IOSUD – UNIVERSITATEA „DUNĂREA DE JOS” DIN GALAȚI

Școala doctorală de Inginerie Mecanică și Industrială

TEZĂ DE DOCTORAT

REZUMAT

SISTEM EXPERT PENTRU RECUNOAȘTEREA IDENTITĂȚII DE CLASĂ A AMFETAMINELOR HALUCINOGENE

Doctorand,

Cătălin NEGOIȚĂ

Conducător științific,

Prof. univ. dr. fiz. Mirela PRAISLER

Seria I4: Inginerie industrială Nr. 72

GALAȚI

2020

IOSUD – UNIVERSITATEA „DUNĂREA DE JOS” DIN GALAȚI

Școala doctorală de Inginerie Mecanică și Industrială

TEZĂ DE DOCTORAT

REZUMAT

SISTEM EXPERT PENTRU RECUNOAȘTEREA IDENTITĂȚII DE CLASĂ A AMFETAMINELOR HALUCINOGENE

Doctorand

Cătălin NEGOIȚĂ

Președinte Prof. univ. dr. ing. Cătălin FETECĂU, Universitatea „Dunărea de Jos” din Galați

Conducător științific, Prof. univ. dr. fiz. Mirela PRAISLER, Universitatea „Dunărea de Jos” din Galați

Referenți stiințifici Prof. univ. dr. ing. Gheorghe NAGÎȚ, Universitatea Tehnică „Gh. Asachi” din Iași

Prof. univ. dr. ing. Remus ZĂGAN, Universitatea Maritimă din Constanța

Conf.dr. Steluta GOSAV, Universitatea „Dunărea de Jos” din Galați

Seria I4: Inginerie industrială Nr. 72

GALAȚI

2020

Seriile tezelor de doctorat susținute public în UDJG începând cu 1 octombrie 2013 sunt:

Domeniul fundamental ȘTIINȚE INGINEREȘTI

Seria I 1**:** **Biotehnologii**
Seria I 2**:** **Calculatoare și tehnologia informației**
Seria I 3: **Inginerie electrică**
Seria I 4**:** **Inginerie industrială**
Seria I 5: **Ingineria materialelor**
Seria I 6**:** **Inginerie mecanică**
Seria I 7**:** **Ingineria produselor alimentare**
Seria I 8: **Ingineria sistemelor**
Seria I 9**:** **Inginerie și management în agicultură și dezvoltare rurală**

Domeniul fundamental ȘTIINȚE SOCIALE

Seria E 1: **Economie**
Seria E 2: **Management**
Seria SSEF: **Știința sportului și educației fizice**

Domeniul fundamental ȘTIINȚE UMANISTE ȘI ARTE

Seria U 1: **Filologie- Engleză**
Seria U 2: **Filologie- Română**
Seria U 3**:** **Istorie**
Seria U 4: **Filologie - Franceză**

Domeniul fundamental MATEMATICĂ ȘI ȘTIINȚE ALE NATURII

Seria C: **Chimie**

Domeniul fundamental ȘTIINȚE BIOLOGICE ȘI BIOMEDICALE

Seria M: **Medicină**

Cuvânt înainte

La finalul acestei importante perioade de studiu și cercetare, aș dori să-mi aduc recunostința tuturor persoanelor care au contribuit la formarea mea profesională și m-au încurajat și sprijinit pe tot parcursul acestor ani.

Doresc să mulțumesc doamnei prof. dr. Mirela Praisler în calitate de coordonator științific, pentru efortul și răbdarea de care a dat dovadă, cât și pentru toate indrumările pe care le-am primit, care m-au ajutat enorm sa finalizez acest demers stiințific.

Mulțumesc doamnelor prof. dr. Luminița Moraru, conf. dr. Steluţa Gosav și domnului prof.dr.habil. Constantin Apetrei de la Universitatea ”Dunărea de Jos” din Galați, membrii ai Comisiei mele de indrumare, atat pentru suportul acordat cat și sugestiile oferite pe parcursul stagiului de studii doctorale.

Galați, Septembrie 2020 Fiz. Cătălin Negoiță

Cuprins

Introducere

În contextul actual al societații moderne în plan economic si social, se poate observa cu usurință că există o continuă provocare pentru combaterea tuturor formelor de crimă organizată, întrucât societatea moderna a ajuns sa fie influențată major de către acestea. Piața drogurilor de mare risc ocupă un mare segment din crima organizată pe plan mondial. Putem considera că, în ce privește depistarea substantelor interzise, există provocari în toate fazele acestui proces, incepand cu sinteza drogurilor sintetice și continuînd cu producerea, transportul, distribuția si / sau consumul acestor substante.

Traficul de droguri reprezintă o gravă amenințare la adresa securității și sănătății publice, atât pe plan național cat si la nivel global. Globalizarea, impreună cu dezvoltarea rapidă a tehnologiei și diversificarea legăturilor comerciale, au facilitat crearea unor noi rute si metode de transport. Eliminarea controalelor la unele frontiere a contribuit la dezvoltarea organizațiilor criminale, ce alimenteaza și controleaza piața ilicită a drogurilor. Din nefericire, România este și ea implicată în acestă rețea mondiala. In ultimele doua decenii, cetățeni români au inceput sa fie cooptati de grupari infracționale, fiind folosiți ca transportatori la nivel internațional, mai ales la nivel european. În anii trecuți a fost inregistrată o crește a încercărilor de introducere a drogurilor sintetice (amfetamine, mentamfetamine, derivați din amfetamină, etc.) din tarile cu tradiție in producerea de medicamente sintetice (Belgia, Olanda, Germania), fie prin sistemul de coletărie, fie prin transport aerian sau terestru [1].

Tehnologia modernă vine, din păcate, si în ajutorul acestor structuri, ducând la un ușor avantaj în favoarea acestor rețele. Metodele moderne de producere si contrabandă în scopul de a evita controalele, au generat necesitatea de a dezvolta noi tehnici moderne de depistare a unei varietati de noi compuși, majoritatea facand parte din familia substanțelor interzise cu efect psihotrop.

Cea mai frecventa modalitate aleasa de retelele de traficanti pentru a incerca sa evite confiscarile si condamnarile este usoara modificarea a structurilor moleculare a compușilor de bază, fie prin adăugarea sau schimbarea unor substituenți din diferitele poziții ale structurilor moleculare de baza. Astfel, noii compuși nu se regaseasc pe listele internaționale de substanțe interzise si in acelasi timp pot fi comercializati pe piata neagra cu acelasi succes, intrucat au o activitate farmacologica similara compusilor – mama (care sunt compusi controlati). Intrucât echipamentele de detectie folosite la ora actuala in vami, porturi, aeroporturi, etc., identifica doar compusii a caror spectre sunt stocate in memoria instrumentului, autoritatile fac apel pentru dezvoltarea de tehnici moderne de recunoaștere a identității de clasă, care sa poata depista *orice substanta cu o structura moleculara similara* cu cea a unei substante interzise. In al doilea rand, autoritatile reliefeaza nevoia de a avea disponibile instrumente cat mai putin voluminoase, portabile, ce permit identificarea *in-situ* a substantelor controlate. Astfel de aplicatii pot fi dezvoltate folosind metode spectrale in combinatie cu metode de inteligență artificială.

Inteligența artificială ocupă din ce în ce mai mult spațiu pe toate palierele, fie ca vorbim de industrie, medicină sau stiințe politice și sociale. **Rezultatele prezentate pe parcursul acestei teze demonstreaza că aceste tehnici sunt adecvate si pentru domeniul toxicologiei si a criminalisticii. Sistemele expert dezvoltate pot fi instalate pe instrumente de control produse la ora actuala nivel industrial, reprezentand astfel o contributie la dezvoltarea ingineriei industriale.**

Rezumatul și structura lucrării științifice

Teza de doctorat cu titlul "Sistem expert pentru recunoașterea identității de clasă a amfetaminelor halucinogene" este structurată in patru capitole, insoțite de o introducere in tema studiată. In final sunt prezentate concluziile generale (deduse pe baza rezultatelor originale prezentate) și direcțiile viitoare de cercetare și dezvoltare.

Această lucrare debutează cu o introducere în contextul necesității unor noi tehnici de indentificare a substanțelor interzise sintetizate, transportate si / sau comercializate de către rețelele de traficanți de droguri sau crima organizată.

Capitolul I descrie atat proprietațile generale ale principalelor amfetamine halucinogene, cât și lista compușilor aleși pentru analiză în această teză.

Capitolul II conține prezentarea metodelor spectrale de caracterizare si identificare a amfetaminelor halucinogene. Pe parcursul acestui capitol sunt descrise conceptele de bază a spectrometriei în infraroșu și componentele principale ale unui astfel de spectrometru.

Capitolul III, denumit "Metode de inteligență artificială aplicabile pentru recunoașterea identității de clasă a unor substanțe organice", cuprinde descrierea celor sase modele de inteligenta artificiala folosite pentru prelucrarea spectelor substantelor de interes, respectiv PLSR, Algoritmii genetici, KNN, Random Forest, SVM și Regresiile Logistice. Pe parcursul acestui capitol au fost descrise conceptele ce stau la baza acestor algoritmi, funcțiile și metodele de interpretare si considerații teoretice în ce privește evaluarea performanței sistemului de atribuire a identitatii de clasa (clasificarii) pe baza erorilor si a intervalelor de confidență.

Capitolul IV, intitulat "Contribuții privind asplicații de inteligență artificială proiectate pentru recunoașterea identitații de clasă a principalelor amfetamine ilicite halucinogene 2C-x si DOx", cuprinde contribuții proprii privind dezvoltarea unor sisteme expert ce permit recunoașterea identității de clasă a amfetaminelor halucinogene descrise în Capitolul I. Sistemele expert descrise au fost construite folosind modelele prezentate teoretic în Capitolul III. Fiecare model a fost analizat în detaliu, în mai multe etape de evaluare, cu variații ale parametrilor ce stabilesc arhitectura sistemelor in cauza.

In capitolul dedicat concluziilor generale, se prezintă a analiza finala comparativa a performantelor modelelor folosite. Sintetizarea rezultatelor a permis identificarea celui mai performant model ce trebuie folosit ca sistem expert de detectie a amfetaminelor halucinogene de interes. **Acesta este sistemul cel mai recomandabil pentru implementare la nivel de inginerie industriala aplicata in productia de serie a instrumentelor portabile folosite de fortele de ordine.**

In finalul tezei se regăsește atat o scurtă descriere a contribuțiilor personale din cadrul acestei lucrări, cât si lista de lucrări publicate și prezentate în cadrul unor conferințe nationale și internaționale.

Motivația alegerii temei de cercetare

În ultimii ani, inteligența artificială este folosită din ce în ce mai mult în mai toate domeniile. Dupa cum s-a aratat mai sus, extinderea performanțelor în ce privește clasificarea pe baza spectrelor FTIR si ATR-FTIR prezentate in aceasta lucrare prezinta un interes practic major. In plus, este de interes si faptul ca modelele studiate si dezvoltate pe parcursul acestei teze pot fi extinse și pentru alte aplicatii din domenii precum medicina si ingineria medicala, farmacia si industria chimica, etc.

Producerea, comercializarea si consumul drogurilor de mare risc reprezintă un pericol din ce în ce mai mare pentru societatea moderna. Sistemele expert, prezentate în această teza de doctorat, ce pot identifica aceste substanțe interzise inclusiv *in situ*, sunt o contributie la efortul depus de toate instituțiile în lupta împotriva rețelelor de traficanti de substanțe de mare risc.

Aceste doua idei cu un foarte mare impact în viitorul societății moderne, respectiv digitalizarea folosind inteligența artificială și destabilizările create de rețelele crima organizată, justifica alegerea temei de cercetare a prezentei lucrari. Evaluarea celor sase modele de inteligenta artificiala studiate arata ca acestea au un potențial mare de creștere al performanței recunoașterii identității de clasă a amfetaminelor halucinogene.

Obiectivele de cercetare urmarite

In cadrul acestei teze am avut ca obiective de cercetare: studiul documentar privind stadiul actual al cercetarilor, identificarea compușilor de studiu, a celor mai adecvate metode spectrale de caracterizare și identificare a amfetaminelor halucinogene, precum si alegerea metodelor de inteligenta artificala cu cel mai mare potential de a genera sisteme expert eficiente pentru operarea instrumentelor analitice folosite in detectia drogurilor de interes.

În urma analizei teoretice a modelelor propuse pentru studiu și cercetare, obiectivele din faza de dezvoltare a modelelor au fost urmatoarele:

- construirea bazei de date cu ajutorul spectrelor compusilor chimici de interes;
- identificarea tuturor parametrilor fiecarui model studiat ce ar putea duce la optimizarea modelului și cresterea implicita a preformantei de clasificarea;
- evaluarea performantelor sistemului expert construit cu fiecare dintre metodele de inteligenta artificiala considerate, prin compararea rezultatelor obtinute pentru setul de date spectrale complet și pentru un set de date selectate (generat de algoritmul genetic) ce contine cele mai reprezentative date (informatii);
- rularea modelelor în mai multe iterații, cu o variatie a parametrilor de fine tuning;
- analiza comparativa a performantelor sistemelor expert construite cu diverse tehnici si pentru cele doua seturi de date spectrale de intrare, pentru identificarea siistemului expert cu cele mai bune performante.

Capitolul I. Amfetamine halucinogene: proprietăţi

I.1 Introducere

Amfetaminele sunt substanțe sintetice cu rol stimulant al sistemului nervos central, intâlnite în mod uzual sub forma unei pudre albe. Compusul – mamă, amfetamina, are denumirea IUPAC *N*,α-methylbenzeneethanamine [1]. Întrucat amfetaminele sunt substanțe controlate, noi compuși, ce nu se află în lista substanțelor interzise, sunt permanent sintetizați în laboratoarele clandestine. Aceștia sunt obținuți prin ușoare modificări ale structurii moleculare(vezi Figura I.1.1), care pastreaza efectul biologic al amfetaminelor și totodată nu intra sub incidenta legii privind substanțele prohibite. Pentru prezenta teza de doctorat am ales spre studiu două noi clase de analogi halucinogeni ai amfetaminei, respectiv 2C-x și DOx.

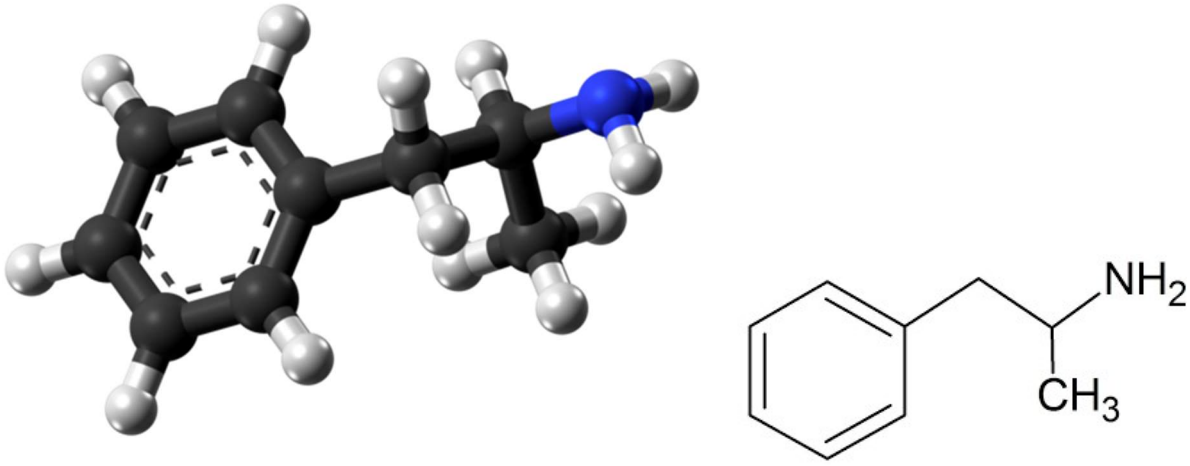
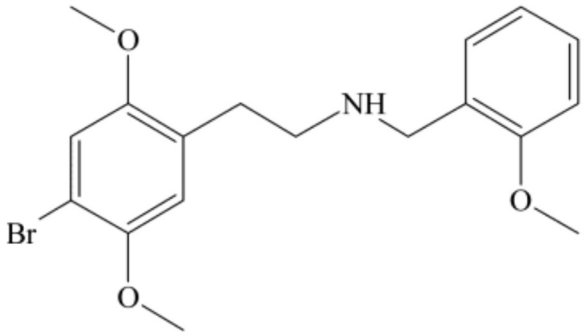


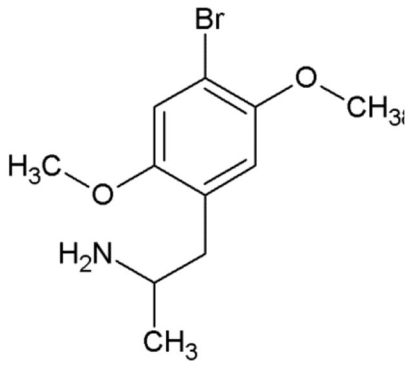
Figura I.1.1. Structura moleculară a amfetaminei [2].

Substanțele din clasa 2C-x conțin grupări metoxi- în pozițiile 2 și 5 ale ciclului aromatic prezent în structura moleculară a amfetaminelor. Substituenții lipofilici de la pozitia 4 a ciclului aromatic amplifică și prelungesc efectele stimulante și halucinogene ale acestor substanțe (vezi Figura I.1.2) [3]. Clasa DOx contine amfetamine halucinogene care prezintă grupări metoxi în pozițiile 2- și 5- ale ciclului aromatic.

Baza de date dezvoltată în aceasta fază de documentare este compusă din spectrele infrarosii (IR), obtinute cu ajutorul unui spectrometru ATR – FTIR (Attenuated total reflectance-Fourier transform infrared spectroscopy) pentru 17 substanțe din clasa 2C-x și 7 substanțe din clasa DOx [4].



a)



b)

Figura I.1.2. Structura moleculară a compusului: a) 25B-NBOMe, substanta din clasa amfetaminelor 2C-x; b) 4-Bromo-2,5-dimethoxyamphetamine (DOB), substanta din clasa amfetaminelor DOx.

Identificarea primară a probelor necunoscute se face automat, folosind aplicația software a instrumentului folosit. Aceasta calculează distanța euclidiana dintre spectrele existente în bază de date proprie și spectrul substanței de analizat.

Capitolul II Metode spectrale de caracterizare și identificare a amfetaminelor halucinogene

II.1. Introducere

În lucrarea de față sunt descrise metodele spectrale folosite în obținerea spectrelor IR ai unor analogi halucinogeni ai amfetaminei, care fac parte din clasele 2C-x și DOx. Întrucat în procesul de identificare al acestor substanțe interzise timpul de prelucrare joacă un rol important, s-a dovedit ca spectroscopia ATR - FTIR poate furniza spectre bogate în informații, fiind necesare pregătiri minime ale probei pentru a fi analiză și chiar recuperarea probei dacă se impune.

II.2. Metode spectrale de caracterizare și identificare

II.2.1 Spectrometria în infraroșu cu transformata Fourier (FTIR)

Spectroscopia în infraroșu cu transformata Fourier (FTIR) se bazează pe interferența radiației între două fascicule, fenomen în urma căruia se obține o interferogramă. Acesta din urmă este un semnal produs ca o funcție a diferenței de drum străbătut de cele două unde. Cele două domenii ale distanței și frecvenței sunt interconvertibile de către metoda matematică a transformatei Fourier.

Componentele de bază ale unui spectrometru FTIR sunt prezentate schematic în Figura II.2.1. Radiația emisă de sursă trece printr-un interferometru pentru a ajunge la proba și apoi la un detector. După amplificarea semnalului, în care contribuțiile de înaltă frecvență au fost eliminate de un filtru, datele sunt convertite în formă digitală printr-un convertor analog-digital și transferate către calculatorul atasat spectrometrului, care efectueaza transformata Fourier [5].

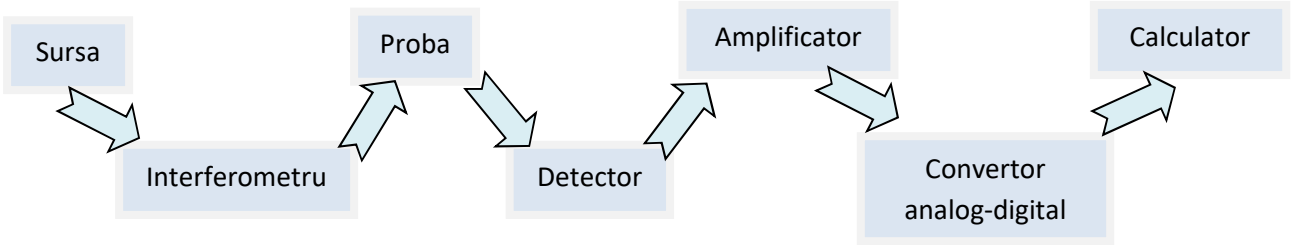


Figura II.2.1 Prezentare schematica spectrometru FTIR.

