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

Scientific and technical report 2021

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

CONTRACT: PCE 92/2021

2021

1. 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 2021. 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 (2021) were disseminated. To summarize, the results obtained within the QuaDeeP project in 2021 are:

- Literature review on Software Defect Prediction (case studies, features, existing approaches), taxonomy of bug types and maintainability evaluation.
- QuaDeeP software architecture.
- Project website (www.cs.ubbcluj.ro/quadeep).
- 7 scientific publications: 2 publications in Web of Science (WoS) journals with impact factor (according to JCR 2020) 1.696 and 2.258; 5 publications in volumes of WoS-indexed international conferences.

The project objectives for 2021 have been achieved, as highlighted by the annual report for 2021. 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 2021 (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 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 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 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 imbalance. 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, a plugin for Integrated Development Environments (IDEs) for assisting software developers, testers and software managers in software maintenance and evolution activities, by notifying the developers when a software entity is likely to be defective.

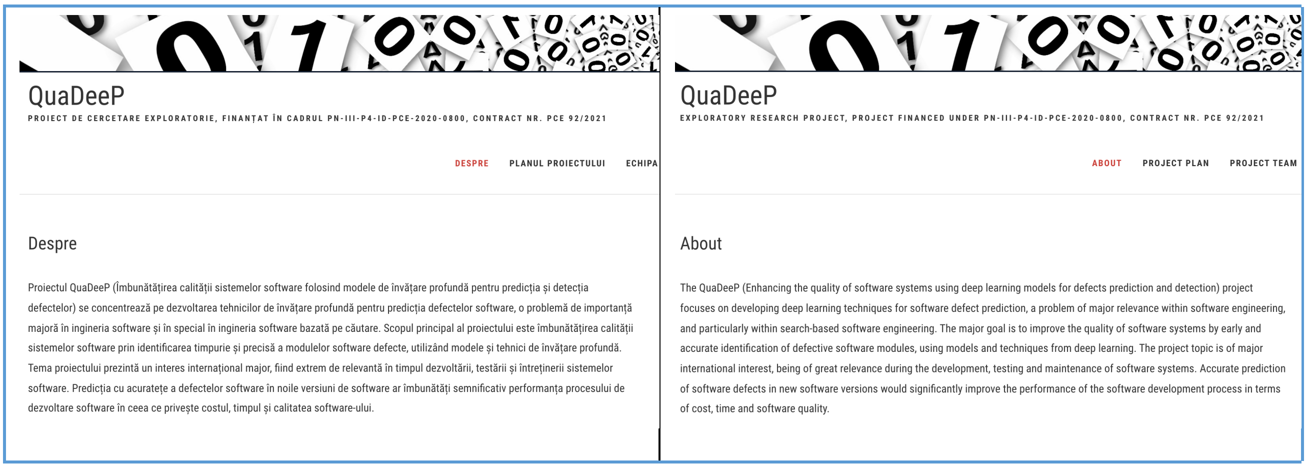
O4. Contribute to the development of scientific knowledge by disseminating the obtained scientific results through scientific publications and the project website.

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.



QuaDeeP

PROIECT DE CERCETARE EXPLORATORIE, FINANȚAT ÎN CADRUL PN-III-P4-ID-PCE-2020-0800, CONTRACT NR. PCE 92/2021

Despre

Proiectul QuaDeeP (Îmbunătățirea calității sistemelor software folosind modele de învățare profundă pentru predicția și detecția defectelor) se concentrează pe dezvoltarea tehnicilor de învățare profundă pentru predicția defectelor software, o problemă de importanță majoră în ingineria software și în special în ingineria software bazată pe căutare. Scopul principal al proiectului este îmbunătățirea calității sistemelor software prin identificarea timpurie și precisă a modulelor software defecte, utilizând modele și tehnici de învățare profundă. Tema proiectului prezintă un interes internațional major, fiind extrem de relevantă în timpul dezvoltării, testării și întreținerii sistemelor software. Predicția cu acuratețe a defectelor software în noile versiuni de software ar îmbunătăți semnificativ performanța procesului de dezvoltare software în ceea ce privește costul, timpul și calitatea software-ului.

QuaDeeP

EXPLORATORY RESEARCH PROJECT, PROJECT FINANCED UNDER PN-III-P4-ID-PCE-2020-0800, CONTRACT NR. PCE 92/2021

About

The QuaDeeP (Enhancing the quality of software systems using deep learning models for defects prediction and detection) project focuses on developing deep learning techniques for software defect prediction, a problem of major relevance within software engineering, and particularly within search-based software engineering. The major goal is to improve the quality of software systems by early and accurate identification of defective software modules, using models and techniques from deep learning. 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.

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 within the QuaDeeP project.

[L1]	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), September 8-10, 2021, Procedia Computer Science 192, pp. 397-406 (B-ranked, indexed WoS)
[L2]	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)
[L3]	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)
[L4]	Diana-Lucia Miholca – "***New Conceptual Cohesion Metrics: Assessment for Software Defect Prediction***", 23rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2021, acceptat spre publicare (D-ranked, indexed WoS)
[L5]	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**)
[L6]	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, 2021, acceptat spre publicare (D-ranked, indexed WoS)
[L7]	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 – special issue on Didactics and Technology in Mathematical Education, 2021, 9(22), 2870 (A-ranked, **2020 IF=2.258**, Q1)

Table 1 - List of scientific publications obtained within the QuaDeeP project

2.3 Presentations

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.
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. **video YouTube**: https://www.youtube.com/watch?v=Ha_kzQkRizI

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), September 8-10, 2021. **video YouTube**: https://www.youtube.com/watch?v=tG65l7wkqWI

Table 2 - Presentations at international conferences.

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 2021 (Phase 1). For each objective provided in the implementation plan for 2021, we indicated the way in which the related activities were performed.

We summarize the results obtained within the project for 2021 as follows: (1) the study of the literature on software defect prediction, taxonomy of bug types and maintainability evaluation; (2) establishing the QuaDeeP software architecture; (3) the web page of the project (http://www.cs.ubbcluj.ro/quadeep); (4) the annual scientific and technical report; (5) scientific articles through which the original results obtained in Phase 1 of the project implementation were disseminated.

The dissemination of the results obtained within the project in 2021 was achieved by publishing 7 scientific articles: 2 publications in ISI listed journals (WoS), with impact factors (according to JCR 2020) of 1.696 and 2.258; 5 publications in volumes of WoS indexed international conferences.

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 2021 have been met, and all associated activities have been completed and carried out in accordance with the project implementation plan.