Enhancing the quality of software systems using deep learning models for defects prediction and detection

Final scientific report
2021 - 2023

SUMMARY

PROJECT CODE: PN-III-P4-ID-PCE-2020-0800

CONTRACT: PCE 92/2021

1. PROJECT SUMMARY

The project topic is *software defect prediction and detection*, a topic of major international interest, being of great relevance during the development, testing and maintenance of software systems. Accurate prediction of software defects in new software versions would significantly improve the performance of the software development process in terms of cost, time and software quality. Prediction of defects in software systems helps in detecting, tracking and solving software anomalies which may have negative effects on human safety and human life, specifically in case of critical software systems. Defect prediction allows modifications early in a software lifecycle, leading to lower costs and improving the software customers' satisfaction.

The project targeted the development of deep learning techniques for software defect prediction, a problem of major relevance within the Software Engineering field, particularly in search-based software engineering. The major goal is to improve the quality of the software systems by early and accurate identification of defective software modules, using deep learning models and techniques. Thus, the main goal is to facilitate software maintenance and evolution activities such as software testing, code review and software quality assessment, through automatically identifying software defects. The major and high-level objective of the project was to improve the quality of software systems using deep learning models for automatic software defects prediction and detection. Our particular target is to increase the accuracy of predicting software defects in a new version of a software system (within-project software defects prediction) and mainly to reduce the proportion of defects which are not detected (false negative rate). We consider two major research directions: (1) improving the feature engineering step by selecting relevant features for specific types of defects (e.g. semantic features, cohesion or conceptual coupling based metrics) and (2) automatically extracting semantic meaningful features from source code representations (other than AST-based).

The project results: (1) original machine learning methods for software defect prediction; (2) software modules implementing the developed machine learning models for predicting faulty software entities; (3) scientific and technical reports containing the original machine learning methods developed for software defect prediction; and (4) scientific publications for disseminating the obtained scientific results. The current report presents the original work carried out during the project implementation (01.01.2021 – 31.12.2023). We will present how the activities from the work plan were accomplished and the means for disseminating the results obtained by the project team.

To summarize, the results obtained within the QuaDeeP project are:

- Machine learning methods and models for learning software characteristics relevant for predicting software defects, cohesion and coupling-based software metrics for defect prediction, deep learning methods for sooftware defect prediction, software modules of QuaDeeP software.
- Project webpage (www.cs.ubbcluj.ro/quadeep).
- **19** scientific papers: **6** publications in ISI (Web of Science, WoS) indexed journals with impact factor (2 from Q1 quartile, 2 from Q2 quartile and 2 from Q3 quartile accrding to the JCR from the publication year); **13** publications in proceedings of international conferences (9 are B-ranked, 1 is C-ranked and 3 are D-ranked according to CORE classification) published/to be published in journals that are WoS indexed or sent for WoS indexing. Among the 19 publications, 3 are currently in presss.
- **8** presentations at international conferences and workshops.

The project objectives have been achieved, as highlighted by the final project report. The planned objectives, together with the related activities have been totally fulfilled and carried out according to the project implementation plan. The minimum performance criteria regarding the results dissemination for each of the project implementation years (2021, 2022, 2023) - at least one paper accepted for publication in an ISI/WoS journal with high impact factor and at least three publications- has been accomplished. All the scientific and

technical results obtained within the project are disseminated through the QuaDeeP project website (www.cs.ubbcluj.ro/quadeep).

1 INTRODUCTION

1.1 QUADEEP PROJECT

The project focuses on developing deep learning techniques for *software defect prediction* (SDP), a problem of major relevance within the Software Engineering field, particularly in search-based software engineering. The major goal is to improve the quality of the software systems by early and accurate identification of defective software modules, using deep learning models and techniques. Thus, the main goal is to facilitate software maintenance and evolution activities such as software testing, code review and software quality assessment, through automatically identifying software defects. The project topic is of major international interest, being of great relevance during the development, testing and maintenance of software systems. Accurate prediction of software defects in new software versions would significantly improve the performance of the software development process in terms of cost, time and software quality. The project will provide a software solution, QuaDeeP, which will integrate novel deep learning methods for software defects identification. For increasing the specificity of the developed learning models, the targeted methods will be specifically tailored for particular types of defects. QuaDeeP will be useful for assisting software developers in accurately predicting software defects and thus, contributing to improving the software quality and to ease the software maintenance and evolution.

