

UNIVERSITATEA BABEȘ-BOLYAI, CLUJ-NAPOCA
FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ
SPECIALIZAREA INFORMATICĂ

Rezumat teză de doctorat

**Sisteme de Diagnoză și Detecție Interpretabile
Bazate pe Învățare Automată
pentru Analiza Datelor Medicale**



CONDUCĂTOR ȘTIINȚIFIC:

PROF. DR. CAMELIA CHIRA

AUTOR:

ADÉL BAJCSI

CLUJ-NAPOCA
2025

Abstract

Progresele în domeniul învățării automate și accesibilitatea tot mai mare a datelor medicale au permis dezvoltarea sistemelor de detecție asistată de calculator (CAD) pentru analiză și clasificare. Această teză propune modele de învățare automată interpretabile și modele de învățare profundă auto-explicative pentru analiza datelor medicale, cu un accent deosebit pe detectarea cancerului de sân utilizând mamografii digitale. Obiectivul principal este îmbunătățirea diagnosticării medicale prin metode precise și interpretabile, asigurând atât încrederea, cât și utilizabilitatea acestora în mediile clinice.

În această lucrare introducem noi sisteme CAD pentru aplicații medicale critice, inclusiv clasificarea cancerului de sân, predicția recurenței polipilor nazali și detectarea bolii Alzheimer prin analiza imaginilor de rezonanță magnetică funcțională (fMRI). Sunt explorate două abordări principale: învățarea automată tradițională și învățarea profundă. Prima abordare integrează preprocesarea, extragerea și selecția caracteristicilor, precum și clasificatori bazați pe arbori de decizie, pentru a gestiona provocările specifice seturilor de date medicale de dimensiuni reduse. Cea de-a doua abordare utilizează modele de învățare profundă interpretabile, cum ar fi rețelele neuronale convoluționale, asigurând transparența proceselor decizionale.

Sunt realizate experimente cuprinzătoare utilizând atât seturi de date publice, cât și date reale. În primul rând, este propus un sistem CAD pentru clasificarea mamografiilor digitale, care combină caracteristici bazate pe textură și formă extrase din mamografii, utilizând metode tradiționale de învățare automată. Înțelegerea procesului decizional al sistemelor utilizate în domeniul asistenței medicale este esențială. Prin urmare, modelele de învățare profundă interpretabile sunt evaluate pentru detectarea cancerului de sân, cu scopul de a compara eficacitatea acestora cu metodele tradiționale de învățare automată. Rezultatele arată că metodele tradiționale de învățare automată oferă o acuratețe competitivă, menținând în același timp o interpretabilitate superioară față de modelele de învățare profundă. Constatările evidențiază eficiența și interpretabilitatea modelelor tradiționale, în special pentru seturile de date de dimensiuni reduse. Metodele propuse sunt, de asemenea, testate pe date medicale numerice reale și pe alte imagini medicale relevante pentru psihologie. Această teză contribuie la domeniul diagnosticării medicale prin furnizarea de metodologii robuste și interpretabile pentru analiza și clasificarea datelor medicale.

Cuprins

Abstract	i
Cuprinsul Tezei de Doctorat	1
1 Introducere	4
1.1 Motivația	4
1.2 Obiective	5
1.3 Contribuții	5
1.4 Lista publicațiilor	8
2 Sistemul Propus Pentru Clasificarea Mamografiilor Digitale Bazat pe Învățare Automată Tradițională	10
2.1 Clasificare în Trei Clase: normal, benign și malign	10
2.2 Clasificare în Două Clase: benign și malign	11
2.3 Experimente	12
2.4 Concluzii	12
3 Sistemul Propus Pentru Clasificarea Mamografiilor Digitale Bazat pe Învățare Profundă	14
3.1 Metoda Propusă	14
3.2 Experimente	15
3.3 Concluzii	15
4 Proposed Systems for Cancer Detection from Numerical Medical Data Analysis	17
4.1 Predicția Cancerului de Sân pe Baza Analizei Datelor Numerice . .	17
4.1.1 Metoda Propusă	17
4.1.2 Experimente	18
4.1.3 Concluzii	18
4.2 Predicția Recidivei Polipilor Nazali După Operația Endoscopică a Sinusurilor	19
4.2.1 Metoda Propusă	19

4.2.2	Experimente	20
4.2.3	Concluzii	20
5	Abordări de Învățare Automată a Analizei fMRI Pentru Detectarea Bolii Alzheimer	21
5.1	Metoda Propusă	21
5.2	Experimente	22
5.3	Concluzii	23
6	Concluzii și Direcții de Cercetare Viitoare	24
6.1	Rezultatele și Contribuțiile Principale	24
6.2	Direcții de Cercetare Viitoare	25
	Bibliografie	27

Cuprinsul Tezei de Doctorat

Abstract	i
List of Tables	v
List of Figures	vii
List of Equations	ix
List of Algorithms	x
List of Symbols	xi
List of Abbreviations	xii
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	2
1.3 Contributions	3
1.4 List of Publications	6
1.5 Outline of the Thesis	8
2 Digital Mammogram Analysis for Breast Cancer Classification	9
2.1 Scientific Problem Definition	9
2.1.1 Mammogram Analysis	10
2.1.2 Proposed General Approach to Mammogram Analysis	11
2.2 Related Work	13
2.2.1 Related Work Using Traditional Machine Learning	14
2.2.2 Related Work Using Deep Learning	19
2.3 Data	21
2.3.1 MIAS	21
2.3.2 DDSM	23
2.4 Evaluation Metrics	23
2.4.1 Segmentation	24
2.4.2 Classification	24
3 Proposed Traditional Machine Learning System for Digital Mammogram Classification	26
3.1 Preprocessing	27

3.1.1	Artifact Removal	28
3.1.2	Pectoral Muscle Removal	31
3.1.3	Enhancement	35
3.2	Feature Extraction	35
3.2.1	Shape Features	36
3.2.2	Texture Features	38
3.3	Feature Selection	40
3.3.1	Principal Component Analysis	40
3.3.2	Genetic Algorithm for Feature Selection	41
3.4	Classification	42
3.4.1	Decision Tree	42
3.4.2	Random Forest	43
3.5	Computational Experiments	43
3.5.1	Experimental Setup	44
3.5.2	Results for Pectoral Removal	44
3.5.3	Results for Three Class Classification: normal, benign and malignant	45
3.5.4	Results for Breast Lesion Classification: benign, and malignant	47
3.6	Discussion	57
3.7	Conclusions	61
4	Proposed Explainable Deep Learning System for Digital Mammograms Classification	62
4.1	Preprocessing	64
4.2	Classification	64
4.2.1	ResNet-50	64
4.2.2	BagNet	66
4.2.3	ProtoPNet	68
4.3	Computational Experiments	69
4.3.1	Data Augmentation and Transfer Learning	69
4.3.2	Experimental Setup	70
4.3.3	Results	72
4.3.4	Comparing Traditional Machine Learning and Deep Learning	74
4.4	Discussion	76
4.5	Conclusions	78
5	Proposed Systems for Cancer Detection From Numerical Medical Data Analysis	79
5.1	Prediction of Breast Cancer based on Numerical Data Analysis . .	80
5.1.1	Scientific Problem Definition	80

5.1.2	Data	81
5.1.3	Proposed Approach	81
5.1.4	Computational Experiments	83
5.1.5	Discussion	85
5.1.6	Conclusions	86
5.2	Prediction of Nasal Polyps Recurrence After Endoscopic Sinus Sur- gery	87
5.2.1	Scientific Problem Definition	87
5.2.2	Data	88
5.2.3	Proposed Approach	89
5.2.4	Computational Experiments	91
5.2.5	Discussion	92
5.2.6	Conclusions	93
6	Machine Learning Approaches to fMRI Analysis for Alzheimer’s Di- sease Detection	96
6.1	Scientific Problem Definition	97
6.2	Related Work	99
6.3	Data	101
6.4	Proposed Approach	101
6.4.1	Preprocessing	102
6.4.2	Feature Extraction	103
6.4.3	Classification	105
6.5	Computational Experiments	107
6.5.1	Experimental Setup	107
6.5.2	Results	107
6.6	Discussion	108
6.7	Conclusions	109
7	Conclusions and Future Work	111
7.1	Main Results and Contributions	111
7.2	Directions for Future Work	112
	Bibliography	114

Capitolul 1

Introducere

Odată cu progresele înregistrate în învățarea automată, aplicarea acesteia în sistemele de diagnostic și detecție asistate de calculator (CAD) a dobândit un interes considerabil. Aceste sisteme asistă profesioniștii din domeniul sănătății în interpretarea imaginilor și a datelor medicale, cu scopul de a îmbunătăți acuratețea și eficiența diagnosticării. Datele medicale includ două tipuri principale: imagini și date numerice. Imaginile medicale surprind structura internă a corpului uman, în timp ce datele numerice includ informații demografice privind pacientul, istoricul clinic și rezultatele testelor de laborator.

