Universitatea Transilvania din Brașov

TEZĂ DE ABILITARE

REZUMAT

Titlu: Calcul Afectiv Aplicat în Tratarea Fobiilor prin Realitate Virtuală și în Procesul de Învățare

Domeniul: Informatică

Autor: Conf. dr. Gabriela Moise

Universitatea Petrol-Gaze din Ploiești

BRAŞOV, 2025

În acest rezumat este prezentată pe scurt o selecție a principalelor contribuții științifice ale autoarei acestei teze de abilitare.

În **Capitolul 1** cu titlul "O Privire asupra Calcului Afectiv" este realizată o scurtă introducere în domeniul inițiat de Prof. Picard (1995) [1]. Pe scurt, domeniul calculului afectiv se ocupă cu dezvoltarea tehnologiilor care sunt conștiente de emoțiile umane și are aplicații în sănătate, jocuri, realitate virtuală, sisteme de recomandare, marketing, interfețe om-mașină, etc. Învățarea afectivă este un subdomeniu al calculului afectiv care consideră includerea tehnologiilor-conștiente de emoții în procesul de învățare [2]. Tehnologiile de învățare conștiente de emoțiile oamenilor sunt acele tehnologii care recunosc și manifestă emoții în situații de învățare. Mello&Graesser (2015) au studiat aceste tehnologii și au accentuat beneficiul utilizării lor în procesul de învățare [3]. Totodată, accentuăm îngrijorările legate de utilizarea calculului afectiv și anume, cele legate de riscurile etice. Aplicațiile care au integrate module de recunoașterea emoțiilor cu ajutorul inteligenței artificiale aparțin sistemelor de clasă de risc mare conform reglementării AI Act (2024) și au cerințe specifice [4].

Calculul afectiv cuprinde două subiecte majore: recunoașterea/detectarea emoțiilor și exprimarea emoțiilor de către mașini. Cercetările realizate de autoarea acestei teze includ recunoașterea emoțiilor în contextul tratării fobiilor utilizând aplicații de realitate virtuală și recunoașterea emoțiilor în contextul procesului de învățare.

În **Capitolul 2** cu titlul "Date pentru recunoașterea automată a emoțiilor" sunt prezentate pe scurt modele ale emoțiilor și tipurile de date care pot fi utilizate pentru recunoașterea acestora. Sunt prezentate modelul discret, modelul dimensional și modelul componențial ale emoțiilor, primele două fiind utilizate în modelele din această teză. Pentru cercetările menționate, am utilizat trei seturi de date ce conțin date biofizice. Două seturi de date reprezintă contribuția noastră și au fost utilizate în cercetări publicate de autoarea acestei teze în calitate de coautor în [5, 6, 7, 8, 9, 10, 11, 12].

Setul de date 1 conține semnale EEG (Electroencephalography), EDA (Electrodermal activity) și HR (Heart Rate) achiziționate de la 4 subiecți atât în mediu virtual cât și în mediul real. Pentru aceasta, am utilizat un dispozitiv Acticap Xpress Bundle cu 16 electrozi uscați pentru achiziționarea semnalelor de la canalele: FP1, FP2, FC5, FC1, FC2, FC6, T7, C3, C4, T8, P3, P1, P2, P4, O1 și O2. Semnalele EDA și HR au fost achizționate cu ajutorul unității GSR a unui dispozitiv Shimmers Multi-Sensory.

Setul de date 2 conține semnalele GSR (Galvanic Skin Response), HR și RR (Respiration Rate) achiziționate de la 5 subiecți în două situații. În prima, am măsurat GSR, HR și RR ca referință în timp ce subiecții efectuau următoarele acțiuni: respirație adâncă, mișcări ale capului spre

stânga, dreapta, sus, jos, clic cu mâna dreapta pe controlerul HTC Vive și ridicarea mâinii drepte. În cea de a doua situație am efectuat aceleași măsurători în timp ce subiecții au jucat un joc video cu realitate virtuală. Pentru măsurarea HR și GSR am utilizat dispozitivul Shimmer3 GSR+ Unit (https://www.shimmersensing.com/product/shimmer3-gsr-unit/). RR a fost calculată pe baza distanței dintre două dispozitive tracker HTC Vive.

Setul de date 3 este un set de referință și anume setul de date DEAP (Database for Emotion Analysis using Physiological signals) prezentat în [13].

Capitolul 3 cu titlul "Sisteme conștiente de emoții cu realitate virtuală de terapie prin expunere pentru tratatarea fobiilor" este bazat pe articolele publicate de autoarea acestei teze în calitate de coautor în [5], [6], [7], [8], [9], [10], [11], [12], [14], [15], [16]. În prima parte a capitolului este prezentată importanța subiectului Terapie prin Expunere cu Realitate Virtuală (TERV) în tratarea fobiilor. Studiile noastre au fost concentrate pe tratarea acrofobiei, care are o incidență mare afectând 1 din 20 de indivizi conform studiului prezentat în [17]. Utilizarea tehnologiile de RV în tratarea fobiilor a fost pentru prima dată remarcată la sfârșitul anilor 1990 [18], de atunci mulți oameni preferând acest tip de terapie.

Subcapitolul "Modele de recunoaștere automată a emoțiilor pe baza setului de date DEAP" prezintă abordările noastre în ceea ce privește dezvoltarea modelelor de Recunoaștere Automată a Emoțiilor (RAE) utilizând setul de date DEAP.

În primul set de modele am recunoscut sentimentul de frică utilizând două paradigme pentru măsurarea intensității fricii: scala cu 2 nivele (absență/prezență) și scala cu 4 nivele (nu/mic/mediu/mare) [14]. Am construit cinci seturi de intrări pentru fiecare paradigmă: Raw, Power Spectral Density, Petrosian Fractal Dimension, Higuchi Fractal Dimension, Approximate Entropy pentru semnalele EEG și înregistrările fiziologice (hEOG, vEOG, zEMG, tEMG, GSR, Respiration, PPG și temperatură). Au fost testate mai multe modele de învățare automată: patru modele de rețele neuronale adânci și Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Random Forest (RF), k-Nearest Neighbors (kNN) cu mai multe tehnici de selecție a atributelor - Principal Component Analysis (PCA), Sequential Feature Selector (SFS), Fisher selection. Modelul cel mai performant a fost clasificatorul RF fără selecția atributelor cu acuratețea – 93.13%, scorul F1– 93.11% în cazul scalei cu 2 nivele și acuratețea – 85.74%, scorul F1 score – 85.33% în cazul scalei cu 4 nivele. Performanțele au fost obținute pe un set de intrare generat din setul de date DEAP la care am aplicat Power Spectral Density (PSD) pentru toate cele 32 de canale EEG, în domeniul frecvențelor alpha, beta și theta și am calculat valori medii pentru alpha, beta și theta PSD în zone ale creierului: prefrontal (FP), AF (între FP și F), frontal (F), FC (între F și C), central (C), temporal (T), P

(parietal), CP (între C și P), O (occipital) și PO (între P și O). De asemenea, setul de intrare a conținut 8 atribute fiziologice din DEAP, hEOG, vEOG, zEMG, tEMG, GSR, Respirația, PPG și temparatura. Abordarea din [14] a fost extinsă la recunoașterea celor 6 emoții discrete definite de Ekman în [19]. Rezultatele sunt prezentate în [16].