II.2.2 FTIR cu Reflexie Totala Atenuată (ATR-FTIR)

Spectroscopia ATR – FTIR derivă din spectroscopia cu reflecția internă și poate fi folosita în analiza probelor solide sau lichide. Bazele ei au fost puse de catre Fahrenfort și Harrick la începutul anilor 1960. Când radiația se propagă dintr-un mediu dens (indicele de refracție n_1) într-un un mediu optic mai puțin dens (indicele de refracție n_2, $n_1 > n_2$), reflexia internă totală va apărea la interfața celor două medii, dacă unghiul de incidență al radiației (θ) depășește

unghiul critic (θc). Unghiul critic poate fi definit ca o funcție ce depinde de indicii de refracție ai celor două medii:

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \qquad (2.1)$$

La fiecare reflexie, un câmp evanescent se extinde în mediul adiacent mai puțin dens. Acest câmp evanescent poate fi descris ca o undă electrică normală la suprafața celor două medii, și rezultă din suprapunerea câmpurilor electrice a undelor incidente și reflectate. Amplitudinea acestei unde electrice (E) scade exponențial cu distanța [7]:

$$E = E_0 e^{-\left(\frac{z}{d_p}\right)} \qquad (2.2)$$

unde E_0 este amplitudinea câmpului electric la suprafața (z=0), z este distanța de la suprafața iar d_p este adâncimea de penetrare a undei, definită ca distanța la care amplitudinea câmpului electric este1/e din E_0:

$$d_p = \frac{\lambda}{2\pi\sqrt{n_1^2 \sin^2\theta - n_2^2}} \qquad (2.3)$$

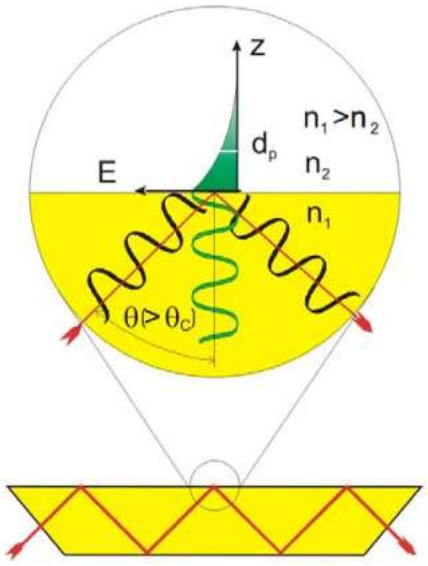


Figura II.2.5 Reprezentarea schematică a ATR [5].

Capitolul III. Metode de inteligență artificială aplicabile pentru recunoașterea identității de clasă a unor substanțe organice

III.1 Modelul PLSR

III.1.1 Considerații generale

Regresia PLS (PLSR) este des intalnită în chemometrie, în special în cazurile în care numărul de variabile independente este semnificativ mai mare decât numărul de observații. Modelul PLS predictiv, folosit în modelarea relației dintre două matrici, X și Y, este considerat mai eficient decât abordarea tradițională în care se foloseste regresie multiplă. PLSR și abordări similare oferă metode de modelare multivariate cantitative, cu posibilități inferențiale similare cu regresia multiplă, ANOVA sau teste *t*.

PLSR poate fi considerată o generalizare a MLR (Multiple Linear Regression), dar prezintă un interes crescut datorită faptului că poate atât analiza seturi de date puternic corelate cu numeroase variabile X, cât și modela numeroase variabile Y, determinând profiluri de performanțe [10-15].

Regresia, adică modul de modelare a uneia sau a mai multor variabile dependente (răspunsuri) Y, cu ajutorul unui set de variabile predictoare X, este una dintre cele mai frecvente probleme de analiză a datelor din știință și tehnologie. Putem considera exemple din chimie, unde proprietățile Y ale eșantioanelor cu compoziția chimică X sunt folosite pentru evaluarea calitații și cantitații de produse fabricate la condițiile X ale procesului de fabricație și proprietățile chimice Y, sau reactivitatea sau activitatea biologică a unui set de molecule cu structura chimică X. Ultimele modele mai sunt denumite și QSPR (Quantitative Structure – Property Relationships) sau QSAR (Quantitative Structure – Activity Relationships). În mod tradițional, această modelare a lui Y prin X se face folosind MLR, care funcționează bine atât timp cât variabilele X sunt destul de puține și destul de necorelate, adică X are un rang complet. În cazul instrumentelor moderne de măsurare, precum spectrometrele sau cromatografele, variabilele X tind să fie și mai multe și mai puternic corelate. Prin urmare, nu le vom numi “independente”, ci „predictori” sau doar variabile X, deoarece acestea sunt de obicei corelate, afectate de zgomot și incomplete. În analiza variabilelor X, cât și a profilurilor de răspuns Y, PLSR ne permite să investigăm probleme mai complexe decât dacă am folosi alte modele.

III.2. Algoritmii genetici

Considerații generale

Algoritmii genetici (GA) reprezintă metode de căutare și optimizare inspirate din principiile naturale de selecție cât și din genetică [26,27,28]. Algoritmii genetici cuprind variabile ce reprezintă soluțile la metodele de căutare într-un set de date cu o strânsă legatură între ele. Vectorii acestor seturi de date inițiale, propuse analizei, mai sunt denumite *cromozomi*, iar elementele vectorilor mai sunt denumite și *gene*. Pentru obținerea unei performanțe ridicate în identificarea celor mai bune soluții, cât și pentru implementarea

selecției naturale, sunt folosite metode de diferențiere a soluțiilor bune fata de cele rele. Metodele pot fi obiective, subiective sau modele matematice, unde arhitectul modelului alege soluțiile bune. În esență, măsura soluției modelului este dată de măsura relativă a soluției fiecărui candidat evaluat, ulterior devenind indicatorul modelului asupra evolutiei performanței acestuia.

Spre deosebire de modelele clasice de căutare, algoritmii genetici se bazează pe populații de potențiale soluții, iar mărimea populației este foarte importantă din perspectiva scalabilității și a performanței modelului. După ce s-au stabilit datele inițiale (cromozomii) și funcția ce determină măsura în care o soluție este considerată bună, se vor iniția iterații de urmatorul tip: Inițializarea. Populația inițială a solutiilor candidate este generată în mod aleatoriu.; Evaluarea. În urma inițializării, se determină dacă măsura soluției funcției ce evaluază soluțiile candidate este considerată bună; Selecția. În aceasta etapă, se rezervă copii ale soluțiilor cu cele mai mari măsuri și se consideră impunerea metodei ”supraviețuirii celor mai buni”; Recombinarea. Două sau mai multe soluții sunt folosite pentru a genera o soluție mai bună.; Mutația. În cadrul acestui pas, se consideră o soluție bună asupra căreia au loc modificări astfel încat un local optim să fie evitat; Înlocuirea. Populația creată în urma selecției, recombinării și mutației înlocuieste populația inițială; Repetiția. Se repetă pașii 2-6 până când soluțiile obținute întrunesc cerințele impuse.

III.3. KNN (K- nearest neighbors)

Clasificarea KNN

În domeniul recunoașterii de pattern-uri (*pattern recognition*), algoritmul KNN reprezintă o metodă de clasificare a obiectelor bazată pe distanța cea mai scurta dintre obiectele folosite la antrenarea modelului, în domeniul caracteristicilor obiectelor (*feature space*). KNN este un model ce presupune învațarea bazată pe instanțe, unde funcția este aproximată local și toate calculele sunt amânate până la clasificare.

KNN poate fi considerat un model fundamental și simplu în ce priveste tehnica de clasificare, mai ales în cazul în care nu se presupune o cunoastere a distribuției datelor înaintea folosirii lui [37]. Această regulă simplă presupune păstrarea întregului set de pregătire și asignarea fiecărei interogări a unei clase reprezentată de majoritatatea etichetelor lui **k**, cei mai apropiați vecini din setul de training.

Modelul NN, cel mai apropiat vecin, este reprezentat în forma cea mai simplă atunci când **K**=1. În acest caz, fiecare eșantion trebuie clasificat în mod similar obiectelor din jur. Daca se considera un obiect neclasificat, el poate fi clasificat pe baza claselor obiectelor cele mai apropiate. Acest lucru se face calculând distanțele dintre obiectul neclasificat și obiectele ce formeaza o clasa cunoscută, distanța cea mai scurtă catre o clasă specifica determinând în acest fel clasificarea obiectului neclasificat [53]. În Figura III.3.1 se poate observa modelul KNN cu K=1 și K=4 pentru un eșantion divizat în două clase. În cazul a), un obiect neclasificat este clasificat folosind doar un singur obiect din eșantionul de training. În situația b), se identifică cele mai apropiate 4 obiecte, determinându-se astfel că face parte din clasa din stânga.

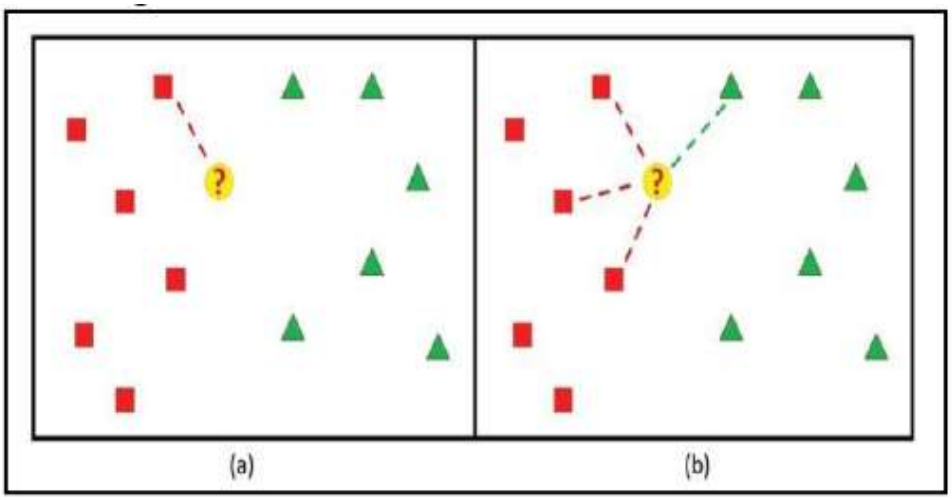


Figura III.3.1 a) Modelul 1-NN, obiectul **?** atribuit clasei din stânga; b) Modelul KNN cu K= 4: obiectul **?** este atribuit clasei din stânga.

Performanța acestui model de clasificare este determinată în principal de alegerea numărului **K** cât și metrica distanțelor [60]. Estimările sunt influențate de senzitivitatea selecțiilor razei de vecinătate a lui **K**, Întrucât raza regiunii locale este determinata de **K** cel mai apropiat vecin, iar valori diferite ale lui **K** pot genera probabilitați de clasificare diferite. Daca valoarea lui **K** este prea mică, estimarile locale pot determina clasificări gresite în cazul seturilor de date cu imprastiere ridicată și cu mult zgomot. Pentru a îmbunatați rata clasificărilor corecte, se marește gradual valoarea lui **K** până când îmbunătățirea nu mai este semnificativă.

III.4. Random Forests

Arborii decizionali reprezintă modele bine cunoscute folosite în clasificare și regresie de foarte mulți ani. Considerând un set de date folosit ca eșantion de antrenare, acest model împarte eșantionul în noduri, luând în considerare valorile obiectelor semnificative. Nodurile sunt create în așa fel încât pe același nivel să se afle caracteristici similare. La testarea modelului, obiectul a carui clasă se dorește a fi identificată avansează în arbore respectând aceeași regulă. Când se ajunge la un element al unui nod, obiectul este clasificat în funcție de clasele setului de date folosit la antrenarea modelului [61,62].

Clasificarea folosind arbori decizionali

Pentru clasificarea unui obiect x, eșantionul parcurge un arbore, iar la fiecare nod, fiecare caracteristică a lui x este comparată cu limita superioară aferentă acelui nod. În funcție de rezultatul evaluării, obiectul avansează spre stânga sau spre dreapta în arbore. Când obiectul ajunge la capatul nodului, îi este atribuită o clasă. În Figura III.4.1 se exemplifică un set de date reprezentat de figuri geometrice, clasele fiind asociate pe baza caracteristicilor considerate determinante, respectiv numărul de unghiuri și luminozitatea. De exemplu, triunghiul din clasa 1 are vectorul caracteristic (3,0.5).

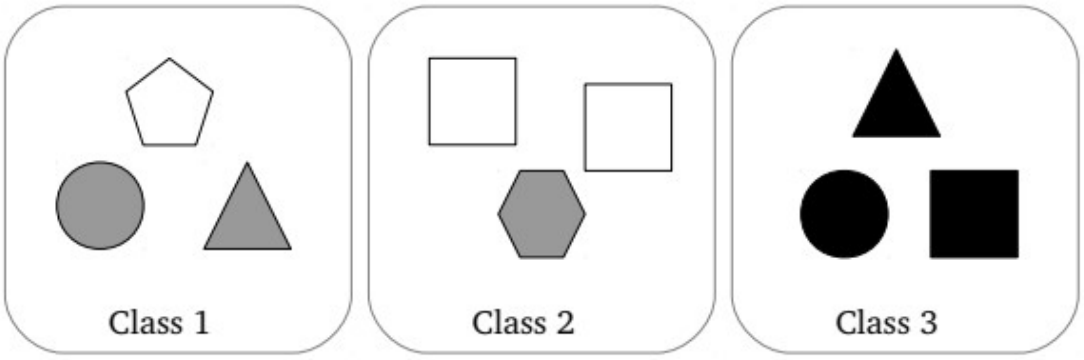


Figura III.4.1 Sunt considerate două caracteristici ce determină apartenența la o anumita clasă, respectiv numărul de unghiuri și luminozitatea (negru=0; gri=0,5; alb=1).

III.5. Clasificarea SVM

Support Vector Machines (SVM) a fost introdus pentru prima oara în 1992 [65] ca fiind un set de metode supervizate de clasificare și regresie. Clasificarea efectuată cu acest model poate fi restransă considerând doar două clase distincte, scopul fiind separarea acestor clase cu o funcție ce este indusă de setul de date folosit. În final se va ajunge la un clasificator ce va reusi sa clasifice setul de test folosind un clasificator liniar ce va maximiza marja, adică distanța dintre cel mai apropiat punct din fiecare clasa (vezi Figura III.5.1).

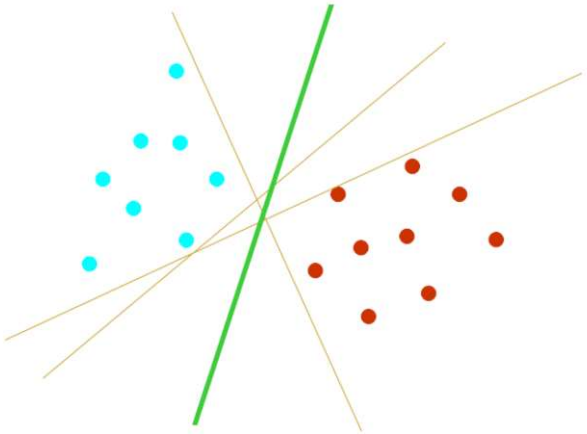


Figura III.5.1 Separarea optimă în hiperplan.

III.6. Clasificare bazata pe regresia logistica (Logistic regression classification model)

Regresia logistică este folosită domeniul clasificărilor binare, ea reprezentând modelarea unui set de variabile independente x și o variabila dependentă binara Y ce poate lua valori doar de 0 sau 1. În acest caz, performanța modelului este dată de valoarea medie a raspunsului modelului în raport cu variabilele independente.

Capitolul IV Contribuții privind aplicații de inteligență artificială proiectate pentru recunoasterea identității de clasă a principalelor amfetamine ilicite halucinogene 2C-x si DOx

IV.1 Introducere

Amfetaminele reprezintă substanțele sintetice cele mai folosite ca droguri. Spre deosebire de celelalte amphetamine, cele cu efect halucinogen nu au niciun fel de intrebuințare medicală datorită efectelor lor adverse severe. Amfetaminele psihedelice, precum cele din clasa 2C-x și Dox, prezintă un nivel ridicat de toxicitate [76]. Structurile moleculare generale ale acestor două familii de halucinogene sunt prezentate în Figura I.1.2.

În ultimul deceniu, noi astfel de substanțe sunt produse și comercializate, atât pe piața neagră din Europa cât și din întreaga lume. Laboratoare clandestine fac eforturi susținute să producă noi substanțe ce nu se află pe lista substanțelor controlate [77]. De aceea se impune dezvoltarea unor metode analitice rapide și eficiente, precum metodele automate de inteligență artificială, care sa fie capabile sa recunoască identitatea individuală sau cea de clasă a acestor noi substanțe interzise.

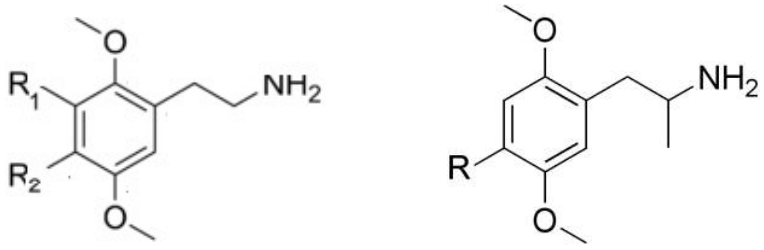


Figura IV.1.1 Structura moleculară a amfetaminelor psihedelice: a) clasa 2C-x; b) clasa DOx. R_1, R_2, R – substituenți.

Tehnicile analitice recomandate pentru identificarea pozitivă (individuală) a amfetaminelor sunt cromatografia în faza gazoasă cuplată cu spectrometria în infraroșu cu transformata Fourier (GC-FTIR) și cromatografia în fază gazoasă cuplată cu spectroscopia de masă (GC-MS). Dintre cele două, GC-FTIR este cea mai potrivită pentru screening-ul (detectarea identității clasei) drogurilor [78].

IV.2. Metode de inteligență artificială

Aplicația de inteligență artificială pe care am dezvoltat-o este concepută să atribuie automat identitatea de clasă nu numai a halucinogenelor de interes (amfetaminele psihedelice), dar și a canabinoidelor JWH, pe bază spectrelor lor ATR-FTIR. Deoarece aceste spectre se caracterizează printr-un raport relativ scăzut semnal / zgomot, numerele de undă la care apar absorbtiile cele mai relevante au fost identificate prin utilizarea algoritmilor genetici (Genetic

Algorithms, GA). Bază de date spectrală selectată a fost utilizată ca intrare pentru construirea unor modele de regresie PLSR (Partial Least Square Regression). Cel mai performant model a fost identificat prin utilizarea a doi indicatori de evaluare a erorilor. Acest model poate fi utilizat cu succes pentru a atribui automat identitatea de clasă a noilor substanțe aparținând claselor vizate de droguri de abuz.

IV.2.1 – Algoritmi Genetici (GA)

Algoritmii genetici sunt o familie de modele computaționale inspirate de evoluția naturală. Acești algoritmi codifică o soluție potențială la o problemă specifică pe o structură de date simplă, asemănătoare cromozomului, aplicând acestor structuri diversi operatori de recombinare pentru a păstra informațiile critice. Algoritmii genetici sunt adesea văzuti ca algoritmi ce optimizeaza funcții. Gama de probleme la care au fost aplicati algoritmi genetici sunt destul de largi. O implementare a GA începe cu o populație de cromozomi (de obicei, aleatorii). Aceste structuri sunt apoi evaluate și oferă oportunități de reproducere în așa fel încât cromozomii care reprezintă o soluție mai bună a problemei țintă sunt alesi astfel incat sa aiba mai multe șanse de a se "reproduce" decât acei cromozomi care reprezinta soluții mai slabe. O caracteristică importantă o reprezintă codificarea variabilelor inițiale ce descriu problema. În mod uzual se foloseste un vector de variabile binare.