1.1 Motivația

Învățarea automată (IA) a fost utilizată pe scară largă în analiza datelor medicale, abordările tradiționale precum arborele de decizie (DT) și pădurile aleatorii (RF) dovedindu-se eficiente în clasificarea bolilor. Recent, învățarea profundă (DL), în special rețelele neuronale convoluționale (CNN), a avut succes în sarcinile medicale bazate pe imagini. Cu toate acestea, seturile de date medicale suferă adesea de dimensiuni mici ale eșantioanelor, dimensionalitate ridicată și dezechilibru între clase, ceea ce face dificilă dezvoltarea modelelor.

Interpretabilitatea este esențială în aplicațiile medicale. Modelele IA tradiționale, cum ar fi DT și RF, sunt în mod inerent interpretabile, oferind o cale de decizie clară. În schimb, modelele DL funcționează adesea ca sisteme de tip “black-box”, necesitând abordări auto-explicative pentru încredere și transparență. Această teză își propune să dezvolte sisteme CAD interpretabile care combină IA tradiționale și DL explicabilă pentru analiza datelor medicale, concentrându-se pe detectarea cancerului de sân utilizând mamografii digitale.

1.2 Obiective

Teza își propune să dezvolte și să evalueze modele IA și DL interpretabile pentru clasificarea datelor medicale, cu un accent principal pe analiza mamografiilor digitale. Obiectivele cheie includ:

1. Dezvoltarea unui sistem CAD pentru clasificarea mamografiilor, care distinge cazurilor normale, benigne și maligne pe baza texturii.
2. Dezvoltarea unui sistem CAD pentru clasificarea mamografiilor, capabil să distingă cazurile benigne și maligne utilizând caracteristici bazate pe textură și formă.
3. Explorarea modelelor de învățare profundă pentru clasificarea leziunilor, asigurând interpretabilitatea prin evidențierea regiunilor relevante ale mamografiei.
4. Analiza datelor medicale numerice reale și aplicarea tehnicilor IA pentru a clasifica boli, precum cancerul de sân și recidiva polipilor nazali.
5. Extinderea metodelor la alte aplicații medicale prin evaluarea capacității de detectare a bolii Alzheimer utilizând date fMRI.

1.3 Contribuții

Teza contribuie la domeniul diagnosticării și detectării asistate de calculator prin dezvoltarea de modele IA și DL pentru analiza datelor medicale prin următoarele inovații:

1. Un *sistem CAD nou pentru clasificarea mamografiilor digitale* în trei clase: (a) normal, (b) benign, și (c) malign. Sistemul utilizează caracteristici de textură extrase din matricea Gray-Level Run-Length (GLRLM) împreună cu clasificarea IA tradițională. Sistemul este evaluat pe setul de date MIAS (Mammographic Image Analysis Society) [SPD1994], demonstrând performanțe competitive în clasificarea cancerului de sân.

Rezultatele acestei contribuții sunt publicate în [BAC2021; Baj2021].

2. Un *sistem CAD nou pentru clasificarea leziunilor* în două clase: (a) benign, și (b) malign. Sistemul utilizează o nouă metodă de extragere a caracteristicilor *combinând caracteristicile bazate pe textură (GLRLM) și pe formă* extrase din mamografii, oferind astfel o imagine cuprinzătoare a acestora și conducând la performanțe mai bune de clasificare. Sistemul este evaluat pe setul

de date Digital Database for Screening Mammography (DDSM) [Hea+1998; Hea+2001], obținând rezultate competitive. În plus, este evaluată combinația de caracteristici extrase din diferite perspective ale mamografiei.

Rezultatele contribuției sunt publicate în [BAC2024; BC2025a].

3. O *analiză cuprinzătoare* care evaluează performanța modelelor IA tradiționale, și anume arbori de decizie și păduri aleatorii, în analiza imaginilor medicale. Analiza include evaluarea (a) caracteristici de textură (GLRLM), (b) caracteristici bazate pe formă (definind reprezentarea unidimensională a unei leziuni) și (c) combinația de caracteristici menționate în (a) și (b). Măștile leziunilor sunt necesare pentru a extrage caracteristicile bazate pe formă; prin urmare, este utilizat setul de date DDSM, deoarece acesta oferă o definiție precisă a limitelor leziunilor.

Rezultatele contribuției sunt publicate în [BC2023; BCA2023].

4. Un sistem inovator și complet automatizat pentru clasificarea cancerului de sân care integrează metode interpretabile de preprocesare, segmentare, extragere și selecție a caracteristicilor, precum și clasificare. Preprocesarea constă în definirea țesutului mamar cu ajutorul pragurilor binare și îmbunătățirea acestuia. Etapele intermediare pot fi prezentate utilizatorului pentru a oferi o perspectivă asupra funcționării sistemului. Pentru segmentare este utilizat algoritmul Threshold-based GrowCut (ThGC) [Mor+2022]. Fiind o metodă iterativă, rezultatele sale pot fi afișate la fiecare iterație pentru o înțelegere mai profundă a procesului de segmentare. Din zona segmentată se extrag caracteristici ușor de înțeles, bazate atât pe textură, cât și pe formă, urmate de un proces de selecție a caracteristicilor, cu scopul de a reduce dimensiunea datelor de intrare prin utilizarea analizei componentelor principale (PCA). Pentru clasificare, se construiesc arbori de decizie și păduri aleatorii. Aceste modele au o structură interpretabilă, facilitând înțelegerea procesului decizional. Sistemul este evaluat pe seturile de date MIAS și DDSM, obținând performanțe comparabile cu sistemele prezentate în literatura de specialitate.

Rezultatele contribuției sunt publicate în [Mor+2022; Mor+2025].

5. Un sistem bazat pe modele de învățare profundă, în care sunt evaluate ResNet-50, BagNet și ProtoPNet, pentru detectarea cancerului de sân din mamografii digitale. Pentru a îmbunătăți performanța modelelor pe seturi de date medicale, se aplică tehnici de învățare prin transfer și augmentare a datelor. În plus, modelele ResNet-50 și ProtoPNet sunt extinse cu straturi suplimentare de “dropout”, separate prin normalizarea loturilor, pentru a îmbunătăți performanța de clasificare. Transparența este esențială în sistemele integrate

în procesul decizional medical. Prin urmare, interpretabilitatea modelelor este esențială. BagNet și ProtoPNet sunt modele autoexplicative bazate pe ResNet-50 cu capacitatea de a evidenția regiunile relevante ale mamografiei utilizate pentru clasificare. Cadrul este evaluat pe seturile de date MIAS și DDSM.

Rezultatele contribuției sunt publicate în [BCS2025; Por+2024].

6. Two novel systems utilizing traditional machine learning methods for *numerical medical data analysis*.

- (i) The first system is designed for breast cancer diagnosis using numerical medical data. The system is based on traditional ML models, using numerical data extracted from two biological fluids: blood (serum) and urine. The system shows high performance and emphasizes the use of this type of numerical analysis for breast cancer screening.
- (ii) The second system aims to predict the recurrence of nasal polyps using numerical medical data. The system utilizes traditional ML models and achieves promising results, offering valuable insights. Its predictions assist healthcare professionals in developing tailored treatment plans for improved outcomes.

The development of the systems involves multiple iterations in cooperation with the field experts from the research group. The results are obtained on real-world data and evaluated both numerically and by medical experts. The achieved results demonstrate the effectiveness and robustness of the proposed systems.

The results of the contribution are published in [Gat+2024; Ian+2022].

7. *Un sistem bazat pe învățarea automată pentru analiza imaginilor de rezonanță magnetică funcțională (fMRI) cu accent pe detectarea bolii Alzheimer.* Din datele fMRI sunt extrase trei rețele de conectivitate funcțională din 116 regiuni cerebrale definite pentru a detecta întreruperile în conectivitate, și anume (1) conectivitate topografică de ordin inferior, (2) conectivitate topografică de nivel înalt, și (3) conectivitate funcțională asociată de ordin înalt. Pentru clasificare sunt utilizate atât modele tradiționale de învățare automată, cât și modele de învățare profundă. Performanța modelelor ML și DL este comparată în ceea ce privește performanța de clasificare a fMRI, evidențiind eficiența modelelor ML tradiționale față de modelele DL în scenarii cu date limitate.