1.2 SCIENTIFIC OBJECTIVES

The major and high-level objective of this project is to improve the quality of software systems using DL models for automatic software defects prediction and detection. Our particular target is to increase the accuracy of predicting software defects in a new version of a software system (within-project SDP) and mainly to reduce the proportion of defects which are not detected (false negative rate). The project is applicative and highly interdisciplinary, having the following scientific and technical objectives.

O1. Development and scientific validation of novel DL based methods for the feature engineering step for SDP. First, existing taxonomies of defect types will be used for identifying relevant features which are specific to particular classes of defects. ML models such as autoencoders (AEs), CNNs and LSTMs are targeted to automatically learn semantic and syntactic features from representations of the source code generated by Doc2Vec, tokens based on the AST of the code, Code2Vec , and their combination. From a manual feature engineering perspective, new cohesion and coupling based software metrics for SDP will be expressed based on existing software metrics and semantic representations of the source code generated by Doc2Vec, Latent Semantic Indexing (LSI) and Graph2Vec.

O2. Development and scientific validation of novel ML based models and techniques for SDP. The ML models will be specifically tailored for particular types of defects (targeted at O1) and thus the specificity of the models will be increased, as they will learn to predict only a particular class of defects. More specifically, one-class classification (OCC) and one-shot learning (OSL) methods are envisaged for handling the main issue of data imbalancement. As one-class classifiers (anomaly detectors) we target to use AEs, Relational Association Rules (RARs), Gradual RARs (GRARs) and a Hybrid classifier based on GRARs (HyGRAR), while OSL with Siamese networks, Bayesian OSL and N-Shot learning are envisaged as one-shot classifiers.

O3. Development and validation of the QuaDeeP software system. Provided as software modules, QuaDeeP will deliver a solution to assist developers, testers, and software managers in software maintenance and evolution activities. It will offer insights that allow stakeholders to identify potential software defects.

O4. Contribute to the development of scientific knowledge by disseminating the obtained scientific results through scientific publications and the project website.

2 DISSEMINATION

2.1 PROJECT WEBSITE

The project website is dedicated to the presentation of the project, the research team and the results obtained. Two versions of the website can be accessed: one in English (http://www.cs.ubbcluj.ro/quadeep/) and one in Romanian (http://www.cs.ubbcluj.ro/quadeep/ro/about-romana/).

The website is organized into several sections, and each of them can be visited at any moment using the tab navigation at the upper right corner of the pages. First, there is the main page with a brief overview of the project (**About/Despre**). Following that, information regarding the project plan (**Project Plan/Planul Proiectului** page) and the project team (**Project Team/Echipa** page) is provided. A section dedicated to the project results is **Project results/Rezultatele proiectului**, while the Dissemination section (**Dissemination/Diseminare**) is divided into three pages: one for project publications (**Publications/Publicații**), another for the annual scientific and technical reports (**Annual Reports/Rapoarte Anuale**), and a third for conference presentation files and video clips (**Presentations/Prezentări**). The project coordinator's contact information is also available on the **Contact** page. The main page of the website (**About/Despre**) includes a brief summary of the project and its objectives, whilst the **Project Plan/Planul Proiectului** page lists the tasks defined within each of the five work packages of the plan. The **Project Team/Echipa** page includes academic biographies and links to Google Scholar profiles for the project team members. The section on **Dissemination/Diseminare** is divided into three pages: (1) **Publications/Publicații**, which contains a list of project publications and a list of related publications, both up to date and the first continuously updated to include the latest works published within the project; (2) **Annual Reports/Rapoarte Anuale**, which will contain all the annual scientific and technical reports; and (3) **Presentations/Prezentări**, which contains conference presentation files and video clips that can be viewed and, in the case of presentation files, downloaded.

2.2 SCIENTIFIC PUBLICATIONS AND PRESENTATIONS

Tables 1 and 2 present the list of scientific publications and presentations during the QuaDeeP project implementation (2021-2023).