În [15], am luat în considerare recunoașterea fricii numai din semnale GSR și HRV (Heart Rate Variability) extrase din setul de date DEAP și am propus un protocol pentru extragerea atributelor. Protocolul a constat în segmentarea fiecărei înregistrări în două moduri: în trei ferestre fără-suprapuneri și în cinci ferestre cu-suprapuneri. Astfel, am extras 33 tipuri de atribute pentru EDA și 7 tipuri de atribute pentru HRV, obținând în total 40 de tipuri de atribute.

Desfășurarea procesului utilizat în [15] este evidențiată în Figura 1.

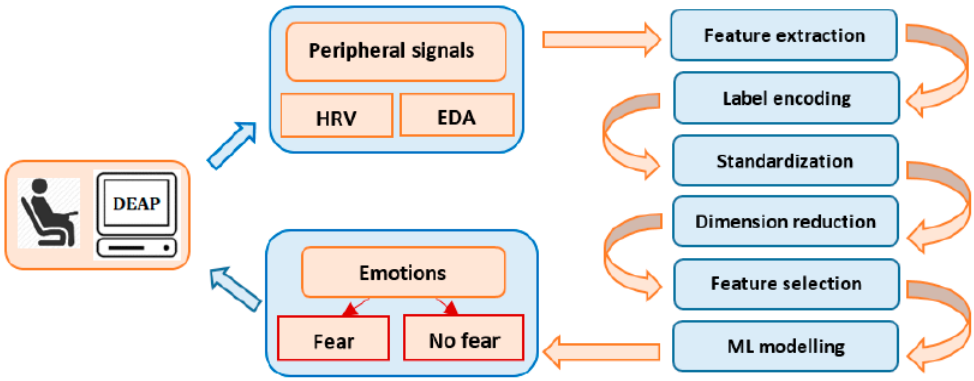

Figura 1. Desfășurarea procesului pentru recunoașterea automată a emoțiilor [15]

Abordarea cu ferestre fără-suprapuneri a constat în divizarea fiecărui segment de 60 de secunde în trei ferestre fără-suprapuneri de câte 20 de secunde fiecare. În abordarea cu ferestre cu-suprapuneri, fiecare segment a fost divizat în cinci ferestre de câte 20 de secunde fiecare și 10 secunde suprapunându-se cu alte segmente.

Pentru ambele abordări, cele mai bune performanțe, peste 89%, au fost atinse cu algoritmii SVM (Support Vector Machine) și GBT (Gradient Boosting Tree). În ceea ce privește scorul ROC AUC, au fost obținute cele mai bune rezultate astfel:

- Pentru setul de date fără-suprapuneri: reducere PCA + SVM – 93.5%.
- Pentru setul de date cu-suprapuneri: GBT – 91.7%.

În subcapitolul "Dezvoltarea de sisteme adaptive TERV cu recunoașterea emoțiilor" este prezentată abordarea noastră pentru tratarea fobiilor utilizând jocuri video cu realitate virtuală.

În primul joc cu RV, jucătorii au fost expuși la diferite niveluri de înălțime din joc în conformitate cu datele fiziologice achiziționate de la aceștia [6], [9], [14]. Arhitectura

sistemului TERV propus este prezentată în Figura 2 [6], [9], [14]. Am proiectat și antrenat două rețele neurale profunde (DNN), care au fost integrate în joc: prima rețea estimează intensitatea fricii jucătorilor, cea de a doua rețea determină nivelul de joc care se va afișa utilizatorului. În acest mod, utilizatorii sunt expuși gradual în timp real la diferite niveluri de înălțime adaptate gradului de intensitate a sentimentului de frică.

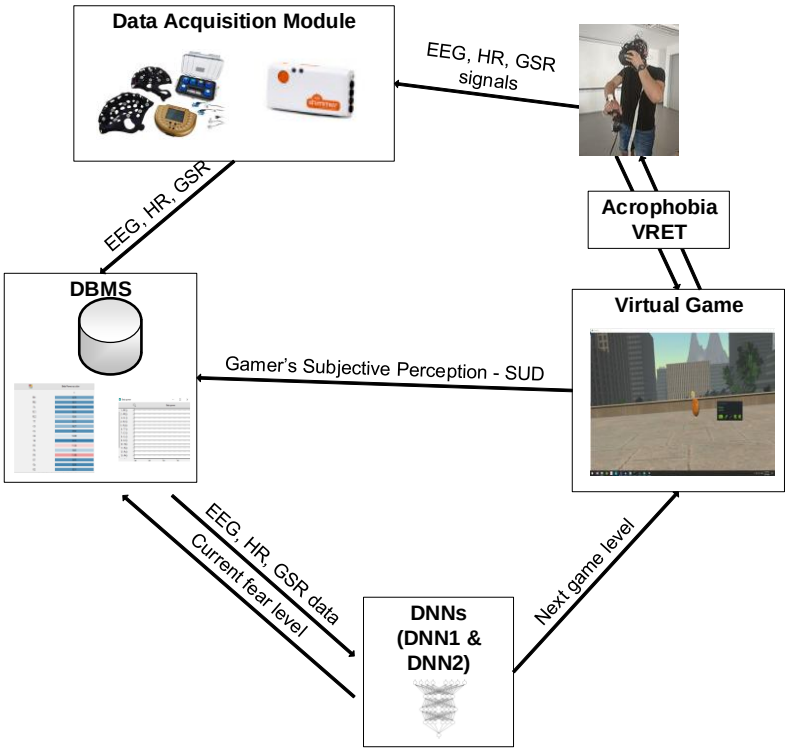

Figura 2. Arhitectura sistemul TERV conștient de emoțiile umane [9]

Sistemul conține un modul pentru achiziția semnalelor EEG, HR și GSR. Semnalele sunt preprocesate și transferate într-un sistem de baze de date, DBMS. Datele extrase din semnalele EEG, HR și GSR reprezintă intrările unui clasificator DNN1 pentru a determina nivelul curent de frică. Utilizând aceleași date și nivelul de frică dorit, cea de a doua rețea DNN2 generează predicția nivelului de joc potrivit unui anumit utilizator.

Scenariul de lucru al jocului este prezentat în Figura 3. Utilizatorul începe jocul cu nivelul l_0 , astfel nivelul curent de joc l_{cr} este l_0. Datele EEG, HR și GSR sunt înregistrate în timpul jocului. Când un utilizator termină un nivel de joc, datele achiziționate sunt transmise primei rețele DNN1, care face predicții asupra nivelului curent de frică. Nivelul următor dorit de frică se calculează în conformitate cu una din cele două paradigme: paradigma cu scala cu 2 nivele sau paradigma cu scala cu 4 nivele.

În cazul scalei cu 2 nivele formula de calcul pentru nivelul dorit de frică este:

if $fl_{cr} == 0$ then $fl_d = 0$

if $fl_{cr} == 1$ then $fl_d = 1$

În cazul scalei cu 4 nivele formula de calcul pentru nivelul dorit de frică este:

if $fl_{cr} == 0$ or $fl_{cr} == 1$ then $fl_d = fl_{cr} + 1$

if $fl_{cr} == 2$ then $fl_d = fl_{cr}$

if $fl_{cr} == 3$ then $fl_d = fl_{cr} - 1$

Nivelul dorit de frică și datele biofizice sunt intrări pentru cea de a doua rețea neurală profundă, DNN2, care generează următorul nivelul de joc (l_{pr}). Utilizatorii joacă nivelul de joc și datele biofizice sunt înregistrate. Algoritmul se termină când este atins un anumit număr de epoci de joc.