Rezultatele contribuției au fost trimise pentru publicare în [BC2025b].

1.4 Lista publicațiilor

Această secțiune oferă o listă a publicațiilor care au rezultat din cercetarea efectuată în cadrul acestei teze. Publicațiile din reviste sunt clasificate în conformitate cu Unitatea Executivă pentru Finanțarea Învățământului Superior, a Cercetării, Dezvoltării și Inovării (UEFISCDI)¹ pe baza anului de publicare și a scorului de influență al articolului. Articolele de conferință sunt clasificate în funcție de Compute Research and Education (CORE)² (categoriile A*, A, B, C, D) în funcție de anul publicării.

1. **Adél Bajcsi**, and Camelia Chira, “Feature combination from different mammo-gram perspectives to improve lesion classification”, în *Logic Journal of the IGPL* (2025), Accepted., [category C, 2 points]
2. **Adél Bajcsi**, Camelia Chira, and Annamária Szenkovits, “Evaluating ResNet-based Interpretable Models for Breast Lesion Classification”, în *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART2025*, SciTePress, 2025, pp. 288–295, ISBN: 978-989-758-737-5, DOI: 10.5220/0013121900003890, [category B, short paper, 2.67 points]
3. Cristiana Moroz-Dubenco, **Adél Bajcsi**, Anca Andreica, and Camelia Chira, “Towards an interpretable breast cancer detection and diagnosis system”, în *Computers in Biology and Medicine* 185 (2025), p. 109520, ISSN: 0010-4825, DOI: 10.1016/j.combiomed.2024.109520, [category A, 4 points]
4. **Adél Bajcsi**, Anca Andreica, and Camelia Chira, “Significance of Training Images and Feature Extraction in Lesion Classification”, în *Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART2024*, SciTePress, 2024, pp. 117–124, ISBN: 978-989-758-680-4, DOI: 10.5220/0012308900003636, [category B, short paper, 2.67 points]
5. Anda Gata, Lajos Raduly, Livița Budișan, **Adél Bajcsi**, Teodora-Maria Ursu, Camelia Chira, Laura Dioșan, Ioana Berindan-Neagoe, and Silviu Albu, “Machine Learning Model Predicts Postoperative Outcomes in Chronic Rhinosinusitis With Nasal Polyps”, în *Clinical Otolaryngology* 49.6 (Aug. 2024), pp. 776–784, ISSN: 1749-4486, DOI: 10.1111/coa.14208, [category B, 0.57 points]
6. Ábel Portik, **Adél Bajcsi**, Annamária Szekovits, and Zalán Bodó, “Exploring the Impact of Backbone Architecture on Explainable CNNs’ Interpretability”, în *Acta Universitatis Sapientiae* 16 (1 2024), pp. 105–123, DOI: 10.2478/ausi-2024-0007, [category D, -]

¹UEFISCDI: <https://uefiscdi.gov.ro/scientometrie-reviste>

²CORE Conference Portal:
[urlhttps://portal.core.edu.au/conf-ranks](https://portal.core.edu.au/conf-ranks)

7. **Adél Bajcsi**, and Camelia Chira, “Textural and Shape Features for Lesion Classification in Mammogram Analysis”, în *Hybrid Artificial Intelligent Systems HAIS2023 (Lecture Notes in Computer Science)*, vol. 14001, Springer Nature Switzerland, 2023, pp. 755–767, ISBN: 9783031407253, DOI: 10.1007/978-3-031-40725-3_64, [category C, 2 points]
8. **Adél Bajcsi**, Camelia Chira, and Anca Andreica, “Extended Mammogram Classification From Textural Features”, în *Studia Universitatis Babeş-Bolyai Informatica* 67 (2 Feb. 2023), pp. 5–20, DOI: 10.24193/subbi.2022.2.01, [category D, 1 point]
9. Stefania D. Iancu, Ramona G. Cozan, Andrei Stefancu, Maria David, Tudor Moisoiu, Cristiana Moroz-Dubenco, **Adél Bajcsi**, Camelia Chira, Anca Andreica, Loredana F. Leopold, Daniela Eniu, Adelina Staicu, Iulian Goidescu, Carmen Socaciu, Dan T. Eniu, Laura Diosan, and Nicolae Leopold, “SERS liquid biopsy in breast cancer. What can we learn from SERS on serum and urine?”, în *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 273 (Mai 2022), p. 120992, ISSN: 1386-1425, DOI: 10.1016/j.saa.2022.120992, [category B, 0.27 points]
10. Cristiana Moroz-Dubenco, **Adél Bajcsi**, Anca Andreica, and Camelia Chira, “An Unsupervised Threshold-based GrowCut Algorithm for Mammography Lesion Detection”, în *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022*, vol. 207, Elsevier BV, 2022, pp. 2096–2105, DOI: 10.1016/j.procs.2022.09.269, [category B, 2 points]
11. **Adél Bajcsi**, “Towards a Support System for Digital Mammogram Classification”, în *Studia Universitatis Babeş-Bolyai Informatica* 66 (2 Dec. 2021), p. 19, ISSN: 2065-9601, DOI: 10.24193/subbi.2021.2.02, [category D, 1 point]
12. **Adél Bajcsi**, Anca Andreica, and Camelia Chira, “Towards feature selection for digital mammogram classification”, în *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021*, vol. 192, Procedia Computer Science, 2021, DOI: 10.1016/j.procs.2021.08.065, [category B, 4 points]

În plus, următorul articol a fost trimis spre publicare la conferința *Artificial Intelligence in Medicine*:

Adél Bajcsi, and Camelia Chira, “fMRI Analysis for Alzheimer’s Disease Detection: Traditional vs. Deep Learning Models”, în *Artificial Intelligence in Medicine AIME2025*, Submitted., 2025, [category B]

Capitolul 2

Sistemul Propus Pentru Clasificarea Mamografiilor Digitale Bazat pe Învățare Automată Tradițională

Cancerul de sân este cel mai frecvent cancer diagnosticat la femei. De asemenea, reprezintă cel mai mare număr de decese cauzate de cancer în Uniunea Europeană [EU 2023]. Detectarea precoce a anomaliilor este crucială pentru un tratament de succes. Mamografia digitală utilizează raze X pentru a surprinde structura țesutului mamar. Mamografiile sunt frecvent utilizate pentru depistarea cancerului de sân și facilitează dezvoltarea sistemelor CAD care ajută medicii în stabilirea diagnosticului și alegerea tratamentului adecvat. Acest capitol prezintă metodele tradiționale de învățare automată (ML) propuse și dezvoltate pentru a aborda problema clasificării mamografiilor.

2.1 Clasificare în Trei Clase: normal, benign și malign

Seturile de date privind mamografiile digitale conțin de obicei trei clase: normale, benigne și maligne. Scopul este de a dezvolta un sistem de clasificare pentru a distinge între aceste clase.

Mamografiile scanate includ adesea artefacte (de exemplu, adnotări) și mușchiul pectoral, care sunt irelevante pentru detectarea tumorilor. Artefactele sunt eliminate prin clasificare binară, reținând cea mai mare componentă conectată ca mască mamară. Propunem două metode pentru eliminarea mușchiului pectoral:

o abordare cu ferestre glisante [Shr+2017] și un algoritm de creștere a regiunii însămânțate [MNB2012]. În plus, egalizarea adaptivă a histogramei cu contrast limitat (CLAHE) îmbunătățește textura imaginii.

Caracteristicile de textură, cum ar fi matricea Gray-Level Run-Length (GLRLM) [CK2019], caracterizează densitatea țesuturilor. GLRLM captează relațiile spațiale prin analizarea execuțiilor la nivel de gri în patru direcții (orizontală, verticală și diagonală). Un total de 44 de caracteristici sunt extrase din imaginile mamografice.

Pentru a reduce dimensionalitatea, utilizăm analiza componentelor principale (PCA) [SGM2010] și un algoritm genetic (GA) [CK2019]. PCA transformă caracteristicile de intrare în componente ortogonale ordonate în funcție de varianța explicată, în timp ce GA identifică subseturile optime de caracteristici prin evaluarea funcției de fitness, bazată pe un clasificator.

Pentru a asigura interpretabilitatea, utilizăm clasificatoare bazate pe arbori de decizie (DT) și păduri aleatorii (RF). Aceste modele oferă transparență și informații privind importanța caracteristicilor, promovând încrederea în sistemele de asistență pentru decizii medicale.