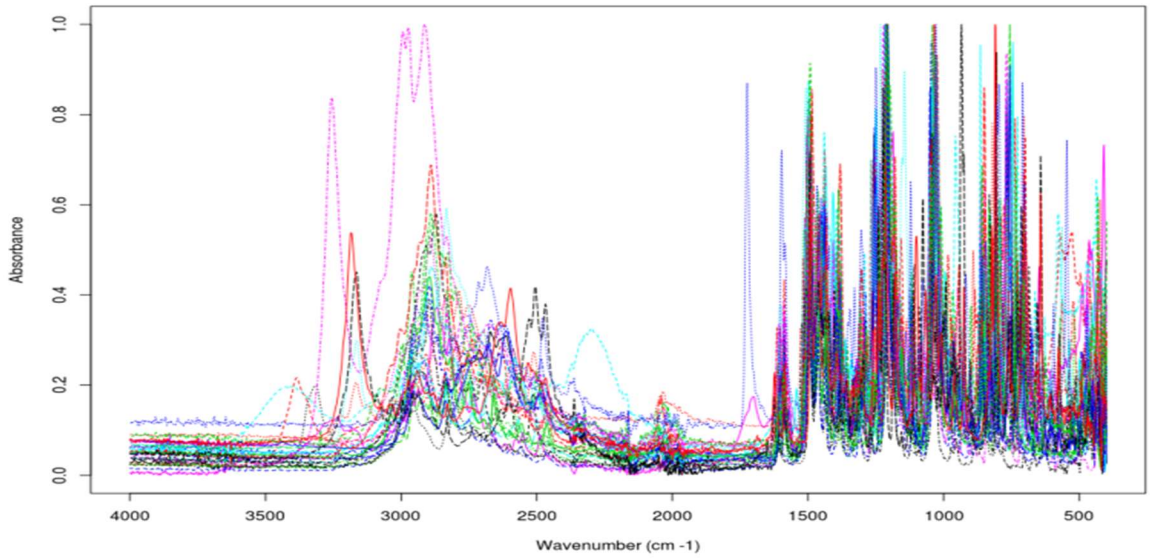


Figura IV.2.1.1 Spectre ATR-FTIR ale halucinogenelor din clasele 2C-x and DOx incluse în bază de date inițială.

Bază de date inițială a constat din 60 de spectre ATR-FTIR normalizate (vezi Tabelul 4.1), care au fost înregistrate între 4000 și 400 cm^{-1}, prin efectuarea unui numar de 1868 scanări efectuate cu o rezoluție de 5 cm^{-1} (vezi Figura IV.2.1.1). În lista compușilor prezentata în Tabelul 4.1 sunt prezentate codurile (ID) corespunzătoare fiecărei amfetamine psihedelice analizate (de la 1 la 28). Compușii având codurile de la 29 la 47 sunt canabinoide JWH (sintetice). Compusii ce au coduri de la 48 la 60 sunt substanțe „negative”, adica o selecție aleatorie de alte substanțe de interes criminalistic.

Tabelul 4.1 Lista compușilor folosiți.

ID	Compus chimic
1	25B-NBOMe HCl (Lot #N18-P1C)
2	25C-NB3OMe HCl (Lot #N17-P72C)
3	25C-NB4OMe HCl (Lot #N17-P73C)
4	25C-NBOMe HCl (Lot #N17-P71D)
5	2,5-Dimethoxy-4-Chloro-amphetamine HCl (Lot #MP193-194)
6	2,5-Dimethoxy-4-ethylamphetamine HCl (Lot #J-1)
7	2,5-Dimethoxy-4-methylamphetamine HCl (Lot #A261B)
8	2,5-Dimethoxyamphetamine HCl (Lot #AKB29A)
9	2,5-Dimethoxyphenethylamine HCl (Lot #MP137-139)
10	25D-NBOMe HCl (Lot #N17-P88C)
11	25E-NBOMe HCl (Lot #N17-P97B)
12	25H-NBOMe HCl (Lot #N16-P81A)
13	25I-NB3OMe HCl (Lot #N17-P74D)
14	25I-NB4OMe HCl (Lot #N17-P75C)
15	25I-NBOMe Base (Lot #SF0003)
16	25I-NBOMe HCl (Lot #N17-P11B)
17	2C-B BZP diHCl (Lot #H-0416)
18	2C-B HCl (Lot #729.1B4.1)
19	2C-E HCl (Lot #H-0407)
20	2C-I HCl (Lot #2TDM-37-04)
21	2C-T-2 HCl (Lot #N1P31)
22	2C-T-7 HCl (Lot #2TDM-198-01)
23	3,4-Dimethoxyamphetamine HCl (Lot #MP150-151 Batch 1)
24	5-Methoxy-alpha-methyltryptamine HCl (Lot #2TDM-97-02)
25	5-Methoxy-Diallyltryptamine HCl (Lot #RM-131001-04)
26	5-Methoxy-N,N-diethyltryptamine HCl (Lot #2DAK-29-05)
27	d,l-DOB HCl (d,l-4-Bromo-2,5-dimethoxyamphetamine HCl) (Lot #274.1B1.1)
28	Methoxetamine HCl (Lot #N16-P100C)
29	AM1220 (Lot #0436983-11)
30	AM1248 (Lot #N1P21)
31	AM2233 (Lot #0436043-18)
32	JWH-018 adamantyl-carboxamide (Lot #0435618-27)
33	JWH-018 Benzimidazole (Lot #RM-131218-04)
34	JWH-018 (Lot #ALB045RC_183-1)
35	JWH-018 N-(5-chloropentyl) analog (Lot #0434099-14)
36	JWH-019 (Lot #K8H81106)
37	JWH-022 (Lot #ALB227-8)
38	JWH-073 (Lot #0409793-37)
39	JWH-081 (Lot #ALB056RC)
40	JWH-122 (Lot #N1P3)
41	JWH-200 (Lot #0424688-3)
42	JWH-203 (Lot #K8H81110)
43	JWH-210 (Lot #N1P36EMG)

44	JWH-250 (Lot #ALB055RC)
45	JWH-307 (Lot #0439287-1)
46	RCS-4 (Lot #N1P38EMG)
47	RCS-8 (Lot #ALB203-21)
48	bk-MDDMA HCl (Lot #0432923-23)
49	Bufotenine Oxalate Hydrate (Lot #042291)
50	Buphedrone HCl(Lot #RM-031113-4)
51	Butylone HCl (Lot #2011DEA003-25A)
52	BZP diHCl (Lot #ALB90-5)
53	Cathine Base (Lot #N11-P-17)
54	Cathine HCl (Lot #284)
55	Cathinone HCl (Lot #113-1191-12)
56	CB-13 (Lot #N1P15)
57	CP 47,497 C8 homologue (Lot #0425163-4)
58	CP 47,497 (Lot #0419860-11)
59	dihydro-PPP HCl (Lot #RM-131218-03)
60	Diisopropyltryptamine HCl (Lot #H-0404)

Probele au fost selectate din cele 3 clase de compuși: Clasa 1 - amfetamine psihedelice; Clasa 2 - canabinoide JWH; Clasa 3 - negative. Baza de date a fost împartită în două seturi, în mod aleatoriu, respectiv un set de instruire care a fost utilizat pentru dezvoltarea modelului și un set de testare pentru validarea modelului. Această secvență de lucru s-a realizat în 8 iterații, cu amestecarea aleatorie a compușilor în setul de instruire și în cel de testare pe fiecare clasă.

Algoritmul genetic a fost dezvoltat în mediul statistic R, utilizând pachetele de librarii *genalg*, *plsVarSel* și *pls*. Algoritmul presupune aplicarea unei funcții de validarea ce utilizează ca model *plsr*, ce presupune folosirea ca metodă de validare încrucișată (LOO) cu o funcție de monitorizare ce realizeaza *Selecția* celor mai bune elemente. *Cromozomul optim*, un vector compus din valori de 0 și 1, este determinat de către funcția de evaluare la momentul atingerii valorii minime.

Prin intermediul functiei *ga_pls* [94], se apeleăza funcția *rbga.bin* [92, 93]. Apelarea funcției s-a facut într-o bucla de iteratie, folosind pe rând clasele stabilite aleator, respectiv seturile de instruire și de testare. În cadrul fiecărei iterații, pe lângă faptul că s-au folosit compuși diferiți aparținând celor 3 clase, s-au folosit și valori diferite ale pragului algoritmului genetic (*GA.threshold* 8:12), numărului de iterații în interiorul modelului (*iters* 40:50) cât și numărul de populații (*popSize* 85:150), date prezentate în Tabelul 4.2.

Tabelul 4.2 Valorile parametrilor folosiți de catre modelul GA în cele 8 iterații.

Iter	GA_threshold	iters	popSize
1	12	48	109
2	11	49	93
3	11	48	102
4	12	43	137
5	12	43	121
6	8	42	86
7	9	43	106
8	12	41	115

Aplicarea GA implică mai mulți pași: a) pentru a selecta populația inițială, o variabilă este setată aleatoriu ca bitul "1", aceasta fiind procedura care selectează variabila corespunzătoare (în timp ce "0" înseamnă non-selecție); b) un model PLSR este dezvoltat la fiecare set de variabile și performanța este evaluată prin utilizarea procedurii de validare încrucișată; c) seturile variabile care produc cele mai bune rezultate sunt selectate pentru a supraviețui până la apariția următoarei generații; d) se realizează încrucișarea și mutația iar apoi se formează noi seturi de variabile; e) seturile de variabile supraviețuitoare și modificate sunt apoi utilizate în următoarea iterație. La finalul parcurgerii modelului, se obține un set vector ce conține valorile numerelor de undă pentru care eroarea de predictie a avut valoarea cea mai mică.

În urma rularii modelului GA, au rezultate 8 seturi de date a cate 30 de vectori cu cele mai semnificative valori ale numerelor de undă, care au fost selectate de către algoritmul genetic ca având relevanța cea mai mare în predicția claselor compușilor selectați.

Întrucât este important să gasim cele mai scăzute valori ale erorii de predicție pe baza setului restrâns de variabile generat de modelul studiat, am analizat și extras seturile cu cele mai bune valori ale erorii medii pătratice a predicției (RMSEP) pentru elementele vectorului rezultat ce contine, în acest caz, 240 de seturi de date.

În urmatoare etapă a evaluarii performanței acestui model de imbunatatire a clasificarii unor substanțe interzise necunoscute, am rulat modelul PLSR de antrenare și predictie cu noile seturi de date. Se poate constata că valorile din etapa de predicție, etapa în care modelul folosește compușii din Tabelul 4.2 din coloanele „Test” pentru fiecare clasă, variază în funcție de parametrii folosiți în modelul GA (Tabelul 4.3), dar și compușii folosiți în antrenarea modelului (Tabelul 4.2, coloanele „Train”).

În urma parcurgerii acestor pași, am constatat că cea mai mică valoare a erorii, în comparație cu cele mai mici valori din cele 30 de valori ai fiecărei iterații, a fost obținută în cazul iteratiei 2 (RMSEP = 0.22198821751905). În raport cu valoarea RMSEP obținută în cazul setului de date inițial, ce cuprinde toate numerele de undă, valoarea cea mai bună a fost identificată în cazul iterației 3:19 (vezi Figura IV.2.1.2), unde eroarea setului GA față de setul inițial este cu 0.18928483 mai mică. (RMSEP set original = 0.497997627, RMSEP set GA = 0.308712797).

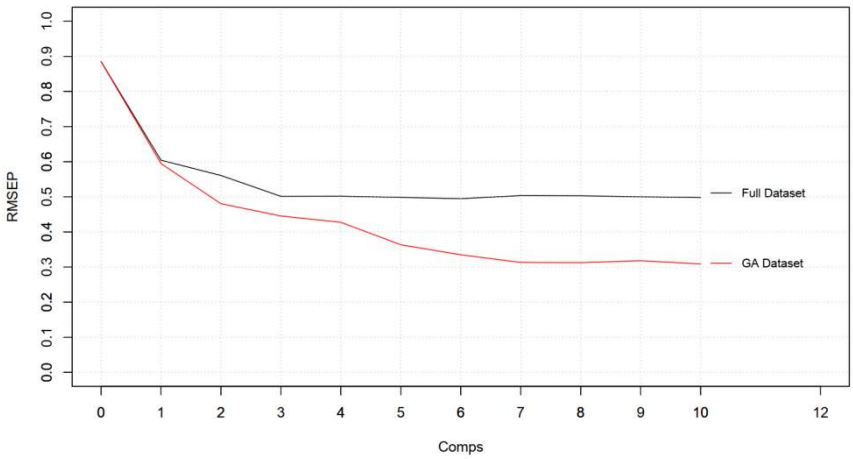


Figura IV.2.1.2 Grafic comparativ cu valorile erorii RMSEP pentru setul de date inițial și setul de date selectat cu GA.

Trebuie observat faptul că RMSEP scade foarte lent (de la 0.313021221117001 la 0.308712797412604) când numărul componentelor este crescut de la 7 la 10, ceea ce indică faptul că majoritatea informațiilor relevante sunt incluse în primele șapte componente. Pe de altă parte, în cazul bazei de date selectate, chiar și aceste ultime componente furnizează informații valoroase: RMSEP variază de la 0.313 pentru 7 componente, la 0.308 pentru 10 componente. Acest comportament poate fi explicat prin selecția semnificativă efectuată de GA, care scade semnificativ (de 10 ori) numărul de variabile de intrare (valori ale absorbanței). În concluzie, analiza RMSEP indică faptul că cele mai bune rezultate sunt obținute pentru bază de date selectată GA și 10 componente.

Selecția celor mai semnificative elemente ale setului de date inițial, ce cuprinde 1868 observatii (numere de undă), este prezentată în Figura IV.2.1.3, ce cuprinde atat spectrele grupelor de substanțe de interes, cât și numerele de undă selectate de GA, care sunt reprezentate prin bare verticale.

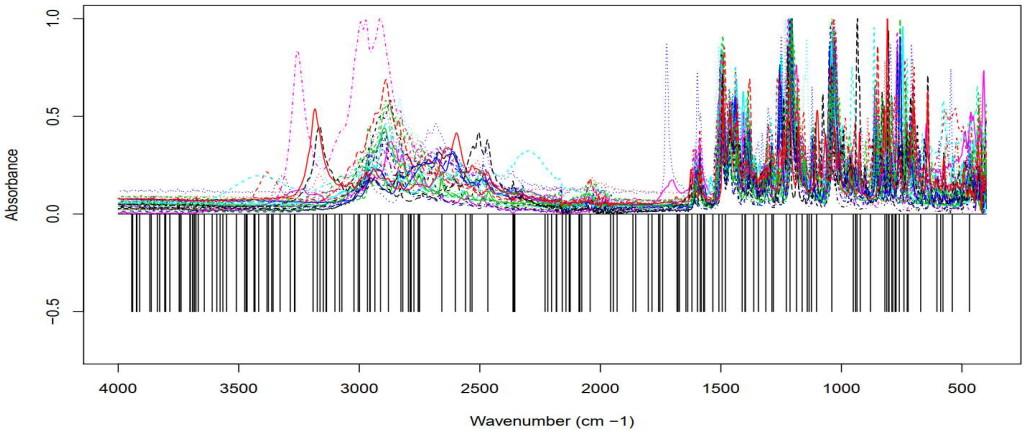


Figura IV.2.1.3 Spectrele ATR-FTIR ale amfetaminelor halucinogene din clasele 2C-x and Dox (numerele de undă selectate de GA sunt evidențiate prin bare verticale sub spectrul ATR-FTIR).

Diferența dintre un obiect optim, ce a dus la o clasificare corectă a compușilor din vectorul de testare (iterația 2:27), și unul cu erori mari la clasificare (iterația 2:14), este ilustrata în Figurile IV.2.1.4 și IV.2.1.5, ce contin spectrele generate cu setul generat de GA.

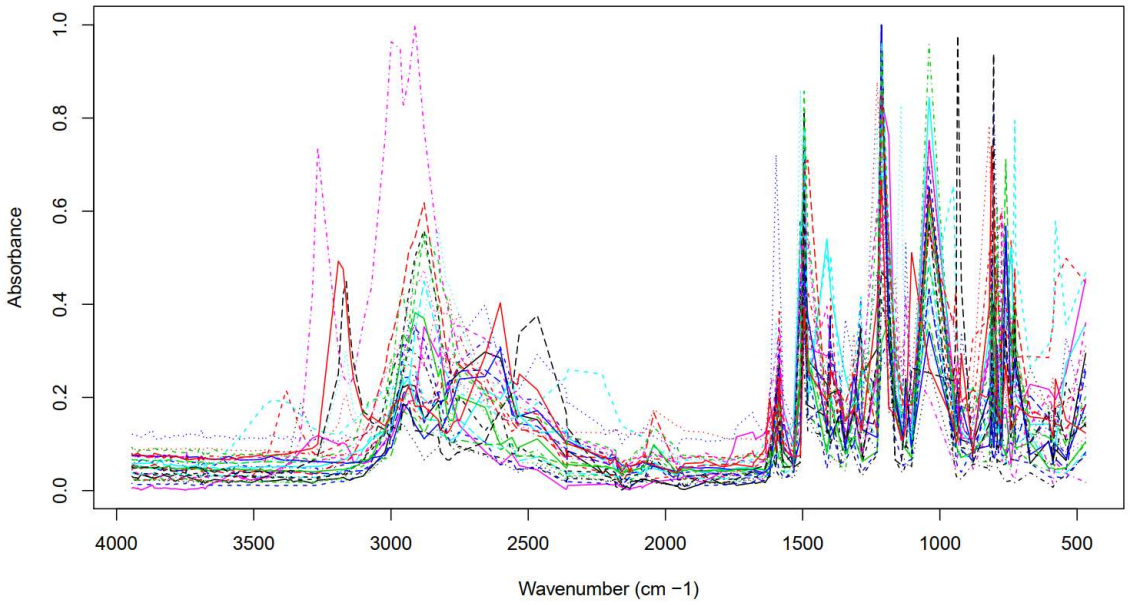


Figura IV.2.1.4 Spectrele ATR-FTIR ale amfetaminelor halucinogene din clasele 2C-x and Dox, trasate pe baza setului GA din iterația 2:27, ce a dus la o clasificare corectă a compușilor din vectorul de testare.

În cazul iterației 2:14 se poate observa o clasificare gresită a compușilor cu indexul 7 (2,5-Dimethoxy-4-methylamphetamine HCl) cu o valoare a predictie de 1.99904747752698 pentru Clasa 1 și compusul cu indexul 13 (25I-NB3OMe HCl) pentru care valoarea predicției a fost de 1.50176435.

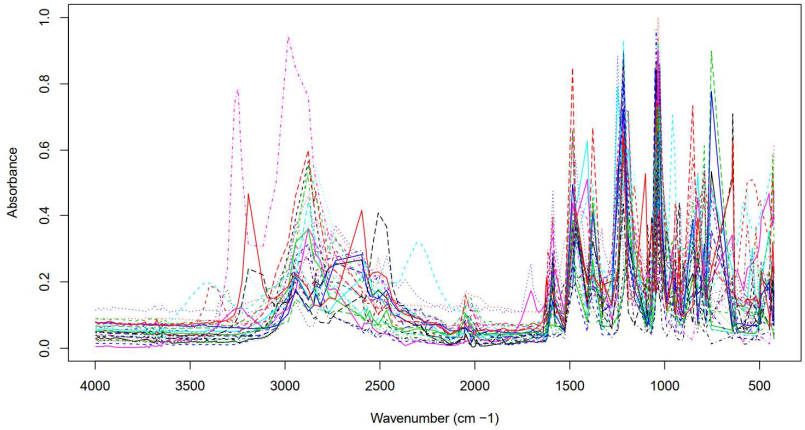
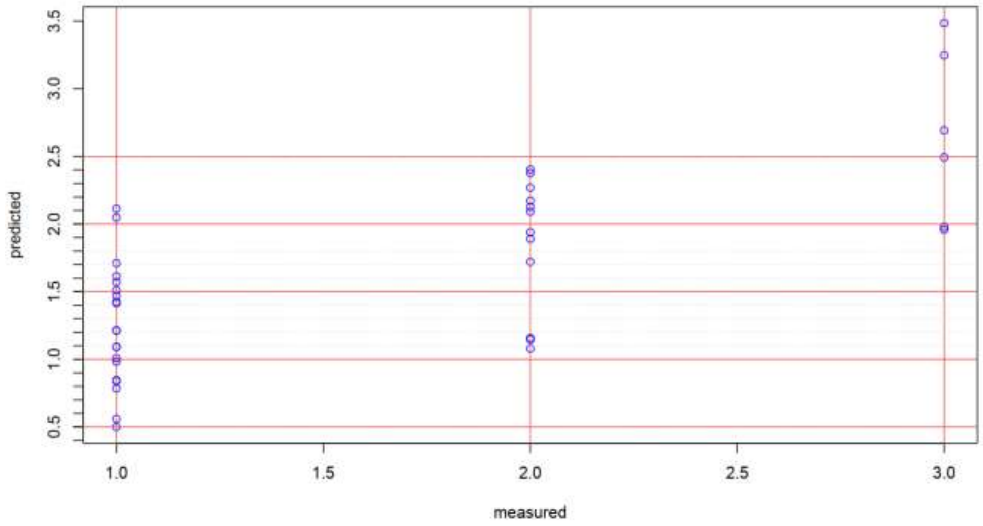


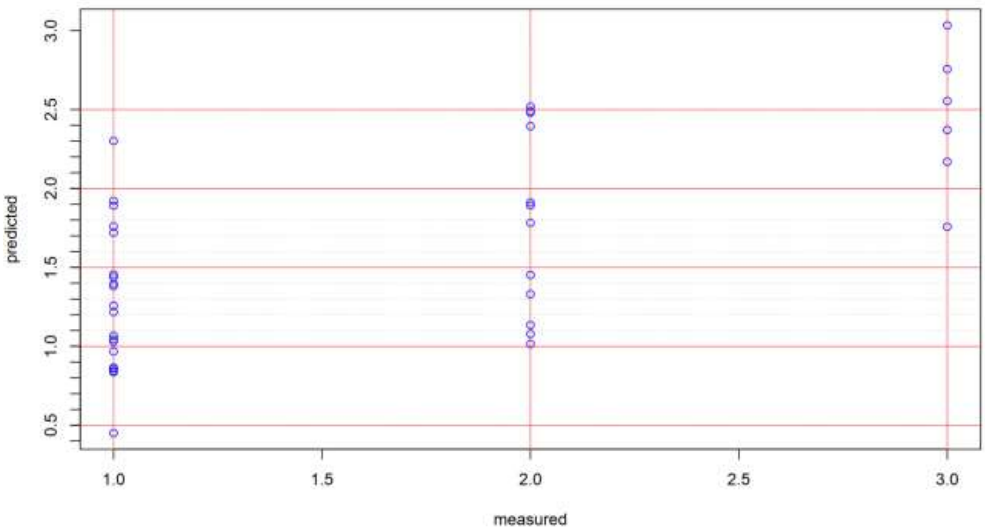
Figura IV.2.1.5 Spectrele ATR-FTIR ale amfetaminelor halucinogene din clasele 2C-x and Dox, trasate pe baza setului GA din iterația 2:14, ce conduce la erori mari la clasificare.

