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

Scientific and technical report 2022

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

CONTRACT: PCE 92/2021

2022

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 2022. 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 (2022) were disseminated. To summarize, the results obtained within the QuaDeeP project in 2022 are:

- Deep learning methods for learning relevant features for software defect prediction
- Cohesion- and coupling-based software metrics for software defect prediction
- Maintaining the project's website up to date (http://www.cs.ubbcluj.ro/quadeep)
- 4 scientific articles: 3 papers published in ISI (Web of Science, WoS) listed journals, with impact factors (according to JCR 2021) 3.476, 2.592 and 3.476, respectively; 1 publication in a WoS indexed international conference volume. Of the three papers published in journals, we note that two of them are in the Q2 quartile and one in the Q1 quartile.

The project objectives for 2022 have been achieved, as highlighted by the annual report for 2022. 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 2022 (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 AND TECHNICAL ACHIEVEMENTS

In the following, we summarize the scientific and technical achievements obtained within Phase 2 (year 2022) - *Establishment of methods based on machine learning to determine the relevant characteristics* - in order to achieve the proposed scientific and technical objectives. The main objective of Phase 2 was to establish machine learning-based methods for determining relevant features for software defect prediction.

1. Extracting features (attributes) for software defect prediction. To mitigate the negative impact of manual feature (attribute) extraction on SDP, we set out to investigate deep learning models that could automatically learn features from semantic and syntactic representations of source code. Unlike many approaches that use traditional metrics, we propose new input features for our models, namely: syntactic and semantic representations of the source code, and new cohesion- and coupling-based metrics for SDP. Moreover, the systematic use of characteristics (attributes) specific to defect types is an original perspective.

We have considered two methods of feature determination: (a) **automatic feature extraction.** Within this type of methods, the goal is to automatically learn/extract both semantic and syntactic features from the semantic and syntactic representations of the source code using deep learning models. The high-dimensional vectors representing the source code of software modules represent the input data to models that will extract meaningful features for SDP; and (b) **manual feature extraction** by defining new software metrics for SDP, specifically cohesion- and coupling-based metrics. These metrics are created based on existing software metrics, semantic and syntactic representations generated by Doc2Vec, LSI, Graph2Vec and Code2Vec and their combination. The relevance of the determined feature set will be evaluated from a software maintenance perspective, on a series of case studies targeting complex open-source software products with a long development history available for study.

2. Machine learning models and techniques for software defect prediction. To handle the unbalanced nature of SDP, we approach the problem from two perspectives that are new to the SDP literature, the one-class classification (OCC) (or anomaly detection) and one-shot learning (OSL) perspectives. Our first approach consists of using models to identify defective instances as anomalies relative to the majority class of non-defective instances. In addition to applying these models in an OCC scenario, we also aim to adapt the OSL methodology mainly used in computer vision. Our goal is to exploit the properties of OSL based models (Few-shot learning, N-shot learning) to be trained using fewer training instances.