2023	
[L1]	George Ciubotariu, Gabriela Czibula, Istvan Gergely Czibula, Ioana-Gabriela Chelaru, *Uncovering Behavioural Patterns of One: And Binary-Class SVM-Based Software Defect Predictors*, In Proceedings of the 18th International Conference on Software Technologies - ICSOFT; ISBN 978-989-758-665-1; ISSN 2184-2833, SciTePress, pages 249-257. (**B-ranked according to CORE classification, indexed WoS**)

[L2]	Anamaria Briciu, Gabriela Czibula, Mihaiela Lupea, *A study on the relevance of semantic features extracted using BERT-based language models for enhancing the performance of software defect classifiers*, 27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2023), Procedia Computer Science, in press (**B-ranked according to CORE classification, indexed WoS**)
[L3]	Gabriela Czibula, Ioana-Gabriela Chelaru, Istvan Gergely Czibula, Arthur Molnar, *An unsupervised learning-based methodology for uncovering behavioural patterns for specific types of software defects*, 27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2023), Procedia Computer Science, in press (**B-ranked according to CORE classification, indexed WoS**)
[L4]	Zsuzsanna Marian-Oneț, Diana-Lucia Miholca*, Source-code embedding-based software defect prediction*, In Proceedings of the 18th International Conference on Software Technologies - ICSOFT; ISBN 978-989-758-665-1; ISSN 2184-2833, SciTePress, pages 185-196. (**B-ranked according to CORE classification, indexed WoS**)
[L5]	Mariana Maier, Gabriela Czibula, Lavinia Delean, *Using unsupervised learning for mining behavioural patterns from data. A case study for the baccalaureate exam in Romania,* Studies in Informatics and Control, vol. 32(2), pp. 73-84, 2023 (**2022 IF=1.6, Q3**)
[L6]	Imre-Gergely Mali, Gabriela Czibula, *Policy-Based Reinforcement Learning in the Generalized Rock-Paper-Scissors Game*, ESANN 2023 proceedings, The 31th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2023), pp. 345-350 (**B-ranked according to CORE classification, indexed WoS**)
[L7]	Alexandra-Ioana Albu. *Temporal ensembling-based deep k-nearest neighbours for learning with noisy labels*. ESANN 2023 proceedings, 31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pp. 483-488 (**B-ranked according to CORE classification, indexed WoS**)
[L8]	Paul-Dumitru Orășan, Gabriela Czibula, *Im2Vide0: A Zero-Shot approach using diffusion models for natural language conditioned Image-to-Video*, 2023 IEEE 19th International Conference on Intelligent Computer Communication and Processing, 2023, in press (**D-ranked according to CORE classification, indexed IEEE**)
2022	
[L9]	Mihaiela Lupea, Anamaria Briciu, Istvan-Gergely Czibula, Gabriela Czibula, *SoftId: An autoencoder-based one-class classification model for software authorship identification*, 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2022), Procedia Computer Science, Volume 207, 2022, Pages 716-725 (**B-ranked according to CORE classification, indexed WoS**)
[L10]	Diana-Lucia Miholca, Vlad-Ioan Tomescu, Gabriela Czibula, *An in-depth analysis of the software features' impact on the performance of deep learning-based software defect predictors*, IEEE Access, 2022, Volume 10, pp. 64801 - 64818 (**B-ranked, indexed WoS, 2021 IF=3.476, Q2**)
[L11]	Gabriela Czibula, Mihaiela Lupea, Anamaria Briciu, *Enhancing the performance of software authorship attribution using an ensemble of deep autoencoders*, Mathematics, Special Issue "Recent Advances in Artificial Intelligence and Machine Learning", 2022, 10(15):2572 (**A-ranked, indexed WoS, 2021 IF=2.592, Q1**)
[L12]	Gabriela Czibula, George Ciubotariu, Mariana Maier, Hannelore-Inge Lisei, *IntelliDaM: A machine learning based framework for enhancing the performance of decision-making processes. A case study for educational data mining*, IEEE Access, 2022, Volume 10, pp. 80651-80666 2 (**B-ranked, indexed WoS, 2021 IF=3.476, Q2**)
2021	