IV.2.2 Algoritmi genetici și regresia prin cele mai mici pătrate parțiale (GA - PLSR)

PLSR a fost aplicat pentru ambele baze de date prin utilizarea a zece componente. Au fost alocate următoarele coduri de clasă: 1 pentru amfetaminele psihedelice, 2 pentru canabinoidele JWH și 3 pentru negative (orice alt compus). Rezultatele obținute pentru baza de date spectrală inițială sunt prezentate în Figura IV.2.2.1 Se observă că valorile prezise obținute pentru unele amfetamine psychedelice sunt mai mari de 1,5, ceea ce înseamnă că sunt clasificate greșit ca fiind canabinoide. În același timp, valorile prezise ale unor canabinoide sunt mai mici de 1,5, ceea ce le include în mod eronat în clasa amfetaminelor psihedelice.



a)



b)

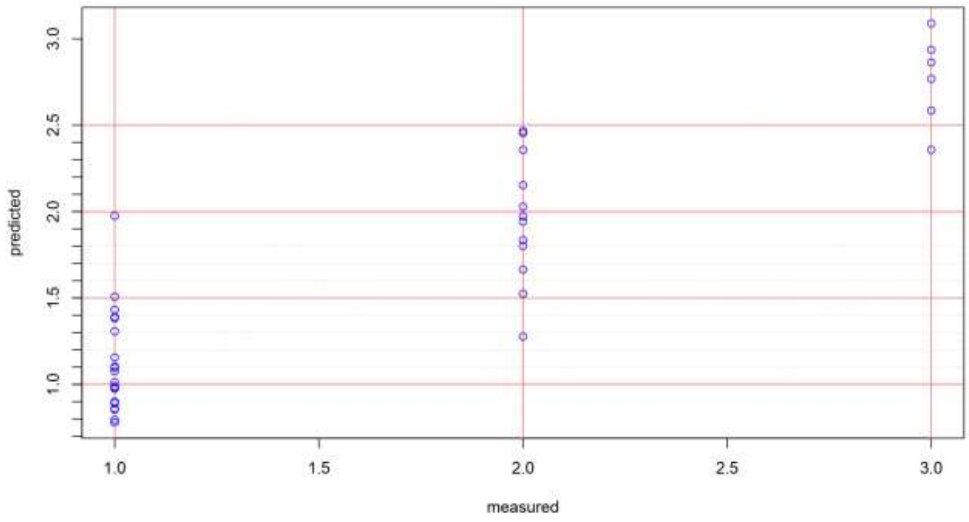
Figura IV.2.2.1 Regresia prin cele mai mici patrate parțiale (PLSR) aplicată pe bază de date ATR-FTIR inițială metode de validare a) CV și b) LOO(Iterația #2). Au fost alocate următoarele coduri de clasă: 1 pentru amfetaminele psihedelice, 2 pentru canabinoidele JWH și 3 pentru negative (orice alt compus).

Un compus ce apartine clasei canabinoizilor este clasificat greșit drept negativ, având o valoare estimată chiar mai mare de 3. În plus, câteva negative sunt clasificate ca canabinoide, deoarece au un scor estimat chiar mai mic de 2 (dar mai mare de 1,5).

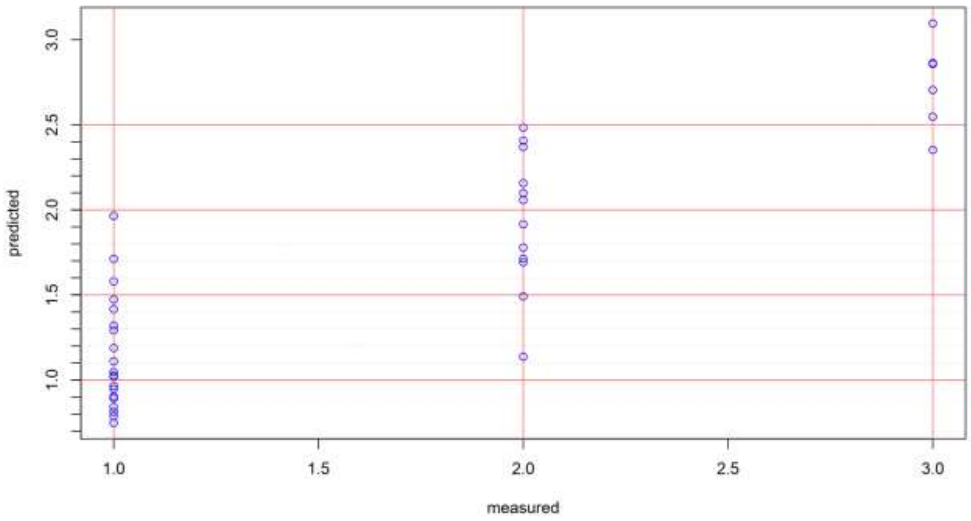
Analizând Figura IV.2.2.1 ce prezintă rezultatele obținute rulând modelul PLSR folosind setul de date ATR-FTIR inițial cu metodele de validare CV și LOO(Iterația #2,vezi Tabel IV.20) se poate observa că în cazul metodei de validare LOO, valorile sunt mai dispersate în afara intervalului asociat clasei, cu mulți compuși ale caror valori de clasificare depășesc valoarea 1,5 ducând la clasificarea incorectă la Clasa 2. De asemenea validarea modelului folosind metoda Cross Validated(CV) duce la clasificarea incorectă a numerosi compuși dar cu valori mai mici ale errorii (vezi Tabel IV.20).

Pe de alta parte, rezultate mult mai bune de clasificare au fost obținute cu baza de date spectrală generată de GA (vezi Figura IV.2.2.2). Toate amfetaminele psihedelice sunt corect clasificate, adică au o valoare estimată între 0,5 și 1,5. Doar trei canabinoide au fost (incorect) atribuite clasei de amfetamine psihedelice și una a fost clasificată ca fiind negativă. În cele din urmă, doar două negative au fost recunoscute incorect ca și canabinoide, deoarece au un scor prognozat mai mic de 2,5 (dar mai mare de 1,5).

În concluzie, sistemul este foarte sensitiv și selectiv din punctul de vedere al amfetaminele psihedelice. Desi rezonabile, rezultatele sunt mai modeste în cazul canabinoidelor, ceea ce sugereaza ca se impune analiza eficienței clasificării prin testarea consecutivă (și nu concomitentă) a celor doua clase de substanțe interzise.



a)



b)

Figura IV.2.2.2 Regresia prin cele mai mici patrate parțiale (PLSR) aplicată pe bază de date generată de catre algoritmul genetic cu metodele de validare a) CV și b) LOO(Iterația #2). Au fost alocate următoarele coduri de clasă: 1 pentru amfetaminele psihedelice, 2 pentru canabinoidele JWH și 3 pentru negative (orice alt compus).

Validarea modelului aplicat pe setul de date GA, s-a realizat, ca și în cazul de date inițial, folosind metodele LOO și CV. Varianța mai mică cat și erori mai mici pot fi observate în cazul validarii folosing CV comparativ cu LOO.

Aplicarea modelului de clasificare pe setul de date inițial ce contine toate cele 1868 de observatii, cu metode de validare Cross Validation (CV) și Leave One Out (LOO) s-a realizat în 8 iterații cu cate 30 de cicluri în care pentru algoritmul genetic au fost variate în mode aleatoriu parametrii de fine tuning, iar pentru modelul de PLSR, a fost schimbată lista compușilor(vezi Tabelul IV.2). Asadar, în cadrul iteratiei #2, considerate în analiza de fată, se poate observa că în cazul validării LOO pentru 6 compuși cu ID-urile 2,1,4,6,10,12 din Tabelul IV.1 clasificarea s-a făcut în mod eronat. În cazul validării CV, pentru compușii cu ID-urile 9,1,15,20,24,10 din Tabelul IV.1, clasificarea a fost greșită de asemenea. Se poate spune că în cazul acestei iterații setul de date inițial duce la o performanță scazută pentru acest model.

Aplicarea modelului menționat mai sus, pe setul de date generat de algoritmul genetic cu parametrii stabiliți în iterația #2, a dus la rezultate considerabil mai bune în ceea ce priveste clasificarea compușilor folosiți în rulare. În Tabelul IV.1, se poate vedea că valorile din iterația #2 cliclul 1, în cazul ambelor metode de validare, au dus la casificarea greșită a doar 2 compuși din clasa 1 (2C-x și DOx). Așa cum se poate vedea și în Figura IV.2.2.2 a) și b), clasificarea cu excepția compușilor cu ID-urile 2 și 10 (vezi Tabelul IV.1) (*25C-NB3OMe HCl Lot #N17-P72C* și *25D-NBOMe HCl (Lot #N17-P88C*) s-a facut în mod corect.

Cele mai bune rezultate observate din cele 8 iterații cu 30 de cicluri, au fost obținute în cazut iteratiei #7 ciclul 4 (vezi Tabelul IV.2), în care toți compușii din Clasa 1 au fost corect clasificați. Analizând valorile preconizate pentru toate cele 3 Clase de compuși pentru această

iterație (#7:4), se poate observa că rezultatele au fost foarte bune comparativ cu cele obținute cu setul de date inițial. Toti compușii sunt corect clasificați, cu o singură excepție, respectiv compusul 31 din Casa 2, canabinoid care a fost clasificat gresit (ca substanța negativă).

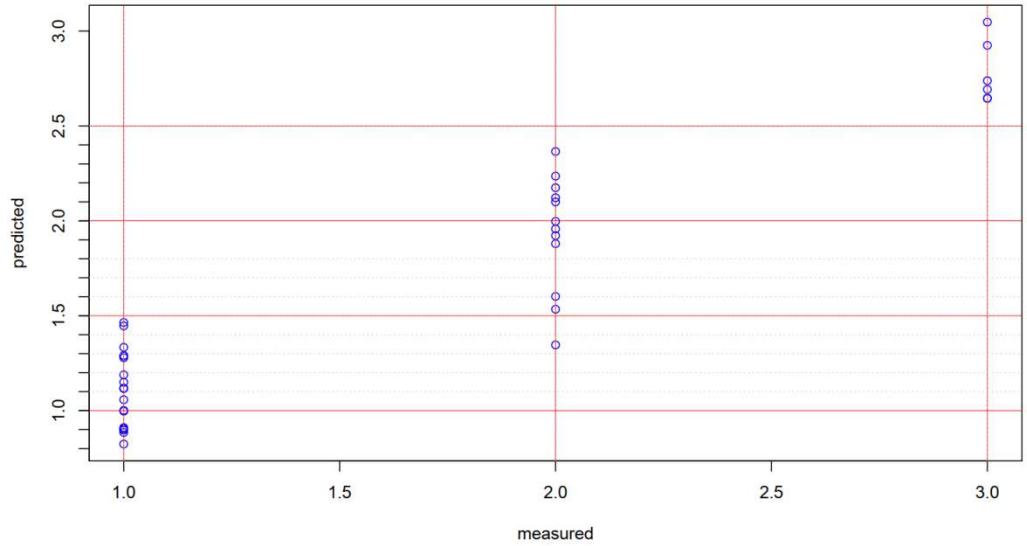


Figura IV.2.2.3 Regresia prin cele mai mici patrate parțiale (PLSR) aplicată pe bază de date generată de catre algoritmul genetic cu metodele de validare: a) CV și b) LOO (Iterația #7:4). Au fost alocate următoarele coduri de clasă: 1 pentru amfetaminele psihedelice, 2 pentru canabinoidele JWH și 3 pentru negative (orice alt compus).

IV.2.3 Random Forest

Modelul Random Forest, bazat pe arborii decizionali descriși în Capitolul III.4, reprezintă un model dezvoltat ca o extensie a arborilor de clasificare și de regresie pentru a îmbunătăți performanța de predicție a modelului [96]. Procesul de construire al acestui modelel constă intr-o repartizare recursivă a setului de date pentru a explora relația dintre variabilele de răspuns și predictor [97]. Arbori generați sunt compuși din subseturi de variabile predictoare și răspunsurile aferente, care sunt apoi agregate pentru a obține valoarea predicției. În timpul generării fiecărui arbore, un eșantion al datelor originale este selectat, iar performanța fiecărui arbore este validată folosind o treime diin setul de date inițial care nu au fost utilizate pentru construirea acelui arbore.

Modelul a fost preocesat în R folosing pachetul „randomForest”, care este o implementare a algoritmului lui Breiman [96, 97, 98]. S-a parcurs modelul, în iterații multiple, cu un număr variat de cicluri, pentru a determina cele mai bune valori ale erorilor și predicțiilor. A fost ales un numar de 500 de arbori iar valoarea *mtry*, reprezentând numărul de variabile eșantionate aleatoriu, care devin candidate pentru fiecare separare, a fost determinată folosind *tuneRF*. Acest modul a ajutat la identificarea valorii optime a variabilelor ce vor fi selectate la fiecare separare în raport cu eroarea Out-of-Bag. Valorile determinate sunt prezentate în Tabelul IV.3.

Tabelul IV.3 Valorile candidate pentru fiecare separare (*mtry*).

mtry	OOBError
14	0.216666667
20	0.216666667
30	0.2
45	0.183333333
67	0.1
100	0.033333333
150	0.033333333

Asa cum se poate observa în Figura IV.2.3.1, valorea optimă este 0,033 și a fost identificată începând cu 100 în ceeea ce priveste variabilele selectate la fiecare separare (*mtry*).

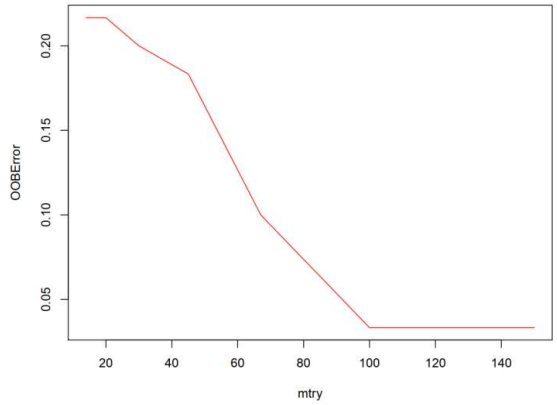


Figura IV.2.3.1 Determinarea numărului de variabile ce vor fi selectate la fiecare separare în funcție de eroarea OOB.

Modelul a fost rulat cu date de intrare din 2 surse, setul de date complet cu toate numere de undă cât și setul de date generat de algoritmul genetic. În ceea ce priveste setul de date GA, a fost considerat în mod aleatoriu drept candidat setul determinat din iterația #7:4, descris în detaliu la capitolul IV.2.1. Antrenarea modelului s-a realizat folosind un eșantion (vezi Tabel IV.4) compus din 42 de substanțe din lista compușilor descriși în Tabel IV.1

Tabelul IV.4 Lista compușilor folosiți în antrenarea și testarea modelului Random Forest.

Eșantion antrenare					Eșantion testare	
1	13	26	39	54	2	34
3	14	28	40	56	7	38
4	15	29	41	57	8	43
5	19	30	42	58	16	45
6	20	31	44	59	17	47
9	21	32	46		18	50
10	22	35	48		25	52
11	23	36	49		27	53
12	24	37	51		33	55

Antrenarea modelului s-a realizat în mai multe iterații, iar cele mai bune valori a erorii de clasificare în cele trei grupe de compuși vizati a determinat selecția pentru etapa de predictie. În ce privește eroarea în raport cu numărul de arbori pentru cele două seturi de date, variația acesteia se poate vedea în Figura IV.2.3.2 pentru setul de date complet și Figura IV.2.3.3 pentru setul de date selectat prin GA.

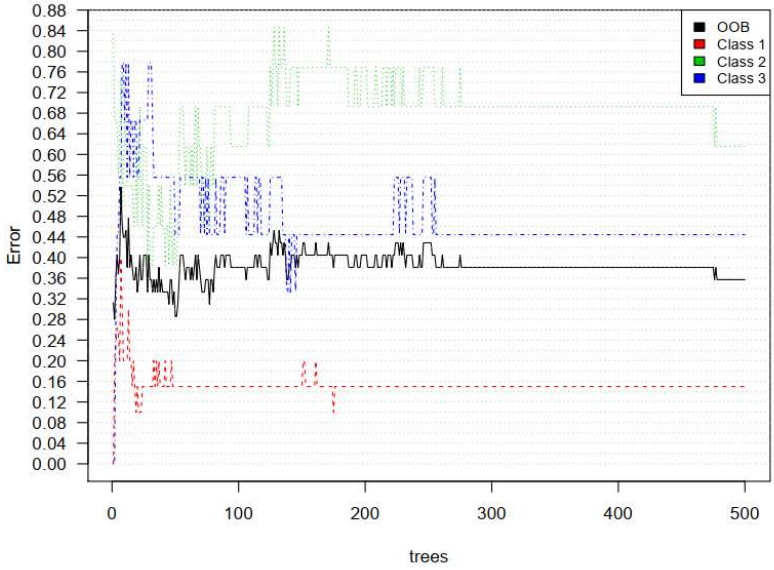


Figura IV.2.3.2 Variația erorii de clasificare în raport cu numărul arborilor, obținută pentru setul de date complet.

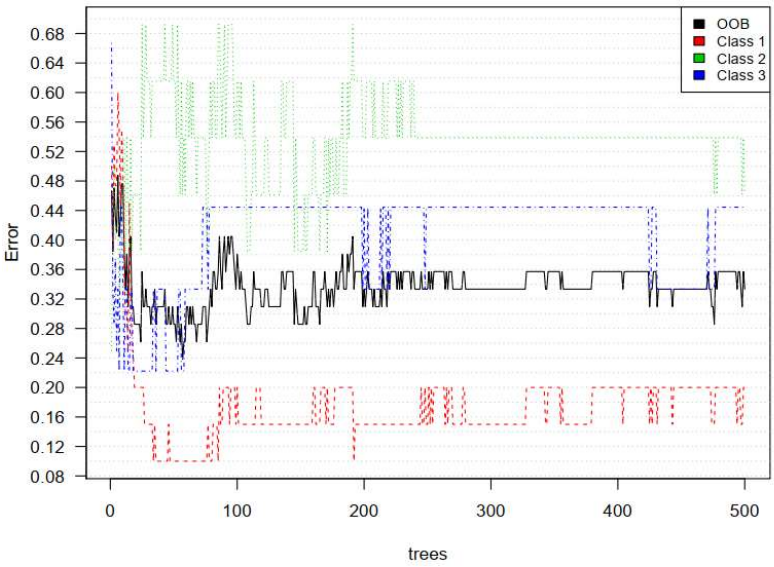


Figura IV.2.3.3 Variația erorii de clasificare în raport cu numărul arborilor, obținută pentru setul de date selectat prin GA.

Testarea acestui modelul s-a realizat pentru esantionul prezentat în Tabelul IV.4. În cazul setului de date complet, compusul 7 (2,5-Dimethoxy-4-methylamphetamine HCl) clasificarea este eronata. În cazul modelului pentru setul de date selectate prin GA, toti compușii ce apartin Clasei 1 au fost corect clasificați.

