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

Scientific and technical report 2023
SUMMARY

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

CONTRACT: PCE 92/2021

2023

PHASE 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. The project targets 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 is 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 estimated results of the project are: (1) scientific and technical reports containing the original machine learning methods developed for software defect prediction; (2) scientific publications for disseminating the obtained scientific results; (3) software modules implementing the developed machine learning models for predicting faulty software entities.

The current report presents the original results obtained during the research carried out within the QuaDeeP project for achieving the scientific and technical objectives proposed in the project plan for 2023. The report highlights the current status of the project implementation, the way in which the activities undertaken in the work plan were carried out and how the results obtained in the current project phase (2023) were disseminated. To summarize, the results obtained within the QuaDeeP project in 2023 are:

- Development of OCC and OSL methods for software defect prediction.
- Design and development of software modules for the QuaDeeP system.
- Updating the project's website at www.cs.ubbcluj.ro/quadeep.
- 8 scientific articles: 1 publication in an ISI-rated journal (Web of Science, WoS) with an impact factor of 1.6 as per JCR 2022, ranked in the Q3 quartile; 6 publications in the proceedings of B-rated international conferences (according to CORE classification) to be submitted for WoS indexing; 1 publication in the proceedings of an IEEE-indexed international conference (to be submitted for WoS indexing).

The project objectives for 2023 have been achieved, as highlighted by the annual report for 2023. 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 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.

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. 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.

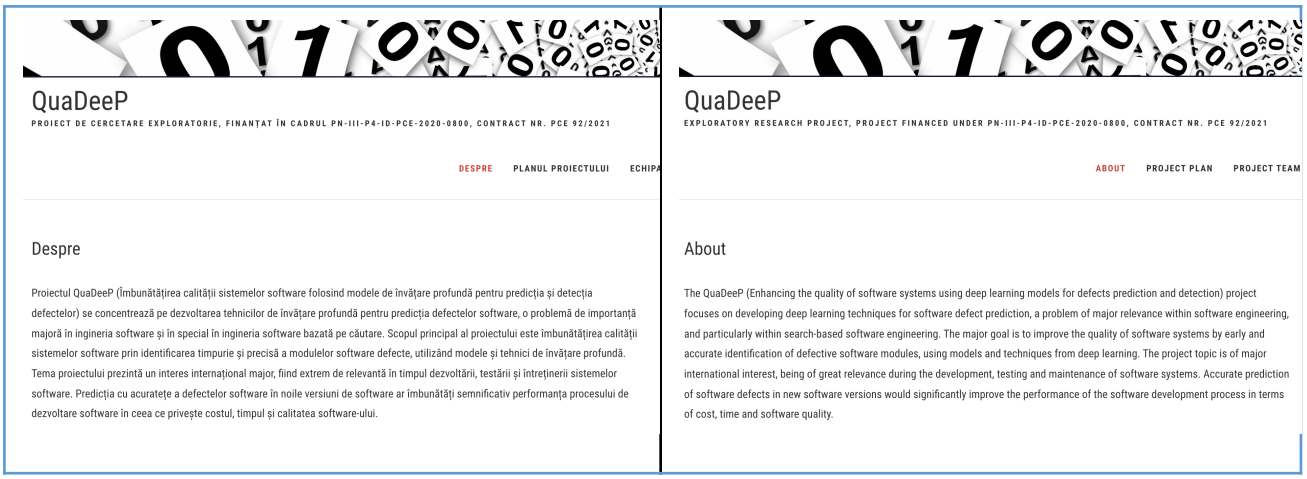

Figure 1 - The website's main page - version in Romanian (left) and version in English (right)

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

Table 1 presents the list of scientific publications obtained during the 3rd Phase (2023) of the QuaDeeP project.

[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. DOI: 10.5220/0012129600003538 (**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**)

Table 1 - List of scientific publications obtained during the 3rd Phase (2023) of the QuaDeeP project.

2.3 PRESENTATIONS

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*, The 27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (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

Table 2 - Presentations at international conferences during the 3rd Phase (2023) of the QuaDeeP project.

3 CONCLUSIONS

This report presented the original results obtained from the research carried out within the project in order to meet the scientific and technical objectives proposed in the implementation plan for 2023 (Phase 3). For each objective provided in the implementation plan for 2023, we indicated the way in which the related activities were performed.

The results obtained within the project for the year 2023 are summarized as follows: (1) Development of deep learning-based methods for software defect prediction; (2) Introduction of software metrics focused on cohesion and coupling for software defect prediction; (3) Annual scientific and technical report; (4) Scientific articles disseminating the original results obtained during Phase 3 of the project's implementation.

The dissemination of results obtained within the project in 2023 was achieved through the publication of 8 scientific articles: One publication in an ISI-rated journal (Web of Science, WoS) with an impact factor of 1.6 as per JCR 2022, ranked in the Q3 quartile; 6 publications in the proceedings of B-rated international conferences (according to CORE classification) set to be submitted for WoS indexing; 1 publication in the proceedings of an IEEE-indexed international conference (to be submitted for WoS indexing).

As a result, the minimum performance criteria provided (at least one paper accepted for publication in an ISI/WoS journal with high impact factor and at least three publications) was met. Furthermore, the project objectives for 2023 have been met, and all associated activities have been completed and carried out in accordance with the project implementation plan.