Raport Științific și Tehnic

Identificare:

Contract de finanțare nr. 34/2021
Contractor: Universitatea de Vest din Timișoara (UVT)
Cod: PN-III-P4-ID-PCE-2020-0407
Etapa 3: Ian - Dec 2023
Nume etapă: Transprecizie in calcul continuu, mediu izolat și sigur de execuție, lansare și validare

Autori ai raportului și membrii ai proiectului:

Dr. Gabriel Iuhasz
Dr. Silviu Panica
Dr. Adrian Florin Spătaru
Drd. Teodora Selea
Drd. Alexandru Munteanu

Director proiect:

Prof. Dr. Dana Petcu

Obiectivul proiectului

Principalul obiectiv științific al CoCo este de a permite calculul continuu Cloud-to-Edge. În special, CoCo intenționează să cupleze mecanismele de învățare (inclusiv învățarea federată) și descoperirea inteligentă a resurselor pentru a realiza un mediu de găzduire adaptiv capabil să ruleze atât pe cloud, cât și aproape de margine.

Principalele obiective tehnice sunt:

TO1. Crearea unui runtime colaborativ Cloud-to-Edge care poate fi implementat pe întregul continuum de calcul folosind generațiile IaC (Infrastructure as a Code). Platforma implementată va fi capabilă de reconfigurare autonomă (tratarea evenimentelor și descoperirii). Va fi dezvoltat un mediu sandbox pentru a testa comportamentul IaC și pentru a simula diferite implementări și scenarii.

TO2. Realizarea achiziția și analiza de date relevante pentru IaC. Ambele ar trebui să poată rula pe hardware eterogen (cloud și edge). De asemenea, sunt vizate descoperirea și adăugarea autonomă a resurselor edge și cloud. Asigurându-vă că toate metodele de analiză dezvoltate pot rula în mod fiabil pe hardware cu specificații foarte diferite, natura lor transparentă va fi explorată.

Rezultatele estimate ale proiectului

CoCo intenționează să cupleze învățarea automată federată și descoperirea inteligentă de resurse pentru a realiza un mediu de găzduire adaptabil, capabil să funcționeze atât în Cloud cât și în Edge. Se preconizează ca o platformă open-source ce demonstrează conceptul să fie rezultatul tangibil al proiectului CoCo.

Pagină web:

http://coco.hpc.uvt.ro

Rezumatul Etapei 3: Transprecizie in calcul continuu, mediu izolat și sigur de execuție, lansare și validare

Având în vedere că obiectivul principal al proiectelor CoCos este de a permite generarea IaC folosind metode care permit un grad mare de adaptabilitate și nu în ultimul rând performanță, am extins în continuare cercetările prezentate până acum. Acest capitol va detalia unele dintre progresele realizate în această perioadă de raportare la componenta Detectare evenimente. Această componentă permite detectarea oricăror evenimente anormale care pot afecta negativ generarea și performanța IaC. Având în vedere că una dintre principalele ținte pentru CoCo sunt dispozitivele Edge/Fog, este evident că această țintă a informat majoritatea cercetărilor și experimentelor efectuate până în acest moment. Acest raport nu este diferit. Configurația și designul experimental general este influențată în mare

măsură de faptul că evenimentele anormale vor avea un impact mai mare asupra dispozitivelor de margine/ceață, astfel încât mecanismele de detectare ar trebui să fie implementate cât mai aproape de dispozitivele de margine. În ultimul raport am detaliat o parte din munca depusă în aplicarea mecanismelor de adaptare transparente care pot adapta sau chiar genera modele predictive cu un compromis cunoscut și realizabil între timpii de inferență, costul de calcul și precizie. În acest raport, explorăm obiective suplimentare de optimizare pe lângă performanța predictivă directă.

Accentul nostru principal a fost crearea de experimente folosind aceleași metode de detectare a anomaliilor incluse în Motorul de detectare a evenimentelor, împreună cu mecanisme de adaptare transparente, adăugând în același timp capacitatea de a măsura consumul de energie nu numai timpii de inferență. Acest lucru ne-a permis să îmbunătățim mecanismele de optimizare a hiper-parametrilor dezvoltate în ultima perioadă de raportare.

După cum sa menționat în raportul anterior, IaC ridică provocări semnificative pentru orice serviciu de analiză și monitorizare. Pe baza concluziilor descrise în raportul anterior, noile noastre experimente sunt concepute pentru a genera un depozit de modele predictive cu un compromis cunoscut între performanța predictivă și consumul de energie. Consumul de energie al fiecărei metode poate fi legat de complexitatea computațională, cel puțin pentru unele modele.

