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

Raport științific final
2021-2023

REZUMAT

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

CONTRACT: PCE 92/2021

REZUMATUL PROIECTULUI

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 și-a propus 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 a fost 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 a fost îmbunătățirea calității sistemelor software folosind modele de învățare profundă pentru predicția și detectarea automată a defectelor software. Scopul specific al proiectului a fost 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). Au fost considerate 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 proiectului sunt: (1) metode originale de învățare pentru predicția defectelor software; (2) module software care implementează modelele de învățare automată dezvoltate pentru predicția entităților software cu defecte; (3) rapoarte științifice și tehnice care conțin metodele originale de învățare automată dezvoltate pentru predicția defectelor software; și (4) publicații științifice pentru diseminarea rezultatelor științifice obținute. 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 toată perioada de implementare a acestuia (1 ianuarie 2021 - 31 decembrie 2023). Vom indica modul în care au fost îndeplinite activitățile asumate în planul de lucru precum și modalitatea în care au fost diseminate rezultate obținute în cadrul proiectului.

Pentru a sumariza, rezultatele obținute în cadrul proiectului (2021-2023) sunt:

- Metode și modele bazate pe învățare profundă dezvoltate pentru învățarea caracteristicilor relevante în vederea predicției defectelor software, metrici software bazate pe coeziune și cuplare pentru predicția defectelor software, metode bazate pe învățare profundă dezvoltate pentru predicția defectelor software, module software ale sistemului QuaDeeP.
- Pagina web a proiectului (www.cs.ubbcluj.ro/quadeep).
- **19** articole științifice: **6** publicații în reviste cotate ISI (Web of Science, WoS), cu factor de impact (2 situate în cuartila Q1, 2 situate în cuartila Q2 și 2 situate în cuartila Q3 conform JCR din anul publicării); **13** publicații în volumul unor conferințe internaționale (9 cotate B, 1 cotată C și 3 cotate D conform clasificării CORE) publicate/în curs de publicare în reviste indexate/trimise în vederea indexării WoS. Dintre cele 19 publicații, 3 sunt în curs de publicare.
- **8** prezentări la conferințe și workshop-uri internaționale.

Considerăm că obiectivele proiectului au fost atinse, lucru dovedit de prezentul raport de cercetare. Obiectivele planificate, activitățile aferente acestora cât și rezultatele asumate în planul de realizare al proiectului au fost realizate în totalitate, și desfășurate conform cu planul de realizare al proiectului. De asemenea, criteriul minim de performanță prevăzut pe cei 3 ani (2021, 2022, 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) au fost îndeplinite. De asemenea, rezultatele științifice și tehnice obținute în cadrul proiectului sunt diseminate și pe pagina web a proiectului QuaDeeP (www.cs.ubbcluj.ro/quadeep).

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 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 şi clasificatori pentru o singură clasă (detectori de anomalii) ne propunem să folosim autoencoderi (AE), reguli de asociere relațională (RAR), RAR graduale (GRAR) și un clasificator hibrid bazat pe GRAR (HyGRAR). Pentru învățarea de tip *one-sho*t (OSL) se vor aplica OSL cu rețele siameze, OSL Bayesian și N-Shot.

O3. 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.

O4. Contribuții la dezvoltarea cunoștințelor științifice prin diseminarea rezultatelor obținute în publicații științifice și pe site-ul web al proiectului.

2 Diseminare

2.1 Pagina Web a Proiectului

Site-ul web al proiectului QuaDeeP (www.cs.ubbcluj.ro/quadeep) 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**), o secțiune dedicată rezultatelor proiectului (**Project results/Rezultatele proiectului**) 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**), rezumatul raportului final al proiectului (**Summary of Final Report/Rezumatul Raportului Final**), rezumatele rapoartelor științifice și tehnice anuale (**Summary of Annual Reports/Rezumatul Rapoartelor 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. Pe pagina **Project results/Rezultatele proiectului** sunt prezentate succint rezultatele obținute în cadrul proiectului.

2.2 Publicații Științifice și Prezentări

Tabelele 1 și 2 prezintă lista publicațiilor științifice și prezentărilor susținute în cadrul proiectului QuaDeep, în perioada 2021-2023.

2023	
[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. DOI: 10.5220/0012052700003538 (**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**)
2022	
[L9]	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**)
[L10]	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**)

[L11]	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**)
[L12]	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**)
2021	
[L13]	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), Procedia Computer Science 192, pp. 397-406 (**B-ranked, indexed Scopus**)
[L14]	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**)
[L15]	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, indexed WoS**)
[L16]	Diana-Lucia Miholca - "*New Conceptual Cohesion Metrics: Assessment for Software Defect Prediction*" *2021 23rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, Timisoara, Romania, 2021, pp. 163-170, doi: 10.1109/SYNASC54541.2021.00036. (**D-ranked, indexed WoS**)
[L17]	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, Q3**)
[L18]	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 (ICCP), Cluj-Napoca, Romania, 2021, pp. 367-374 (**D-ranked, indexed Scopus**)
[L19]	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, 2021, 9(22), 2870 (A-ranked, **2020 IF=2.258**, **Q1**)

Tabel 1 - Lista publicațiilor științifice din cadrul proiectului

2023	
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
2022	
7	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.
2021	
8	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 2021.

Tabel 2 - Prezentările susținute în cadrul proiectului

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 anii 2021-2023. Sintetizăm rezultatele obținute în cadrul proiectului ca fiind următoarele: (1) rapoarte științifice și tehnice care conțin metodele originale de învățare automată dezvoltate pentru predicția defectelor software și învățarea caracteristicilor relevante; (2) publicații științifice pentru diseminarea rezultatelor științifice obținute; (3) module software care implementează modelele de învățare automată dezvoltate pentru predicția entităților software cu defecte.

Diseminarea rezultatelor obținute în cadrul proiectului a fost realizată prin publicarea a **19** articole științifice și **8** prezentări la conferințe și workshop-uri internaționale. Dintre publicații, **6** sunt în reviste cotate ISI (Web of Science, WoS), cu factor de impact (2 situate în cuartila Q1, 2 situate în cuartila Q2 și 2 situate în cuartila Q3 conform JCR din anul publicării); **13** publicații în volumul unor conferințe internaționale (9 cotate B, 1 cotată C și 3 cotate D conform clasificării CORE) publicate/în curs de publicare în reviste indexate/trimise spre indexare WoS. Dintre cele 19 publicații, 3 sunt în curs de publicare.

Ca urmare, criteriul minim de performanță prevăzut pe anii 2021, 2022, 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. De asemenea, toate obiectivele planificate și activitățile aferente acestora au fost realizate în totalitate, și desfășurate conform cu planul de realizare al proiectului.