IV.2.4 KNN

Modelul KNN a fost dezvoltat în mediul de lucru R [99,100] folosind libraria *class*. Performanța acestui model a fost evaluata variind valoarea lui K, numărul celor mai apropiați vecini. Într-o perspectivă comparativă, antrenarea și predicția modelului s-a realizat în mai multe iterații, atat pentru setul de date ce cuprinde spectrele complete (cu toate numerele de undă), cât și pentru setul de date generat de GA. Modelararea s-a realizat pe doua esantioane de antrenare și testare, cu o proportie de 70% (vezi Tabelul IV.5)

Tabelul IV.5 Lista Compușilor selectați pentru etapa de antrenare și testare a modelului KNN.

Set Compusi Antrenare Model					Set Compuși Testare Model	
1	12	27	40	50	**4**	**29**
2	14	28	41	52	**8**	33
3	15	30	42	53	**9**	34
5	17	31	44	54	**13**	35
6	18	32	45	55	**16**	38
7	19	36	46	56	**21**	43
10	20	37	47	57	**23**	49
11	22	39	48	59	**26**	51

Se poate observa în Figura IV.2.4.1 că, pentru setul de date inițial (complet), valorile maxime ale acurateții sunt mici atunci cand numărul K este mai mic de 14. În cazul setului de date generat de GA, ce cuprinde cele mai relevante variabile, acuratețea variază relativ puțin, luând valori între 90 și 100.

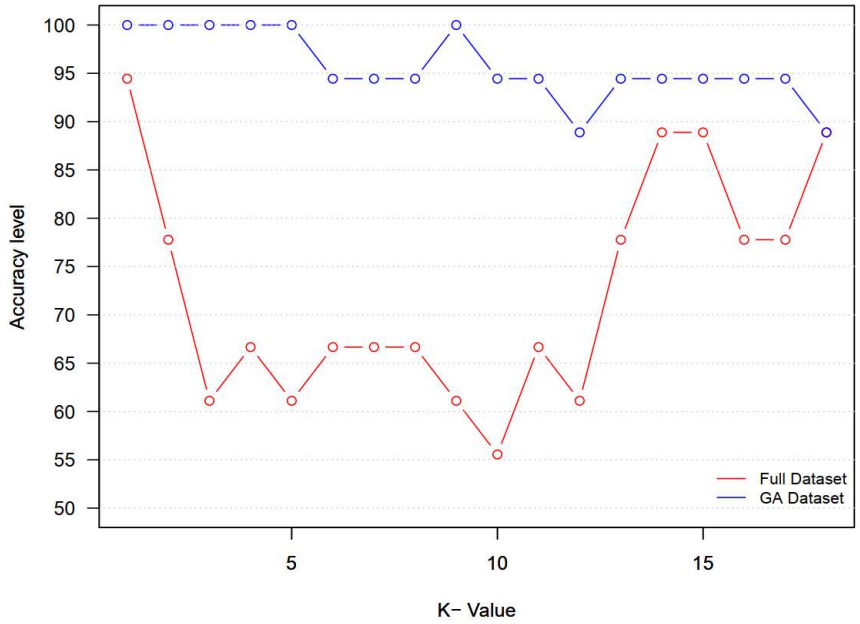


Figura IV.2.4.1 Variația acurateții în raport cu numărul de vecini K.

Analizând valorile predicțiilor în ce privește clasificarea compușilor din eșantionul de testare, putem observa în Tabelul IV.6 și Tabelul IV.7 că în cazul setului de date complet sunt greșit clasificați un număr de 3 compuși, ce provin din Clasa 1. În schimb, pentru setul de date selectat prin GA, clasificarea s-a facut în mod corect pentru toți compușii din Clasa 1.

În ce priveste clasificarea compușilor din clasele 2 și 3, se poate observa o performanță crescută a modelului în ceea ce priveste eroarea de clasificare în cazul celor doua seturi de date (vezi Tabelul IV.8), setul de date GA, ducând la valori foarte mari ale acurateții comparativ cu setul inițial.

Tabelul IV.6 Valori ale predicțiilor pentru setul de date inițial. Valorile evidențiate în galben reprezintă cazurile in care modelul a atribuit compusului analizat o identitate de clasă greșită.

Class / Pred	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1	1	1	1	2	1	2	2	2	2	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	2	2	2	2	2	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	3	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1
2	3	3	3	3	2	2	2	2	2	2	2	2	2	2	2	3	3	2	3
2	2	2	2	2	2	2	2	2	2	2	2	3	2	2	2	2	3	2	2
3	3	3	3	3	3	2	2	2	2	2	3	3	3	3	3	3	3	3	3
3	3	2	2	3	3	3	3	3	3	2	2	2	2	3	3	2	3	3	3
3	3	2	2	2	2	2	2	3	2	2	2	2	3	3	3	3	3	3	3
3	3	2	2	3	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3

Tabelul IV.7 Valorile predicțiilor pentru setul de date selectat prin GA. Valorile evidențiate în galben reprezintă valori de clasificare unde modelul a atribuit o altă clasă compusului analizat, clasificarea fiind greșită.

Class / Pred	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2

Valorile acurateții pentru seria de iterații pentru ambele seturi de date sunt prezentate în Tabelul IV.8. și în Tabelul IV.9. Acestea prezintă valorile sensitivității și specificității pentru seriile de iterații, defalcate pentru fiecare clasa de compuși.

Tabelul IV.8 Indicii de evaluare a performanței modelului KNN.

Set de Date Inițial						
	1	2	3	4	5	6
Accuracy	0.944444444	0.777778	0.611111	0.666667	0.611111	0.666667
Kappa	0.914285714	0.647059	0.38835	0.480769	0.38835	0.470588
AccuracyLower	0.7270564	0.523627	0.357451	0.409925	0.357451	0.409925
AccuracyUpper	0.998594444	0.935908	0.827014	0.866573	0.827014	0.866573
AccuracyNull	0.444444444	0.444444	0.444444	0.444444	0.444444	0.444444
AccuracyPValue	0.00001075	0.00427	0.118141	0.048664	0.118141	0.048664
Set de Date GA						
	1	2	3	4	5	6
Accuracy	1	1	1	1	1	0.944444
Kappa	1	1	1	1	1	0.912621
AccuracyLower	0.814698032	0.814698	0.814698	0.814698	0.814698	0.727056
AccuracyUpper	1	1	1	1	1	0.998594
AccuracyNull	0.444444444	0.444444	0.444444	0.444444	0.444444	0.444444
AccuracyPValue	4.5784090	4.5784099	0.0000005	0.0000005	0.0000005	0.0000108

IV.2.5 - Clasificarea SVM

Support Vector Machines (SVM) reprezinta un model cu preformanțe ridicate ce poate fi folosit cu success în clasificarea generală sau în regresie. În clasificarea binară, obiectivul principal îl reprezintă predicția claselor obiectelor **y**, (-1,+1) dintr-un set de date cu m dimensiuni reprezentat printr-un vector scris $X = (x_1, x_2, \ldots, x_m)$. În cazul de fata, **m** este indicele numărului de undă corespunzator absorbanțelor din setul de date prezentat în capitolul anterior. Utilizarea acestui model pentru clasificarea claselor de compuși de interes, respectiv

amfetaminele psihedelice 2C-x și DOx (Clasa 1), canabinoidele JWH (Clasa 2) și negative (Clasa 3), presupune o etapa de instruire a modelului folosind eșantionul de antrenare din Tabelul IV.9

Tabelul IV.9 Lista compușilor selectați pentru etapa de instruire și testare a modelului SVM.

Esantion antrenare model					Esantion testare model	
3	16	27	42	52	1	30
4	17	28	43	54	2	36
5	19	31	44	55	9	37
6	21	32	45	57	10	39
7	22	33	46	58	13	40
8	23	34	47	60	14	49
11	24	35	48		18	53
12	25	38	50		20	56
15	26	41	51		29	59

Modelul a fost realizat în mediul de lucru R, folosind libraria *e1071*. În scop comparativ, s-au folosit ca seturi de date de antrenare și testare atât setul complet de date (ce cuprinde 1868 de scanări cu rezolutia de 5 cm^{-1} de la 4000 la 400cm^{-1}), cât și setul de date selectat prin GA. Performanța modelului depinde printre alți parametri și de selecția celor mai semnificative elemente, reducând sau eliminând în acest fel overfitt-ingul. Modelul SVM a fost rulat folosind funcția kernel liniara [101].

Valorile probabilităților obținute în urma clasificării cu ajutorul modelului SVM pentru setul de date complet sunt prezentate în Tabelul IV.10. Putem remarca ca 3 compuși din Clasa 1 sunt clasificați gresit, respectiv sunt atribuiți claselor 2 și 3. De asemenea, unele substanțe din Clasa 2, sunt clasificate în mod eronat ca aparținând claselor 1 sau 3, iar unele din Clasa 3 sunt atribuite eronat claselor 2.

Tabelul IV.10 Predicțiile obținute cu ajutorul modelului SVM pentru setul de date complet. Valorile evidențiate în galben reprezintă cazurile in care modelul a atribuit compusului analizat o identitate de clasă greșită.

FULL								
Index		Clasa	Valori decizionale			Probabilitati		
Compus	Clasa	Predict.	1.2	1.3	2.3	Clasa 1	Clasa 2	Clasa 3
1	1	1	0.158369	2.315982	2.796912	0.6818	0.31594	0.00226
2	1	1	0.368545	0.843619	0.144417	0.687949	0.214398	0.097653
9	1	2	-0.43154	1.096326	0.911375	0.454764	0.489576	0.055659
10	1	2	-0.81892	0.663276	0.927262	0.321095	0.598647	0.080258
13	1	3	-0.71498	-0.00702	-0.41116	0.306281	0.316294	0.377425
14	1	1	0.541453	1.062966	0.512503	0.750231	0.193084	0.056686
18	1	1	0.82264	1.624528	1.918238	0.843087	0.147265	0.009648
20	1	1	0.286188	0.782534	1.215637	0.676816	0.26601	0.057174
29	2	1	0.656881	0.5139	-1.32387	0.718478	0.090152	0.19137
30	2	1	0.085215	1.855435	1.851623	0.65347	0.337297	0.009233
36	2	1	-0.31185	0.284426	-0.32608	0.443757	0.292685	0.263558

37	2	1	-0.17235	0.838247	0.833489	0.534845	0.395439	0.069716
39	2	3	-0.68484	-0.8245	-0.04206	0.147264	0.405419	0.447317
40	2	2	-1.14314	-0.38894	1.015961	0.161968	0.696371	0.141661
49	3	3	-1.65297	-0.53853	-0.42796	0.11532	0.334476	0.550204
53	3	2	-0.93107	0.653981	1.807352	0.288976	0.674564	0.03646
56	3	2	-1.55819	-0.97003	0.085518	0.070449	0.525114	0.404437
59	3	3	-0.69663	-0.65261	-1.14642	0.15244	0.137401	0.710159

Antrenarea și testarea modelului SVM studiat în cazul setului de date generat de GA a dus la obținerea unor rezultate mult mai bune. Tabelul IV.11 arata că, în acest caz, clasificarea compușilor din Clasa 1 a fost realizată corect pentru tot esantionul de testare, cu excepția compusului 9. Totodată, toți compușii din clasele 2 sunt corect clasificați, cu excepția substanțelor 29 și 37. Pe de altă parte insă, multe negative (clasa 3) au fost clasificate incorect ca fiind substanțe apartinand Clasei 2. Se poate deduce deci ca selecția variabilelor de intrare prin GA imbunatateste semnificativ senzitivitatea și selectivitatea sistemului de screening din punctul de vedere al amfetaminelor (Clasa 1), dar nu reuseste sa distinga eficient canabinoidele (Clasa 2) de negative (Clasa 3).

Tabelul IV.11 Predicțiile obținute cu ajutorul modelului SVM pentru setul de date selectate prin GA. Valorile evidențiate în galben reprezintă cazurile in care modelul a atribuit compusului analizat o identitate de clasă greșită.

GA								
		Clasa	Valori decizionale			Probabilitati		
Compus	Clasa	Predict.	1.2	1.3	2.3	Clasa 1	Clasa 2	Clasa 3
1	1	1	0.346012	2.53357	3.45121	0.723163	0.273401	0.003436
2	1	1	0.59443	2.806744	1.819817	0.79869	0.194765	0.006546
9	1	2	-0.38951	1.860823	1.620672	0.411327	0.564667	0.024006
10	1	1	0.111278	1.441701	2.594055	0.622194	0.357307	0.020499
13	1	1	1.165823	1.742053	2.059693	0.901575	0.081397	0.017028
14	1	1	1.202964	1.88695	1.497504	0.907611	0.076755	0.015634
18	1	1	0.969905	1.277444	2.435506	0.854744	0.110921	0.034336
20	1	1	0.858019	1.076589	1.265034	0.811265	0.132637	0.056098
29	2	1	-0.11678	0.246135	-0.74829	0.447698	0.226031	0.32627
30	2	2	-0.50186	0.986232	2.246244	0.35782	0.611957	0.030223
36	2	2	-0.63879	0.1427	0.762419	0.264559	0.578493	0.156948
37	2	3	-0.76995	0.228591	-0.59369	0.302427	0.33199	0.365584
39	2	2	-0.29599	0.45305	0.232302	0.396319	0.41232	0.191362
40	2	2	-0.95851	0.490418	1.158985	0.197102	0.717442	0.085456
49	3	2	-1.93684	-0.78537	-0.00415	0.06819	0.503353	0.428457
53	3	2	-1.30171	-0.37723	0.510798	0.114019	0.646618	0.239363
56	3	2	-1.41472	0.262078	0.515748	0.129634	0.684408	0.185958
59	3	3	-1.46104	-2.15594	-1.76074	0.015461	0.062169	0.92237

Evolutia valorilor FPR (False Positive Rate) în raport cu sensitivitatea este reprezentata pentru cele două seturi de date în Figura IV.2.5.1 și Figura IV.2.5.2. Se observă

că, în cazul setului de date selectat prin GA, valoarea inițială a senzitivității este semnificativ mai mare decât cea obținută pentru setul de date complet.

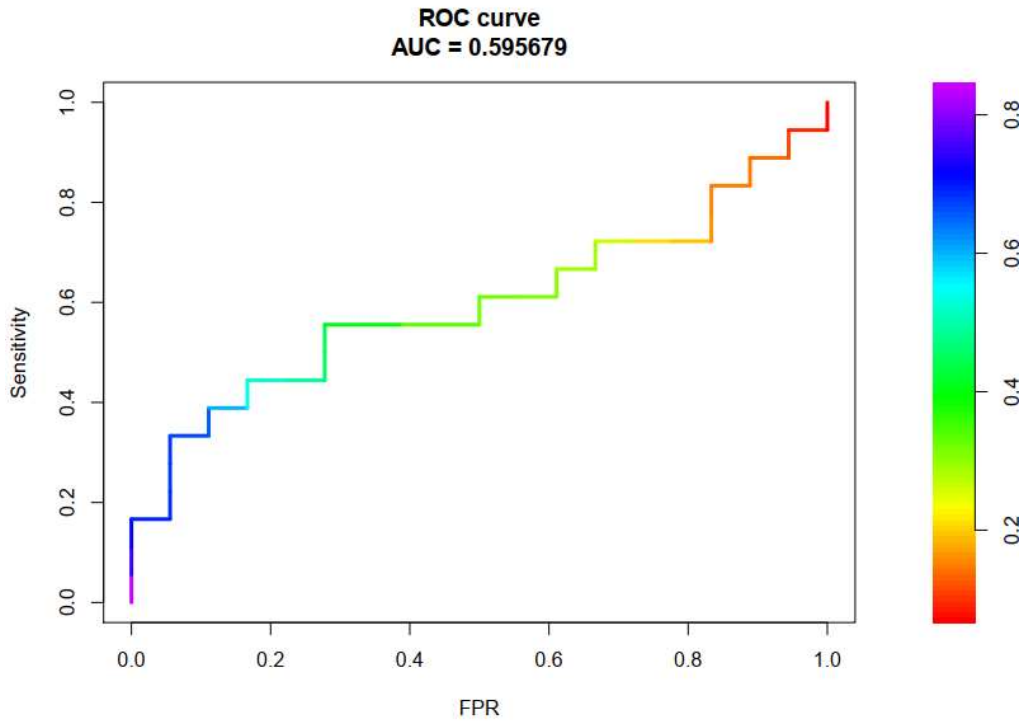


Figura IV.2.5.1 Rata falselor pozitive (False Positive Rate, FPR) în raport cu Sensitivitatea, pentru setul de date complet.

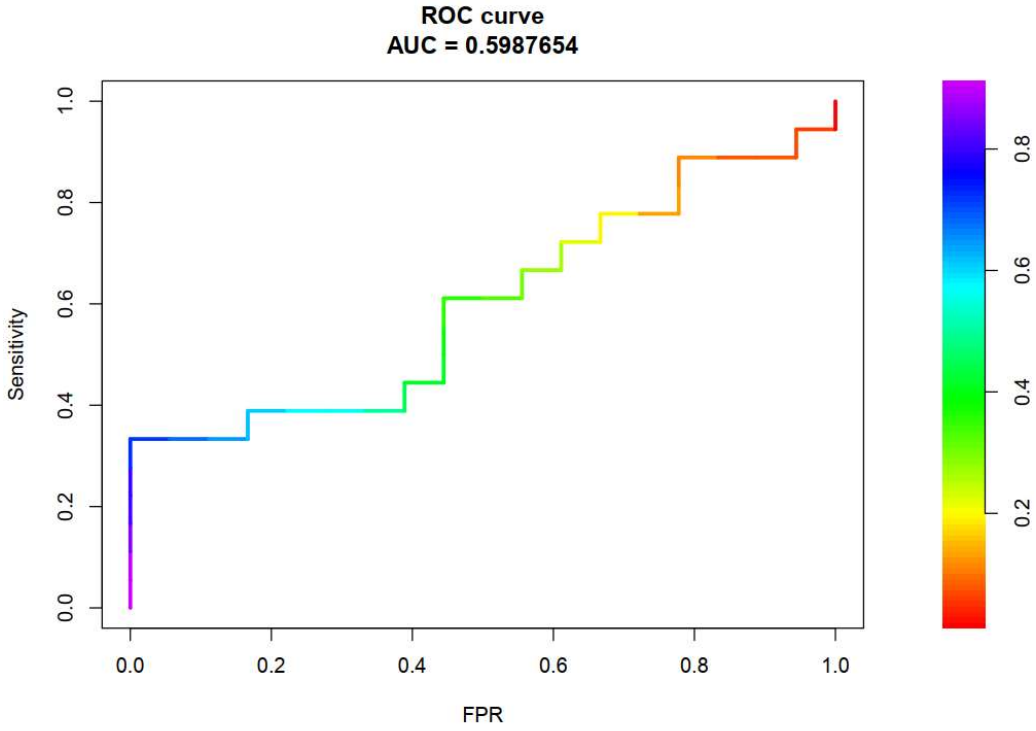


Figura IV.2.5.2 Valorile False Positive Rate (FPR) în raport cu Sensitivitatea, pentru setul de date selectat prin GA.

Pentru o analiză mai detaliată a comportarii modelului SVM din punctul de vedere al tuturor claselor de substanțe analizate, atât în cazul setul de date complet, cât și pentru setul de date selectate cu ajutorul GA, a fost analizata performanța modelelor SVM pentru cele doua seturi de date pe bază unei serii de indici precum senzitivitatea, specificitatea, precizia, rata de detecție, prevalența detecției, etc. Rezultatele sunt prezentate în Tabelul IV.12.

Se observă că modelul SVM are performanțe mai bune în cazul tuturor substanțelor de abuz pozitive. În cazul drogurilor din Clasa 1, toți indicii sunt net superiori în cazul setului de date selectate prin GA, cu excepția prevalenței (care se menține la același nivel) și a prevalenței de detecție (care scade ușor). În cazul drogurilor din Clasa 2 se observă însă o imbunatățire chiar și mai importantă decât în cazul Clasei 1: toți indicii sunt net superiori în cazul setului de date selectate prin GA, cu excepția specificitații și prevalenței, care se mentin la acelasi nivel.

În schimb, rezultatele arată că aceste îmbunătățiri se obțin cu prețul unei deteriorari a performanțelor de clasificare a negativelor (compușilor din Clasa 3). Deși unii indici, precum âspecificitatea, se imbunătățesc, sau alții rămîn la același nivel (e.g. precizia), mulți indici scad în cazul negativelor (Clasa 3).

Tabelul IV.12 Performanța modelelor SVM pentru setul de date complet și pentru setul de date selectate cu ajutorul GA.