Gradul de îndeplinire a obiectivelor etapei

Obiectivul științific și tehnic al etapei a fost îndeplinit în totalitate, cum vom prezenta în cele ce urmează.

Descriere științifică și tehnică

Reamintim pachetele de lucru și sarcinile stabilite pentru etapa 3 a proiectului:

WP2 Federated analysis and discovery
T2.4 Transprecision Machine Learning in Continuum Computing
WP3 Runtime and sandboxing
T3.3 Sandbox Runtime
T3.4 Final solution step, deployment and validation
WP4 Management, dissemination, exploitation and communication
T4.1 Day-to-day management, quality assurance, periodic reporting
T4.2 Dissemination, communication and networking
T4.3 Exploitation strategy, sustainability & innovation management

Design

Vom descrie în cele ce urmează designul general și scopul experimentelor noastre privind adaptarea algoritmică transprecisă. Am executat mai multe experimente al căror obiectiv

principal a fost să găsim, folosind Hyper-Parameter Optimization (HPO) ghidat, cei mai buni hiper-parametri pentru fiecare model, în scopul de a maximiza performanța predictivă.

Odată ce am avut cei mai performanti hiper-parametri pentru fiecare metodă ML, am proiectat un set de experimente pentru a evalua practic impactul energetic al diferitelor scenarii de antrenament și inferență pentru fiecare model. Pentru măsurarea utilizării energiei am folosit interfața RAPL (Running Average Power Limit) dezvoltată de Intel. Este utilizat pentru raportarea consumului de energie acumulat pentru diferite domenii de putere. Sunt acceptate procesoarele de calitate server, în mare parte din familia Intel Xeon (arhitectura Sandy Bridge după 2010). Aceste domenii depind în mare măsură de CPU utilizat. Figura 1 arată cum sunt organizate fiecare domeniu RAPL monitorizat. Domeniul pachetului conține informații despre nucleele CPU, memoria cache și GPU-ul integrat (dacă este disponibil). Domeniul DRAM adună informații despre memoria de lucru. Măsurătorile CPU și GPU integrate pot fi preluate folosind domeniile Core și Uncore în mod respectabil. Un domeniu suplimentar privind GPU-urile Nvidia poate fi obținut (Nvidia Volta 2018) folosind Biblioteca de management NVIDIA (NVML).