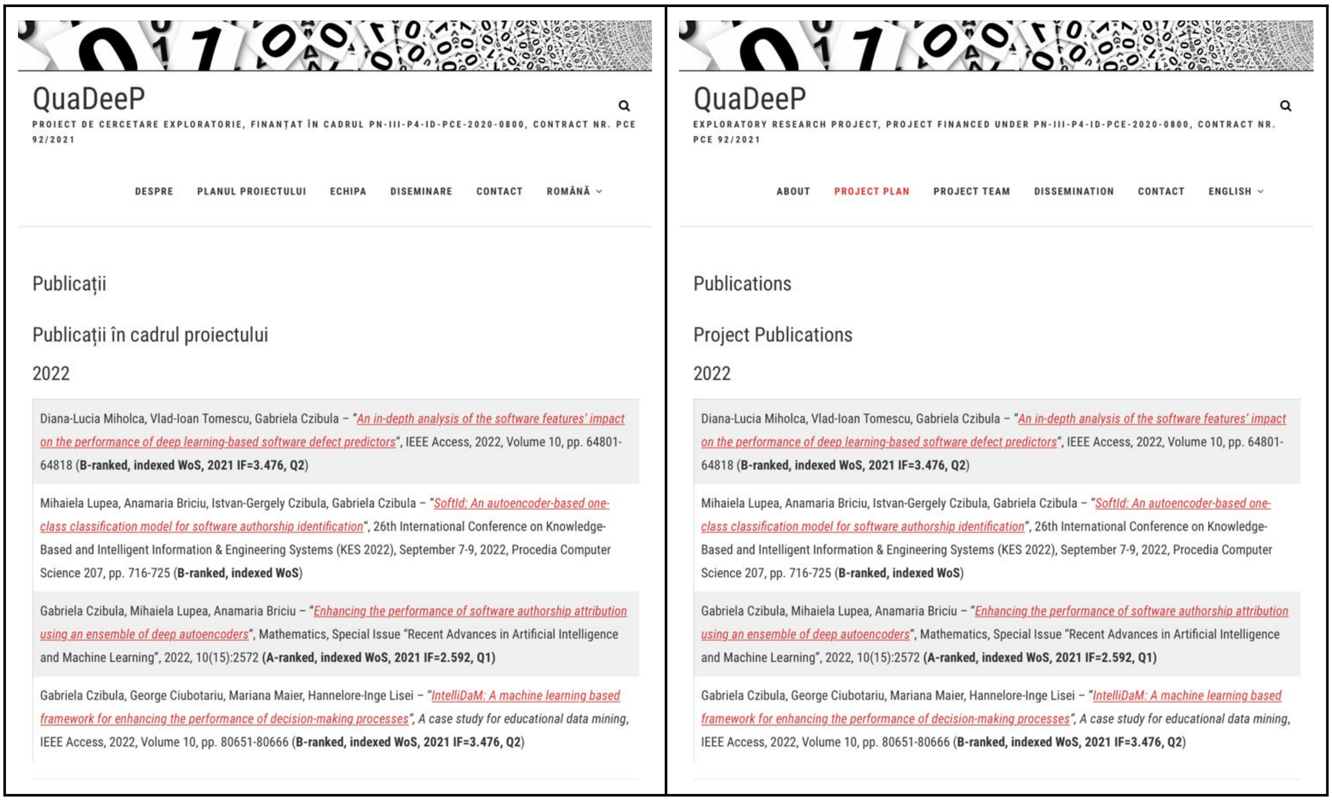
2 DISSEMINATION

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.

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.



QuaDeeP

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

Publicații

Publicații în cadrul proiectului

2022

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**)

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 (KES 2022), September 7-9, 2022, Procedia Computer Science 207, pp. 716-725 (**B-ranked, indexed WoS**)

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)**

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 (**B-ranked, indexed WoS, 2021 IF=3.476, Q2**)

QuaDeeP

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

Publications

Project Publications

2022

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**)

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 (KES 2022), September 7-9, 2022, Procedia Computer Science 207, pp. 716-725 (**B-ranked, indexed WoS**)

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)**

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 (**B-ranked, indexed WoS, 2021 IF=3.476, Q2**)

Figure 1 - Scientific publications obtained within the reported phase of the project on the project's website, in the Romanian version (left) and the English version (right)

2.2 SCIENTIFIC PUBLICATIONS

The table below presents the list of scientific publications obtained within Phase 2 of the QuaDeep project (2022).

[L1]	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**)
[L2]	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**)
[L3]	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**)
[L4]	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**)

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

2.3 PRESENTATIONS

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.

Table 2 - Conference presentations corresponding to publications in the previous section

3 CONCLUSIONS

This report presents the original results obtained as a result of the research carried out within the project in order to achieve the scientific and technical objectives proposed in the implementation plan for 2022 (Phase 2). For each objective provided in the implementation plan for 2022, we indicated the way in which the related activities were carried out. We summarize the results obtained within the project for the year 2022 as follows: (1) the development of methods based on deep learning for learning the relevant characteristics for software defect prediction; (2) introducing software metrics based on cohesion and coupling for software defect prediction; (3) the annual scientific and technical report; (4) scientific articles through which the original results obtained in Phase 2 of the project implementation were disseminated.

The dissemination of the results obtained within the project in 2022 was achieved by publishing 4 scientific articles: 3 publications in Web of Science (WoS) listed journals, with an impact factor (according to JCR 2021) of 3.476, 2.592 and 3.476, respectively; 1 publication in a WoS indexed international conference volume. Among the three publications in WoS indexed journals, we note that two are in the Q2 quartile and one publication is in the Q1 quartile.

As a result, the minimum performance criteria provided (at least one paper accepted for publication in an ISI journal with a high impact factor and at least 3 publications) was met. Furthermore, the project objectives for the year 2022 have been met, and all associated activities have been completed and carried out in accordance with the project implementation plan.