2.2 Clasificare în Două Clase: benign și malign

Selectarea tratamentului adecvat pentru cancerul de sân este o provocare, deoarece depinde de tipul de cancer, de stadiu și de starea de sănătate a pacientului. Propunem un sistem binar de clasificare pentru a diferenția tumorile benigne de cele maligne.

Masca leziunii izolează regiunea de interes. Dreptunghiul de delimitare din jurul anomaliei este extins cu 25 de pixeli pentru a surprinde țesutul înconjurător și microcalcificările potențiale, care pot indica malignitatea.

Leziunile maligne prezintă forme neregulate, în timp ce leziunile benigne sunt mai circulare [Dav2022]. Extragem caracteristicile formei, inclusiv aria, perimetrul și compactitatea. În plus, caracteristicile bazate pe contur sunt calculate prin compararea leziunii segmentate cu o elipsă înconjurătoare [Li+2017].

Mamografiile sunt realizate din două perspective: (1) mediolateral oblic (MLO) (vedere laterală) (2) craniocaudal (CC) (vedere de sus în jos). Caracteristicile extrase din ambele perspective sunt concatenate într-un singur vector pentru a îmbunătăți clasificarea.

Utilizăm două metode de selecție a caracteristicilor (PCA, GA) și două clasificatoare (DT, RF), descrise în Section 2.1, pentru a clasifica mamografiile ca fiind benigne sau maligne.

2.3 Experimente

Metodele propuse sunt evaluate utilizând patru metrici: acuratețe, precizie, recall și scorul F_1 . Acești parametri sunt frecvent utilizați pentru a evalua sarcinile de clasificare dezechilibrate.

Eficiența metodelor de eliminare a mușchiului pectoral este evaluată utilizând setul de date MIAS [SPD1994]. Deoarece acest set de date nu conține adnotări explicite ale mușchilor pectorali, gruparea k-means [Has+2021] este utilizată pentru a genera măști pectorale. Rezultatele indică faptul că metoda de creștere a regiunilor înșămânțate are performanțe superioare față de abordarea cu ferestre glisante, obținând o precizie de 95,72% și o recall de 70,14%.

Sistemul CAD pentru distingerea țesutului mamar normal, benign și malign este evaluat pe setul de date MIAS. Aceasta se bazează pe caracteristicile de textură GLRLM extrase din mamografii. Cea mai bună performanță este obținută utilizând PCA pentru selectarea caracteristicilor și păduri aleatorii pentru clasificare, atingând o acuratețe de 72,84%.

Pentru clasificarea binară a leziunilor mamare, sistemul este evaluat pe DDSM [Hea+1998; Hea+2001], care oferă segmentare manuală pentru fiecare leziune. Experimentele preliminare arată că presupunerea asupra formei leziunii nu este întotdeauna valabilă, deoarece unele leziuni benigne au forme neregulate, în timp ce unele leziuni maligne prezintă modele circulare. Pentru a aborda această problemă, sunt create subseturi stratificate ale setului de date pe baza neregularității leziunilor. Rezultatele demonstrează că caracteristicile de formă depășesc caracteristicile de textură în distingerea anomaliilor mamare. Cea mai mare acuratețe, 96,12%, este obținută utilizând clasificarea Random Forest fără selectarea caracteristicilor pe subsetul care exclude valorile aberante. În plus, o analiză a caracteristicilor extrase din diferite perspective ale mamografiei indică faptul că cele derivate din perspectiva MLO depășesc ușor caracteristicile combinate din ambele perspective.

2.4 Concluzii

Acest capitol a prezentat un sistem CAD bazat pe învățarea automată tradițională pentru clasificarea cancerului de sân utilizând mamografii digitale. Sistemul propus urmează o structurată bine definită: preprocesare, extragerea caracteristicilor, selectarea caracteristicilor și clasificarea.

Rezultatele experimentale demonstrează că arborele de decizie (DT) și pădurile aleatorii (RF) ating performanțe competitive de clasificare, menținând totodată interpretabilitatea acestora. Metodele de extragere a caracteristicilor bazate pe textură (GLRLM) și formă contribuie semnificativ la clasificarea precisă a leziunilor. În plus, tehnicile de selecție a caracteristicilor, precum analiza componentelor

principale (PCA) și algoritmi genetici (GA), îmbunătățesc eficiența modelului prin reducerea dimensionalității. Constatările evidențiază avantajele învățării automate tradiționale față de modelele de învățare profundă în scenarii cu date limitate, demonstrând că metodele interpretabile pot obține o precizie ridicată, oferind în același timp transparență în luarea deciziilor.

Capitolul 3

Sistemul Propus Pentru Clasificarea Mamografiilor Digitale Bazat pe Învățare Profundă

Modelele de învățare profundă sunt frecvent utilizate pentru sarcinile de clasificare a imaginilor. În general, construirea unei rețele neuronale profunde necesită o cantitate considerabilă de date, datorită numărului mare de parametri optimizați în timpul antrenării.

Seturile de date medicale conțin, de obicei, cel mult câteva mii de cazuri, din cauza numărului limitat de subiecți și a preocupărilor legate de confidențialitatea datelor. În consecință, augmentarea datelor a devenit o tehnică esențială pentru extinderea artificială a dimensiunii și varianței acestor seturi de date.

În domenii critice, cum ar fi asistența medicală, transparența modelului este esențială. Explicabilitatea acestuia joacă un rol crucial în acceptarea sa de către cadrele medicale. Prin urmare, în acest capitol investigăm două rețele neuronale explicabile, și anume ProtoPNet [Che+2018] și BagNet [BB2019].

3.1 Metoda Propusă

Imaginile sunt preprocesate similar cu Section 2.2. În plus, acestea sunt normalizate pentru a avea o medie zero și o varianță unitară și redimensionate la 224×224 .

În experimente, sunt incluse trei rețele: ResNet-50 [He+2015], BagNet [BB2019] și ProtoPNet [Che+2018]. ResNet-50 [He+2015] servește drept baza pentru BagNet și ProtoPNet. Acesta este o rețea de tip “black-box” utilizat pe scară largă,

conceput pentru a atenua problema gradientului care dispare, prin blocuri reziduale. Pentru a spori performanța, sunt introduse straturi suplimentare de renunțare înainte de stratul complet conectat, combinate cu normalizarea loturilor și activarea ReLU.

BagNet [BB2019] modifică structura modelului ResNet-50 prin: (1) înlocuirea convoluției inițiale de 7×7 cu o convoluție de 3×3 și (2) reducerea numărului de convoluții de 3×3 , păstrând doar primul strat al unei convoluții reziduale de 3×3 , în timp ce restul sunt înlocuite cu convoluții de 1×1 .

ProtoPNet [Che+2018] constă din trei module: (1) extragerea caracteristicilor, (2) generarea prototipurilor și (3) clasificarea. Pentru a asigura comparabilitatea, extragerea caracteristicilor este realizată utilizând un model ResNet-50. Similar celorlalte modele, ProtoPNet este extins cu straturi de normalizare a loturilor și de renunțare, înainte de stratul responsabil pentru clasificarea.

3.2 Experimente

Pentru a evalua sistemul CAD bazat pe învățarea profundă propus, au fost efectuate mai multe experimente pe seturi de date publice de mamografii (MIAS [SPD1994] și DDSM [Hea+1998; Hea+2001]). Experimentele analizează eficiența ResNet-50, BagNet și ProtoPNet, concentrându-se atât pe acuratețea clasificării, cât și pe interpretabilitate (prin evaluare vizuală).

Având în vedere dimensiunea limitată a seturilor de date medicale, a fost aplicată augmentarea datelor (de exemplu, rotație, răsturnare, ajustări de contrast) pentru a îmbunătăți capacitatea de generalizare a modelului. Învățarea prin transfer a fost utilizată prin ajustarea fină a modelelor pre-antrenate pentru a valorifica cunoștințele din seturi de date mai mari.

Rezultatele au arătat că BagNet și ProtoPNet au îmbunătățit interpretabilitatea prin generarea de imagini care evidențiază regiunile relevante din mamografii, în timp ce ResNet-50 a obținut o acuratețe ridicată, dar fără transparență. Atunci când se compară învățarea automată tradițională cu modelele de învățare profundă, metodele de învățare profundă ofer performanțe puternice de clasificare. Totuși, metodele tradiționale de învățare automată au rămas competitive, în special pentru seturile de date mai mici în care interpretabilitatea joacă un rol crucial.