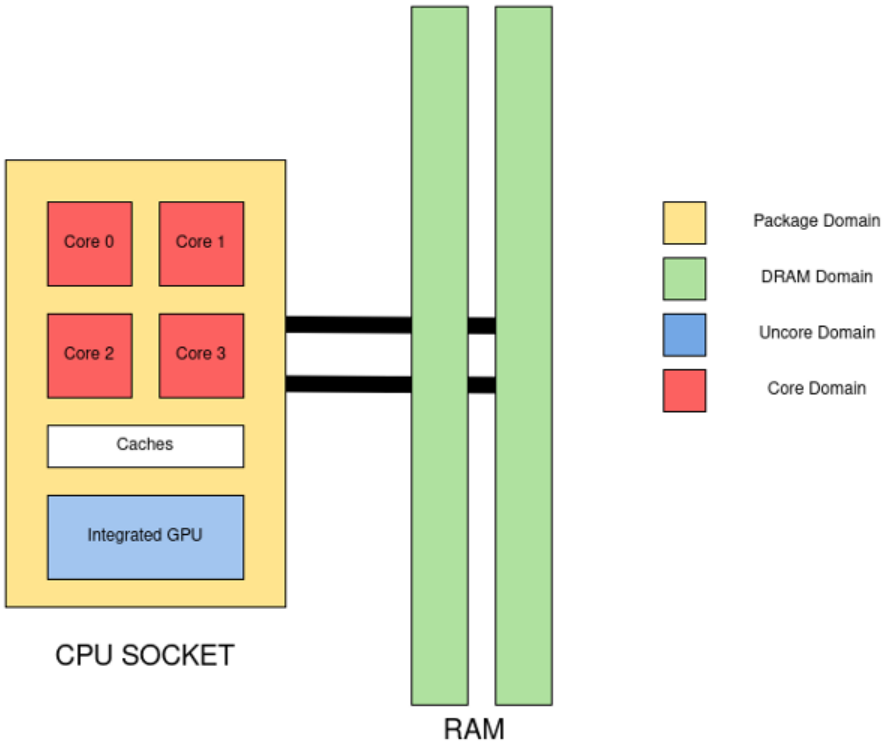


Figura 1 Domenii RAPL

Deoarece majoritatea cluster-ului nostru HPC disponibil se bazează pe procesoare AMD (în mare parte AMD EPYC 7702), nu toate domeniile RAPL sunt acceptate. De fapt, pentru a utiliza măsurătorile RAPL, au trebuit făcute unele modificări la kernel-ul Linux. Cu toate acestea, chiar și așa am reușit să culegem date de consum de energie doar pentru domeniul pachetului. Dacă există două socluri CPU pe hardware-ul fizic, atunci există două domenii de pachete care trebuie monitorizate (PKG 0, PKG 1). În cazul nostru, avem un singur soclu pentru

serverul nostru de testare. Specificațiile complete ale infrastructurii de testare pot fi găsite mai jos (Tabelul 1).

Tabelul 1 Infrastructura hardware 16 x HPE Proliant DL385 Gen10

Specification	Description
CPU	128x EPYC 7702 2,0 Ghz/core
RAM	1024 GB
Storage	2x 480 GB SSD
Inter-connect (storage and communications)	2x25GbE adaptors
Operating system	Ubuntu 22.04 LTS (custom kernel)

Experimentele inițiale au fost efectuate folosind instrumentul perf Linux utilizat pentru analiza performanței. Acceptă contoare de performanță hardware, puncte de urmărire, contoare de performanță software și sonde dinamice. Un exemplu de scenariu de utilizare poate fi găsit mai jos:

```
perf stat -e ,power/energy-gpu/,power/energy-pkg/ ede.py -f config.yaml
```

Listing 1 Scenariul de utilizare perf

Exemplul anterior colectează consumurile de energie (în Jouli) pentru domeniile GPU și pachet. În timp ce utilizăm perf pentru măsurătorile de energie, am descoperit că oferă capabilități limitate de măsurare pentru cazul nostru de utilizare. De exemplu, atunci când antrenăm/validăm un model predictiv, de obicei executăm o formă de validare încrucișată, în timp ce utilizând perf ar fi necesar să împărțim fiecare divizare în apeluri de perf separate, nu există o metodă fiabilă pentru a împărți măsurătorile altfel. Acesta este motivul principal pentru care am decis să instrumentăm codul nostru pentru a oferi măsurare cât mai precisă posibil. Pentru aceasta am folosit pyJoules. Această bibliotecă ne permite să măsurăm anumite fragmente de cod. În cazul nostru am dorit să măsurăm doar antrenamentul și predicția modelului, ignorând operatorii de preprocesare. Pentru fiecare plica de validare încrucișată, colectăm antrenament, consumul de energie de inferință și durata. Măsurăm în continuare consumul total de energie folosind perf ca formă de verificare a stării de spirit.

Experimente și rezultate

În special pentru aceste experiențe, am folosit cei mai buni parametri de metodă obținuți după o optimizare extensivă, ghidată, hiper-parametrică a fiecărei metode de detectare a metodei descrise în perioada de raportare anterioară. Experimentele actuale pot fi împărțite în 3 faze distincte. Fiecare dintre aceste faze este concepută pentru a descrie pe deplin compromisul dintre consumul de energie și performanța predictivă.

În prima fază, executăm pentru fiecare algoritm o validare încrucișată de 10 ori folosind metoda stratificată pentru împărțirea seturilor de antrenament și validare. Acest lucru ne permite să menținem aceeași distribuție de clasă pentru aceste seturi ca și setul original. Dacă am împărți datele aleatoriu, este posibil ca unele clase să fie subreprezentate sau chiar să lipsească din seturile de instruire sau validare. Acest lucru va duce la o evaluare a performanței distorsionate. Scopul principal al acestei faze este de a obține o linie de bază a performanței, văzând că cei mai buni hiper-parametri sunt utilizați în acest pas.

În a doua fază, executăm o eliminare recursivă a caracteristicilor pentru fiecare metodă. Antrenarea modelelor predictive inițial pe o caracteristică, apoi adăugând la fiecare iterație o caracteristică suplimentară până când obținem spațiul complet de caracteristici. Acest lucru ne permite să verificăm modul în care dimensiunea datelor de intrare afectează atât consumul de energie, cât și performanța predictivă.

