



**Îmbunătățirea calității sistemelor software folosind modele de învățare
profundă pentru predicția și detecția defectelor**

Raport științific și tehnic 2023

REZUMAT

COD PROIECT: PN-III-P4-ID-PCE-2020-0800

CONTRACT: PCE 92/2021

2023

REZUMATUL ETAPEI

Tema proiectului este aceea de predicție și detecție a defectelor în sisteme software și prezintă un interes internațional major, fiind de mare relevanță în timpul dezvoltării, testării și întreținerii sistemelor software. Predicția exactă a defectelor în versiuni noi de software ar îmbunătăți semnificativ performanța procesului de dezvoltare a software-ului în ceea ce privește costul, timpul și calitatea acestuia. Predicția defectelor în sisteme informatice, ajută la detectarea, urmărirea și rezolvarea anomaliilor din sistem care ar putea avea efecte negative asupra siguranței și vieții umane, în special în cazul sistemelor software critice. Predicția defectelor permite efectuarea modificărilor în stadii incipiente ale ciclului de viață a sistemului, ducând astfel la costuri mai mici și îmbunătățind satisfacția clienților sistemului software. Proiectul își propune dezvoltarea de tehnici de învățare profundă pentru predicția defectelor software, o problemă de importanță majoră în domeniul Ingineriei Software, în special în ceea ce privește ingineria software bazată pe căutare. Scopul principal este îmbunătățirea calității sistemelor software prin identificarea timpurie și precisă a modulelor software defecte, folosind modele și tehnici de învățare profundă. Astfel, obiectivul principal al acestui proiect este de a facilita activitățile de întreținere și evoluție a software-ului, cum ar fi testarea, revizuirea codului și evaluarea calității software-ului, prin identificarea automată a defectelor.

Obiectivul major al proiectului este îmbunătățirea calității sistemelor software folosind modele de învățare profundă pentru predicția și detectarea automată a defectelor software. Scopul specific al proiectului este creșterea acurateții în predicția defectelor software într-o nouă versiune a unui sistem software (predicția defectelor în cadrul aceluiași proiect software) și în principal reducerea proporției de defecte neidentificate (rata de rezultate fals negative). Considerăm două direcții principale de cercetare: (1) îmbunătățirea etapei de creare a reprezentărilor prin selectarea caracteristicilor măsurabile relevante pentru tipuri specifice de defecte (de exemplu, proprietăți semantice, metrice bazate pe coeziune sau cuplare conceptuală) și (2) extragerea automată a caracteristicilor semantice semnificative din reprezentările codului sursă (altele decât cele bazate pe AST).

Rezultatele estimate ale proiectului sunt: (1) rapoarte științifice și tehnice care conțin metodele originale de învățare automată dezvoltate pentru predicția defectelor software; (2) publicații științifice pentru diseminarea rezultatelor științifice obținute; (3) module software (incluse în sistemul software QuaDeep) care implementează modelele de învățare automată dezvoltate pentru predicția entităților software cu defecte.

În prezentul raport vom prezenta rezultatele originale obținute în urma cercetărilor efectuate în cadrul proiectului în scopul îndeplinirii obiectivelor științifice și tehnice propuse în planul de realizare a proiectului pe anul 2023. Vom indica stadiul curent al implementării proiectului, modul în care au fost îndeplinite activitățile asumate în planul de lucru precum și modalitatea în care au fost diseminate rezultate obținute în cadrul etapei 2023. Pentru a sumariza, rezultatele obținute în cadrul proiectului pe anul 2023 sunt:

- Dezvoltarea unor metode de tip OCC și OSL pentru predicția defectelor software.
- Dezvoltarea modulelor software pentru sistemul QuaDeep.
- Actualizarea paginii web a proiectului (www.cs.ubbcluj.ro/quadeep).
- 8 articole științifice: 1 publicație în revistă cotate ISI (Web of Science, WoS), cu factor de impact (conform JCR 2022) 1.6 situată în cuartila Q3; 6 publicații în volumul unor conferințe internaționale cotate B (conform clasificării CORE) care vor fi trimise în vederea indexării WoS; 1 publicație în volumul unei conferințe internaționale indexate IEEE (va fi trimisă spre indexare WoS).