3.3 Concluzii

Acest capitol a prezentat un sistem CAD bazat pe învățare profundă pentru clasificarea cancerului de sân, concentrându-se pe modele explicabile pentru a spori interpretabilitatea. Au fost evaluate trei arhitecturi: ResNet-50, BagNet și Pro-

toPNet. Rezultatele arată că modelele de învățare profundă interpretabilă oferă informații valoroase în procesul de luare a deciziilor prin evidențierea regiunilor relevante ale mamografiei. Deși învățarea profundă atinge o acuratețe ridicată, rămâne mai complexă și mai puțin interpretabilă în comparație cu modelele tradiționale de învățare automată. BagNet și ProtoPNet oferă o transparență mai bună prin generarea imaginilor care evidențiază regiunile relevante, ceea ce le face mai potrivite pentru aplicațiile medicale care necesită explicabilitate.

Capitolul 4

Proposed Systems for Cancer Detection from Numerical Medical Data Analysis

Acest capitol se concentrează pe dezvoltarea sistemelor CAD prin utilizarea metodelor tradiționale de învățare automată bazate pe analiza datelor numerice. În primul rând, analiza Spectroscopiei Raman amplificată de suprafață a biofluidelor este utilizată pentru predicția cancerului de sân, urmată de clasificarea rapoartelor de analiză a țesuturilor pentru a prezice recidiva polipilor nazali. Ambele studii de cercetare sunt rezultatul unei colaborări cu experți din domeniul medical și sunt realizate pe date reale.

4.1 Predicția Cancerului de Sân pe Baza Analizei Datelor Numerice

Detectarea cancerului de sân cu ajutorul datelor medicale numerice reprezintă o alternativă valoroasă la abordările bazate pe imagini. Printre acestea, Spectroscopie Raman amplificată de suprafață (SERS) a stârnit un interes semnificativ ca tehnică promițătoare. Odată cu dezvoltarea inteligenței artificiale, sistemele CAD bazate pe învățare automată au fost studiate pe scară largă. Această secțiune prezintă rezultatele unei colaborări cu un grup multidisciplinar de cercetători.

4.1.1 Metoda Propusă

Primul pas al sistemului constă în preprocesarea spectrelor SERS. Pentru a reduce zgomotul se utilizează filtrare. Ulterior, spectrele sunt normalizate și netezite.

Propunem utilizarea a cinci metode de învățare automată pentru clasificarea informațiilor SERS rezultate din analiza diferitelor specimene biologice: analiza discriminantă liniară – LDA [Zha+2024], arbori de decizie – DT [Bur2019], păduri aleatorii – RF [Bur2019], Bayes naiv – GNB [Rah+2021], și mașinile cu vectori suport – SVM [An+2023]. Scopul acestor modele este de a distinge pacientele din grupul de control de cele cu cancer mamar.

DT și RF au fost alese pentru capacitatea lor de a gestiona date cu dimensiuni ridicate și a interpretabilității pe care o oferă. LDA este un clasificator liniar utilizat pe scară largă în domeniul diagnosticării medicale. GNB este un clasificator probabilistic simplu, ușor de implementat și eficient din punct de vedere computațional. SVM este un clasificator puternic care poate gestiona date înalt-dimensionale și este utilizat pe scară largă în diagnosticarea medicală.

4.1.2 Experimente

Sistemul este evaluat pe un set de date reale, colectate de la paciente diagnosticate cu cancer mamar și subiecți de control. Setul de date conține spectre SERS obținute din probe de ser și urină.

Performanța sistemului este evaluată utilizând următorii parametri: acuratețe, precizie, reamintire și scor F_1 . Utilizarea spectrelor extrase din ser obține cea mai bună performanță cu o acuratețe de 83.33% și un scor F_1 de 88% utilizând clasificatorul LDA. Pe de altă parte, cel mai performant model care utilizează datele SERS din urină este clasificatorul DT, cu o acuratețe de 88.89%, o precizie, un recall și un scor F_1 de 91.67%. Rezultatele indică faptul că spectrele SERS extrase din urină și ser conțin informații diferite, ceea ce poate duce la o performanță mai bună a clasificării. Combinația spectrelor SERS ale serului și urinei este, de asemenea, investigată și prezintă performanțe similare cu clasificarea care utilizează date SERS ale serului.

În plus, a fost investigat efectul selecției caracteristicilor (PCA), iar rezultatele arată că reducerea dimensionalității datelor nu îmbunătățește semnificativ performanța modelelor. Rezultatele sugerează că spectrele SERS conțin informații valoroase pentru clasificarea pacientelor diagnosticate cu cancer mamar și a subiecților de control.

În plus față de evaluarea numerică, rezultatele sunt evaluate și de experții medicali ai grupului de cercetare.

4.1.3 Concluzii

În această secțiune, am propus un sistem CAD pentru detectarea cancerului de sân utilizând spectrele SERS. Sistemul utilizează cinci clasificatoare diferite (analiza discriminantă liniară, arborele de decizie, păduri aleatorii, Bayes naiv și mașinile

cu vectori suport) pentru a distinge pacientele diagnosticate cu cancer mamar de subiecții de control pe baza spectrelor SERS extrase din ser și urină. Rezultatele arată că se obțin precizii de clasificare ușor mai mari utilizând probe de urină comparativ cu probe de ser, în ciuda variabilității mai mari a spectrelor SERS din urină. În plus, a fost investigat efectul combinării datelor SERS obținute din ser și urină, demonstrând o performanță similară cu cea a clasificării utilizând date SERS extrase din ser.

4.2 Predicția Recidivei Polipilor Nazali După Operația Endoscopică a Sinusurilor

Chirurgia sinusală endoscopică este cel mai frecvent utilizat tratament pentru rinossinuzita cronică cu polipi nazali atunci când tratamentele medicale nu sunt suficient de eficiente [Tao+2018]. Rata de recidivă a polipilor este îngrijorătoare, apărând dezvoltarea unui sistem CAD care să prezică probabilitatea de recidivă și să ajute medicii în selectarea celui mai adecvat tratament. Experimentele prezentate în această secțiune sunt rezultatul unei colaborări cu experți medicali, care au colectat date din lumea reală.

4.2.1 Metoda Propusă

În această secțiune, propunem un sistem CAD pentru prezicerea recidivei polipilor nazali din două tipuri de date: (1) neinvaziv, și (2) invaziv. Datele sunt normalizate, iar corelația dintre caracteristici este analizată.

Scopul sistemului este de a prezice trei clase: (1) control, (2) control parțial, și (3) recidivă. Interpretabilitatea modelelor este crucială în aplicațiile medicale, motiv pentru care clasificatorii pe bază de arbori sunt utilizați datorită interpretabilității lor inerente și a rezultatelor promițătoare obținute în analiza datelor medicale. Arborii de decizie (DT) sunt clasificatori simpli care iau decizii prin evaluarea unei serii de condiții aplicate la fiecare nod al arborelui, conducând la o decizie de clasificare la nivelul frunzelor. Această structură face ca DT-urile să fie ușor de interpretat și utile pentru înțelegerea modului în care anumite caracteristici influențează predicțiile. Pădurile aleatorii (RF) sunt modele de ansamblu formate dintr-un număr de DT, unde eticheta finală este determinată prin vot majoritar. Acest mecanism îmbunătățește precizia, surprinde o gamă mai largă de modele de decizie și sporind astfel robustețea, reducând supraajustarea.

4.2.2 Experimente

Sistemul este evaluat pe un set de date reale, obținut în urma unei perioade de urmărire de 18 luni, cu evaluări semestriale ale pacienților care au suferit o intervenție chirurgicală endoscopică a sinusurilor. Scopul sistemului este de a prezice starea pacientului la 18 luni după operație. În primul rând, performanța clasificatorilor este evaluată în funcție de capacitatea lor de a prezice recidiva la fiecare vizită de monitorizare. Rezultatele indică faptul că clasificatorii păduri aleatorii obțin rezultate mai bune.

În ultimii ani, utilizarea microARN-urilor a devenit interesantă ca posibil biomarker pentru rinosinuzita cronică [Son+2022]. Setul de date conține două tipuri de microARN: microARN 125b și microARN 203a-3p, extrase din probe de țesut prelevate în timpul intervenției chirurgicale. Este analizat efectul diferitelor combinații de microARN asupra performanței de predicție. Rezultatele arată că utilizarea microARN 125b îmbunătățește performanța clasificatoarelor, obținând o acuratețe de 84.62%, o precizie de 87.28%, un recall de 84.62% și scor F_1 de 83.79%.

Utilizarea clasificatorului RF permite analiza importanței caracteristicilor, oferind o perspectivă valoroasă asupra relevanței acestora pentru sarcina de predicție. Rezultatele indică faptul că microARN 125b este cea mai importantă caracteristică pentru predicția recidivei polipilor nazali.

