.2017.2770124

Wroclaw University of Science and Technology

References (3/4)

49. Project Management Institute, “The Standard for Project Management and a Guide To The Project Management Body Of Knowledge,” PMI, 2021, https://www.pmi.org/pmbok-guide-standards/foundational/pmbok
50. T. Pyzdek, The Six Sigma Handbook: A Complete Guide for Green Belts, Black Belts, and Managers at All Levels, McGraw-Hill Companies, New Your, USA, 2003, https://10.1036/0071415963

Wrocław University of Science and Technology

Thank You

This research was carried out in partnership with the NOKIA Corporation and was financed by the Polish Ministry of Science and Higher Education 'Implementation Doctorate' program ID: DWD/5/0178/2021.

Ministerstwo Nauki
i Szkolnictwa Wyższego

NOKIA

Wrocław University of Science and Technology

Recognitions and awards

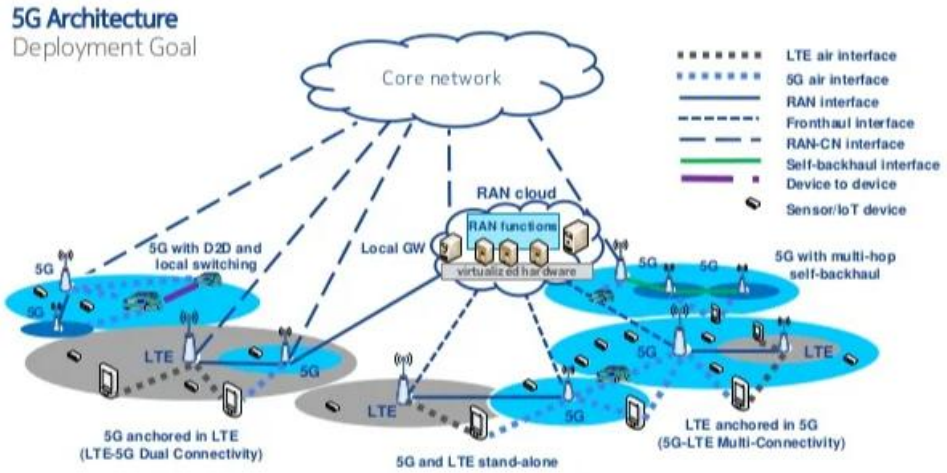
During this dissertation effort, related research outcomes have been recognized with the following awards:

- A distinction in the ”Studencki Program Stypendialny - scholarship of Marian Suski” in the field of engineering and technical sciences awarded by the mayor of Wrocław in 2024.
- Three Wrocław University of Science and Technology ”PRIMUS” awards for publications that contribute to the development of particular scientific disciplines in 2024 and 2025.
- Two Wrocław University of Science and Technology ”rector’s awards” for outstanding scientific achievements related to the doctoral dissertation in the academic years 2022/2023, 2023/2024, and 2024/2025.
- Two Wrocław University of Science and Technology ”scholarships from the own fund” for active and creative research in the academic years 2023/2024 and 2024/2025.
- Two Nokia ”recognize excellence” awards for machine learning software defect prediction implementation efforts in 2022 and 2023.
- Also, a testimony of related PhD experience was presented during the Unite! Research Week (part of the Unite! Research School) in October 2024 in Grenoble, France.
- Finally, the publications that constitute this work have amassed 1260 MEIN[1] points and aggregated Impact Factor of 19.9[2], as well as 130+ citations in Google Scholar[3], 120+ citations in Research Gate[4], and 80+ citations in Scopus[5], at the time of dissertation submission in September 2025.

[1] https://www.gov.pl/web/nauka/ujednolicony-wykaz-czasopism-naukowych
[2] https://research.com/
[3] https://scholar.google.com/citations?user=sykq35cAAAAJ&hl
[4] https://www.researchgate.net/profile/Szymon-Stradowski
[5] https://www.scopus.com/authid/detail.uri?authorId=57899636200

What is 5G?

5G is the fifth generation of cellular networks. With a transfer speed 100-times faster than 4G, 5G creates never-seen-before possibilities for businesses and people.



5G Purpose:

- 5G is the first mobile technology designed for machines as well as people and to enable very high transmission speed, low latency, and reduced error rate.
- The gNodeB (gNB), which is the main focus of our study, is the ‘‘Next Generation Node B’’ 5G base transceiver station, compliant with strict 3GPP standards [12].