Considerăm că obiectivele proiectului aferente anului 2023 au fost atinse, lucru dovedit de prezentul raport de cercetare. Obiectivele planificate pe anul 2023, cât și activitățile aferente acestora au fost realizate în totalitate, și desfășurate conform cu planul de realizare al proiectului. De asemenea, criteriul minim de

performanță prevăzută pe anul 2023 în ceea ce privește diseminarea rezultatelor (cel puțin un articol acceptat pentru publicare într-un jurnal ISI/WoS cu factor mare de impact și cel puțin 3 publicații) a fost îndeplinit.

1 INTRODUCERE

1.1 PROIECTUL QUADDEEP

Proiectul se concentrează pe dezvoltarea de tehnici de învățare profundă pentru *predicția defectelor software* (eng. *Software defect prediction - SDP*), o problemă de importanță majoră în domeniul Ingineriei Software, în special în ceea ce privește ingineria software bazată pe căutare. Scopul principal este îmbunătățirea calității sistemelor software prin identificarea timpurie și precisă a modulelor software defecte, folosind modele și tehnici de învățare profundă. Astfel, obiectivul principal al acestui proiect este de a facilita activitățile de întreținere și evoluție a software-ului, cum ar fi testarea, revizuirea codului și evaluarea calității software-ului, prin identificarea automată a defectelor. Tema proiectului prezintă un interes internațional major, fiind de mare relevanță în timpul dezvoltării, testării și întreținerii sistemelor software. Predicția exactă a defectelor în versiuni noi de software ar îmbunătăți semnificativ performanța procesului de dezvoltare a software-ului în ceea ce privește costul, timpul și calitatea acestuia. Proiectul prevede o soluție software, QuaDeep, care va integra noi metode de învățare profundă pentru identificarea defectelor software. Pentru a crește specificitatea modelelor, metodele de învățare vizate vor fi dezvoltate specific pentru tipuri de defecte. QuaDeep va oferi asistență dezvoltatorilor de software în predicția cu exactitate a defectelor software, contribuind astfel la îmbunătățirea calității software-ului și la facilitarea întreținerii și evoluției acestuia.

1.2 OBIECTIVE ȘTIINȚIFICE

Obiectivul principal al acestui proiect este îmbunătățirea calității sistemelor software folosind modele de învățare profundă pentru predicția și detectarea automată a defectelor software. Scopul specific este de a crește acuratețea predicției defectelor software într-o nouă versiune a unui sistem software și, în principal, de a reduce proporția de defecte care nu sunt detectate (rata fals negative). Proiectul este aplicativ și interdisciplinar, având următoarele obiective științifice și tehnice:

O1. Dezvoltarea și validarea științifică a unor metode originale bazate pe învățarea profundă pentru determinarea caracteristicilor relevante în predicția defectelor software. Taxonomii existente ale tipurilor de defecte (de exemplu, ODC, CWE, CVE) vor fi utilizate pentru identificarea caracteristicilor (atributelor) relevante care sunt specifice anumitor clase de defecte. Modelele de învățare profundă, cum ar fi *autoencoderii (AE)*, *rețelele neuronale convoluționale (CNN)* și *rețele Long Short Term Memory (LSTM)*, vor fi aplicate pentru a învăța automat caracteristicile semantice și sintactice din reprezentările codului sursă generate de Doc2Vec, *token-urile AST* (eng. *abstract syntax tree* - arbore de analiză abstractă a codului), Code2Vec și combinații ale acestora. Noi metrici software pentru predicția defectelor, bazate pe coeziune și cuplare, vor fi definite folosind metrici software existente și reprezentări semantice ale codului sursă, reprezentări generate de Doc2Vec, Latent Semantic Indexing (LSI) și Graph2Vec.

O2. Dezvoltarea și validarea științifică a unor modele și tehnici originale de învățare automată pentru predicția defectelor software. Modelele de învățare automată vor fi adaptate pentru anumite tipuri de defecte (vizate în O1) și astfel specificitatea modelelor va fi crescută, deoarece vor învăța să prezică doar o

anumită clasă de defecte. Mai precis, metodele de clasificare într-o singură clasă (OCC) și de învățare de tip *one-shot* (OSL) sunt avute în vedere pentru a gestiona problema principală a datelor dezechilibrate. Ca și clasificatori pentru o singură clasă (detectorii de anomalii) ne propunem să folosim autoencodere (AE), reguli de asociere relațională (RAR), RAR graduale (GRAR) și un clasificator hibrid bazat pe GRAR (HyGRAR). Pentru învățarea de tip *one-shot* (OSL) se vor aplica OSL cu rețele siameze, OSL Bayesian și N-Shot.