4.2.3 Concluzii

În această secțiune, sunt prezentate rezultatele unei colaborări cu experți medicali de la Departamentul de Otorinolaringologie al Universității de Medicină și Farmacie “Iuliu Hațieganu” având ca obiectiv predicția reformării polipilor nazali după operație. Pentru clasificarea datelor reale, colectate la Spitalul Clinic CF Cluj-Napoca, sunt propuse modele bazate pe arbori de decizie cu scopul de a estima severitatea postoperatorie a polipilor nazali. Cercetarea noastră arată relevanța microARN-urilor, în special microARN 125a, în predicția recurenței bolii. În plus, RF obține cea mai bună performanță, cu o precizie de 84.62% și un scor F_1 de 85.93%. Deoarece RF este un model interpretabil prin natură sa, acesta oferă o perspectivă valoroasă asupra importanței caracteristicilor, permițând o mai bună înțelegere și transparență în procesul decizional medical.

Capitolul 5

Abordări de Învățare Automată a Analizei fMRI Pentru Detectarea Bolii Alzheimer

Imaginile medicale sunt utilizate pe scară largă și sunt interdisciplinare. De exemplu, analiza imaginilor de rezonanță magnetică funcțională (fMRI) poate ajuta psihologii să prezică efectul unei traume sau neurologii să prezică stadiul demenței. Scopul acestui capitol este de a explora performanța metodelor propuse anterior pentru analiza fMRI pentru boala Alzheimer.

În prezenta teză, propunem un sistem CAD care utilizează învățarea automată pentru analiza și clasificarea datelor fMRI. fMRI sunt utilizate pe scară largă pentru detectarea bolii Alzheimer prin înregistrarea nivelului de oxigen din sânge, permițând analiza conectivității creierului.

5.1 Metoda Propusă

fMRI în stare de repaus (rs-fMRI) este cea mai frecvent utilizată metodă pentru detectarea Alzheimer. rs-fMRI sunt imagini cvadridimensionale: volumul creierului este înregistrat pe o perioadă de timp. Aceste imagini captează nivelul de oxigen din sângele cerebral, reprezentând activitatea regiunii respective. Prin extragerea și analizarea semnalelor din aceste imagini, putem deduce conectivitatea dintre diferite regiuni ale creierului. Matricea de conectivitate rezultată este apoi utilizată pentru antrenarea unui model de învățare automată pentru clasificare.

Preprocesarea datelor rs-fMRI începe cu eliminarea primelor 10 volume pentru a asigura stabilizarea câmpului magnetic. Datele sunt apoi corectate pentru sincronizarea feliilor și pentru reducerea artefactelor de mișcare. Imaginile sunt coregistrate la imaginile anatomice și normalizate la șablonul MNI. Ulterior, datele

sunt netezite și filtrate pentru a elimina zgomotul. Regiunile de interes (ROI) ale creierului sunt determinate utilizând atlasul AAL. Semnalul mediu din fiecare ROI este extras și utilizat pentru a calcula matricea de conectivitate.

Propunem utilizarea a trei tipuri de rețele de conectivitate funcțională (FC): (1) topografică de ordin inferior, (2) topografică de ordin superior și (3) asociată de ordin superior. FC de ordin inferior reprezintă corelația dintre ROI-uri care formează o subrețea. FC topografică de ordin superior descrie corelația dintre subrețele, iar FC asociată de ordin superior reflectă corelația dintre ROI-uri și subrețele. Utilizând o fereastră glisantă, semnalul poate fi împărțit în segmente mai mici, care sunt folosite pentru a calcula rețelele FC menționate anterior, obținând astfel caracteristici dinamice. Pe baza experimentelor prezentate în [Zha+2017], lungimea ferestrei este setată la 70, iar pasul la 1, rezultând 61 de subserii. În plus, caracteristicile statice sunt definite utilizând o fereastră de lungime egală cu durata totală a semnalului.

Pentru a clasifica caracteristicile extrase, sunt utilizați atât algoritmi tradiționali de învățare automată, cât și modele de învățare profundă. Algoritmii tradiționali de învățare automată includ: mașinile cu vectori suport (SVM), arbori de decizie (DT), păduri aleatorii (RF) și k-Nearest Neighbors (kNN). Modelele de învățare profundă utilizate sunt ResNet-18 și ResNet-50. DT, RF și SVM au fost descriși în Section 4.1. kNN este un algoritm simplu și eficient, care clasifică datele pe baza majorității celor mai apropiați k vecini. ResNet-18 și ResNet-50 sunt rețele neuronale convoluționale utilizate pe scară largă pentru sarcini de clasificare a imaginilor, diferențiindu-se prin complexitatea modelului, determinată de numărul de straturi.

5.2 Experimente

Pentru a evalua performanța metodelor propuse, am efectuat experimente pe setul de date Alzheimer’s Disease Neuroimaging Initiative (ADNI) [ADNI2024]. Acest set de date conține o colecție extinsă de date clinice și de neuroimagistică, concentrându-se în special pe boala Alzheimer (AD) și stadiile sale incipiente. Pentru experimentele actuale, am selectat indivizi etichetați ca pacienți CN (cognitiv normali) sau AD, care au scanări MRI structurale și rs-fMRI. De asemenea, am luat în considerare o distribuție echilibrată a cazurilor și o vârstă medie similară (aproximativ 75 de ani). Ca urmare, au fost selectate 93 de cazuri din fiecare clasă, rezultând un total de 186 de cazuri.

În primul rând, se selectează cea mai eficientă filtrare pentru analiza fMRI. Comparăm performanța algoritmilor tradiționali de învățare automată utilizând următoarele filtre: (1) slow5 (0.01 Hz – 0.027 Hz), (2) slow4 (0.027 Hz – 0.08 Hz), and (3) full-band (0.01 Hz – 0.08 Hz). În general, clasificatoarele obțin performanțe

mai bune cu filtrul slow5. De asemenea, comparăm performanța algoritmilor tradiționali de învățare automată cu cea a modelelor de învățare profundă. Rezultatele indică faptul că metodele tradiționale depășesc modelele de învățare profundă, ceea ce poate fi atribuit dimensiunii reduse a setului de date. Cele mai bune rezultate obținute cu învățarea automată tradițională sunt atinse cu SVM, care obține o acuratețe de testare de 97.37%, în timp ce cele mai bune rezultate din învățarea profundă sunt obținute cu ResNet-18, care atinge o acuratețe de testare de 71.05%. ResNet-18 oferă cele mai bune rezultate atunci când este antrenat pe conectivitate funcțională (FC) dinamică, atingând 71.05%. ResNet-50 este mai puțin eficient decât ResNet-18, ceea ce ar putea fi atribuit numărului mai mare de parametri ai ResNet-50 și dimensiunii limitate a setului de date de antrenare.

5.3 Concluzii

Boala Alzheimer afectează viața a milioane de oameni din întreaga lume. Detectarea la timp a bolii și tratamentul adecvat pot încetini evoluția demenței. În acest capitol, am prezentat un sistem bazat pe analiza fMRI utilizând modelele de învățare automată propuse anterior pentru a detecta boala Alzheimer. După preprocesarea și extragerea conectivității funcționale, am construit patru modele tradiționale de învățare automată (DT, RF, kNN și SVM) și două modele de învățare profundă (ResNet-18 și ResNet-50). Rezultatele arată că SVM are cea mai bună performanță dintre metodele tradiționale de învățare automată, obținând o precizie a testului de 97.37%. De asemenea, concluzionăm că ResNet-18 are performanțe mai bune decât ResNet-50, obținând o acuratețe a testului de 71.05%. Diferența de performanță ar putea fi cauzată de numărul limitat de date disponibile pentru antrenarea modelelor de învățare profundă.

În experimente viitoare, ne propunem să investigăm impactul creșterii datelor și al straturilor suplimentare utilizând normalizarea și renunțarea la loturi pentru a spori performanța învățării profunde. De asemenea, vom extinde analiza noastră și la alte modele de învățare profundă. În plus, în experimentele viitoare, vom include stări intermediare ale AD, cum ar fi insuficiența cognitivă ușoară (MCI), și vom evalua performanța modelelor propuse în ceea ce privește (1) clasificare binară (CN vs. MCI, MCI vs. AD) și (2) clasificare multclasă.