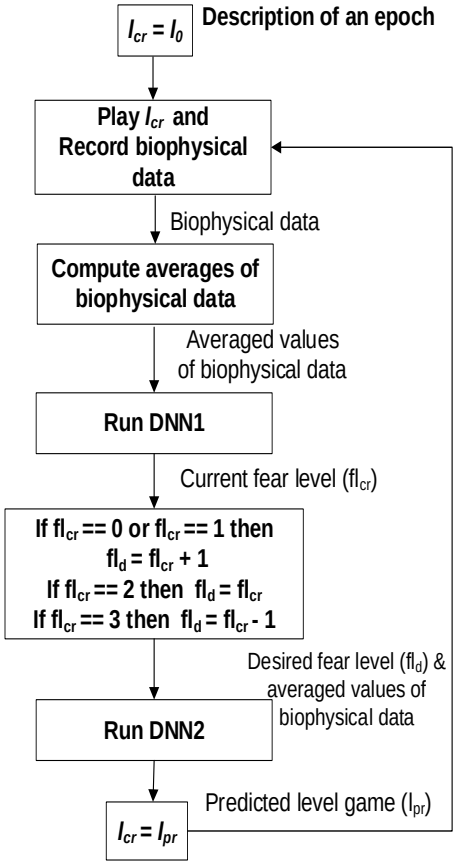

Figura 3. Algoritmul jocului TERV în cazul scalei cu 4 nivele [9]

Am testat trei tipuri de scale pentru evaluarea fricii, scale cu 2, 4 și 11 nivele, obținând următoarele performanțe pentru clasificarea fricii, în termeni de acuratețe:

- Scala cu 2 nivele - 95.51% utilizând un DNN cu 3 straturi ascunse – 300 de neuroni pe fiecare strat, activare RELU, stratul de ieșire – activare Sigmoid, funcția de pierdere Binary cross-entropy.
- Scala cu 4 nivele - 90.49% utilizând un DNN cu 3 straturi ascunse – 300 de neuroni pe fiecare strat, activare RELU, stratul de ieșire – funcția de activare Softmax, funcția de pierdere Logarithmical categorical cross-entropy.

- Scala cu 11 nivele – 85.09% utilizând un DNN cu 3 straturi ascunse – 150 de neuroni pe fiecare strat, activare RELU, stratul de ieșire – activare Softmax, funcția de pierdere Logarithmical categorical cross-entropy.

Pentru clasificarea nivelului de joc, am obținut următoarele valori pentru acuratețe:

- Scala cu 2 nivele – 98.72% utilizând un DNN cu 3 straturi ascunse – 300 de neuroni pe fiecare strat, activare RELU, stratul de ieșire cu 5 neuroni, activare Softmax, funcția de pierdere Logarithmical categorical cross-entropy.
- Scala cu 4 nivele – 98.67% utilizând un DNN cu 3 straturi ascunse – 150 de neuroni pe fiecare strat, activare RELU, stratul de ieșire cu 5 neuroni, activare Softmax, funcția de pierdere Logarithmical categorical cross-entropy.
- Scala cu 11 nivele – 98.75% utilizând un DNN cu 3 straturi ascunse – 150 de neuroni pe fiecare strat, activare RELU, stratul de ieșire cu 5 neuroni, activare Softmax, funcția de pierdere Logarithmical categorical cross-entropy.

Ulterior, am adus modificări la abordarea noastră și am adăugat clasificatori kNN, SVM, RF și LDA atât pentru determinarea nivelului de frică, cât și pentru determinarea următorului nivel de joc [10]. Am utilizat două metode pentru calculul acurateței clasificatorilor, dependent de utilizator/jucător și independent de utilizator/jucător.

Pentru clasificatorul de predicție a nivelului de frică, am obținut cea mai mare acuratețe de peste 98% cu validare încrucișată cu unul din algoritmi kNN sau RF, atât pentru modalitatea independent de jucător, cât și pentru modalitatea dependent de jucător. Aceeași tendință o găsim și în cazul clasificatorului pentru predicția nivelului următor de joc, unde cea mai mare acuratețe cu validare încrucișată a fost obținută cu RF, peste 99.75%.

Cele mai importante atribute pentru recunoașterea emoției au fost GSR, HR și valorile EEG în domeniul beta. Pentru predicția următorului nivel de joc, parametrul "nivelul de frică dorit" a influențat ieșirile clasificatorului.

Am evidențiat necesitatea dezvoltării de sisteme TREV cu recunoașterea emoțiilor utilizând o abordare centrată pe om în [6], [7]. Acest subiect este tratat în secțiunea "Abordarea centrată pe om pentru sisteme TREV", unde sunt prezentate metodologii pentru construirea de astfel de sisteme având în centru terapeuții și pacienții.

Complexitatea sistemelor TERV poate fi abordată printr-o arhitectura a sistemelor bazată pe holoni [11]. Secțiunea "O Arhitectură bazată pe holoni pentru sisteme TREV" descrie propunerea noastră pentru sisteme pentru tratarea fobiilor, numită PhoVRET.

Modul în care ar trebui achiziționate biosemnalele pentru a obține date valide este descris în subcapitolul "Provocări privind integrarea recunoașterii automată a emoțiilor în sisteme

TERV". Este definit conceptul de artefact ca fiind orice alterare în datele fiziologice ca urmare a unor acțiuni externe realizate de om, precum mișcări ale capului, mâinii, corpului, modificări fără nicio legătură cu efectele emoționale generate de anumiți stimului și am propus o metodă în trei pași pentru recunoașterea artefactelor [8]. Metoda constă în:

- Pasul I Măsurarea de referință a artefactelor.
- Pasul II Măsurarea artefactelor în timpul jocului.
- Pasul III Evaluarea potrivirii artefactelor.

Pentru validarea metodei propuse, am utilizat un al doilea joc de RV dezvoltat de echipa proiectului. La studiu au participat 5 subiecți, care au jucat jocul de mai multe ori și cărora le-am achiziționat semnalele fiziologice (GSR, HR și RR). Pentru HR și GSR am utilizat un dispozitiv Shimmer3 GSR+ și valorile RR au fost calculate în funcție de distanța dintre două dispozitive tracker HTC Vive.

Am verificat potrivirea între segmentele de referință cu artefacte și segmentele din timpul jocului prin calculul unor valori bias, eroarea medie absolută, eroarea procentuală absolută medie.

Rezultatele obținute au arătat că bias-ul este mai mic, dar nesemnificativ, pe segmentele de date aliniate față de segmentele înainte și după momentele cu artefacte. În cazul RR, valorile măsurate au fost identice în sesiunile de referință și de joc.

Având în vedere experiența acumulată în timpul experimentelor în laborator, am proiectat un protocol pentru achiziția semnalelor biofizice în medii de realitatea virtuală în scopul tratării fobiilor. Acest protocol a fost validat printr-un experiment la care au participat 7 subiecți. Am determinat cea mai influentă combinație de atribute extrase din EDA/HRV, dintr-un total de 32 de atribute. Valorile pentru SSE (Sum Squared Error – suma pătratelor erorilor) descresc după cum urmează: 3,827 (în cazul 1 atribut), 3,059 (2 atribute), 2,663 (3 atribute), 2,041 (4 atribute), 1,656 (5 atribute), 1,286 (6 atribute), and 1,031 (7 atribute).