Full Dataset			
	Clasa 1	Clasa 2	Clasa 3
Sensitivity	0.625	0.166667	0.5
Specificity	0.6	0.666667	0.857143
Pos Pred Value	0.555556	0.2	0.5
Neg Pred Value	0.666667	0.615385	0.857143
Precision	0.555556	0.2	0.5
Recall	0.625	0.166667	0.5
F1	0.588235	0.181818	0.5
Prevalence	0.444444	0.333333	0.222222
Detection Rate	0.277778	0.055556	0.111111
Detection Prevalence	0.5	0.277778	0.222222
Balanced Accuracy	0.6125	0.416667	0.678571

Ga Dataset			
	Clasa 1	Clasa 2	Clasa 3
Sensitivity	0.875	0.666667	0.25
Specificity	0.9	0.666667	0.928571
Pos Pred Value	0.875	0.5	0.5
Neg Pred Value	0.9	0.8	0.8125
Precision	0.875	0.5	0.5
Recall	0.875	0.666667	0.25
F1	0.875	0.571429	0.333333
Prevalence	0.444444	0.333333	0.222222
Detection Rate	0.388889	0.222222	0.055556
Detection Prevalence	0.444444	0.444444	0.111111
Balanced Accuracy	0.8875	0.666667	0.589286

Totuși, aceste scăderi nu sunt de natură să afecteze performanțele generale ale sistemului de clasificare. Caracteristicile generale ale performanțelor modelelor SVM construite pentru setul de date complet cât și pentru setul de date selectat cu ajutorul GA sunt prezentate în Tabelul IV.13. Acesta indică faptul că, în cazul modelului SVM construit cu ajutorul setul de date complet, acuratetea totala este de 0,44. În acelasi timp, în cazul modelului SVM construit cu ajutorul setul de date selectate cu ajutorul GA, aceasta creste cu 50%, ajungand sa fie egala cu 0,66.

Tabelul IV.13 Caracteristici generale ale performanțelor modelelor SVM construite pentru setul de date complet și pentru setul de date selectat cu ajutorul GA.

FULL			GA	
Accuracy	0.444444444		Accuracy	0.666666667
Kappa	0.126213592		Kappa	0.470588235
AccuracyLower	0.215301507		AccuracyLower	0.409925238
AccuracyUpper	0.692428341		AccuracyUpper	0.866572597
AccuracyNull	0.444444444		AccuracyNull	0.444444444
AccuracyPValue	0.5899888		AccuracyPValue	0.048664156
McnemarPValue	0.572406704		McnemarPValue	NA

Se poate deci trage concluzia că și în cazul acestei combinatii de tehnici multivariate, Selecția variabilelor de intrare prin GA imbunatățeste semnificativ performanțele sistemului de clasificare în ceea ce priveste eficiența recunoașterii amfetaminelor halucinogene 2C-x și DOx. Pe de altă parte, sistemul nu reușește sa distingă eficient canabinoidele de substanțele negative, ceea ce sugereaza testarea eficientei clasificarii substanțelor în cascadă, adică analiza acestor două tipuri de substanțe printr-un modul de clasificare separat.

IV.2.6 - LRCM (Logistic Regression Classification Model) - metodele de regularizare LASSO și Ridge

Modelul descris în acest capitol se bazează pe regresia logistică, utilizată pentru determinarea apartenenței la o clasă de compuși cunoscute, prin modelarea probabilităților. Funcția de regresie este o relație neliniară de combinații liniare ale elementelor obiectelor. Pentru clasificarea substanțelor pe baza spectrelor lor ATR-FTIR, au fost alocate urmatoarele valori pentru rezultatul acestui model de regresie logistică: 1 daca compusul aparține clasei formate din amfetamine 2C-x and DOx, respectiv 0 pentru alte categorii de substanțe testate.

Ca și în cazul modelelor prezentate anterior, spectrele ATR-FTIR, obținute în urma a 1686 de scanări efectuate pentru fiecare substanța în parte, reprezintă setul de date complet, inițial. Construcția acestui model are drept scop determinarea celor mai relevanți predictori și eliminarea elementelor irelevante pentru o clasificare corectă. Aceste elemente ar putea să conducă la o eficiență scazută a clasificării, Întrucât numărul de covariate creste. Overfitting-ul poate apărea datorită numărului mare de elemente ce compun obiectul,în cazul de fata, spectrul compusului analizat, pe masură ce se incearcă obținerea unor valori minime pentru esantion. Această situație poate fi evitată penalizând modelul, respectiv prin atingerea unui echilibru între complexitate și potrivire (fitting). Penalizarea reprezintă magnitudinea coeficienților regresiei scalați de un factor λ. Valoarea acestui factor s-a ales prin validare incrucișată (cross-validation).

Setul de date a fost compus din spectrele a doua seturi, amfetaminelor psihedelice 2C-x and DOx (cod atribuit 1) și o serie de negative, respectiv 12 compuși selectați în mod aleator (cod atribuit 0). Acest set de date reprezintă vectori cu 1868 de elemente, ce reprezintă absorbțiile masurate la 1868 numere de undă echidistante de-a lungul intregului spectru infrarosu. Al doilea set de date contine vectori ce contin absorbtiile masurate la un numar de 186 de numere de undă, variabile ce au fost selectate din cele inițiale cu ajutorul GA.

Modelul a fost dezvoltat în mediul R, folosind librariile *glmnet* și *kernlab*. Penalizările Lasso și Ridge s-au realizat în mai multe iterații, în care factorul λ a variat. S-a urmărit obținerea unor valori maxime în ce priveste performanța predicțiilor pentru clasele de compuși testati. Seturile de date au fost împărțite în eșantioane de antrenare și testare. Lista compușilor testați este prezentată în Tabelul IV.14.

Tabelul IV.14 Lista compușilor selectați pentru etapa de testare a modelelor de regresie logistică.

Index	Compus	Clasa
1	25B-NBOMe HCl	1
2	25C-NB3OMe HCl	1
5	2,5-Dimethoxy-4-Chloro-amphetamine HCl	1
12	25H-NBOMe HCl	1
14	25I-NB4OMe HCl	1
16	25I-NBOMe HCl	1
24	5-Methoxy-alpha-methyltryptamine HCl	1
27	l-4-Bromo-2,5-dimethoxyamphetamine HCl	1
34	JWH-018	0
35	JWH-018 N-(5-chloropentyl)	0
36	JWH-019	0
38	JWH-073	0
41	JWH-200	0
42	JWH-203	0
43	JWH-210	0
47	RCS-8	0
50	Buphedrone HCl	0
52	BZP diHCl	0
55	Cathinone HCl	0
56	CB-13	0

Modelul a fost rulat în mai multe iterații, subseturile compușilor fiind selectate în mod aleatoriu. S-a pornit cu valoarea $\alpha = 0$, iar modelul a fost dezvoltat pe baza regresiei Ridge și a unui model multinomial. Dupa antrenarea modelului și obținerea valorilor erorilor prin validare încrucișată, au fost deduse două valori pentru parametrul λ. Ținând cont că sunt cazuri în care valoarea minimă a parametrului λ poate duce la overfitting [114], pentru a obține cele mai bune performanțe în predicție folosind cross validarea incrucișată K fold, s-a ales valoarea lui λ ca fiind egală cu 1SE (standard error) [115].

Rata clasificărilor corecte a fost analizată pentru ambele seturi de date, valorile predicțiilor obținute fiind prezentate în Tabelul IV.15 și IV.16. În aceste tabele sunt prezentate

clasele compușilor cat și clasificarea facută de acest model pentru ambele seturi de date, în cazul penalizărilor Ridge, respectiv în cazul penalizărilor Lasso.

Se poate observa în Tabelul IV.15 ca toate amfetaminele (ce apartin clasei 1) au fost corect clasificate. În schimb doi compuși negativi (aparținând clasei 0) au fost clasificați în mod eronat ca facând parte din clasa 1. Se poate spune că regresia Ridge este foarte senzitivă, dar mai putin selectivă. Aceleași negative, compușii 35 și 36, au fost însă clasificați greșit în ambele seturi de date. În concluzie, în cazul clasificării efectuate cu penalizări Ridge, selecția datelor prin GA imbunatățește acuratețea detecției amfetaminelor.

Trebuie subliniat faptul că rata clasificărilor corecte obținută prin regresia Ridge este similară cu cea obținută folosind setul de date complet (spectre obținute cu un spectrometru CG-FTIR) procesat cu tehnici precum Analiza Componentelor Principale (Principal Component Analysis, PCA), SIMCA [82,83,109,110] sau cu descriptori molecular procesati cu PCA combinata cu Retele Neuronale Artificiale (Artificial Neural Networks, ANN) [83,111].

Tabelul IV.15 Valori ale predicțiilor în cazul penalizărilor Ridge. Valorile evidențiate în galben reprezintă cazurile in care modelul a atribuit compusului analizat o identitate de clasă greșită.

		Dataset inițial			Dataset GA		
Index	Clasa	Cl. Pred.	Coef. Cl. 0	Coef. Cl. 1	Cl. Pred.	Coef. Cl. 0	Coef. Cl. 1
1	1	1	0.277	0.722	1	0.355	0.644
2	1	1	0.301	0.698	1	0.318	0.681
5	1	1	0.261	0.738	1	0.289	0.71
12	1	1	0.264	0.735	1	0.209	0.79
14	1	1	0.404	0.595	1	0.401	0.598
16	1	1	0.325	0.674	1	0.373	0.626
24	1	1	0.28	0.719	1	0.293	0.706
27	1	1	0.311	0.688	1	0.339	0.66
34	0	0	0.879	0.12	0	0.795	0.204
35	0	1	0.43	0.569	1	0.447	0.552
36	0	1	0.33	0.669	1	0.401	0.598
38	0	0	0.767	0.232	0	0.682	0.317
41	0	0	0.873	0.126	0	0.836	0.163
42	0	0	0.64	0.359	0	0.66	0.339
43	0	0	0.818	0.181	0	0.779	0.22
47	0	0	0.766	0.233	0	0.689	0.31
50	0	0	0.762	0.237	0	0.648	0.351
52	0	0	0.78	0.219	0	0.806	0.193
55	0	0	0.707	0.292	0	0.599	0.4
56	0	0	0.833	0.166	0	0.788	0.211

Aplicând metoda Lasso aceluiași model, se aplică o constrângere asupra sumei valorilor absolute ale parametrilor modelului. Mai exact această suma trebuie sa aibă o valoare mai mică decat o limită superioară considerate. În acest scop, coeficienții variabilelor de regresie sunt penalizați prin alegerea și reducerea unora la valoarea zero. Procesul de selecție

a variabilelor ce au coeficienti mai mari ca zero, fiind luați în considerare de model după procesul de restrângere. Această procedură este aplicată pentru a reduce la minim eroarea de predicție. Elementele selectate a fi cele mai importante pentru model sunt prezentate în Figura IV.2.6.1.

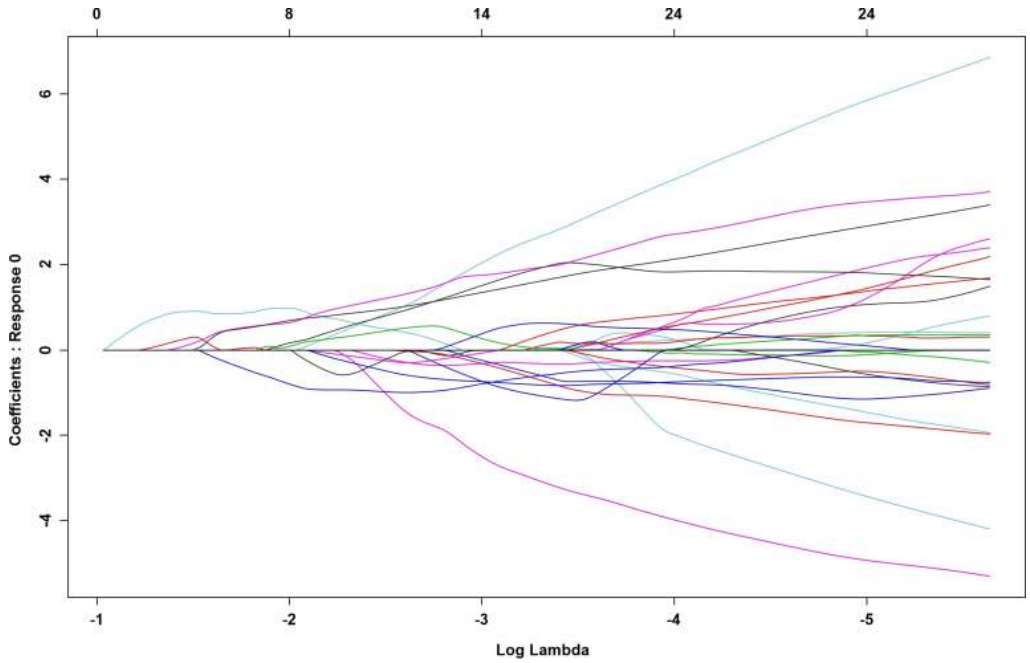


Figura IV.2.6.1 Selecția celor mai importante elemente la rularea modelului cu regresia Lasso pentru setul de date inițial.

Rezultatele obținute pentru α = 0 în cazul regularizării LASSO sunt prezentate în Tabelul IV.16 atât pentru baza de date inițială, cât și pentru baza de date selectată GA. Rezultatele arată că: a) nu toate amfetaminele au fost recunoscute ca atare; b) rata corectă de clasificare a pozitivelor este mai bună în cazul setului de date selectat GA; c) eșantioanele clasificate greșit nu sunt aceleași pentru ambele baze de date; d) același număr de negative au fost clasificate greșit ca pozitive pentru ambele seturi de date; e) negativele clasificate greșit nu sunt neapărat aceleași în cele două seturi de date; f) numărul negativelor clasificate greșit este (de două ori) mai mare decât în cazul metodei de regularizare a penalizărilor Ridge.

Tabel IV.16 Valori ale predicțiilor în cazul penalizărilor Lasso. Valorile evidențiate în galben reprezintă cazurile in care modelul a atribuit compusului analizat o identitate de clasă greșită.

		Dataset inițial			Dataset GA		
Index	Clasa	Cl. Pred.	Coef. Cl. 0	Coef. Cl. 1	Cl. Pred.	Coef. Cl. 0	Coef. Cl. 1
1	1	1	0.28	0.719	1	0.389	0.61
2	1	1	0.328	0.671	1	0.396	0.603
5	1	0	0.589	0.41	1	0.456	0.543
12	1	0	0.606	0.393	1	0.401	0.598
14	1	1	0.329	0.67	0	0.516	0.483
16	1	1	0.384	0.615	1	0.419	0.58
24	1	1	0.356	0.643	1	0.447	0.552
27	1	1	0.304	0.695	1	0.42	0.579
34	0	0	0.831	0.168	0	0.727	0.272
35	0	1	0.459	0.54	1	0.393	0.606
36	0	1	0.332	0.667	1	0.452	0.547
38	0	1	0.469	0.53	0	0.578	0.421
41	0	0	0.854	0.145	0	0.629	0.37
42	0	0	0.833	0.166	0	0.509	0.49
43	0	0	0.854	0.145	0	0.705	0.294
47	0	0	0.839	0.16	1	0.486	0.513
50	0	0	0.848	0.151	0	0.647	0.352
52	0	1	0.418	0.581	0	0.664	0.335
55	0	0	0.851	0.148	1	0.494	0.505
56	0	0	0.813	0.186	0	0.765	0.234

Modelele au fost construite folosind două metode de regularizare, adică regresia Ridge și LASSO. Rezultatele indică faptul că cele mai bune rezultate sunt obținute cu aceasta din urmă metodă. Aplicația construită pe bază regresiei de penalizare Ridge se caracterizează printr-o sensibilitate remarcabilă, deoarece toate amfetaminele 2C-x și DOx testate au fost recunoscute ca atare (rata pozitivelor adevărate = 100, rata negativelor false = 0%).

Din punct de vedere criminalistic, sensibilitatea instrumentului de screening este crucială, întrucât niciun eșantion pozitiv nu trebuie confundat cu un negativ. Aplicația de regresie Lasso este mai puțin selectivă (rata negativelor adevărate 83,33%, rata pozitivelor false 16,67%), deoarece câteva (2) negative au fost clasificate greșit ca pozitive. Cu toate acestea, acest lucru este acceptabil, întrucât cererea este menită să acționeze doar ca instrument de screening *în situ* (aeroport și controlul frontierei portuare etc.). Legea impune ca toate eșantioanele clasificate ca pozitive să fie supuse în continuare unor analize de laborator mai puternice. În cazul amfetaminelor, eșantionul trebuie identificat pozitiv prin spectroscopie GC-FTIR și GC-MS, analize realizate în laborator, luandu-se în consideratie spectrele complete. În timpul acestui proces, orice negativ clasificat greșit va fi identificat fara dubiu.

Preprocesarea spectrelor ATR-FTIR prin selectarea variabilelor GA îmbunătățește ratele de clasificare corecte. Cu toate acestea, această procedură de selecție variabilă reduce timpul necesar de calcul și capacitatea de memorie. În plus, îmbunătățește echilibrul dintre numărul de variabile și numărul de eșantioane incluse în setul de instruire. Deoarece numărul

acestor compuși nu poate fi crescut semnificativ datorită naturii lor, considerăm că selecția variabilă GA este utilă și trebuie păstrată ca parte a metodei de regresie cu penalizări Ridge.

IV.3 Concluzii

Studiul prezent demonstrează că GA-PLSR reprezintă o combinație valoroasă de metode de inteligență artificială, deoarece permite testarea eficientă a unei mari varietăți de medicamente de abuz. Selectarea GA a numărului de valori relevante determină o comprimare semnificativă a hiperspațiului și, prin urmare, îmbunătățește echilibrul dintre numărul de variabile de intrare și setul de antrenament PLSR. Acest aspect este foarte important, deoarece numărul de droguri de abuz disponibile pentru instruire și validarea modelului GA-PLSR este limitat datorită naturii lor (criminalistice).

În ciuda raportului semnal-zgomot relativ scăzut, modelul GA-PLSR identifică cu succes clasa de membru al amfetaminelor psihedelice, canabinoide și le distinge de o mare varietate de negative. Modelul GA-PLSR identifică în mod clar care sunt absorbțiile care caracterizează asemănarea structurii moleculare a compușilor aparținând fiecăreia dintre clasele de compuși modelate. Cum numerele de undă selectate acoperă practic întregul spectru, putem concluziona că nu există o regiune preferată a spectrului infraroșu.

Analizând rezultatele algoritmului de clasificare KNN ce a fost rulat folsind bază de date inițială cat și cea generată de algoritmul genetic putem remarca cu usurință valori mult mai bune în cazul bazei de date generate de GA.

Procesul de validare a indicat faptul că GA-PLSR este mai selectiv decât senzitiv. Capacitatea sa de a distinge între diversele clase de substanțe este mai mică decât cea obținută cu spectrele complete înregistrate cu un spectrometru GC-FTIR (în laborator), dar este comparabila cu senzitivitate obținută cu alte spectrometre portabile, de exemplu spectrometre laser (QCL) în infrarosu. Pe de altă parte, selectivitatea sa remarcabilă recomandă modelul GA-PLSR ca o metoda valoroasă de identificare a drogurilor pe bază spectrului lor ATR-FTIR.

În plus, metoda este foarte prietenoasa cu utilizatorii. Trebuie să ținem cont de faptul că acest sistem este conceput pentru a ajuta ofițerii de politie și vamesii sa aplice legea și să monitorizeze substanțele controlate. Probele confiscate sunt trimise la laboratoarele de medicină legală pentru o analiză ulterioară. Acolo, identitatea lor individuală este stabilită prin metode mai puternice. Cu alte cuvinte, deși se dorește o sensibilitate ridicată, cea mai importantă caracteristică a aplicației medico-legale este selectivitatea acesteia.

CAPITOLUL V. Concluzii generale, contribuții originale si directii viitoare de cercetare si dezvoltare

Concluzii generale

În cadrul acestei teze de doctorat, activitatea de cercetare a avut ca scop principal explorarea, validarea și compararea performantelor unor modele cât mai variate care au fost dezvoltate pentru recunoasterea identitatii de clasa a amfetaminelor 2C-x si DOx. Principala provocare a fost modelarea acestor doua clase de amphetamine halucinogene, si discriminarea lor de orice alte substante, pe baza spectrelor lor ATR-FTIR, care sunt foarte putin intense.

In acest scop, au fost folosite urmatoarele tehnici de inteligenta artificiala: regresia prin cele mai mici patrate parțiale (GA-PLS), algoritmul Random Forest, Regresia KNN, Clasificarea SVM, respectiv LRCM (Logistic Regression Classification Model) - metodele de regularizare LASSO și Ridge.