5G Characteristics:

- The gNodeB (gNB) connects the 5G User Equipment (UE) with 5G core using 5G air interface. The air interface, defined by the 3GPP specification, is divided into two frequency bands, FR1 (below 6 GHz) and FR2 (24–54 GHz), each having different propagation characteristics, requiring specific approaches to develop and test [17].
- Importantly, advanced techniques such as massive MIMO (Multiple Input Multiple Output, using multiple antennas in the transmitter and receiver) and beamforming (sending signals a particular angles to utilize constructive and destructive interference) are used to achieve performance requirements [17], but also add increased complexity to the testing process.

5G Challenges:

- Need for extensive testing in Over-the-air (OTA) conditions [18, 41], as well as conducted mode, due to the characteristics of employed frequency ranges.
- Each cellular phone company brings its own specific needs and requirements, translating to hundreds of features and thousands of software and hardware configurations.
- The 5G system is comprised of a multitude of features, and each new software release introduces new ones incrementally. Due to the complexity and size of the system, it is extremely hard to predict all interactions on the specification level.

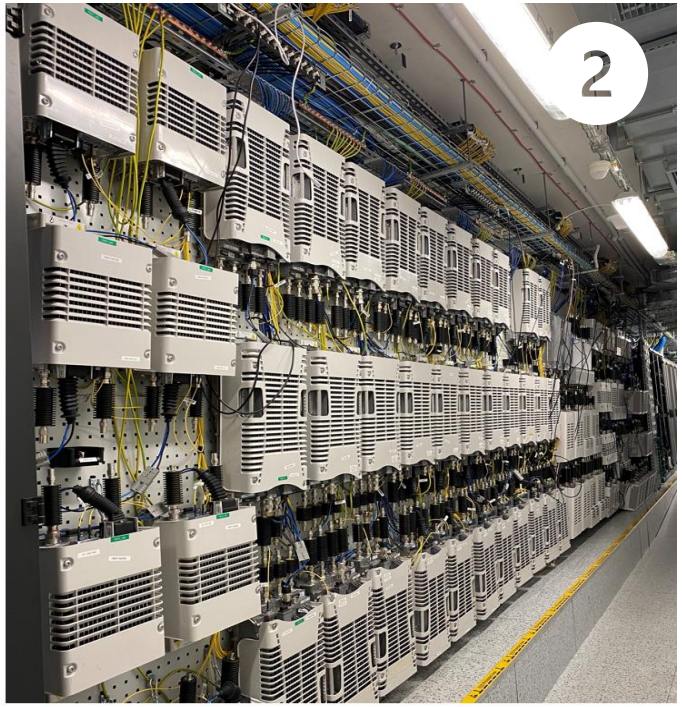
5G test process

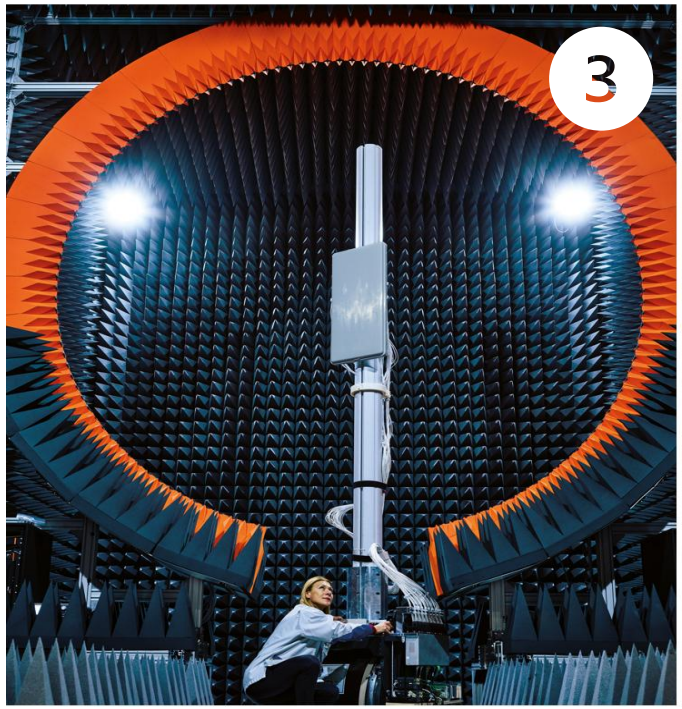
Nokia 5G test process visualization [1,4]:

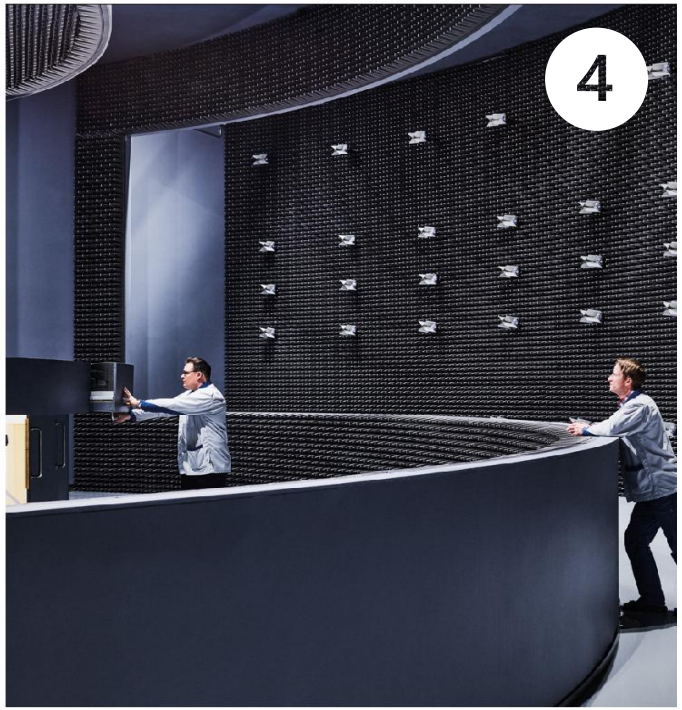
1. low-level code tests run on simulators,
2. real RF HW tested in Nokia Laboratory in Wrocław,
3. “stargate portal” [13],
4. anechoic “walls” in Oulu [13].

Each stage is more expensive to run as it executes more code, benchmarks over more extended periods of time, increases the number of repetitions, or replaces simulators with actual hardware to be more equivalent to the real-life environment.

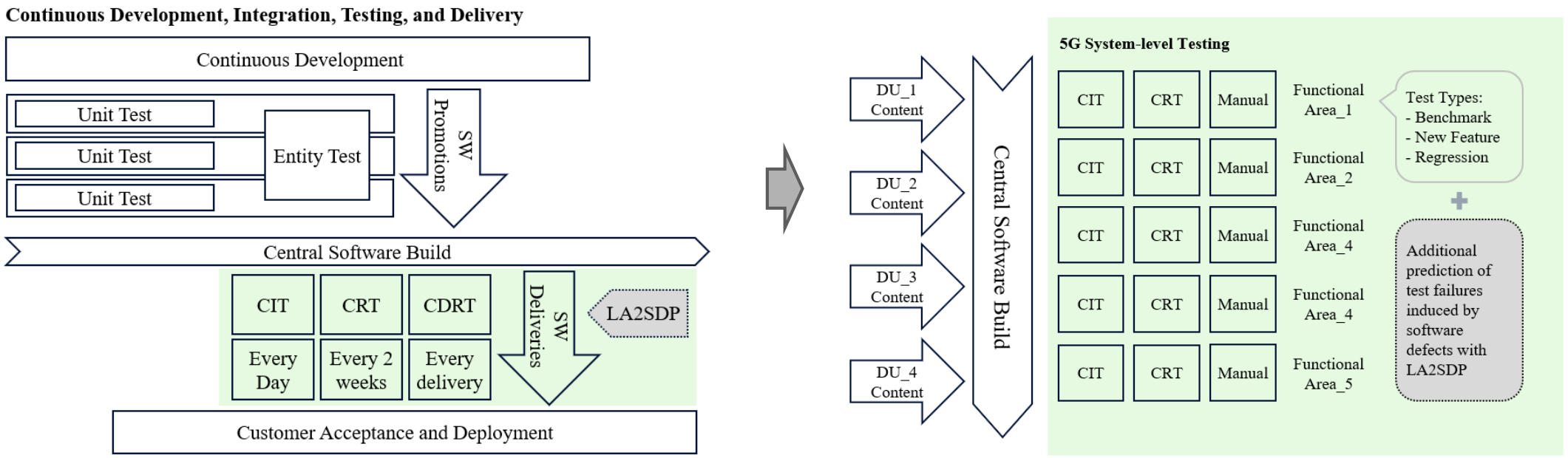
Any improvement to the process, can have a significant upside potential.