Capitolul 4 cu titlul "Construirea de grupuri creative în învățarea colaborativă utilizând tehnici de inteligență artificială" se bazează pe articolele publicate de autoarea acestei teze de abilitare în calitate de coautor în [20], [21], [22], [23]. Cercetarea a plecat de la o problemă reală întâlnită în procesul educațional și anume, cum putem defini echipe de studenți de la profilul informatică pentru a dezvolta cele mai creative soluții software la problemele din lumea reală. În prima parte a capitolului este prezentată importanța subiectului, accentuând necesitatea lucrului studenților în echipe. Întrebarea de cercetare care a ghidat studiul din capitolul 4 al tezei este: Poate fi accentuată creativitatea grupului de studenți prin asigurarea

unui mediu instrucțional adecvat și prin organizarea optimă a studenților în grupuri utilizând tehnici ale inteligenței artificiale?

În secțiunea "Algoritm Q-Learning pentru construirea grupurilor creative de studenți", este prezentată o metodă bazată pe algoritmul Q-learning pentru a construi în mod optimal cele mai creative grupuri de învățare (GC-Q-Learning) introdusă în [20], [22], [23].

Algoritmul constă în:

1. Se construiește o matrice bi-dimensională Q pentru toate perechile posibile $\langle state, action \rangle$:

   $(c_1, c_2, \dots c_m, id\_group, action\_number, q)$

   O valoare pentru $action\_number$ egală cu *i* înseamnă că dacă un tip particular de student (descris prin vectorul său de creativitate $(c_1, c_2, \dots c_m)$) va fi mutat în grupul cu valoarea id_group egală cu *i*, atunci contribuția sa la creativitatea grupului este cuantificată prin q (în această etapă). Toate elementele din coloana q pot fi inițializate cu 0 sau cu o valoare random mică. Pe fiecare linie a matricei sunt incluse datele care corespund fiecărui tip de student, valorile pentru atributele studenților, numărul curent al grupului, numărul acțiunii, valoarea calculată pentru q *(care cuantifică un potential pentru creativitate).* Un tip particular de student poate avea mai multe linii corespondente, câte una pentru fiecare combinație $\langle current\ id\_group, action \rangle$

2. Se inițializează *optimal_policy* cu o politică inițială. În cazul nostru, politica optimală este reprezentată de gruparea optimală a studenților care maximizează creativitatea grupului. Gruparea inițială este setată de instructor și studenți, experiența noastră arată că studenții tind să se grupeze pe baza afinităților inter-personale..

3. Se grupează studenții și se realizează o sesiune de lucru, în care este evaluată creativitatea fiecărui grup și o valoare este asignată recompensei $R(s, a)$. Valorile pentru $R(s, a)$ sunt obținute cu ajutorul experților umani. Recompensa descrie potențialul creativității grupurilor. Matricea Q este actualizată pentru fiecare sesiune de lucru cu procedura de mai jos.

   procedure working_session_computation
   select action of (optimal_policy) /* student grouping*/
   compute $R(s, a)$
   compute table Q

4. Se analizează creativitatea fiecărui grup având în vedere un obiectiv global (politica de grupare optima), care se apropie de valoarea maximă posibilă pentru R, pentru fiecare grup sau pentru toate grupurile. Se reia pasul 3 dacă este necesar.

Politica optimală este definită prin tupluri de forma $(c_1, c_2, \dots c_m, id\_group)$.

Următoarele notații au fost utilizate:

n – număr de studenți

$c = (c_1, c_2, \dots c_m)$ - vector de creativitate, c_i reprezintă un atribut al studenților ce influențează creativitatea grupului, m – număr de atribute individuale

id_group – identificator al grupului

k – numărul de grupuri

$(c_1, c_2, \dots c_m, id\_group)$ – o stare (*s*) compusă din vectorul de creativitate și identificarea grupului

acțiunea (*a*) – acțiunea de a muta un student în alt grup la a cărei creativitate ar putea contribui cel mai mult

Q – exprimă calitatea asocierii dintre stare și acțiune, în sensul scopului nostru, de a construi cele mai creative k grupuri

R - recompensa este valoarea creativității grupului și variază între 1 și 5.

Algoritmul a fost testat prin executarea mai multor cazuri de utilizare. Pentru studenți am considerat două atribute, nivelul individual de creativitate si nivelul de motivație. Pentru grupurile de studenți am calculat valorile Q. Rezultatele au arătat că algoritmul este o soluție validă pentru a grupa studenții asigurând o creativitate ridicată a grupurilor.

În subcapitolul "Un sistem multiagent pentru construirea grupurilor creative de studenți", descriem propunerea noastră pentru un sistem multiagent, numit GC-MAS și publicat în [20], [23]. Sistemul propus integrează algoritmul GC-Q-Learning. Cei cinci agenți care compun sistemul GC-MAS sunt: agentul de comunicare (CommGC); constructorul de grupuri creative (BuildGC); agentul de evaluare a creativității grupurilor (EvalGC); agentul stimulator (EnvrGC) și agentul facilitator (FcIGC).

Subcapitolul cu titlul "Clasificatori Bayes pentru construirea grupurilor creative de studenți" are ca scop prezentarea unui model și metode de a grupa studenții într-un mod optim în situații de învățare colaborativă utilizând clasificatori Bayes [21]. Ideea principală a abordării noastre este considerarea caracteristicii grupului cea mai relevantă pentru scopul propus și repetarea grupării studenților pe baza valorilor unor atribute individuale. Modelul nostru conține 3 stagii prezentate în Figura 3.

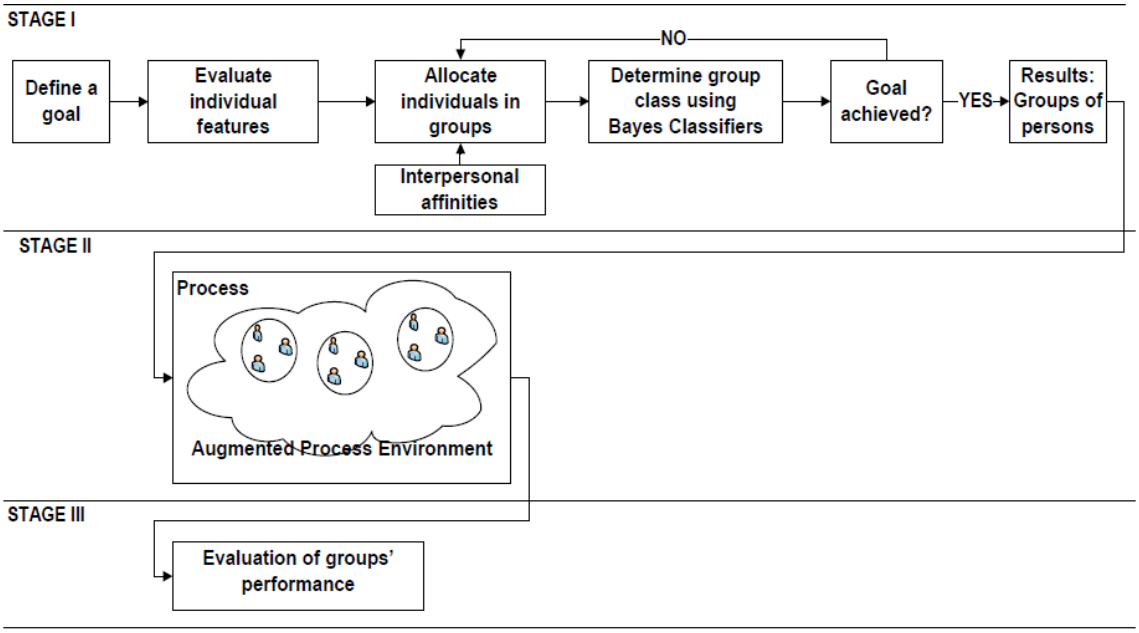

Figura 3. Model pentru construirea celor mai performante grupuri colaborative [21]

Modelul a fost validat într-un scenariu din lumea reală: scopul nostru a fost să grupăm 20 de studenți ai programului de studiu informatică în cele mai creative echipe. Am utilizat clasificatori Bayes, unul pentru predicția clasei de creativitate pentru fiecare student și altul pentru predicția clasei de creativitate pentru fiecare grup. Rezultatele obținute au confirmat o creștere a performanțelor la învățare a studenților prin construirea de echipe conform modelului nostru.