Fiecare model a fost realizat într-o perspectivă comparative, pentru două seturi de date input. Primul set cuprinde spectrele ATR-FTIR complete a principalilor compuși din cele doua clase de amphetamine ilicite studiate și a unei largi diversitati de negative (substante ce nu apartin celor doua clase positive, respectiv 2C-x si DOx). Al doilea set a fost generat aplicand un algoritm genetic (GA), care a permis selectarea unui număr restrans de numere de undă, unde au apar benzile cele mai reprezentative pentru clasificare. Aceasta selectie a variabilelor a permis eliminarea variabilelor de intrare redundante si deci construirea unor sisteme de clasificare care necesita capacitati si timp de calcul mai reduse, adecvate operarii unor spectrometre portabile precum cele ATR-FTIR.

Rezultatele au indicat ca cel mai performant sistem expert pentru recunoașterea identității de clasă a amfetaminelor halucinogene 2C-x si DOx este cel construit cu ajutorul algoritmului genetic și regresia prin cele mai mici pătrate parțiale (GA - PLSR) folosind baza de date spectrala selectata cu GA. Rezultate apropiate au fost obtinute pentru sistemul expert construit cu ajutorul modelului bazat pe Random Forest.

Contributii orginale

Prezenta lucrare este structurata in cinci capitole. În primele două capitole au fost descrise atat substanțele analizate, cât și metodele spectrale folosite pentru caracterizarea si identificarea amfetaminelor halucinogene de interes. În capitolul III sunt descrise metodele de inteligență artificiale folosite pentru constructia sistemelor expert (modelelor) descrise in capitolul IV.

Contributiile originale sunt prezentate eminamente in Capitolului IV, începând cu dezvoltarea unui model bazat pe algoritmii genetici. Acest algoritm a fost folosit pentru generarea unui set de date restrâns, ce a fost folosit pentru dezvoltarea sistemelor expert ce au fost construite prin cele cinci modele de clasificare diferite mentionate mai sus.

Astfel, in cadrul sectiunii IV.2.2 am analizat modelul GA-PLS, analizand efectul variației parametrilor acestui model. Analiza a condus la rezultate încurajatoare în ce privește folosirea

acestui model insoțit de setul restrâns GA. Analiza modelului bazat pe arborii decizionali este prezentata in cadrul capitolului IV.2.3. Aceasta nu a evidențiat diferențe notabile intre cele doua seturi de date folosite, dar este important de mentionat că rularea modelului s-a realizat in mai multe iteratii. Rezultatele au indicat faptul ca dacă în cazul setului de date inițial au fost cazul de clasificări incorecte, pentru setul de date GA, in toate rulările modelului, clasificările eșantionului de test au fost corecte. Analiza rezultatelor obtinute cu sistemul expert construit cu ajutorul algoritmului KNN, prezentata in sectiunea IV2.4, a indicat ca aceasta tehnica conduce la obținerea unor valori ale acurateții clasificarilor usor crescută in cazul folosirii setului de date GA fata de cazul setul initial. Rezultatele obtinute prin folosirea SVM si a regresiei logistice este prezentata in sectiunile IV2.5 si IV2.6. Asemenea modelelor studiate anterior, folosirea setului de date restrâns poate conduce la erori mai mici la clasificarea compușilor chimici ce fac parte din familia substanțelor interzise.

Directii viitoare de cercetare si dezvoltare

Sistemele expert de identificare automată a identitatii de clasa a amfetaminelor halucinogene 2C-x si DOx dezvoltate si prezentate în această teză evidențiază fezabilitatea automatizarii procesului de recunoaștere a clasei de substanțe căreia îi aparține un nou compus. Atribuirea identitatii de clasa are loc într-un timp foarte scurt, in conditii de senzitivitate și selectivitate foarte bune.

Avem în vedere următoarele direcții viitoare de cercetare și dezvoltare:

- Crearea și testarea altor sisteme expert construite pentru detecția halucinogenelor 2C-x si DOx folosind ca vectori de intrare alte tipuri de spectre;

- Dezvoltarea bazei de date curente prin adăugarea de noi halucinogene 2C-x si DOx, pe măsură ce noi substanțe din această familie de droguri, concepute în laboratoare clandestine, vor fi identificate; actualizarea continuă a bazei de date va crește continuu capacitatea de detecție a sistemelor expert descries; folosirea acestor sisteme va permite ca autorităților competente (politie, vama, etc) să fie mai eficiente în aplicarea legii in ceea ce priveste persoanele care produc, dețin, comercializează și/sau consumă ilegal substanțe ilicite;

- Dezvoltarea unor sisteme expert pentru detectia halucinogenelor 2C-x si DOx bazate pe alte tehnici de inteligență artificială, cum ar fi Analiza Componentelor Principale (Principal Component Analysis, PCA) sau Analiza Clusterelor (hierarchical cluster Analysis, HCA);

- Testarea eficienței unor combinații de astfel de tehnici, cum ar fi PCA și HCA, sau GA si HCA, având în vedere că GA si PCA sunt tehnici folosite mai ales pentru reducerea dimensionalității unei baze de date prin selectarea variabilelor de intrare relevante si eliminarea celor care contin informatie redundanta;

- Dezvoltarea de sisteme expert analoage celor descries in teza pentru alte familii de substanțe de abuz, naturale sau sintetice, cum ar fi canabinoizii.

Lista lucrărilor publicate și prezentate

Lucrări publicate ISI

- **indexate Thomson Reuters Web of Knowledge – Web of Science (WoS)**

1. **Negoita, C.**, Praisler M., Ion, A., Artificial Intelligence Application Designed to Screen for New Psychoactive Drugs Based on their ATR-FTIR spectra, *Proceedings of the 10th Jubilee International Conference of the Balkan Physical Union BPU10*, 26-30 August 2018, Sofia, Bulgaria. Mishonov, T. M., Varonov, A.M. (Eds), *AIP Conference Proceedings*, vol. 2075, issue 1, Article number 170026 (2019). DOI: 10.1063/1.5091391 **WOS:000472653800274**
ISSN: 0094-243X ISBN: 978-0-7354-1803-5
https://aip.scitation.org/doi/abs/10.1063/1.5091391
https://bpu10.balkanphysicalunion.com/

2. Praisler, M., Ciochina, S., Coman, M. M., **Negoita, C.**, Chemometric tool for doping control based on Hierarchical Cluster Analysis, in Miclea, L and Stoian, I. (Eds), *Proceedings of 2018 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR 2018)*, 24-26 May 2018, Cluj – Napoca, Romania, IEEE, p. 127. DOI: 10.1109/AQTR.2018.8402722 **WOS:000450065900025**
Electronic ISBN: 978-1-5386-2205-6 USB ISBN: 978-1-5386-2203-2 Print on Demand(PoD) ISBN: 978-1-5386-2206-3
https://ieeexplore.ieee.org/document/8402722/

3. Praisler, M. Ciochina, S., **Negoita, C.**, Improved Selectivity in Detecting Controlled Amphetamines and their Main Precursors based on Laser Infrared Spectra, *Proceedings of the 6th IEEE International Conference on E-Health and Bioengineering (EHB)*, 28 July 2017, Sinaia, Romania, pp. 233-236. Article number 7995404. DOI: 10.1109/EHB.2017.7995404 **WOS: 000445457500059**
Electronic ISBN: 978-1-5386-0358-1 Print on Demand(PoD) ISBN: 978-1-5386-1514-0 http://ieeexplore.ieee.org/abstract/document/7995404/

LUCRARE CITATA IN

1	Jones, N. S., Comparin, J. H., Interpol review of controlled substances review 2016–2019, Forensic Science International: Synergy, Available online 24 March 2020, DOI: 10.1016/j.fsisyn.2020.01.019 ISSN 2589-871X https://www.sciencedirect.com/science/article/pii/S2589871X2030019X?via%3Dihub

4. Ciochina, S., Praisler, M., **Negoita, C.**, Cluster Analysis Evaluating the Automated Detection of Drugs of Abuse with a New Hollow Fiber based Quantum Cascade Laser Infrared Spectrometer, *Proceedings of the 6th IEEE International Conference on E-Health and Bioengineering (EHB)*, 28 July 2017, Sinaia, Romania, pp. 237-240. Article number 7995405. DOI:10.1109/EHB.2017.7995405, WOS: 000445457500060 Electronic ISBN: 978-1-5386-0358-1 Print on Demand(PoD) ISBN: 978-1-5386-1514-0 http://ieeexplore.ieee.org/abstract/document/7995405/

LUCRARE CITATA IN

1	Jones, N. S., Comparin, J. H., Interpol review of controlled substances review 2016–2019, Forensic Science International: Synergy, Available online 24 March 2020, DOI: 10.1016/j.fsisyn.2020.01.019 ISSN 2589-871X

	https://www.sciencedirect.com/science/article/pii/S2589871X2030019X?via%3Dihub
2	Altulea, A. H., Jalab, H. A., Ibrahim, R. W., Fractional Hölder mean-based image segmentation for mouse behavior analysis in conditional place preference test, *Signal, Image and Video Processing* 14 (2020) 135-142. DOI: 10.1007/s11760-019-01533-1 Print ISSN 1863-1703 Online ISSN 1863-1711 **ISI Impact factor 1.894 (2018)** https://link.springer.com/article/10.1007/s11760-019-01533-1#citeas

- **în curs de indexare Thomson Reuters Web of Knowledge – Web of Science (WoS)**

1. **Negoita, C.**, Praisler, M., Logistic regression classification model identifying drugs of abuse based on their ATR- FTIR spectra, *6th International Symposium on Electrical and Electronics Engineering - ISEEE 2019*, 18 - 20 October 2019, Galaţi, Romania. DOI: 10.1109/ISEEE48094.2019.9136133 IEEE, Electronic ISBN: 978-1-7281-2906-8 USB ISBN: 978-1-7281-2905-1 Print on Demand(PoD) ISBN: 978-1-7281-2907-5 https://ieeexplore.ieee.org/abstract/document/9136133

2. Ion, A., Gosav, S., Praisler, M., **Negoita**, **C.**, Artificial neural network designed to identify NBOMe hallucinogens based on 3D-MoRSE descriptors and topological descriptors, in Precup R.-E. (Ed.), *2019 23rd International Conference on System Theory, Control and Computing, ICSTCC 2019 – Proceedings*, Article number 8885908, Pages 872-876. DOI: 10.1109/ICSTCC.2019.8885908 ISSN: 2372-1618 . ISBN: 978-172810699-1 https://ieeexplore.ieee.org/document/8885908

Lucrări publicate în reviste BDI

1. **Negoita, C.,** Praisler, M., Evolutionary algorithm applied for improving the accuracy of the automated detection of psychedelic amphetamines, *Annals of "Dunarea de Jos" University of Galati, Mathematics, Physics, Theoretical Mechanics, Fascicle II*, Year X (XXXXI) 2018, No. 1, p. 70-75. ISSN 2067-2071 http://www.phys.ugal.ro/Annals_Fascicle_2/

2. **Negoita, C.**, Praisler, M., Identification of Functional Groups in the ATR-FTIR Spectra of 2C-x and DOx Amphetamine Analogues, *Annals of "Dunarea de Jos" University of Galati, Mathematics, Physics, Theoretical Mechanics, Fascicle II*, Year IX (XXXX) 2017, No. 1, p. 25-30. ISSN 2067-2071, http://www.phys.ugal.ro/Annals_Fascicle_2/

Lucrări comunicate la conferinţe internaţionale

1. **Negoita, C.,** Praisler, M., Improved detection of 2C-x and Dox amphetamines – an analytical tool mitigating the environmental impact of their illicit manufacturing, consumption and disposal, *International Conference "Environmental Challenges in the Black Sea Basin: Impact on Human Health"*, Galati, Romania, September 23-26, 2020. https://blacksea-cbc.net/projects-newsevents/monitox-bsb27-international-conference-environmental-challegens-in-the-black-sea-basin-impact-on-hunam-health/

2. **Negoita, C.**, Praisler, M., Logistic regression classification model identifying drugs of abuse based on their ATR- FTIR spectra, *6th International Symposium on Electrical and Electronics Engineering - ISEEE 2019*, 18 - 20 October 2019, Galaţi, Romania. DOI: 10.1109/ISEEE48094.2019.9136133 IEEE, Electronic ISBN: 978-1-7281-2906-8 USB ISBN: 978-1-7281-2905-1 Print on Demand(PoD) ISBN: 978-1-7281-2907-5 http://www.iseee.ugal.ro/2019/

3. Ion, A., Gosav, S., Praisler, M., **Negoita**, **C.**, Artificial neural network designed to identify NBOMe hallucinogens based on 3D-MoRSE descriptors and topological descriptors, in Precup R.-E. (Ed.), *2019 23rd International Conference on System Theory, Control and Computing, ICSTCC 2019 – Proceedings*, Article number 8885908, Pages 872-876. DOI: 10.1109/ICSTCC.2019.8885908 ISSN: 2372-1618 . ISBN: 978-172810699-1 http://www.iseee.ugal.ro/2019/

4. Praisler, M., Ciochină, S., **Negoiţă, C**., Expert system operating a new portable infrared spectrometer designed to scan for synthetic drugs of abuse, *UGAL International Conference "Multidisciplinary HUB for the Higher Education Internationalization by Means of Innovative Interaction with the Labour Market and Society"*, 26 -27 October 2018, Galati, Romania.
http://fdi.ugal.ro/index.php/ro/conference-home

5. Ciochină, S., Praisler, M., **Negoiţă, C**., Automatic detection of illicit psychoactive drugs based on laser infrared absorption spectrometry (IRAS), *UGAL International Conference "Multidisciplinary HUB for the Higher Education Internationalization by Means of Innovative Interaction with the Labour Market and Society"*, 26 -27 October 2018, Galati, Romania.
http://fdi.ugal.ro/index.php/ro/conference-home

6. **Negoiţă, C.**, Praisler, M., SVM classification model based on a Coarse/Fine Grained Parallel Genetic Algorithm applied for the detection of hallucinogenic amphetamines based on their ATR-FTIR spectra, *UGAL International Conference "Multidisciplinary HUB for the Higher Education Internationalization by Means of Innovative Interaction with the Labour Market and Society"*, 26 -27 October 2018, Galati, Romania. http://fdi.ugal.ro/index.php/ro/conference-home

7. **Negoita, C.**, Praisler M., Ion, A., Artificial Intelligence Application Designed to Screen for New Psychoactive Drugs Based on their ATR-FTIR spectra, *Proceedings of the 10th Jubilee International Conference of the Balkan Physical Union BPU10*, 26-30 August 2018, Sofia, Bulgaria. Mishonov, T. M., Varonov, A.M. (Eds), *AIP Conference Proceedings*, vol. 2075, issue 1, Article number 170026 (2019). DOI: 10.1063/1.5091391 **WOS:000472653800274**
ISSN: 0094-243X ISBN: 978-0-7354-1803-5
https://bpu10.balkanphysicalunion.com/

8. Praisler, M., Ciochina, S., Coman, M. M., **Negoita, C.**, Chemometric tool for doping control based on Hierarchical Cluster Analysis, in Miclea, L and Stoian, I. (Eds), *Proceedings of 2018 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR 2018)*, 24-26 May 2018, Cluj – Napoca, Romania, IEEE, p. 127. DOI: 10.1109/AQTR.2018.8402722 **WOS:000450065900025**
Electronic ISBN: 978-1-5386-2205-6 USB ISBN: 978-1-5386-2203-2 Print on Demand(PoD) ISBN: 978-1-5386-2206-3
http://www.aqtr.ro/

9. Praisler, M. Ciochina, S., **Negoita, C.**, Improved Selectivity in Detecting Controlled Amphetamines and their Main Precursors based on Laser Infrared Spectra,

Proceedings of the 6th IEEE International Conference on E-Health and Bioengineering (EHB), 28 July 2017, Sinaia, Romania, pp. 233-236. Article number 7995404. DOI: 10.1109/EHB.2017.7995404 **WOS: 000445457500059** Electronic ISBN: 978-1-5386-0358-1 Print on Demand(PoD) ISBN: 978-1-5386-1514-0 http://www.ehbconference.ro/2017/Home.aspx

10. Ciochina, S., Praisler, M., **Negoita, C.**, Cluster Analysis Evaluating the Automated Detection of Drugs of Abuse with a New Hollow Fiber based Quantum Cascade Laser Infrared Spectrometer, *Proceedings of the 6th IEEE International Conference on E-Health and Bioengineering (EHB)*, 28 July 2017, Sinaia, Romania, pp. 237-240. Article number 7995405. DOI:10.1109/EHB.2017.7995405, WOS: 000445457500060 Electronic ISBN: 978-1-5386-0358-1 Print on Demand(PoD) ISBN: 978-1-5386-1514-0 http://www.ehbconference.ro/2017/Home.aspx

Lucrări comunicate la conferinţe naţionale

1. **Negoita, C**., Praisler, M., Evaluation of the accuracy in classifying synthetic hallucinogens based on the Random Forest Model, *Book of abstracts SCDS-UDJG 2020*, 8th Edition, 18 - 19 June 2020, Galati, Romania. http://www.cssd-udjg.ugal.ro/

2. **Negoita, C**., Praisler, M., Improving the classification accuracy of psychedelic amphetamines based on the use of Stochastic Descendant Gradient, *Book of abstracts SCDS-UDJG 2019*, 7th Edition, 13 - 14 June 2019, Galati, Romania. http://www.cssd-udjg.ugal.ro/

3. **Negoita, C**., Praisler, M., Combining Bootstrap Aggregation and Random Forest Algorithms for refining the automated detection of psychedelic amphetamines, *Book of abstracts SCDS-UDJG 2019*, 7th Edition, 13 - 14 June 2019, Galati, Romania. http://www.cssd-udjg.ugal.ro/

4. Coman, M., **Negoita, C**., Praisler, M., Improving the classification accuracy of synthetic cannabinoids by combining evolutionary and classification algorithms, *Book of abstracts SCDS-UDJG 2019*, 7th Edition, 13 - 14 June 2019, Galati, Romania. http://www.cssd-udjg.ugal.ro/

5. Coman, M., **Negoita, C**., Praisler, M., Automated detection of synthetic cannabinoids based on Genetic Algorithms and Partial Least Square Regression, *Book of abstracts SCDS-UDJG 2019*, 7th Edition, 13 - 14 June 2019, Galati, Romania. http://www.cssd-udjg.ugal.ro/

6. **Negoita, C**., Praisler, M., Evolutionary algorithm applied for improving the accuracy of the automated detection of psychedelic amphetamines, *Book of abstracts SCDS-UDJG 2018*, 6th Edition, 7 - 8 June 2018, Galati, Romania. http://www.cssd-udjg.ugal.ro/

7. **Negoita, C**., Praisler, M., Parameter fine-tuning for improving Partial Least Squares Regression models built for the automated recognition of psychedelic amphetamines, *Book of abstracts SCDS-UDJG 2018*, 6th Edition, 7 - 8 June 2018, Galati, Romania. http://www.cssd-udjg.ugal.ro/

8. **Negoita, C**., Praisler, M., Automatic detection of 2C-x and DOx hallucinogenic amphetamines based on genetic algorithms and ATR-FTIR spectroscopy,

Workshop National Workshop National Creşterea Relevanţei Învăţământului Universitar Tehnic in Relaţie cu Dezvoltarea Industrialã Regionalã, 24 Noiembrie 2017, Galati, Romania. ISBN 978-606-696-098-4
http://www.ing.ugal.ro/Resurse/PagInit/IM_Program_workshop.pdf

9. **Negoita, C**., Praisler, M., Identification of Functional Groups in the ATR-FTIR Spectra of 2C-x and DOx Amphetamine Analogues, *Book of abstracts CSSD-UDJG 2017*, 5th Edition, 8 - 9 June 2017, Galati, Romania, p. 59.
http://www.cssd-udjg.ugal.ro/index.php/abstracts-2017

10. **Negoita, C**., Praisler, M., Detection of Illicit Amphetamines Performed with Artificial Neural Networks (ANN) Running on Small Scale Portable Hardware with Custom Embedded Operating System, *Book of abstracts CSSD-UDJG 2017*, 5th Edition, 8 - 9 June 2017, Galati, Romania, p. 60.
http://www.cssd-udjg.ugal.ro/index.php/abstracts-2017

Premii:

Premiul I, sectiunea 4: Negoita, C., Praisler, M., Identification of Functional Groups in the ATR-FTIR Spectra of 2C-x and DOx Amphetamine Analogues, Book of abstracts CSSD-UDJG 2017, 5th Edition, 8 - 9 June 2017, Galati, Romania, p. 59.
http://www.cssd-udjg.ugal.ro/index.php/abstracts-2017

Premiul III, sectiunea 2: Negoiţă, C., Praisler, M, Evolutionary Algorithm Applied for Improving the Accuracy of the Automated Detection of Psychedelic Amphetamines Conferinta Scolii Doctorale Universitatea “Dunarea de Jos“ Galati 2017 Book of abstracts CSSD-UDJG 2018, 6th Edition, 7 - 8 June 2018, Galati, Romania, p.56 .
http://www.cssd-udjg.ugal.ro/files/2018/05_Program_detaliat_al_conferintei_2018.pdf
http://www.cssd-udjg.ugal.ro/index.php/abstracts-2018

Bibliografie

[1] European Monitoring Centre for Drugs and Drug Addiction, www.emcdda.europa.eu

[2] ***https://en.wikipedia.org/wiki/Amphetamine

[3] Shulgin A., Manning T., Daley P.F. (2011), *The Shulgin Index. Volume 1. Psychedelic Phenethylamines and Related Compounds*. Transform Press.