03. Dezvoltarea și validarea modulelor software QuaDeep. Furnizat sub formă de module software, QuaDeep va oferi o soluție pentru a asista dezvoltatorii, testerii și managerii de software în activitățile de întreținerea și evoluția software-ului, oferind informații care permit părților interesate să identifice posibilele defecte ale software-ului.

04. Contribuții la dezvoltarea cunoștințelor științifice prin diseminarea rezultatelor obținute în publicații științifice și pe site-ul web al proiectului.

2 DISEMINARE

2.1 SITE-UL WEB AL PROIECTULUI

Site-ul web al proiectului este dedicat prezentării proiectului, a echipei de cercetare și a rezultatelor obținute, putând fi accesate două versiuni: una în limba engleză (<http://www.cs.ubbcluj.ro/quadeep/>) și una în limba română (<http://www.cs.ubbcluj.ro/quadeep/ro/about-romana/>).

În ceea ce privește structura site-ului, acesta este împărțit după cum urmează: o pagină de prezentare a proiectului (**About/Despre**), o descriere succintă a planului de lucru (**Project Plan/Planul Proiectului**), o pagină de prezentare a echipei de cercetare (**Project Team/Echipa**) și o secțiune dedicată diseminării rezultatelor științifice și tehnice obținute (**Dissemination/Diseminare**), împărțită la rândul ei în pagini care conțin lista de publicații din cadrul proiectului (**Publications/Publicații**), rapoartele științifice și tehnice anuale (**Annual Reports/Rapoarte Anuale**) și prezentările din cadrul conferințelor (**Presentations/Prezentări**). De asemenea, pe site sunt incluse detaliile de contact pentru coordonatorul proiectului (pagina **Contact**).

Pe prima pagină a site-ului (**About/Despre**) se regăsește o scurtă descriere a proiectului și o prezentare a obiectivelor definite în cadrul acestuia. Pagina **Project Plan/Planul Proiectului** detaliază planul de lucru al proiectului, fiind precizate task-urile din cadrul fiecăruia din cele cinci pachete de lucru în care este împărțit planul. Prezentarea echipei de cercetare se regăsește pe pagina **Project Team/Echipa**, unde este inclusă o scurtă biografie academică pentru fiecare membru al echipei și link-ul către profilul său Google Scholar.

Secțiunea dedicată diseminării cuprinde: (1) o listă a publicațiilor din cadrul proiectului și a publicațiilor conexe, permanent actualizată cu cele mai recente publicații (pagina **Publications/Publicații**), (2) o pagină în care vor fi introduse rapoartele științifice și tehnice anuale (**Annual Reports/Rapoarte Anuale**) și (3) materiale utilizate pentru prezentările din cadrul conferințelor (documente și clipuri video, unde sunt disponibile - pagina **Presentations/Prezentări**).

2.2 PUBLICAȚII ȘTIINȚIFICE

Tabelul 1 prezintă lista publicațiilor științifice obținute în cadrul Etapei 3 (2023) a proiectului QuaDeep.

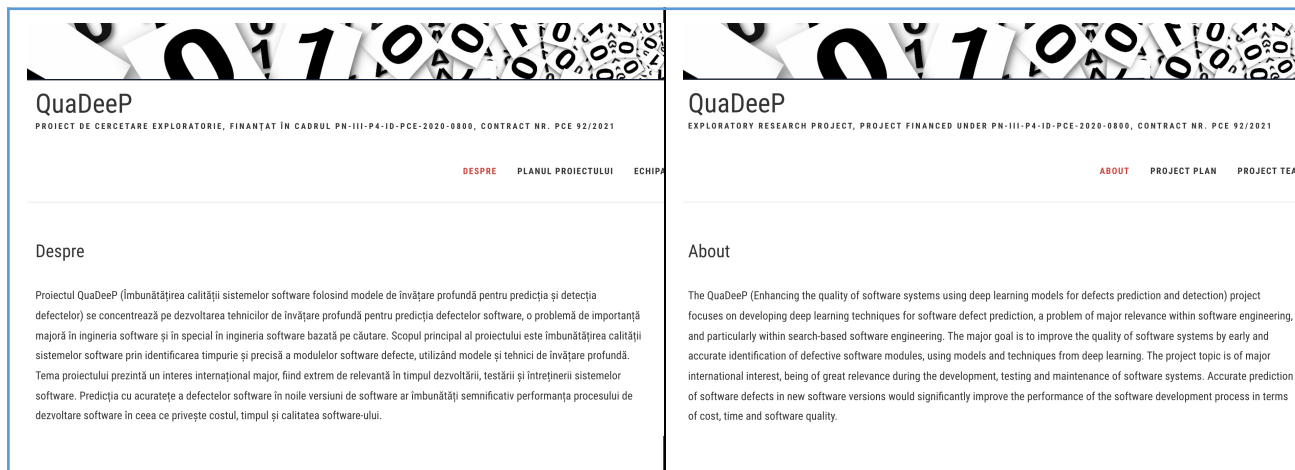


Figura 3 - Prima pagină a site-ului proiectului, versiunea în limba română (stânga) și engleză (dreapta)