Capitolul 5, cu titlul "Recunoașterea automată a emoțiilor în procesul de învățare" prezintă rezultatele cercetării începută în anul 2023 în proiectul cu titlul Învățare Afectivă: Beneficii și Riscuri Etice în Învățământul Superior (Affective Learning: Benefits and Ethical Risks in Superior Education - ALBER) coordonat de autoarea acestei teze. Rezultatele cercetărilor au fost publicate în articolele [24], [25]. Capitolul începe cu o scurtă introducere privind importanța domeniului Învățare Afectivă. Am evidențiat rolurile emoțiilor în context academic și multitudinea lor: anxietate, speranță, lipsă de speranță, ușurare, plăcerea învățării, mândria succesului, furie, rușine, plictiseală, surpriză, tristețe, frustrare, confuzie, fericire, frică, bucurie, dezgust, interes, curiozitate, dispreț, încântare și entuziasm [26], [27], [28].

În subcapitolul "Roluri ale recunoașterii automată a emoțiilor în educație" sunt prezentate 6 categorii de roluri ale sistemelor RAE în educație [24]:

- Dezvoltarea Sistemelor Inteligente de Tutorat cu abilități emoționale, capable să detecteze emoțiile și să reacționeze adecvat la acestea.
- Sprijinirea angajării și motivării studenților și profesorilor.
- Evaluarea învățării (detectarea încercărilor de fraudă academică de către studenți).

- Evaluarea predării (profesorii trebuie să fie conștienți de impactul emoțiilor asupra procesului de predare-învățare).
- Construirea de medii de învățare adecvate procesului.
- Sprijinirea studenților cu nevoi speciale (suferind de ADHD, anxietate, etc.).

Pe baza setului de date DEAP, am construit un model 1D-CNN pentru recunoașterea a șapte emoții des întâlnite în context academic, plictiseală, confuzie, frustrare, curiozitate, entuziasm, concentrare și anxietate. Contribuția noastră a constat în obținerea unul model performant pe baza a numai a 5 canale EEG [25]. Modelul și rezultatele sunt descrise în subcapitolul "Model 1D-CNN pentru recunoașterea emoțiilor pe baza a 5 canale EEG – setul de date DEAP". Am plecat de la modelul RAE propus de Akter și alții (2022) care utilizează 14 canale EEG (FP1, FP2, AF3, Fz, F3, F4, F7, F8, FC1, C4, P3, P4, PO3, PO4) cu performanțele în ceea ce privește acuratețea: pentru valență - 99.89%, și pentru excitare – 99.83% [100].
Pentru a obține modelul RAE am folosit procedura din Figura 4.

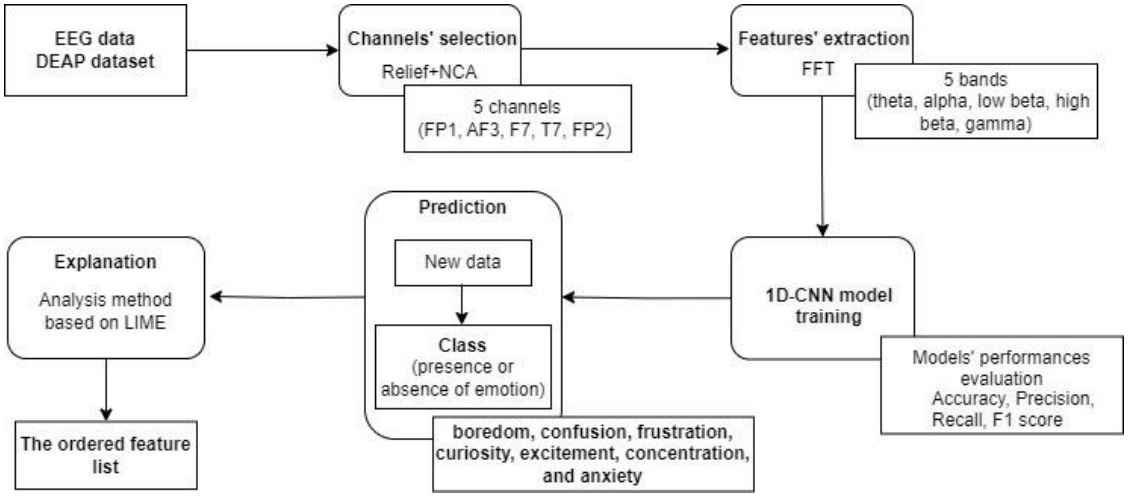

Figura 4. Procedura pentru obținerea modelului RAE în procesul de învățare [25]

Modelul nostru utilizează 5 canale din semnalele EEG și anume FP1, AF3, F7, T7, FP2. Am efectuat extragerea caracteristicilor utilizând FFT (Fast Fourier Transform), după care am aplicat standardizare.
Performanța modelului 1D-CNN bazat pe 5 canale EEG este prezentată în Tabelul 1. Timpul de antrenare a fost 9342.58 secunde și valorile pentru acurateță peste 99.21%. Scorul F1 variază de la 91.76%, in cazul anxietății, la 99.07%, în cazul concentrării.

Tabelul 1. Performanțele modelului 1D-CNN bazat pe 5 canale EEG [25]

Emotion	Performanță (%)				Timp de antrenare (secunde)
	Acuratețe test	Precizie	Rechemare	Scorul F1	
Plictiseală	99.64	94.62	97.41	95.97	8281.34
Confuzie	99.70	99.17	96.19	97.63	6004.65
Frustrare	99.66	98.98	95.13	96.97	5690.35
Curiozitate	99.80	99.74	96.83	98.24	8675.36
Entuziasm	99.91	98.74	97.75	98.24	5790.36
Concentrare	99.70	99.16	98.98	99.07	9342.58
Anxietate	99.21	95.03	88.96	91.76	6655.65

Am dezbătut subiectul eticii privind utilizarea RAE în procesul de învățare în subcapitolul cu titlul "Model etic pentru RAE în procesul de învățarea online". Modelul nostru etic, prezentat în [24] consideră 16 clase de riscuri etice asociate utilizării RAE în procesul de învățare:

- „Prejudecăți și discriminare,
- Rezultate nefiabile, incerte, nesigure sau slabe.
- Rezultate netransparente, inexplicabile, nejustificate sau total neprevizibile.
- Încălcarea vieții private prin (1) proprietatea și gestionarea incorectă a datelor cu caracter personal, (2) neacordarea și retragerea consimțământului și (3) supraveghere internă.
- Nedreptate și diviziune digitală.
- Înșelăciune.
- Manipularea și construirea de relații autoritare.
- Schimbări în percepția umană asupra realității, înțelegerii, expertizei și comportamentului natural.
- Portretizare eronată a ființelor umane și a emoțiilor.
- Negarea sau ocolirea autonomiei și drepturilor individuale (restricționarea capacității utilizatorilor de a-și exercita voința sau libertatea de exprimare, decizii nelibere și neinformate cu privire la utilizatori).
- Utilizare dual.
- Izolarea indivizilor, dezintegrarea conexiunilor sociale și dezumanizarea relațiilor interumane prin interacțiunea emoțională și socială cu sisteme de inteligență artificială de înaltă performanță, dar lipsite de conștientizare de sine.
- Dependență de o mașină.

