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

Raport științific și tehnic 2022

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

CONTRACT: PCE 92/2021

2022

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 acurateței 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, metrici 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 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 2022. 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 2022. Pentru a sumariza, rezultatele obținute în cadrul proiectului pe anul 2022 sunt:

- Metode bazate pe învățare profundă pentru învățarea caracteristicilor relevante în vederea predicției defectelor software.
- Metrici software bazate pe coeziune și cuplare pentru predicția defectelor software.
- Actualizarea paginii web a proiectului (www.cs.ubbcluj.ro/quadeep).
- 4 articole științifice: 3 publicații în reviste cotate ISI (Web of Science, WoS), cu factor de impact (conform JCR 2021) 3.476, 2.592 și respectiv 3.476; 1 publicație în volumul unei conferințe internaționale indexate WoS. Dintre cele trei publicații în reviste indexate WoS, menționăm faptul că două sunt situate în quartila Q2 și o publicație este situată în quartila Q1.

Considerăm că obiectivele proiectului aferente anului 2022 au fost atinse, lucru dovedit de prezentul raport de cercetare. Obiectivele planificate pe anul 2022, 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 prevăzut pe anul 2022 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 QUADEEP

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 REALIZĂRI ŞTIINŢIFICE ŞI TEHNICE

În cele ce urmează, sintetizăm realizările pe plan ştiinţific şi tehnic obţinute în cadrul Etapei 2 (anul 2022) - *Stabilirea unor metode bazate pe învățare automată pentru determinarea caracteristicilor relevante* - în vederea atingerii obiectivelor ştiinţifice şi tehnice asumate. Obiectivul principal al Etapei 2 a fost stabilirea unor metode bazate pe învățare automată pentru determinarea caracteristicilor relevante în vederea predicției defectelor software.

1. Extragerea caracteristicilor (atributelor) pentru predicția defectelor software. Pentru a atenua impactul negativ al extragerii manuale a caracteristicilor (atributelor) asupra SDP, ne-am propus să investigăm modele de învățare profundă pentru a învăța automat caracteristici din reprezentările semantice și sintactice ale codului sursă. Spre deosebire de multe abordări care folosesc metrici tradiționale, propunem noi caracteristici de intrare pentru modelele noastre, şi anume: reprezentări sintactice și semantice ale codului sursă și noi metrici bazate pe coeziune și cuplare pentru SDP. Mai mult, utilizarea sistematică a caracteristicilor (atributelor) specifice tipurilor de defecte este o perspectivă originală.

Am vizat două metode de determinare a caracteristicilor: (a) **determinarea automată a caracteristicilor**. Scopul este de a învăța/extrage automat, folosind modele învățare profundă, atât caracteristici semantice, cât și sintactice din reprezentările semantice și sintactice ale codului sursă. Vectorii de dimensiuni mari care reprezintă codul sursă al modulelor software sunt date de intrare în modele care vor extrage caracteristici semnificative pentru SDP; și (b) **determinarea manuală a caracteristicilor** prin definirea unor noi metrici software pentru SDP, metrici bazate pe coeziune și cuplare. Aceste metrici sunt exprimate pe baza metricilor software existente, reprezentărilor semantice și sintactice generate de Doc2Vec, LSI, Graph2Vec și Code2Vec și combinația acestora. Relevanța setului de caracteristici determinat va fi evaluată dintr-o perspectivă a mentenanței produselor software, pe o serie de studii de caz care vizează produse software complexe de tip open-source cu un istoric lung de dezvoltare disponibil pentru studiu.

2. Modele și tehnici de învățare automată pentru predicția defectelor software. Pentru gestionarea naturii dezechilibrate a SDP, abordăm problema din două perspective care sunt noi în literatura SDP, perspectiva OCC (eng. *one-class* classification, sau detectarea anomaliilor) și OSL (eng. *one-shot learning*). Prima noastră abordare este să folosim modele pentru identificarea instanțelor defecte ca anomalii în raport cu clasa majoritară de instanțe care nu sunt defecte. Pe lângă aplicarea acestor modele într-un scenariu OCC, ne propunem și să adaptăm metodologia OSL utilizată în principal în viziunea artificială. Scopul nostru este de a exploata caracteristicile modelelor bazate pe OSL (Few-shot learning, N-shot learning) pentru a fi antrenate folosind mai puține instanțe de antrenament.

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**).