Figura 2 și Figura 3 arată rezultatele privind consumul de energie în timpul antrenamentului și, respectiv, inferența. Putem vedea că AdaBoost are o creștere aproape liniară a consumului de energie, în timp ce RandomForest și într-o oarecare măsură CatBoost prezintă aproape aceeași utilizare a datelor, independent de spațiul de intrare. XGBoost și LightGBM au un consum de energie relativ zgomotos. Dintre toate metodele, cel mai mare consum de energie a fost obținut de XGBoost, în timp ce cel mai scăzut este CatBoost. Trebuie menționat că pentru fiecare iterație se folosește aceeași metodologie de validare încrucișată de 10 ori, ceea ce înseamnă că în total au fost antrenate 10 * 89 de modele în timpul acestui proces. Figura 4 prezintă scorurile obținute pentru fiecare iterație. Putem vedea că ambele metode funcționează bine cu un spațiu de caracteristici relativ mic, LightGBM performând de fapt mai bine în comparație cu modelul CNN. Acest lucru este de așteptat, deoarece modelele Deep Lerning sunt renumite pentru că necesită o cantitate substanțială de date de antrenament.

Figura 2 Eliminarea caracteristicilor recursive - Antrenament

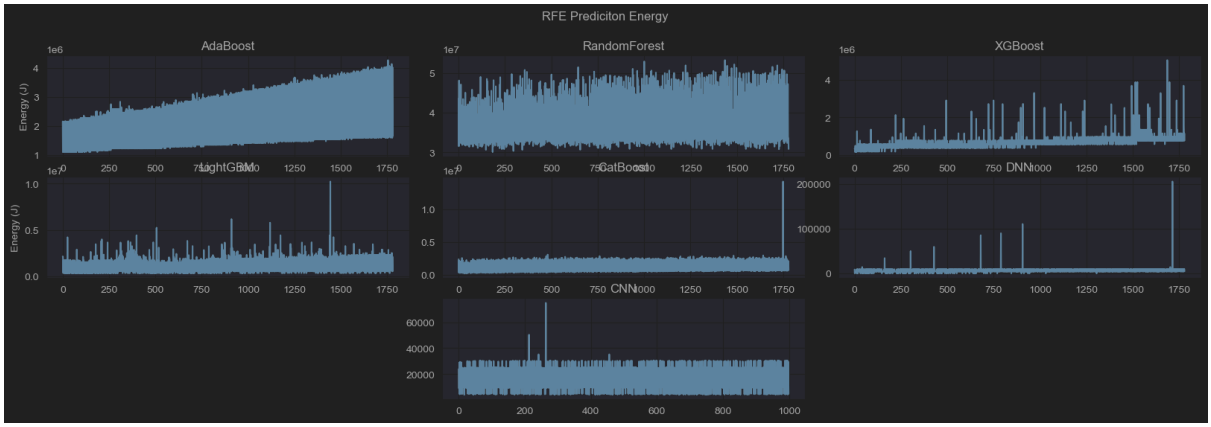


Figura 3 Eliminarea caracteristicilor recursive - Inferență

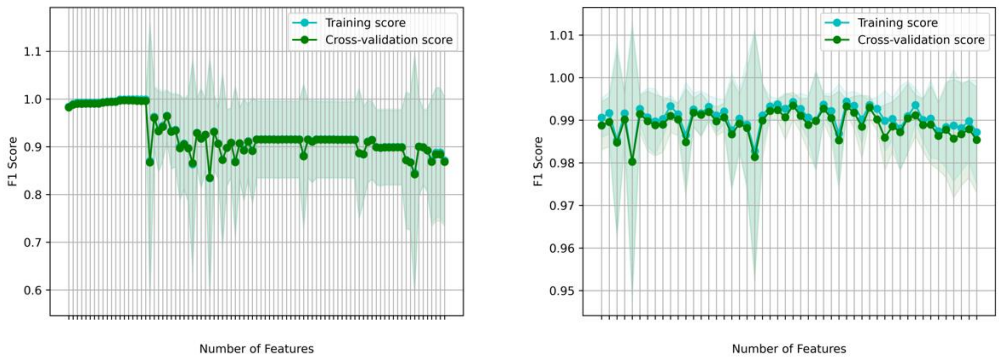


Figura 4 Scoruri RFE pentru LightGBM și CNN

În faza 3 luăm spațiul de caracteristici definit în timpul HPO și vedem ce impact are fiecare valoare a parametrului asupra performanței predictive și a consumului de energie. Figura 5 arată impactul valorilor parametrilor asupra consumului de energie. Putem vedea clar că unii parametri au un impact mult mai mare decât alții, un model clar poate fi observat pentru toate metodele. Aparent, modelele de învățare profundă sunt valori aberante. Acest lucru poate fi explicat oarecum din cauza inițializării GPU și a grupării datelor din memoria de lucru în GPU.