- Riscul de a pierde simțul identității individuale.
- Înlocuirea profesorilor.
- Lipsa sustenabilității energetice."

Modelul etic (Figura 5) urmărește fluxul dedicat dezvoltării unui model de învățare automată cu considerarea în plus a eticii și a celor trei nivele definite în framework-ul lui Leslie din [29].

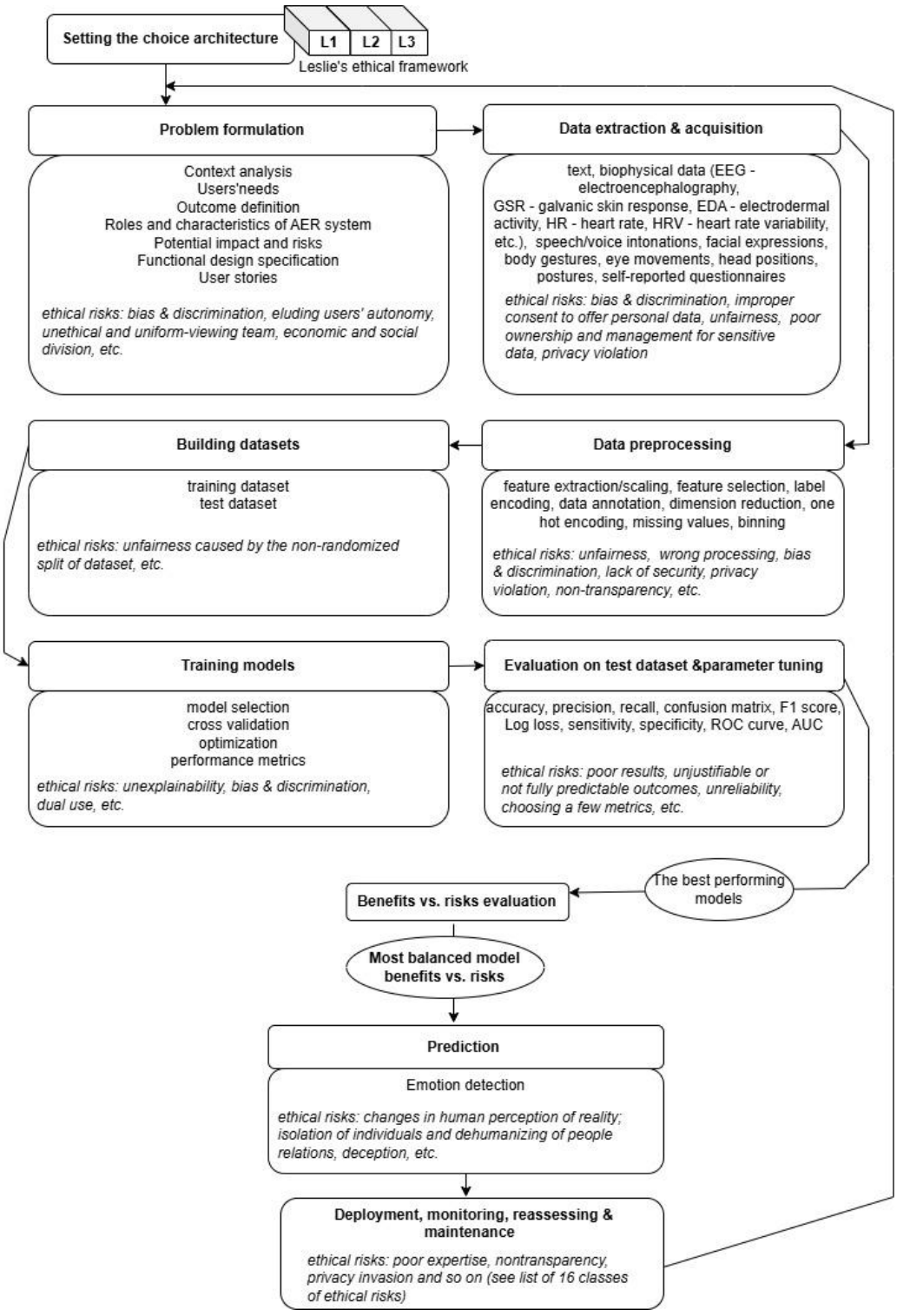

Figura 5. Model etic pentru RAE în învățarea online [24]

Capitolul 6 cu titlul "Explicabiliate în Învățarea Automată" este bazat pe articolele publicate de autoarea acestei teze de abilitare în calitate de coautor în [30] și [25]. În prima parte a capitolului este investigată importanța subiectului în peisajul aplicațiilor cu inteligență artificială. Accentuăm faptul că modelul conceptual FAT (fairness, accountability, and transparency – echitate, responsabilitate și transparență) din [31] trebuie operaționalizat atât în industrie, cât și în mediul academic. Ieșirile diferiților explicatori aplicați pentru aceeași instanță pot fi diferite. Mai mult, un explicator aplicat de mai multe ori unei instanțe poate genera rezultate diferite. Problema dezacordului în învățarea automată constă în obținerea de explicații contradictorii pentru același eșantion și model [32], [33], [34].

Propunem în [25] o metodă bazată pe LIME pentru a furniza explicații, care este descrisă în subcapitolul "O metodă de analiză cu LIME pentru explicații". Metoda cuprinde următorii pași:

1. "LIME este executat de mai multe ori (în experimentele noastre, de 20 ori).
2. La fiecare execuție se determină influența atributului asupra predicției (unele atribute sprijină prezența unei emoții specific, altele sprijină absența acelei emoții).
3. Pentru fiecare atribut, este calculat numărul de apariții ale atributului în subsetul de atribute care sprijină prezența unei emoții specifice, precum și numărul de apariții ale atributului în subsetul de atribute care sprijină absența unei emoții specifice.
4. Pentru fiecare atribut, se calculează valoarea absolută a diferenței dintre cele două frecvențe calculate anterior.
5. Ulterior, atributele sunt sortate în ordine descendentă pe baza diferenței absolute. Atributul cu cea mai mare valoare contribuie cel mai mult la prezența sau absența emoției. Dacă numărul de apariții ale unui atribut în subsetul de atribute care sprijină prezența emoției este mai mare decât numărul de apariții ale aceluiaș atribut în subsetul de atribute care sprijină absența emoției, atunci se consideră că acel atribut sprijină prezența emoției, altfel invers. [25]"

Metoda a fost validată pe predicțiile obținute cu modelul 1D-CNN cu 5 canale. Rezultatele arată funcționalitatea metodei în fiecare caz studiat.

De exemplu, într-un caz al predicției absenței plictiselii, sunt obținute următoarele rezultate pentru fiecare atribut rulând LIME de 20 de ori: culoarea orange marchează atributele care sprijină mai mult prezența emoției, culoarea albastră le marchează pe cele care sprijină absența emoției, iar atributele fără nicio culoare au diferența absolută egală cu 0.

19	6	23	7	4	11	12		2	5	13	20	24
14	12	12	10	8	8	8		6	6	6	6	6
1	9	14	18	21	0	8	15	16	17	22	3	10
4	4	4	4	4	2	2	2	2	2	2	0	0

Adunând diferențele absolute, putem concluziona că atributele influențează mai mult absența emoției (72) decât prezența acesteia (62).