Capitolul 6

Concluzii și Direcții de Cercetare Viitoare

Această teză a explorat dezvoltarea și evaluarea sistemelor interpretabile de învățare automată și de detectare asistată de calculator (CAD) pentru analiza și clasificarea datelor medicale, cu un accent deosebit pe detectarea cancerului de sân utilizând mamografii digitale, printre alte aplicații. Obiectivul principal a fost acela de a avansa metodologiile în diagnosticarea medicală, asigurând atât acuratețea, cât și interpretabilitatea pentru a spori încrederea și capacitatea de utilizare în mediile clinice.

6.1 Rezultatele și Contribuțiile Principale

Această teză contribuie la domeniul clasificării mamografiilor digitale prin propunerea unui nou sistem CAD care combină caracteristicile de textură și formă extrase din mamograme utilizând o abordare tradițională de învățare automată (ML). Sistemul a fost evaluat pe setul de date Digital Database for Screening Mammography (DDSM). Sistemul a clasificat cu succes leziunile mamografice în două categorii: benigne și maligne, cu o acuratețe de 96.12% pe setul de date DDSM. În plus, am propus un sistem complet automatizat de clasificare a cancerului de sân care integrează metode interpretabile. Sistemul a fost evaluat pe seturile de date Mammographic Image Analysis Society (MIAS) și DDSM, obținând o acuratețe de 95%, respectiv 97%, pentru clasificarea cancerului de sân, subliniind robustețea abordării. Sistemele propuse evidențiază eficiența și interpretabilitatea modelelor ML tradiționale, în special pentru seturile de date mai mici.

Cercetarea a subliniat importanța explicabilității modelelor în domeniul asistenței medicale. Am propus utilizarea unor modele de învățare profundă (DL)

auto-explicative pentru detectarea cancerului de sân, esențiale pentru crearea încredere în sistem și pentru a înțelege procesul decizional. Am evaluat trei modele de învățare profundă: ResNet-50, BagNet și ProtoPNet, pe seturile de date MIAS și DDSM. Pentru a aborda limitările seturilor de date medicale, tehnici precum învățarea prin transfer și augmentarea datelor au fost utilizate în mod eficient, îmbunătățind semnificativ performanța modelului, păstrând în același timp integritatea diagnosticului. Modelele au clasificat cu succes leziunile, cu o acuratețe de 93.38%, 92.39% și 78.11% pe MIAS și, respectiv, 95.26%, 95.25% și 91.23% pe setul de date DDSM, utilizând modelele ResNet-50, BagNet și ProtoPNet.

Metodologiile propuse au fost aplicate în trei domenii medicale: detectarea cancerului de sân, predicția recurenței polipilor nazali și detectarea bolii Alzheimer utilizând date fMRI, în colaborare cu experți. Diversitatea aplicațiilor demonstrează flexibilitatea și robustețea metodelor dezvoltate. Pe lângă analiza mamografiilor, a fost dezvoltat un sistem de detectare a cancerului de sân utilizând date medicale numerice (biopsie lichidă SERS). Acest sistem a obținut o acuratețe de 88.89% și un scor F_1 de 91.67% utilizând clasificatorul unui arbore de decizie. Pentru predicția reformării polipilor nazali, au fost utilizate date numerice reale, iar performanța clasificatoarelor bazate pe arbori a fost evaluată. Rezultatele arată superioritatea pădurilor aleatorii față de arbore de decizie. În analiza fMRI, au fost comparate abordările ML și DL. Cel mai performant model ML (SVM) a obținut o acuratețe de 97.37%, în timp ce cel mai performant model DL (ResNet-18) a obținut o acuratețe de 71.05%. Aceste rezultate evidențiază avantajul ML în comparație cu DL atunci când datele sunt limitate.

6.2 Direcții de Cercetare Viitoare

O colaborare interdisciplinară cu radiologi și oncologi de la “Ion Chiricuță” Institutul Oncologic din Cluj-Napoca, România, lucrează la colectarea și adnotarea noului set de date de mamografii digitale. În viitor, vom aplica metodologiile propuse la noul set de date pentru a evalua performanța sistemelor pe țesutul mamar dens.

O direcție de cercetare promițătoare pentru clasificarea cancerului de sân este extinderea sistemului într-un model multimodal, care să combine imaginile mamografice cu detalii demografice și antecedentele medicale ale pacientului. Această abordare ar oferi o perspectivă mai cuprinzătoare asupra stării de sănătate a pacientului, conducând la un diagnostic mai precis și mai personalizat. Cu toate acestea, din cunoștințele noastre, nu există în prezent un astfel de set de date disponibil.

În experimente viitoare, ne propunem să realizăm un studiu cuprinzător al utilizatorilor pentru a evalua utilitatea și eficacitatea sistemelor propuse în contexte

clinice. Studiul va implica radiologi, oncologi și alți profesioniști din domeniul sănătății, care vor analiza interpretabilitatea și fiabilitatea acestor sisteme. Feedbackul obținut va fi utilizat pentru îmbunătățirea și perfecționarea modelelor. În plus, vor fi explorate abordări post-hoc privind interpretabilitatea rețelelor neuronale, precum Grad-CAM [Sel+2019], SHAP [LL2017] și LIME [RSG2016], pentru a furniza informații mai detaliate despre procesul decizional al modelelor de învățare profundă.

Bibliografie

- [ADNI2024] The Image & Data Archive, *ADNI: Alzheimer's Disease Neuroimaging Initiative*, <https://adni.loni.usc.edu/about/> (accessed on 20/10/2024), Accessed on 20/10/2024.
- [An+2023] Qi An, Saifur Rahman, Jingwen Zhou și James Jin Kang, "A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges", în *Sensors* 23.9 (Apr. 2023), p. 4178, ISSN: 1424-8220, DOI: 10.3390/s23094178.
- [BAC2021] Adél Bajcsi, Anca Andreica și Camelia Chira, "Towards feature selection for digital mammogram classification", în *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021*, vol. 192, Procedia Computer Science, 2021, DOI: 10.1016/j.procs.2021.08.065.
- [BAC2024] Adél Bajcsi, Anca Andreica și Camelia Chira, "Significance of Training Images and Feature Extraction in Lesion Classification", în *Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART2024*, SciTePress, 2024, pp. 117–124, ISBN: 978-989-758-680-4, DOI: 10.5220/0012308900003636.
- [Baj2021] Adél Bajcsi, "Towards a Support System for Digital Mammogram Classification", în *Studia Universitatis Babeş-Bolyai Informatica* 66 (2 Dec. 2021), p. 19, ISSN: 2065-9601, DOI: 10.24193/subbi.2021.2.02.
- [BB2019] Wieland Brendel și Matthias Bethge, "Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet", în *arXiv* (2019), DOI: 10.48550/arXiv.1904.00760.
- [BC2023] Adél Bajcsi și Camelia Chira, "Textural and Shape Features for Lesion Classification in Mammogram Analysis", în *Hybrid Artificial Intelligent Systems HAIS2023 (Lecture Notes in Computer Science)*, vol. 14001, Springer Nature Switzerland, 2023, pp. 755–767, ISBN: 9783031407253, DOI: 10.1007/978-3-031-40725-3_64.
- [BC2025a] Adél Bajcsi și Camelia Chira, "Feature combination from different mammogram perspectives to improve lesion classification", în *Logic Journal of the IGPL* (2025), Accepted.