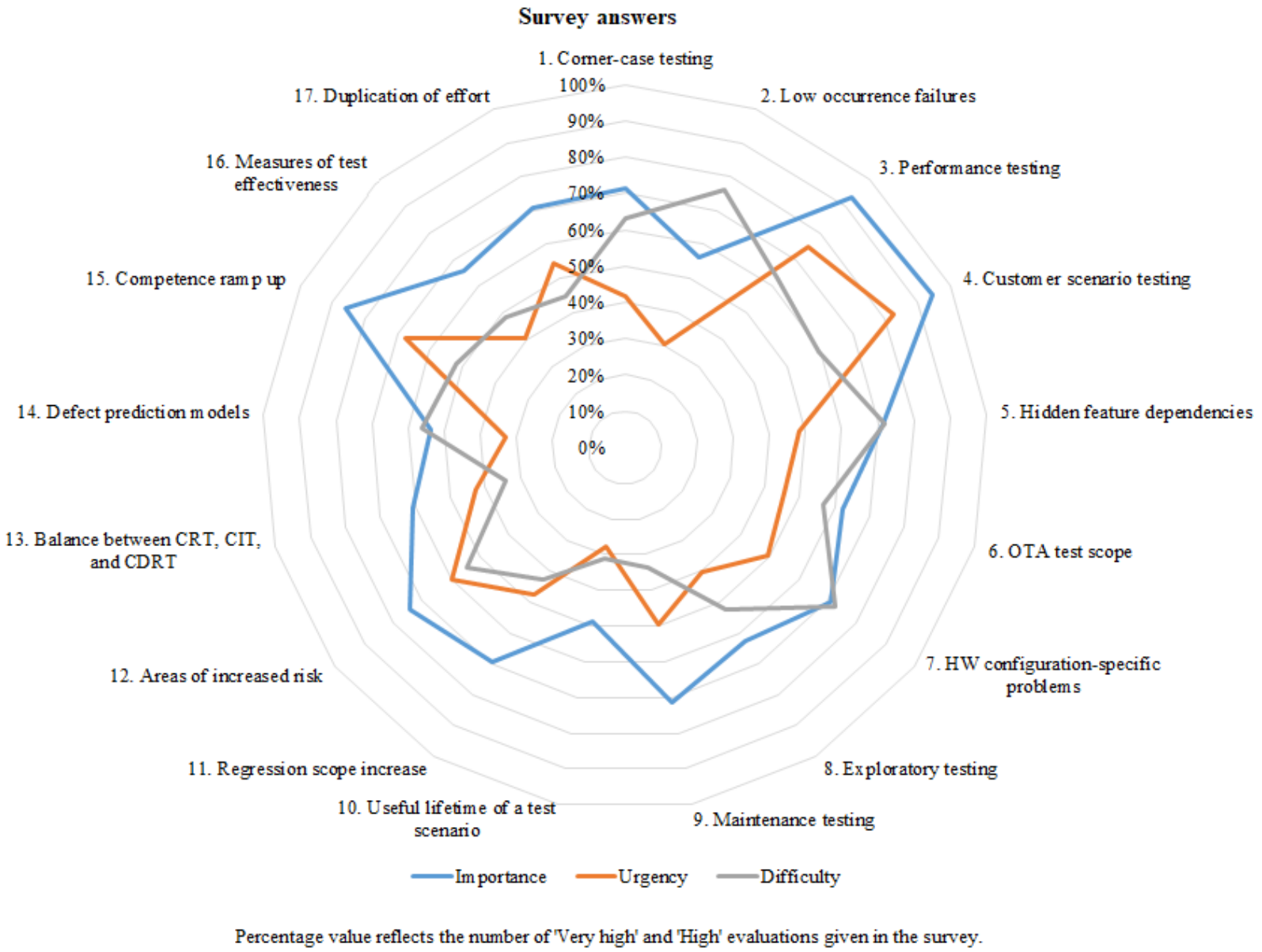
CDIT real-world data set



Real-world data set:

- collection of historical test process metrics from the main test case repository for the Nokia 5G quality assurance process,
- almost 800,000 unique results for more than 100,000 test cases,
- a period of five and a half months from the beginning of January 2021 until the middle of June 2021,
- executed into two-week feature builds (allowing expanding / sliding window prediction approaches),
- available in ART5 Supplementary Material https://doi.org/10.6084/m9.figshare.28263290.

Exploring the challenges in software testing of the 5G system at Nokia: A Survey - results



Percentage value reflects the number of 'Very high' and 'High' evaluations given in the survey.

The questionnaire was completed by 312 out of 2935 (10.63%) possible respondents.