Metoda este costisitoare din punct de vedere al timpului necesar, așa că intenționăm să oferim un nou algoritm bazat pe LIME.

În subcapitolul "O metodă de agregare bazată pe clustere pentru alinierea explicațiilor", este prezentată o metodă inspirată din raționamentul bazat pe cazuri pentru agrearea diferitelor explicații [30]. În elaborarea acestei metode a fost utilizată abordarea lui Pirie și alții (2023) pentru agrearea explicațiilor, care a aplicat o aliniere locală și a propus o metrică de încredere a alinierii între explicatori. De asemenea, aceștia au dezvoltat un framework pentru agrearea explicatorilor, numit AGREE (AGgregation for Robust Explanations) [35]. Metoda conține două etape: în prima etapă, sunt generate explicații și în etapa a doua este aplicată o metodă de agreare bazată pe clustere inspirată din Case-Based Reasoning.

Metoda a fost evaluată pe șase seturi de date des utilizate (Pima Indian Diabetes Dataset [36], Indian Liver Patient Dataset [37], Hepatitis Dataset [38], Fetal Dataset [39], Abalone Dataset [40], Water Quality Dataset [41]) prin furnizarea vectorilor cu ponderi ale explicațiilor agregate spațiului de caracteristici al unui clasificator k-NN ponderat și compararea performanțelor de predicție cu cele obținute cu un algoritm k-NN neponderat. Explicațiile au fost generate cu LIME, Anchors, Kernel SHAP, and Tree SHAP.

Rezultatele obținute validează strategia propusă (Tabelele 2 și 3).

Tabelul 2. Acuratețea algoritmului k-NN ponderat vs. acuratețea algoritmului k-NN neponderat [30]

	Diabetes	Liver	Hepatitis	Abalone	Water	Fetal
Average accuracy for weighted k-NN	**74.3**	**67.93**	**94.85**	**53.55**	**61.97**	**92.87**
Non-weighted k-NN	72.58	65.65	92.43	53.62	60.86	91.74

Tabelul 3. Scorul F1 pentru algoritmul k-NN ponderat vs. scorul F1 pentru algoritmul k-NN neponderat [30]

	Diabetes	Liver	Hepatitis	Abalone	Water	Fetal
Average F1 score for weighted k-NN	**60.67**	**78.08**	**71.51**	**46.44**	**72.35**	**83.04**
Non-weighted k-NN	55.74	76.45	55.01	47.78	42.81	79.95

Strategia este costisitoare din punct de vedere al resurselor și timpului de calcul. Intenționăm să ne dedicăm cercetării în XAI, dezvoltând metode fiabile cu costuri de calcul reduse.

Resurse bibliografice

[1] Picard, R., Affective Computing, M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321, Cambridge, 1995.

[2] Picard, R.W., Papert, S., Bender, W. Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy D., and Strohecker, C.: Affective Learning – A Manifesto, BT Technology Journal, 22: 253, 2004. https://doi.org/10.1023/B:BTTJ.0000047603.37042.33.

[3] . D'Mello, S. K., & Graesser, A. C., Feeling, thinking, and computing with affect-aware learning technologies. In R. A. Calvo, S. K. D'Mello, J. Gratch, & A. Kappas (Eds.), The Oxford handbook of affective computing (pp. 419–434). Oxford University Press. 2015. https://doi.org/10.1093/oxfordhb/9780199942237.013.032

[4] Artificial Intelligence Act, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828.

[5] Bălan, O., **Moise, G.**, Moldoveanu, A., Moldoveanu, F. and Leordeanu. M., Does automatic game difficulty level adjustment improve acrophobia therapy? differences from baseline. In Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology (VRST '18). Association for Computing Machinery, New York, NY, USA, Article 78, 1–2. 2018., https://doi.org/10.1145/3281505.3281583.

[6] Bălan, O., **Moise, G.**, Moldoveanu, A., Leordeanu, M., Moldoveanu. F., Challenges for ML-based Emotion Recognition Systems in Medicine. A Human-Centered Approach. CHI'19 Extended Abstracts, May 4-9, 2019, Glasgow, Scotland, UK. ACM CHI Conference on Human Factors in Computing Systems Workshop on Emerging Perspectives in Human-Centered Machine Learning.

[7] Bălan, O., Cristea, Ș., Moldoveanu, A., **Moise, G.**, Leordeanu, M., Moldoveanu, F. (2020). Towards a Human-Centered Approach for VRET Systems: Case Study for Acrophobia. In: Siarheyeva, A., Barry, C., Lang, M., Linger, H., Schneider, C. (eds) Advances in Information Systems Development. ISD 2019. Lecture Notes in Information Systems and Organisation, vol 39. Springer, Cham. https://doi.org/10.1007/978-3-030-49644-9_11.

[8] Bălan, O., Moldoveanu, A., Petrescu, L., **Moise, G.**, Cristea, Ș., Petrescu, C., Sensors system methodology for artefacts identification in Virtual Reality games, 2019 International

Symposium on Advanced Electrical and Communication Technologies (ISAECT), Rome, Italy, 2019, pp. 1-6, doi: 10.1109/ISAECT47714.2019.9069719.

[9] Balan, O., **Moise, G.**, Moldoveanu, A., Moldoveanu, F. and Leordeanu, M., Automatic Adaptation of Exposure Intensity in VR Acrophobia Therapy, Based on Deep Neural Networks. In Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden, June 8-14, 2019. ISBN 978-1-7336325-0-8 Research Papers. https://aisel.aisnet.org/ecis2019_rp/52.

[10] Bălan, O.; **Moise, G.**; Moldoveanu, A.; Leordeanu, M.; Moldoveanu, F. An Investigation of Various Machine and Deep Learning Techniques Applied in Automatic Fear Level Detection and Acrophobia Virtual Therapy. Sensors 2020, 20, 496. https://doi.org/10.3390/s20020496.

[11] Bălan, O., **Moise, G.**, Moldoveanu, A., Moldoveanu, F., Leordeanu, M., Classifying the Levels of Fear by Means of Machine Learning Techniques and VR in a Holonic-Based System for Treating Phobias. Experiments and Results. In: Chen, J.Y.C., Fragomeni, G. (eds) Virtual, Augmented and Mixed Reality. Industrial and Everyday Life Applications. HCII 2020. Lecture Notes in Computer Science, vol 12191. Springer, Cham. 2020. https://doi.org/10.1007/978-3-030-49698-2_24

[12] Petrescu, L.; Petrescu, C.; Mitruț, O.; **Moise, G.**; Moldoveanu, A.; Moldoveanu, F.; Leordeanu, M. Integrating Biosignals Measurement in Virtual Reality Environments for Anxiety Detection. Sensors 2020, 20, 7088. https://doi.org/10.3390/s20247088.

[13] Koelstra, S.; Muehl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. DEAP: A Database for Emotion Analysis using Physiological Signals. IEEE Trans. A_ect. Comput. 2012. 3, 18–31, doi: 10.1109/T-AFFC.2011.15.

[14] Bălan, O.; **Moise, G.**; Moldoveanu, A.; Leordeanu, M.; Moldoveanu, F. Fear Level Classification Based on Emotional Dimensions and Machine Learning Techniques. Sensors 2019, 19, 1738. https://doi.org/10.3390/s19071738

[15] Petrescu, L.; Petrescu, C.; Oprea, A.; Mitruț, O.; **Moise, G.**; Moldoveanu, A.; Moldoveanu, F. Machine Learning Methods for Fear Classification Based on Physiological Features. Sensors 2021, 21, 4519. https://doi.org/10.3390/s21134519.