Din motive de concizie și simplitate, nu vom enumera rezultatele pentru fiecare metodă și parametrii acesteia. În schimb ne vom concentra pe LightGBM din simplul motiv că are o performanță destul de robustă atât din punct de vedere al predicției, cât și din punct de vedere al consumului de energie.

Ar trebui să menționăm în acest moment că toate experimentele au fost efectuate folosind o sămânță aleatorie fixă pentru a ajuta la reproductibilitatea experimentului. Această sămânță fixă a fost folosită pentru generarea diviziunilor de instruire și validare pentru validarea încrucișată și pentru inițializarea metodei. Un efect secundar interesant este modul în care unele perechi de seturi de date de antrenament și validare în timpul validării încrucișate provoacă creșteri ale consumului de energie. Acest lucru are loc în special în timpul antrenamentului și este legat de unele caracteristici de bază ale datelor. Cauza exactă nu este încă pe deplin înțeleasă și necesită cercetări suplimentare. În acest moment, cea mai probabilă cauză a acestui lucru ar putea fi legată de evenimente specifice (adică rânduri) și distribuția datelor și/sau entropia.

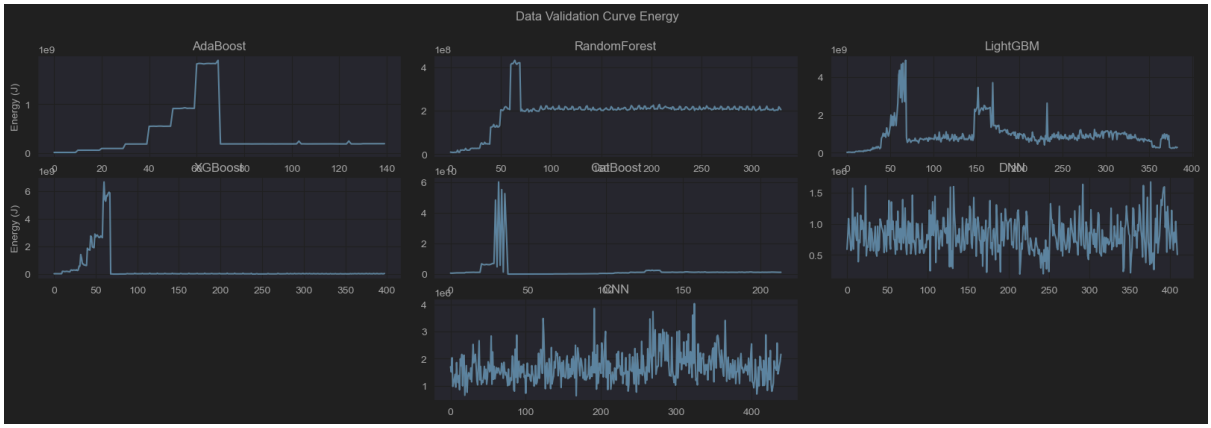
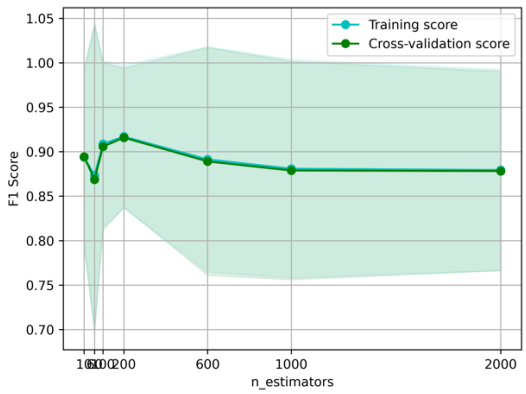
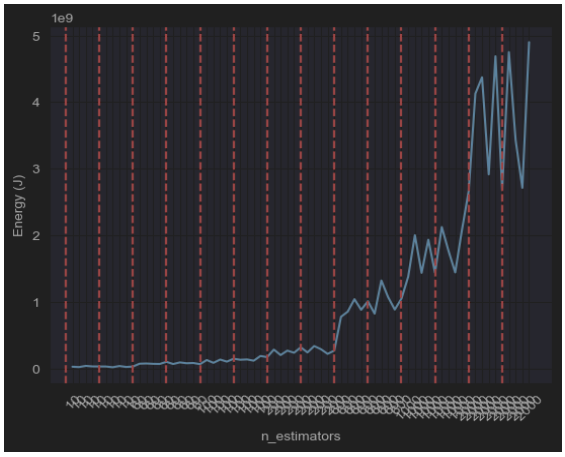