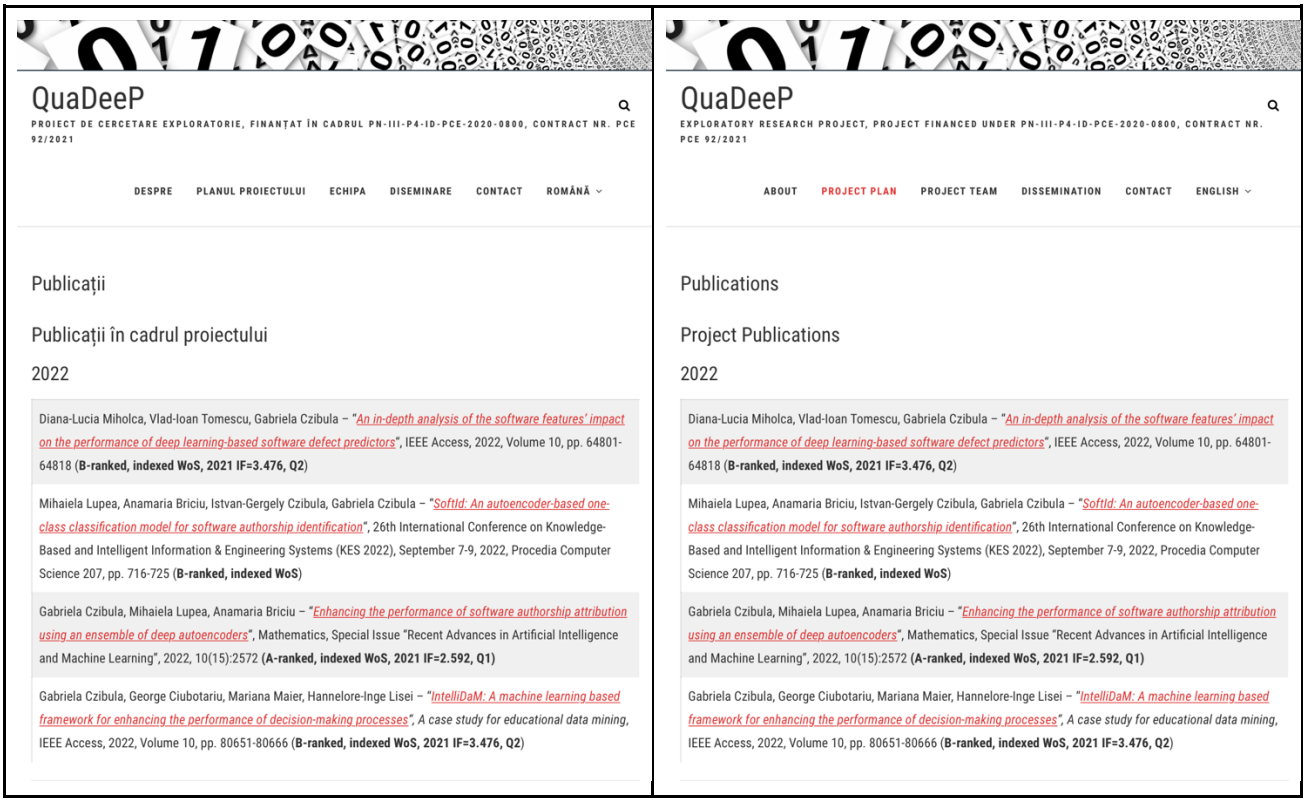

Figura 1 - Publicațiile din cadrul etapei raportate pe site-ului proiectului, versiunea în limba română (stânga) și engleză (dreapta)

2.2 PUBLICAȚII ȘTIINȚIFICE

Tabelul de mai jos prezintă lista publicațiilor științifice în cadrul proiectului QuaDeep, în cadrul Etapei 2 (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**)

Tabel 1 - Lista publicațiilor științifice în cadrul proiectului QuaDeep

2.3 PREZENTĂRI

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.

Tabel 2 - Prezentările din cadrul conferințelor aferente publicațiilor din secțiunea anterioară.

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 2022 (Etapa 2). Pentru fiecare obiectiv prevăzut în planul de realizare pe anul 2022, am indicat modul în care au fost îndeplinite activitățile aferente. Sintetizăm rezultatele obținute în cadrul proiectului pe anul 2022 ca fiind următoarele: (1) dezvoltarea unor metode bazate pe învățare profundă pentru învățarea caracteristicilor relevante în vederea predicției defectelor software; (2) introducerea unor metrici 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 2 de implementare a proiectului.

Diseminarea rezultatelor obținute în cadrul proiectului în anul 2022 a fost realizată prin publicarea a 4 articole științifice: 3 publicații în reviste cotate Web of Science (WoS), cu factor de impact (calculat conform JCR 2021) 3.476, 2.592 și respectiv 3.476; 1 publicație în volumul unei conferințe internaționale indexate WoS. Dintre cele trei publicații în reviste indexate WoS, menționăm faptul că două sunt situate în cuartila Q2 și o publicație este situată în cuartila Q1.

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 2022, cât și activitățile aferente acestora au fost realizate în totalitate, și desfășurate conform cu planul de realizare al proiectului.