The challenges are seen as the most important and urgent:

- customer scenario testing,
- performance testing, and
- competence ramp-up.

Challenges seen as the most difficult to solve:

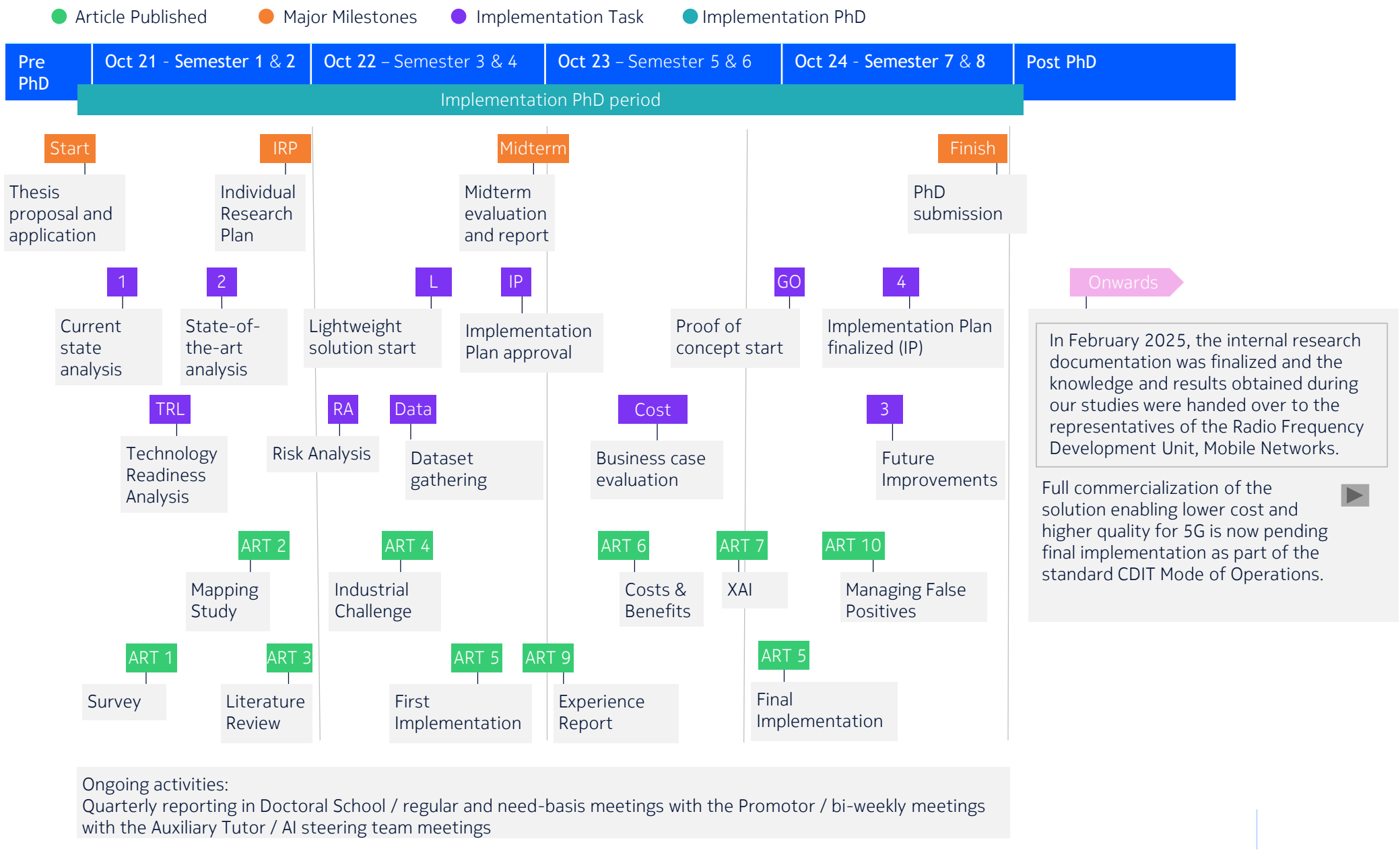
- low occurrence failures,
- hidden feature dependencies, and
- hardware configuration-specific problems.

Other challenges in system level testing:

- deficient test infrastructure,
- time pressure, and
- missing specification.

Also, respondents see SDP as under-utilized within the company.

Implementation Timeline



Ongoing activities:
Quarterly reporting in Doctoral School / regular and need-basis meetings with the Promotor / bi-weekly meetings with the Auxiliary Tutor / AI steering team meetings