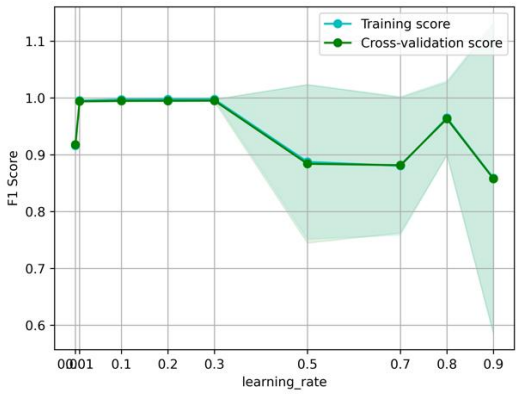
Figura 5 Curbele de validare pentru toate valorile hiper-parametrului

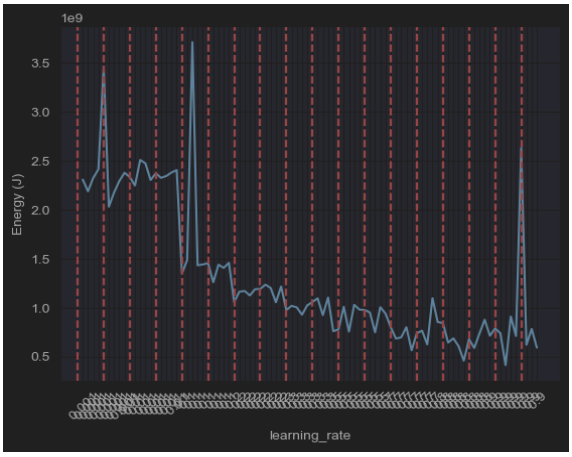
Figura 6 arată consumul de energie (partea stângă) și performanța predictivă folosind scorul F1 (partea dreaptă).

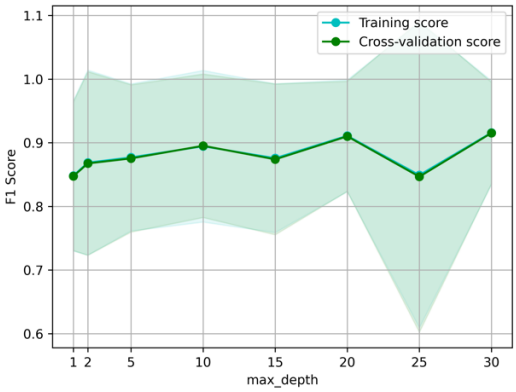
Tabelul 2 prezintă un raport de clasificare în cazul LightGBM, acesta reprezintă o defalcare detaliată folosind mai multe măsuri de punctare, inclusiv scorul F1 separat pentru fiecare clasă parțial. Coloana „suport” arată, de asemenea, numărul de apariții ale fiecărei clase în datele de validare. Aceste tipuri de rapoarte sunt generate pentru fiecare rând de antrenament în timpul validării încrucișate.











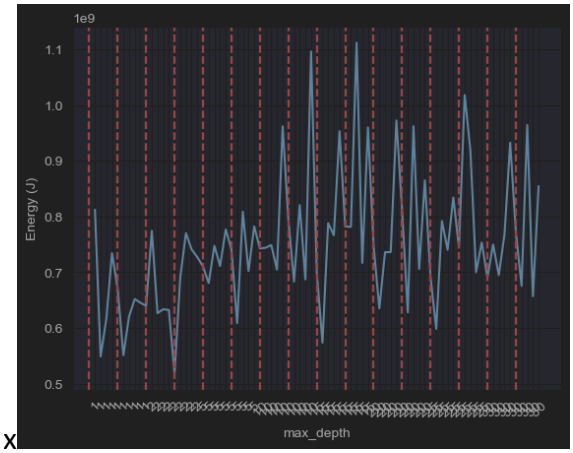


Figura 6 Consumul de energie LightGBM pe valoarea parametrului și Scorul F1 pe parametru