[16] Bălan, O.; **Moise, G.**; Petrescu, L.; Moldoveanu, A.; Leordeanu, M.; Moldoveanu, F. Emotion Classification Based on Biophysical Signals and Machine Learning Techniques. Symmetry 2020, 12, 21. https://doi.org/10.3390/sym12010021.

[17] Todorovska, E., 22 Astonishing Phobia Statistics for 2024, 22 Astonishing Phobia Statistics for 2024, https://medalerthelp.org/blog/stats-and-facts/phobia-statistics/, Accessed March 2025.

[18] North, M.M., North, S.M., and Joseph R. Coble, J. R., Virtual Reality Therapy: An Effective Treatment for Psychological Disorders. Virtual Reality in Neuro-Psycho-Physiology, Giuseppe Riva (Ed.), Ios Press: Amsterdam, Netherlands, 1997. PMID: 10175343

[19] Ekman, P., Universals and cultural differences in facial expressions of emotion, Nebraska Symposium on Motivation, 19, p. 207–283, 1971.

[20] **Moise, G.**, Vladoiu, M., Constantinescu, Z., GC-MAS – A Multiagent System for Building Creative Groups Used in Computer Supported Collaborative Learning. In: Jezic, G., Kusek, M., Lovrek, I., J. Howlett, R., Jain, L. (eds) Agent and Multi-Agent Systems: Technologies and Applications. Advances in Intelligent Systems and Computing, vol 296. Springer, Cham. 2014. https://doi.org/10.1007/978-3-319-07650-8_31.

[21] **Moise, G.**, Vladoiu, M., Constantinescu, Z., Building the Most Creative and Innovative Collaborative Groups Using Bayes Classifiers. In: Panetto, H., et al. On the Move to Meaningful Internet Systems. OTM 2017 Conferences. International Conference on Cooperative Information Systems (CoopIS) 2017, Lecture Notes in Computer Science(), vol 10573. Springer, Cham. https://doi.org/10.1007/978-3-319-69462-7_17.

[22] Vladoiu, M., **Moise, G.**, & Constantinescu, Z., Towards Building Creative Collaborative Learning Groups Using Reinforcement Learning. In B. Andersson, B. Johansson, S. Carlsson, C. Barry, M. Lang, H. Linger, & C. Schneider (Eds.), Designing Digitalization (ISD2018 Proceedings). Lund, Sweden: Lund University. ISBN: 978-91-7753-876-9. 2018. http://aisel.aisnet.org/isd2014/proceedings2018/Education/9

[23] **Moise, G.**, Vladoiu, M., & Constantinescu, Z. (2018). Towards Construction of Creative Collaborative Teams Using Multiagent Systems. In B. Andersson, B. Johansson, S. Carlsson, C. Barry, M. Lang, H. Linger, & C. Schneider (Eds.), Designing Digitalization (ISD2018 Proceedings). Lund, Sweden: Lund University. ISBN: 978-91-7753-876-9. 2018. http://aisel.aisnet.org/isd2014/proceedings2018/Education/10.

[24] **Moise, G.**, Nicoară, E., S., Chapter 4 - Ethical aspects of automatic emotion recognition in online learning, Editor(s): Santi Caballé, Joan Casas-Roma, Jordi Conesa, In Intelligent Data-Centric Systems, Ethics in Online AI-based Systems, Academic Press, 2024, Pages 71-95, ISBN 9780443188510, https://doi.org/10.1016/B978-0-443-18851-0.00003-2.

[25] **Moise, G.**, Dragomir, E.G., Șchiopu, D. et al. Towards Integrating Automatic Emotion Recognition in Education: A Deep Learning Model Based on 5 EEG Channels. Int J Comput Intell Syst 17, 230, 2024. https://doi.org/10.1007/s44196-024-00638-x.

[26] Pekrun, R., Goetz, T., Titz W. & Perry R. P., Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program of Qualitative and Quantitative Research, Educational Psychologist, 37:2, 91-105, 2002. DOI: 10.1207/S15326985EP3702_4

[27] Yadegaridehkordi E., Noor N.F.B.M., Ayub M.N.B., Affal H.B. & Hussin N.B., Affective computing in education: A systematic review and future research, Computers & Education, 2019. doi: https://doi.org/10.1016/j.compedu.2019.103649.

[28] D'Mello, S., A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. Journal of Educational Psychology, 105(4), 1082–1099. 2013. https://doi.org/10.1037/a0032674

[29] Leslie, D., Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute, 2019. https://doi.org/10.5281/zenodo.3240529.

[30] Mitruț, O., Moise, G., Moldoveanu, A. et al. Clarity in complexity: how aggregating explanations resolves the disagreement problem. Artif Intell Rev 57, 338, 2024. https://doi.org/10.1007/s10462-024-10952-7

[31] Shin, D., Park, Y. J., Role of fairness, accountability, and transparency in algorithmic affordance, Computers in Human Behavior, Volume 98, 2019, Pages 277-284, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2019.04.019.

[32] Brughmans, D., Melis L, Martens D., Disagreement amongst counterfactual explanations: How transparency can be deceptive. 2023. arXiv [csAI]. http://arxiv.org/abs/2304.12667

[33] Krishna, S., Han, T., Gu, A., Jabbari, S., Wu, Z.S., Lakkaraju, H., The disagreement problem in explainable machine learning: A practitioner's perspective. Research Square. 2023. http://arxiv.org/abs/2202.01602

[34] Müller, S., Toborek, V., Beckh, K., Jakobs, M., Bauckhage, C., Welke, P., An Empirical Evaluation of the Rashomon Effect in Explainable Machine Learning. In: Koutra, D., Plant, C., Gomez Rodriguez, M., Baralis, E., Bonchi, F. (eds) Machine Learning and Knowledge Discovery in Databases: Research Track. ECML PKDD 2023. Lecture Notes in Computer Science(), vol 14171. Springer, Cham. 2023. https://doi.org/10.1007/978-3-031-43418-1_28.

[35] Pirie, C., Wiratunga, N., Wijekoon, A., Moreno-Garcia, C.F., AGREE: a feature attribution aggregation framework to address explainer disagreements with alignment metrics. In Proceedings of the Workshops at the 31st International Conference on Case-Based Reasoning (ICCBR-WS 2023), pp184–199. CEUR. 2023. https://ceur-ws.org/Vol-3438/paper_14.pdf Accessed April 2025.

[36] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., Johannes, R.S., Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Proc Annu Symp Comput Appl Med Care. 1988 Nov 9:261–5. PMCID: PMC2245318.

[37] Ramana, B., Venkateswarlu, N., ILPD (Indian Liver Patient Dataset). UCI Machine Learning Repository. 2012. https://doi.org/10.24432/C5D02C. Accessed January 2024.

[38] Hepatitis, 1988, UCI Machine Learning Repository. https://doi.org/10.24432/C5Q59J. Accessed January 2024.

[39] Campos, D., Bernardes, J., Cardiotocography. 2010. UCI Machine Learning Repository. https://doi.org/10.24432/C51S4N. Accessed January 2024.

[40] Nash, W., Sellers, T., Talbot, S., Cawthorn, A., & Ford, W., Abalone [Dataset]. 1994. UCI Machine Learning Repository. https://doi.org/10.24432/C55C7W.

[41] Kadiwal, A., Water quality dataset. 2021. https://www.kaggle.com/datasets/adityakadiwal/water-potability. Accessed January 2024.