- [BC2025b] Adél Bajcsi și Camelia Chira, “fMRI Analysis for Alzheimer’s Disease Detection: Traditional vs. Deep Learning Models”, în *Artificial Intelligence in Medicine AIME2025*, Submitted., 2025.
- [BCA2023] Adél Bajcsi, Camelia Chira și Anca Andreica, “Extended Mammogram Classification From Textural Features”, în *Studia Universitatis Babeș-Bolyai Informatica* 67 (2 Feb. 2023), pp. 5–20, DOI: 10.24193/subbi.2022.2.01.
- [BCS2025] Adél Bajcsi, Camelia Chira și Annamária Szenkovits, “Evaluating ResNet-based Interpretable Models for Breast Lesion Classification”, în *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART2025*, SciTePress, 2025, pp. 288–295, ISBN: 978-989-758-737-5, DOI: 10.5220/0013121900003890.
- [Bur2019] Andriy Burkov, *The hundred-page machine learning book*, vol. 1, Andriy Burkov Quebec City, QC, Canada, 2019, ISBN: 9781999579517, URL: <https://books.google.hu/books?id=0jbxwQEACAAJ>.
- [Che+2018] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin și Jonathan K Su, “This looks like that: deep learning for interpretable image recognition”, în *Advances in neural information processing systems* 32 (Jun. 2018), DOI: 10.48550/arXiv.1806.10574.
- [CK2019] Ramzi Chaieb și Karim Kalti, “Feature subset selection for classification of malignant and benign breast masses in digital mammography”, în *Pattern Analysis and Applications* 22.3 (Aug. 2019), pp. 803–829, ISSN: 1433-755X, DOI: 10.1007/s10044-018-0760-x.
- [Dav2022] Lauren Evoy Davis, *Breast Masses: Cancerous Tumor or Benign Lump?*, <https://www.verywellhealth.com/breast-cancer-tumors-or-benign-masses-430277>, Published on 31/01/2022. Accessed on 20/10/2024., 2022.
- [EU 2023] EU Science Hub, Joint Research Centre, *Cancer cases and deaths on the rise in the EU*, https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/cancer-cases-and-deaths-rise-eu-2023-10-02_en, Published on 02/10/2023. Accessed on 20/10/2024., 2023.
- [Gat+2024] Anda Gata, Lajos Raduly, Livița Budișan, Adél Bajcsi, Teodora-Maria Ursu, Camelia Chira, Laura Dioșan, Ioana Berindan-Neagoe și Silviu Albu, “Machine Learning Model Predicts Postoperative Outcomes in Chronic Rhinosinusitis With Nasal Polyps”, în *Clinical Otolaryngology* 49.6 (Aug. 2024), pp. 776–784, ISSN: 1749-4486, DOI: 10.1111/coa.14208.
- [Has+2021] Noor Salah Hassan, Adnan Mohsin Abdulazeez, Diyar Qader Zeebaree și Dathar A. Hasan, “Medical Images Breast Cancer Segmentation Based on K-Means Clustering Algorithm: A Review”, în *Asian Journal of Research in Computer Science* (Mai 2021), pp. 23–38, ISSN: 2581-8260, DOI: 10.9734/ajrcos/2021/v9i130212.
- [He+2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren și Jian Sun, “Deep Residual Learning for Image Recognition”, în *Proceedings of the IEEE Conference on Com-*

- puter Vision and Pattern Recognition (CVPR), Iun. 2015, pp. 770–778, DOI: 10.1109/CVPR.2016.90.
- [Hea+1998] Michael Heath, Kevin Bowyer, Daniel Kopans, Philip Kegelmeyer, Richard Moore, Kyong Chang și S. Munishkumaran, “Current Status of the Digital Database for Screening Mammography”, în *Digital Mammography: Nijmegen, 1998*, Dordrecht: Springer Netherlands, 1998, pp. 457–460, ISBN: 978-94-011-5318-8, DOI: 10.1007/978-94-011-5318-8_{_}75.
- [Hea+2001] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore și Philip Kegelmeyer, “The digital database for screening mammography”, în *Proceedings of the Fifth International Workshop on Digital Mammography*, ed. de M.J. Yaffe, Medical Physics Publishing, 2001, pp. 212–218, ISBN: 1-930524-00-5.
- [Ian+2022] Stefania D. Iancu, Ramona G. Cozan, Andrei Stefancu, Maria David, Tudor Moisoiu, Cristiana Moroz-Dubenco, Adél Bajcsi, Camelia Chira, Anca Andreica, Loredana F. Leopold, Daniela Eniu, Adelina Staicu, Iulian Goidescu, Carmen Socaciu, Dan T. Eniu, Laura Diosan și Nicolae Leopold, “SERS liquid biopsy in breast cancer. What can we learn from SERS on serum and urine?”, în *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 273 (Mai 2022), p. 120992, ISSN: 1386-1425, DOI: 10.1016/j.saa.2022.120992.
- [Li+2017] Haixia Li, Xianjing Meng, Tingwen Wang, Yuchun Tang și Yilong Yin, “Breast masses in mammography classification with local contour features”, în *Bio-Medical Engineering OnLine* 16.1 (Apr. 2017), p. 44, ISSN: 1475-925X, DOI: 10.1186/s12938-017-0332-0.
- [LL2017] Scott M Lundberg și Su-In Lee, “A Unified Approach to Interpreting Model Predictions”, în *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017, DOI: 10.48550/arXiv.1705.07874.
- [MNB2012] I. K. Maitra, S. Nag și S. K. Bandyopadhyay, “Technique for preprocessing of digital mammogram”, în *Computer Methods and Programs in Biomedicine* 107.2 (2012), pp. 175–188, ISSN: 0169-2607, DOI: 10.1016/j.cmpb.2011.05.007.
- [Mor+2022] Cristiana Moroz-Dubenco, Adél Bajcsi, Anca Andreica și Camelia Chira, “An Unsupervised Threshold-based GrowCut Algorithm for Mammography Lesion Detection”, în *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022*, vol. 207, Elsevier BV, 2022, pp. 2096–2105, DOI: 10.1016/j.procs.2022.09.269.
- [Mor+2025] Cristiana Moroz-Dubenco, Adél Bajcsi, Anca Andreica și Camelia Chira, “Towards an interpretable breast cancer detection and diagnosis system”, în *Computers in Biology and Medicine* 185 (2025), p. 109520, ISSN: 0010-4825, DOI: 10.1016/j.combiomed.2024.109520.
- [Por+2024] Ábel Portik, Adél Bajcsi, Annamária Szekovits și Zalán Bodó, “Exploring the Impact of Backbone Architecture on Explainable CNNs’ Interpretability”, în *Acta Universitatis Sapientiae* 16 (1 2024), pp. 105–123, DOI: 10.2478/ausi-2024-0007.

- [Rah+2021] Amir Masoud Rahmani, Efat Yousefpoor, Mohammad Sadegh Yousefpoor, Zahid Mehmood, Amir Haider, Mehdi Hosseinzadeh și Rizwan Ali Naqvi, “Machine Learning (ML) in Medicine: Review, Applications, and Challenges”, în *Mathematics* 9.22 (Nov. 2021), p. 2970, ISSN: 2227-7390, DOI: 10.3390/math9222970.
- [RSG2016] Marco Tulio Ribeiro, Sameer Singh și Carlos Guestrin, ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier”, în *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Aug. 2016, pp. 1135–1144, DOI: 10.1145/2939672.2939778.
- [Sel+2019] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh și Dhruv Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”, în *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359, ISSN: 1573-1405, DOI: 10.1007/s11263-019-01228-7.
- [SGM2010] Fengxi Song, Zhongwei Guo și Dayong Mei, “Feature Selection Using Principal Component Analysis”, în *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization*, IEEE, Nov. 2010, pp. 27–30, DOI: 10.1109/icsem.2010.14.
- [Shr+2017] A. Shrivastava, A. Chaudhary, D. Kulshreshtha, V. Prakash Singh și R. Srivastava, “Automated digital mammogram segmentation using Dispersed Region Growing and Sliding Window Algorithm”, în *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Iun. 2017, pp. 366–370, DOI: 10.1109/ICIVC.2017.7984579.
- [Son+2022] Li Song, Xi Wang, Xiangyang Qu și Chao Lv, “Transcription Factor Specificity Protein 1 Regulates Inflammation and Fibrin Deposition in Nasal Polyps Via the Regulation of microRNA-125b and the Wnt/ -catenin Signaling Pathway”, în *Inflammation* 45.3 (Ian. 2022), pp. 1118–1132, ISSN: 1573-2576, DOI: 10.1007/s10753-021-01605-w.
- [SPD1994] J. Suckling, J. Parker și D.R. Dance, “The mammographic image analysis society digital mammogram database”, în *International Congress Series*, vol. 1069, Ian. 1994, pp. 375–378.
- [Tao+2018] Xiaoyao Tao, Fenghong Chen, Yueqi Sun, Shulian Wu, Haiyu Hong, Jianbo Shi și Rui Xu, “Prediction models for postoperative uncontrolled chronic rhinosinusitis in daily practice”, în *The Laryngoscope* 128.12 (Sept. 2018), pp. 2673–2680, ISSN: 1531-4995, DOI: 10.1002/lary.27267.
- [Zha+2017] Yu Zhang, Han Zhang, Xiaobo Chen, Seong-Whan Lee și Dinggang Shen, “Hybrid High-order Functional Connectivity Networks Using Resting-state Functional MRI for Mild Cognitive Impairment Diagnosis”, în *Scientific Reports* 7 (1 Iul. 2017), p. 6530, ISSN: 2045-2322, DOI: 10.1038/s41598-017-06509-0.

[Zha+2024] Shuping Zhao, Bob Zhang, Jian Yang, Jianhang Zhou și Yong Xu, “Linear discriminant analysis”, în *Nature Reviews Methods Primers* 4.1 (Sept. 2024), ISSN: 2662-8449, DOI: 10.1038/s43586-024-00346-y.