Tabel 2 Raport de clasificare pentru LightGBM (Fold 6)

classes	precision	recall	specificity	F1	geo	iba	support
normal	0.998	0.998	0.991	0.998	0.995	0.991	959.0
ddot	1.0	1.0	1.0	1.0	1.0	1.0	18.0
mem	1.0	0.923	1.0	0.960	0.960	0.915	26.0
cpu	0.928	1.0	0.999	0.962	0.999	0.99	13.0
copy	0.984	1.0	0.999	0.992	0.999	0.99	64.0
avg/total	0.997	0.997	0.997	0.997	0.997	0.997	0.997

Concluzii și discuții

Rezultatele prezentate în această secțiune ne permit să creăm un depozit atât de modele predictive pre-antrenate, cât și un set de hiper-parametri viabili, care oferă un compromis posibil între performanța predictivă și consumul de energie. În plus, am observat o corelație puternică între timpii de antrenament și de inferență în ceea ce privește consumul de energie. Acest lucru ne conduce la concluzia preliminară că cel puțin când vine vorba de metodele pe care le-am testat, un consum mai mic de energie duce la timpi de inferență mai mici. Cu toate acestea, aceasta este doar o ipoteză în această etapă, deoarece am folosit doar o singură configurație hardware pentru testare. Trebuie efectuat un experiment extins care să ia în considerare toate arhitecturile majore actuale ale CPU; X86_64, ARM, Putere.

În cazul modelelor bazate pe Deep Learning, optimizarea transprecisă suplimentară după antrenament poate fi aplicată și sub formă de cuantificare a greutății, grupare de greutate. Aceste optimizări nu au fost aplicate în această perioadă de raportare și vor necesita cercetări și experimente suplimentare.

Aplicarea calculului valorii Shapely pentru fiecare predicție oferă o clasare a cât de influentă este fiecare caracteristică în determinarea/detectarea oricăror evenimente anormale. Deși acest lucru nu este neapărat în conformitate cu paradigma inițială Belief Desire Intention (BDI)

prezentată pentru prima dată în perioadele anterioare de raportare, pe baza experienței noastre empirice, considerăm că este suficient pentru planificarea unor acțiuni semnificative de remediere a codului IaC.

Aplicații

Setul de date AgriSen-COG propus subliniază volumul semnificativ de date de observare a Pământului și provocările în procesarea lor eficientă, o preocupare centrală pentru obiectivele proiectului COCO. Articolul se concentrează pe tehnici avansate de procesare a datelor combinate cu modele de învățare profundă se aliniază cu obiectivul COCO de a regândi paradigmele de calcul, în special în manipularea eficientă a seturilor de date la scară largă în mediile de calcul cloud-to-edge.

Crearea unui set de date pe scară largă, multi-temporale și multi-țari pentru cartografierea culturilor prezintă capacitatea tehnică care se aliniază cu obiectivul științific al proiectului COCO de a permite calcularea eficientă a continuumului cloud-to-edge. Metodologia articolului pentru manipularea seturilor complexe de date reflectă ambiția COCO de a integra mecanisme de învățare și descoperire inteligentă a resurselor în platformele cloud și edge computing.

Activitatea desfășurată abordează provocări precum gestionarea seturilor de date la scară largă și asigurarea calității datelor, reflectând concentrarea proiectului COCO pe abordarea unor probleme similare în cloud-to-edge computing. Abordarea de a crea un set de date de referință și de a folosi metode de detectare a anomaliilor contribuie cu informații valoroase la proiectul COCO în gestionarea eterogenității resurselor și orchestrarea infrastructurii.

Mai mult, investigația privind detectarea culturilor folosind modele de învățare profundă din articol are implicații de anvergură pentru politicile agricole și cadrele socio-economice. Acest lucru rezonează cu scopul proiectului COCO de a influența domeniile științifice și socio-economice. Diseminarea constatărilor articolului și a setului de date poate servi drept model pentru proiectul COCO în angajarea cu comunitatea științifică mai largă.

Metodologia detaliată prezentată în articol pentru crearea setului de date AgriSen-COG, care implică colectarea datelor, procesarea și detectarea anomaliilor, se aliniază cu abordarea proiectului COCO pentru a depăși provocările tehnologice. Metoda articolului de manipulare și procesare a seturilor de date la scară largă poate informa strategia COCO în cloud-to-edge computing, în special în achiziția și analiza datelor.