[L1]	George Ciubotariu, Gabriela Czibula, Istvan Gergely Czibula, Ioana-Gabriela Chelaru, <i>Uncovering Behavioural Patterns of One: And Binary-Class SVM-Based Software Defect Predictors</i> , In Proceedings of the 18th International Conference on Software Technologies - ICSOFT; ISBN 978-989-758-665-1; ISSN 2184-2833, SciTePress, pages 249-257. DOI: 10.5220/0012052700003538 (B-ranked according to CORE classification, indexed WoS)
[L2]	Anamaria Briciu, Gabriela Czibula, Mihaiela Lupea, <i>A study on the relevance of semantic features extracted using BERT-based language models for enhancing the performance of software defect classifiers</i> , 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, <i>An unsupervised learning-based methodology for uncovering behavioural patterns for specific types of software defects</i> , 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, <i>Source-code embedding-based software defect prediction</i> , 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, <i>Using unsupervised learning for mining behavioural patterns from data. A case study for the baccalaureate exam in Romania</i> , Studies in Informatics and Control, vol. 32(2), pp. 73-84, 2023 (2022 IF=1.6, Q3)
[L6]	Imre-Gergely Mali, Gabriela Czibula, <i>Policy-Based Reinforcement Learning in the Generalized Rock-Paper-Scissors Game</i> , 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. <i>Temporal ensembling-based deep k-nearest neighbours for learning with noisy labels</i> . 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, <i>Im2Vid0: A Zero-Shot approach using diffusion models for natural language conditioned Image-to-Video</i> , 2023 IEEE 19th International Conference on Intelligent Computer Communication and Processing, 2023, in press (D-ranked according to CORE classification, indexed IEEE)

Tabel 1 - Lista publicațiilor științifice obținute în cadrul Etapei 3 (2023) a proiectului QuaDeep.

2.3 PREZENTĂRI

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=iR8D2FIG9W8
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

Tabel 2 - Prezentări la conferințe internaționale în cadrul Etapei 3 (2023) a proiectului QuaDeep..

3 CONCLUZII

În prezentul raport au fost prezentate rezultatele originale obținute în urma cercetărilor efectuate în cadrul proiectului în scopul îndeplinirii obiectivelor științifice și tehnice propuse în planul de realizare pe anul 2023 (Etapa 3). Pentru fiecare obiectiv prevăzut în planul de realizare pe anul 2021, am indicat modul în care au fost îndeplinite activitățile aferente.

Sintetizăm rezultatele obținute în cadrul proiectului pe anul 2023 ca fiind următoarele: (1) dezvoltarea unor metode bazate pe învățare profundă pentru predicția defectelor software; (2) introducerea unor metrice software bazate pe coeziune și cuplare pentru predicția defectelor software; (3) raport științific și tehnic anual; (4) articole științifice prin care s-au diseminat rezultatele originale obținute în cadrul Etapei 3 de implementare a proiectului.

Diseminarea rezultatelor obținute în cadrul proiectului în anul 2023 a fost realizată prin publicarea a 8 articole științifice: 1 publicație în revistă cotate ISI (Web of Science, WoS), cu factor de impact (conform JCR 2022) 1.6 situată în cuartila Q3; 6 publicații în volumul unor conferințe internaționale cotate B (conform clasificării CORE) care vor fi trimise în vederea indexării WoS; 1 publicație în volumul unei conferințe internaționale indexate IEEE (va fi trimisă spre indexare WoS).

Ca urmare, criteriul minim de performanță prevăzut (cel puțin un articol acceptat pentru publicare într-un jurnal ISI cu factor mare de impact și cel puțin 3 publicații) a fost îndeplinit. De asemenea, obiectivele planificate pe anul 2023, cât și activitățile aferente acestora au fost realizate în totalitate, și desfășurate conform cu planul de realizare al proiectului.