[4] SWGDRUG Infrared Library http://www.swgdrug.org/ir.htm

[5] Stuart B. (2004), *Infrared Spectroscopy: Fundamentals and Applications*. Analytical Techniques în the Sciences (AnTs), Wiley.

[6] Fahrenfort J. (1961), *Attenuated total reflection. A new principle for the production of useful infrared reflection spectra of organic compounds*, Spectrochim. Acta **17**: 698-709.

[7] Harrick N.J. (1967), *Internal reflection spectroscopy*. Interscience Publishers, NewYork.

[8] *** http://lab-training.com/landing/free-hplc-training-programme-5/

[9] Hamdam R., Hassan N. (2015), *Characterisation of Seized Clandestine Methamphetamine în MalaysiaMalaysian*, J. Forensic Sci. **6(1)**: 20-29

[10] Wold H. (1982), Soft *modelling, The basic design and some extensions, în: K.-G. Joreskog*, H. Wold Eds. , Systems Under direct Observation, vol. I and II, North-Holland, Amsterdam.

[11] Wold S., Ruhe A., Wold H., Dunn W.J. (1984), *The collinearity problem în linear regression, The partial least squares approach to generalized inverses*, SIAM J. Sci. Stat. Comput. **5**: 735–743.

[12] Hoskuldsson A. (1988), *PLS regression methods*, J. Chemom. **2**: 211–228.

[13] Hoskuldsson A. (1996), *Prediction Methods în Science and Technology*, vol. 1, Thor Publishing, Copenhagen.

[14] Wold S., Johansson E., Cocchi M. (1993), *PLS partial least squaresprojections to latent structures, în: H. Kubinyi,* Ed. 3D QSAR în Drug Design, Theory, Methods, and Applications, ESCOM Science Publishers, Leiden, pp. 523–550.

[15] Tenenhaus M. (1998), *La Regression PLS: Theorie et Pratique*, Technip, Paris.

[16] Gerlach R.W., Kowalski B.R., Wold H. (1979), *Partial least squaresmodelling with latent variables*, Anal. Chim. Acta **112**: 417–421.

[17] Jackson J.E. (1991), *A User's Guide to Principal Components*, Wiley, New York.

[18] Burnham A.J., MacGregor J.F., Viveris R. (1999), *Latent variableregression tools*, Chemom. Intell. Lab. Syst. **48**: 167–180.

[19] Burnham A., Viveros R., MacGregor J. (1996), *Frameworks for latent variable multivariate regression*, J. Chemom. **10**: 31–45.

[20] Manne R. (1987), *Analysis of two partial least squares algorithms formultivariate calibration*, Chemom. Intell. Lab. Syst. **1**: 187–197.

[21] Lindgren F., Geladi P., Wold S. (1993), *The kernel algorithm for PLSI, Many observations and few variables*, J. Chemom. **7**: 45–59.

[22] Rannar S., Geladi P., Lindgren F., Wold S. (1994), *The kernel algorithm for PLS II, Few observations and many variables*, J.Chemom. **8**: 111–125.

[23] Denham M.C. (1997), *Prediction intervals în partial least squares*, J.Chemom. **11**: 39–52.

[24] Efron B., Gong G. (1983), *A leisurely look at the bootstrap, thejackknife, and cross-validation*, Am. Stat. **37**: 36–48.

[25] Wold S., Sjöström M., Eriksson L. (2001), *PLS-regression: a basic tool of chemometrics*. Chemometr. Intel. Lab. **58(2)**: 109-130.

[26] Fraser A.S. (1957), *Simulation of genetic systems by automatic digital computers. II: Effects of linkage on rates under selection*, Austral. J. Biol. Sci. **10(4)**: 492 - 500

[27] Bremermann H.J. (1958), *The evolution of intelligence. The nervous system as a model of its environment*, Technical Report No. 1, Department of Mathematics, University of Washington, Seattle, WA.

[28] Holland J.H. (1975), *Adaptation în Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI.

[29] Baker J.E. (1985), *Adaptive selection methods for genetic algorithms*, Proc. Int. Conf. on Genetic Algorithms and Their Applications, pp. 101–111.

[30] Grefenstette J.J., Baker J.E. (1989), *How genetic algorithms work: A critical look at implicit parallelism*, în: Proc. 3rd Int. Conf. on Genetic Algorithms, pp. 20–27.GENETIC ALGORITHMS 121.

[31] Goldberg D.E. (1989), *Genetic Algorithms în Search Optimization and Machine Learning*, Addison-Wesley, Reading, MA.

[32] Booker L.B., Fogel D.B., Whitley D., Angeline P.J. (1997)*, Recombination, în: The Handbook of Evolutionary Computation*, T. B¨ ack, Fogel D.B. and Michalewicz Z., eds, chapter E3.3, pp. C3.3:1–C3.3:27, IOP Publishing and Oxford University Press, Philadelphia, PA.

[33] Spears W. (1997), *Recombination parameters, în: The Handbook of Evolutionary Computation*, T. B¨ ack, Fogel D. B. and Michalewicz Z., eds, Chapter E1.3, IOP Publishing and Oxford University Press, Philadelphia, PA, pp. E1.3:1–E1.3:13.

[34] Goldberg D.E., Sastry K. (2001), *A practical schema theorem for genetic algorithm design and tuning*, în: Proc. of the Genetic and Evolutionary Computation Conf., pp. 328–335.

[35] Syswerda G. (1989), *Uniform crossover în genetic algorithms*, în: Proc. 3rd Int. Conf. on Genetic Algorithms, pp. 2–9.

[36] Spears W.M., De Jong K.A. (1994), *On the virtues of parameterized uniform crossover*, în: Proc. 4th Int. Conf. on Genetic Algorithms.

[37] Devroye L., Wagner T.J. (1982) *Nearest neighbor methods în discrimination, În Classification, Pattern Recognition and Reduction of Dimensionality*, Handbook of Statistics, North-Holland, Amsterdam, **2**: 193–197.

[38] Bailey T., Jain A. (1978) *A note on distance-weighted k-Nearest Neighbor rules*, IEEE Trans. Systems, Man, Cybernetics **8**: 311-313.

[39] Baoli L., Shiwen Y., Qin L. (2003) *An Improved k-Nearest Neighbor Algorithm for Text Categorization*, ArXiv Computer Science e-prints.

[40] Bauer M.E., Burk T.E., Ek A.R., Coppin P.R., Lime S.D., Walsh T.A., Walters D.K., Befort W., Heinzen D.F. (1994), *Satellite Inventory of Minnesota's Forest Resources*, Photogramm. Eng. Rem. S. **60(3)**: 287–298.

[41] Bax E. (2000), *Validation of nearest neighbor classifiers*, IEEE Trans. Inform. Theory **46**: 2746–2752.

[42] Benetis R., Jensen C., Karciauskas G., Saltenis S. (2006), *Nearest and Reverse Nearest Neighbor Queries for Moving Objects*, VLDB J. **15(3)**: 229–250.

[43] Bermejo T., Cabestany J. (2000), Adaptive *soft k-Nearest Neighbor classifiers*, Pattern Recogn. **33**: 1999-2005.

[44] Chitra A., Uma S. (2010), *An Ensemble Model of Multiple Classifiers for Time Series Prediction,* Int. J. Comput. Theory Eng. **2(3)**: 1793-8201.

[45] Cover T.M. (1968), *Rates of convergence for nearest neighbor procedures*, În Proceedings of the Hawaii International Conference on System Sciences, Univ. Hawaii Press, Honolulu, 413–415.

[46] Cover T.M., Hart P.E. (1967), *Nearest neighbor pattern classification*, IEEE Trans. Inf. Theory **13**: 21–27.

[47] Devroye L. (1981), *On the asymptotic probability of error în nonparametric discrimination,* Ann. Statist. **9**: 1320–1327.

[48] Devroye L. (1981), *On the equality of Cover and Hart în nearest neighbor discrimination*, IEEE Trans. Pattern Anal. Mach. Intell. **3**: 75-78.

[49] Devroye L., Gyorfi L., Krzyzak A., Lugosi G. (1994), *On the strong universal consistency of nearest neighbor regression function estimates,* Ann. Statist. **22**: 1371–1385.

[50] Devroye L., Wagner T.J. (1977), *The strong uniform consistency of nearest neighbor density estimates*, Ann. Statist. **5**: 536–540.

[51] Audibert J.Y., Tsybakov A.B. (2007), *Fast learning rates for plug-în classifiers under the margin condition*, Ann. Statist. **35**: 608–633.

[52] Domeniconi C., Peng J., Gunopulos D. (2002), *Locally adaptive metric nearestneighbor classification*, IEEE Transactions on Pattern Analysis and Machine Intelligence **24(9)**: 1281–1285.

[53] Dudani S.A. (1976), *The distance-weighted k-nearest neighbor rule*, IEEE Transactions on System, Man, and Cybernetics **6**: 325-327.

[54] Eldestein H.A. (1999), *Introduction to Data Mining and Knowledge Discovery*, Two Crows Corporation, USA.

[55] Enas G.G., Choi S.C. (1986), *Choice the smoothing parameter and efficiency of KNearest Neighbor classification*, Comp & Maths with Apps **12(2)**: 235-244.

[56] Fix E., Hodges J.L. (1951), *Nonparametric Discrimination: Consistency Properties*, Randolph Field, Texas, Project 21-49-004, Report No. 4.

[57] Fritz J. (1975), *Distribution-free exponential error bound for nearest neighbor pattern classification*, IEEE Trans. Inform. Theory **21**: 552–557.

[58] Fukunaga K., Hostetler L. (1975), *K nearest-neighbor Bayes risk estimation*, IEEE Trans. Information Theory **21(3)**: 285-293.

[59] Gil-Garcia R., Pons-Porrata A. (2006), *A New Nearest Neighbor Rule for Text Categorization*, Lecture Notes în Computer Science 4225, Springer, New York, 814–823.

[60] Gou J., Du L., Zhang Y., Xiong T. (2012), *A New Distance-weighted k-nearest Neighbor Classifier*, J. Inf. Comput. Sci. **9(6)**: 1429-1436.

[61] Widmaier F. (2015), *Robot Arm Tracking with Random Decision Forests*. Eberhard-Karls-Universität Tübingen.

[62] Antonio C., Jamie S., Konukoglu E. (2012), *Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning*, Found. Trends Comput. Graph. Vis. **7(2-3)**: 81–227.

[63] Criminisi A., Shotton J. (2013), *Decision Forests for Computer Vision and Medical Image Analysis*. Springer Science & Business Media.

[64] T. K. Ho (1995), *Random decision forests.* Proceedings of the Third International Conference on Document Analysis and Recognition **1**: 278–282.

[65] Vapnik V. (1995), *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

[66] Minoux M (1986), *Mathematical Programming: Theory and Algorithms*. John Wiley and Sons.

[67] Gunn S. (1998), *R. Support vector machines for classification and regression*. ISIS technical report **14.1**: 5-16.

[68] Cortes C., Vapnik V. (1995), *Support Vector Networks,* Mach. Learn. **20**: 273-297.

[69] Mukherjee S., Osuna E., Girosi F. (1997), *Nonlinear Prediction of Chaotic Time Series using Support Vector Machines*. To appear în Proc. of IEEE NNSP'97,Amelia Island, FL, Image Speech and Intelligent Systems Group, 24-26.

[70] Stitson M.O., Weston J.A.E. (1996), *Implementational Issues of Support Vector Machines*. Technical Report CSD-TR-96-18, Computational Intelligence Group, Royal Holloway, University of London.

[71] Stitson M.O., Weston J.A.E., Gammerman A., Vovk V., Vapnik V. (1996) *Theory of Support Vector Machines*. Technical Report CSD-TR-96-17, Computational Intelligence Group, Royal Holloway, University of London.

[72] Vapnik V., Golowich S., Smola A. (1997), *Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing*. În: M. Mozer, M. Jordan, and T. Petsche (eds.): Neural Information Processing Systems, Vol. 9. MIT Press, Cambridge, MA.

[73] Hosmer Jr., David W., Stanley L., Rodney X.S. (2013), *Applied logistic regression*. Vol. 398. John Wiley & Sons.

[74] Agresti A. (2010), *Analysis of ordinal categorical data*. Vol. 656. John Wiley & Sons.

[75] Feinberg S.E. (1980), *The Analysis of Cross-Classified Categorical Data*. 2nd Edit.

[76] Karch S. (2007), *Drug of Abuse Handbook*, 2nd ed. CRC Press, Boca Raton.

[77] United Nations Office on Drugs and Crime, 2014 *Global Synthetic Drugs Assessment: Amphetamine-type stimulants and new psychoactive substances* (UNODC, New York, 2014).

[78] Gosav S., Dinica R., Praisler M. (2008), J. Mol. Struct. **887:** 269-278.

[79] Praisler M., Ciochina S. (2014), *Global clustering quality coefficient assessing the efficiency of PCA class identity assignment*, J. Anal. Methods Chem., Article ID 342497.

[80] Ciochina S., Praisler M. (2013), *Optimization of amphetamines multivariate detection by GC-FTIR spectra preprocessing,* 2013 17th International Conference on System Theory, Control and Computing ICSTCC, IEEE Conference Proceedings, (IEEE, Sinaia, Romania, 2013), 125-130.

[81] Praisler M., Ciochina S. (2013), *Intelligent screening for designer drugs: a signal analysis*, 2013 17th International Conference on System Theory, Control and Computing ICSTCC, IEEE Conference Proceedings, (IEEE, Sinaia, Romania, 2013), 428-433.

[82] Praisler M., Van Bocxlaer J., De Leenheer A., Massart D.L. (2002), *Automated recognition of ergogenic aids using Soft Independent Modeling of Class Analogy (SIMCA),* Turk. J. Chem. **26**: 45-58.

[83] Gosav S., Praisler M., Dorohoi D.O., Popa G. (2006), *Structure – Activity Correlations for Illicit Amphetamines Using ANN and Constitutional Descriptors,* Talanta **70**: 922-928.

[84] Gosav S., Praisler M. (2009), *Artificial Neural Networks Built for the Recognition of Illicit Amphetamines Using a Concatenated Database*, Rom. Rep. Phys. **54**: 929–935.

[85] Ciochina S., Praisler M. (2013), *Pattern Recognition Techniques Applied for the Detection of Amphetamines Based on Infrared Laser Spectroscopy,* 2013 E-Health and Bioengineering Conference EHB, IEEE Conference Proceedings, (IEEE, Iasi, Romania, 2013), 1-4.

[86] Praisler M., Ciochina S., Negoita C. (2017), *Improved Selectivity în Detecting Controlled Amphetamines and their Main Precursors based on Laser Infrared Spectra*, 2017 E-Health and Bioengineering Conference EHB, IEEE Conference Proceedings, (IEEE, Sinaia, Romania, 2017), 233-236.

[87] Praisler M., Ciochina S. (2013), *PCA Evaluation of Quantum Cascade Lasers as Radiation Sources for Portable IRAS Systems Detecting Amphetamines*, 2013 E-Health and Bioengineering Conference EHB, IEEE Conference Proceedings, (IEEE, Iasi, Romania, 2013), 1-4.

[88] Ciochina S., Praisler M., Negoita C. (2017), *Cluster Analysis Evaluating the Automated Detection of Drugs of Abuse with a New Hollow Fiber based Quantum Cascade Laser Infrared Spectrometer,* 2017 E-Health and Bioengineering Conference EHB, IEEE Conference Proceedings, (IEEE, Sinaia, Romania, 2017), 237-240.

[89] ***http://www.swgdrug.org Scientific Working Group for the Analysis of Seized Drugs - Drug Enforcement Administration (DEA)

[90] ***http://lab-training.com/landing/free-hplc-training-programme-5/

[91] Gosav S., Dinica R., Praisler M. (2008), *Choosing between GC-FTIR and GC-MS spectra for an efficient intelligent identification of illicit amphetamines*, J. Mol. Struct. **887**: 269-278.

[92] Lucasius C.B., Kateman G. (1993), *Understanding and using genetic algorithms - Part 1. Concepts, properties and context*, Chemometr. Intell. Lab. **19**: 1-33.

[93] Lucasius C.B., Kateman G. (1994), *Understanding and using genetic algorithms - Part 2. Rep-resentation, configuration and hybridization*, Chemometr. Intell. Lab. **25**: 99-145.

[94] Hasegawa K., Miyashita Y., Funatsu K. (1997), *GA strategy for variable selection în QSAR studies: GA-based PLS analysis of calcium channel antagonists*, J. Chem. Inf. Comput. Sci. **37**: 306-310.

[95] Negoita C., Praisler M., Ion A. (2019), *Artificial Intelligence Application Designed to Screen for New Psychoactive Drugs Based on their ATR-FTIR spectra*, Proceedings of the 10th Jubilee International Conference of the Balkan Physical Union BPU10, 26-30 August 2018, Sofia, Bulgaria. Mishonov, T. M., Varonov, A.M. (Eds), AIP Conference Proceedings, vol. 2075, issue 1, Article number 170026.

[96] Breiman L. (2001), *Random forests*. Mach. Learn. **45**: 5–32.

[97] Breiman L. (2017), *Classification and Regression Trees*, Routledge: New York City, USA.

[98] Breiman L. (2002), *Manual On Setting Up, Using, And Understanding Random Forests V3.1*

[99] Ripley B.D. (1996), *Pattern Recognition and Neural Networks*. Cambridge.

[100] Venables W.N., Ripley B.D. (2002), *Modern Applied Statistics with S*. Fourth edition. Springer.

[101] Devos O., Downey G., Duponchel D. (2014), Simultaneous *data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils*. Food Chem. **148**: 124–130.

[102] Koza J.R., Andre D. (1995), *Parallel genetic programming on a network of transputers. În Proceedings of the Workshop on Genetic Programming: From Theory to Real-World Applications*, Rochester, UK, Rosca J. (Ed.), 111 - 120.

[103] Whitley D., Rana S., Heckendorn R.B. (1997), *Island Model Genetic Algorithms and Linearly Separable Problems*. În Proceedings of the AISB Workshop on Evolutionary Computing, Springer, Berlin, Corne D., Shapiro J.L (Eds.), 109-125.

[104] Lin S.-C., Punch W.F., Goodman E.D. (1994), *Coarse-grain Genetic Algorithms, Categorization and New Approaches.* În Proceedings of 1994 6th IEEE Symposium on Parallel and Distributed Processing, 26-29 October 1994, 28-37.

[105] European Monitoring Centre for Drugs and Drug Addiction (2018), European *Drug Report 2018 - Trends and Developments*, Luxembourg: Publications Office of the European Union.

[106] Esposito V., Chin W.W., Henseler J., Wang H. (2010), *Handbook of Partial Least Squares - Concepts, Methods and Applications*. Berlin: Springer Handbooks of Computational Statistics.

[107] Lee L.C., Liong C.Y., Jemain A.A. (2018), *Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps*, Analyst **143**: 3526-3539.

[108] Lee Y., Han S.-H., Nam S.-H. (2017), *Soft Independent Modeling of Class Analogy (SIMCA) Modeling of Laser-Induced Plasma Emission Spectra of Edible Salts for Accurate Classification*, Appl. Spectrosc. **71(9)**: 2199-2210.

[109] Praisler M., Dirinck I., Van Bocxlaer J., A. De Leenheer, and D.L. Massart (2000), *Identification of Novel Illicit Amphetamines from Vapor-Phase FTIR spectra - a Chemometrical Solution*, Talanta **53**: 155-170.

[110] Hassoun M. (2003), *Fundamentals of Artificial Neural Networks*. US: Bradford.

[111] Gosav S., Praisler M., Dorohoi D.O. (2007), *ANN Expert System Screening for Illicit Amphetamines using Molecular Descriptors*, J. Mol. Struct. **834-836**: 188-194.

[112] Tibshirani R. (1996), *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. B **58**: 267-288.

[113] Hastie T., Tibshirani R., Friedman J. (2009), *The elements of statistical learning: prediction, inference and data mining*. New York: Springer-Verlag.

[114] Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984), *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks.

[115] Friedman J., Hastie T., Tibshirani R. (2010), *Regularization paths for generalized linear models via coordinate descent*, J. Stat. Softw. **33**: 1-22.

[116] Fonti V., Belitser E. (2017), *Feature selection using LASSO*, Vrije Universiteit Amsterdam.

[117] van Wieringen W.N., Wessel N. (2015), *Lecture notes on Ridge regression*, arXiv:1509.09169.