Utilizarea resurselor Centrului de calcul de înaltă performanță UVT pentru prelucrarea datelor se aliniază cu obiectivele de utilizare a resurselor proiectului COCO. Acest lucru subliniază rolul vital al resurselor de calcul avansate și exemplifica aplicarea practică a calculului de înaltă performanță. O astfel de integrare nu este doar pertinentă, ci și susține cerințele de infrastructură a proiectului COCO, întărind argumentul pentru resurse de calcul robuste în gestionarea sarcinilor de prelucrare a datelor la scară largă. Integrarea unor astfel de resurse în metodologia articolului demonstrează o aplicare practică a calculului de înaltă performanță în prelucrarea datelor la scară largă, relevantă pentru cerințele de infrastructură COCO.

Diseminarea rezultatelor

1. Articole

Rezultatele activităților proiectului au fost publicate în următoarelor lucrări în volume ale unor conferințe:

- *D. Petcu*, Cognitive Cloud Continuum : Plenary Talk, 2023 IEEE 17th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, 2023, doi: 10.1109/SACI58269.2023.10158559.
- E. Chondrogiannis, E. Karanastasis, V. Andronikou, *A. Spătaru*, A. Nanos, A. Kretsis, P. Kokkinos (2023). Intent-Based AI-Enhanced Service Orchestration for Application Deployment and Execution in the Cloud Continuum. In: Papadopoulos, G.A., Rademacher, F., Soldani, J. (eds) Service-Oriented and Cloud Computing. ESOCC 2023. Lecture Notes in Computer Science, vol 14183. Springer, doi: 10.1007/978-3-031-46235-1_17
- M. Neagul, I. Nedelcu, *A. Munteanu*, Building a National Spatio-Temporal Datacube, IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, IEEE, doi: 10.1109/IGARSS52108.2023.10281405

în reviste indexate ISI în prima categorie:

- *T. Selea*, 2023. AgriSen-COG, a Multicountry, Multitemporal Large-Scale Sentinel-2 Benchmark Dataset for Crop Mapping Using Deep Learning, Remote Sensing 15, no. 12: 2980. https://doi.org/10.3390/rs15122980

Prezentări la conferințe:

- *D. Petcu*, Extreme Data Processing in Exascale Systems, Invited Keynote, 16th International Symposium on International Symposium on High-Level Parallel Programming and Applications (HLPP 2023), Cluj-Napoca, Iunie 2023
- *G. Iuhasz*, Performance and Compliance Anomaly Detection, 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing SYNASC 2023), Nancy, Septembrie 2023

Sunt în curs de publicare:

- A. Tuns, *A. Spătaru*, Cloud Service Failure Prediction on Google's Borg Cluster Traces Using Traditional Machine Learning, 2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), IEEE, in print

Articole în evaluare:

- *A. Spătaru*, Configuration of Container deployments on the Compute Continuum using Alien4Cloud, Scalable Computing: Practice and Experience[1], doi: 2024

[1] Indexată în WOS și SCOPUS, factor de impact 1.1 în iunie 2023, categoria ESCI

- *G. Iuhasz*, F. Fortis, *S. Panica*, A. Duma, Distributed platform for Cycle detection and analysis in Production Cyber-Physical Systems, AINA 2024

2. Organizare de evenimente științifice

A fost organizat un evenimente științific legat de tema proiectului CoCo: DBB 2023[2], The 4th Joint International Conference on Deep Learning, Big Data and Blockchain, August 2023, Marrakesh, chairs: *Dana Petcu* & Abdeslam En-Nouaary. Postproceeding-ul este publicat:

- M. Younas, I. Awan, S. Benbernou, D. Petcu, The 4th Joint International Conference on Deep Learning, Big Data and Blockchain, DBB 2023, August 2023, Springer, Lecture Notes in Networks and Systems (LNNS), volume 768, Softcover ISBN 978-3-031-42316-1, eBook ISBN 978-3-031-42317-8, http://doi.org/10.1007/978-3-031-42317-8

Buget

Costurile de personal sunt conforme cu efortul depus și estimările inițiale.
Costurile legate de conferințe internaționale sunt cele pentru câte o participare la HLPP 2023 și DBB 2023 (*Dana Petcu*). Costul auditului este cel din contractul stabilit în 2021.

Concluzii

Obiectivele etapei au fost îndeplinite în totalitate. Implementarea conceptelor a fost finalizată. Rezultatele etapei au fost parțial publicate în reviste si volume de conferințe.

Director
Prof. Dr. Dana Petcu

[2] https://www.ficloud.org/dbb2023/