[L13]	Anamaria Briciu, Gabriela Czibula, Mihaiela Lupea – "*AutoAt: A deep autoencoder-based classification model for supervised authorship attribution*", 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2021), Procedia Computer Science 192, pp. 397-406 (**B-ranked, indexed Scopus**)
[L14]	Vlad-Ioan Tomescu, Gabriela Czibula, Ștefan Nițică – "*A study on using deep autoencoders for imbalanced binary classification*", 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2021), September 8-10, 2021, Procedia Computer Science 192, pp. 119-128 (**B-ranked, indexed WoS**)
[L15]	George Ciubotariu, Vlad-Ioan Tomescu, Gabriela Czibula – "*Enhancing the performance of image classification through features automatically learned from depth-maps*", 13th International Conference on Computer Vision Systems, September 22-24, 2021, LNCS 12899, pp. 68-81 (**C-ranked, indexed WoS**)
[L16]	Diana-Lucia Miholca - "*New Conceptual Cohesion Metrics: Assessment for Software Defect Prediction*" *2021 23rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, Timisoara, Romania, 2021, pp. 163-170, doi: 10.1109/SYNASC54541.2021.00036. (**D-ranked, indexed WoS**)
[L17]	Zsuzsanna Oneț-Marian, Gabriela Czibula, Mariana Maier – "*Using self-organizing maps for comparing students' academic performance in online and traditional learning environments*", Studies in Informatics and Control (SIC) journal, 30(4), 2021, pp. 17-28 (C-ranked, indexed WoS, **IF 2020=1.649, Q3**)
[L18]	Maria-Mădălina Mircea, Rareș Boian, Gabriela Czibula – "*A machine learning approach for data protection in virtual reality therapy applications*" 2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 2021, pp. 367-374 (**D-ranked, indexed Scopus**)
[L19]	Mariana-Ioana Maier, Gabriela Czibula, Zsuzsanna Oneț-Marian – "*Towards Using Deep Autoencoders for Comparing Traditional and Synchronous Online Learning in Assessing Students' Academic Performance*", Mathematics, Engineering Mathematics, 2021, 9(22), 2870 (A-ranked, **2020 IF=2.258**, **Q1**)

Table 1 - List of scientific publications

2023	
1	George Ciubotariu, *Comparing one- and binary-class SVM-based software defect predictors*, WeADL worksop, 2023 (**video YouTube**: https://www.youtube.com/watch?v=dO-gPupAJyU)
2	Anamaria Briciu, *Enhancing the performance of software authorship attribution using deep autoencoders*, WeADL worksop, 2023 (**video YouTube:** https://www.youtube.com/watch?v=VzKJ3Jum4uo)
3	George Ciubotariu, *Uncovering Behavioural Patterns of One: And Binary-Class SVM-Based Software Defect Predictors*, The 18th International Conference on Software Technologies - ICSOFT 2023
4	Anamaria Briciu, *A study on the relevance of semantic features extracted using BERT-based language models for enhancing the performance of software defect classifiers*, The 27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2023), **video Youtube:** https://www.youtube.com/watch?v=iR8D2FlG9W8
5	Ioana-Gabriela Chelaru, *An unsupervised learning-based methodology for uncovering behavioural patterns for specific types of software defects*, KES2023, **video YouTube:** https://www.youtube.com/watch?v=cTYoSbCu4Vw
6	Diana-Lucia Miholca, *Source-code embedding-based software defect prediction*, The 18th International Conference on Software Technologies - ICSOFT 2023
2022	
7	Mihaiela Lupea, Anamaria Briciu, Istvan-Gergely Czibula, Gabriela Czibula – "*SoftId: An autoencoder-based one-class classification model for software authorship identification*", 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2022), September 7-9, 2022.
2021	
8	George Ciubotariu, Vlad-Ioan Tomescu, Gabriela Czibula,"*Enhancing the performance of image classification through features automatically learned from depth-maps*", 13th International Conference on Computer Vision Systems, September 2021.

Table 2 - Presentations at international conferences during the QuaDeeP project (2021-2023).

3 CONCLUSIONS

The current report presented the original results obtained during the project implementation (01.01.2021 – 31.12.2023) for achieving the scientifical and technical objectives from the project workplan. We will present how the activities from the work plan were accomplished and the means for disseminating the results obtained by the project team. We summarize the main project achievements: (1) original machine learning methods for software defect prediction; (2) software modules implementing the developed machine learning models for predicting faulty software entities; and (3) scientific and technical reports containing the original machine learning methods developed for software defect prediction.

The obtained results were disseminated through **19** scientific papers and **8** presentations at international conferences and workshops. Among the publication, **6** are in ISI (Web of Science, WoS) indexed journals with impact factor (2 from Q1 quartile, 2 from Q2 quartile and 2 from Q3 quartile accrding to the JCR from the publication year); **13** publications in proceedings of international conferences (9 are B-ranked, 1 is C-ranked and 3 are D-ranked according to CORE classification) published/to be published in journals that are WoS indexed or sent for WoS indexing. Among the 19 publications, 3 are currently in presss.

The minimum performance criteria regarding the results dissemination for each of the project implementation years (2021, 2022, 2023) - at least one paper accepted for publication in an ISI/WoS journal with high impact factor and at least three publications- has been accomplished. The planned objectives, together with the related activities have been totally fulfilled and carried out according to the project implementation plan.