

UNIVERSITATEA BABEȘ BOLYAI, CLUJ NAPOCA, ROMÂNIA
FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ

Sistem interpretabil automat pentru detectia și diagnoza cancerului de sân

– Rezumatul tezei de doctorat –

Coordonator științific

Prof. Dr. Anca Andreica

Student doctorand

Cristiana Moroz-Dubenco

2025

Cuvinte cheie: Analiza mamografiilor, Segmentarea sânului, Clasificarea leziunilor, Detectie și diagnoză asistate de calculator, Învățare automată, Interpretabilitate

Cuprins

1	Introducere	2
1.1	Motivație	2
1.2	Obiective	3
1.3	Contribuții originale	4
2	Detecția și diagnoza leziunilor din mamografii	6
2.1	Pre-procesare	8
2.2	Segmentare	8
2.3	Extracția caracteristicilor	10
2.4	Selecția caracteristicilor	11
2.5	Clasificare	12
3	Un nou sistem de detecție și diagnoză a leziunilor din mamografii	14
3.1	Pre-procesare	14
3.2	Segmentare	15
3.2.1	Threshold-based GrowCut	15
3.2.2	Generarea automată a punctelor inițiale	15
3.2.2.1	Generarea punctelor inițiale de fundal	16
3.2.2.2	Generarea punctelor inițiale de prim-plan	16
3.3	Extracția caracteristicilor	17
3.4	Selecția caracteristicilor	17
3.5	Clasificare	17
4	Evaluarea sistemului de detecție și diagnoză a leziunilor din mamografii propus	18
4.1	Detecție	18
4.1.1	Mini-MIAS	18
4.1.2	Mini-DDSM	19
4.2	Diagnoză	19
4.2.1	Mini-MIAS	20
4.2.2	Mini-DDSM	21
4.2.3	RDBMC	22
4.3	Un punct de vedere clinic	22
5	Către un sistem multi-modal de detecție și diagnoză a cancerului de sân	24
5.1	Bio-fluide	24
5.1.1	Rezultate experimentale	25
5.2	Rapoarte medicale	25
5.2.1	Rezultate experimentale	26
6	Concluzii și extinderi	28
	Bibliography	31

Rezumat

Conform Organizației Mondiale a Sănătății, cancerul de sân devine fatal doar dacă se răspândește în organism. Prin urmare, screening-ul regulat este esențial. Deși mamografia este cea mai utilizată metodă de depistare, interpretarea acesteia poate fi dificilă și consumatoare de timp. Din acest motiv, sistemele de detecție și diagnoză asistate de calculator sunt utilizate din ce în ce mai frecvent ca o a doua opinie. Totuși, pentru ca medicii să aibă încredere în astfel de sisteme, este esențial să înțeleagă modul în care sunt luate deciziile. Propunem un sistem automat și interpretabil pentru detecția și diagnoza cancerului de sân, bazat pe cinci etape. După o preprocesare robustă și o segmentare nesupravegheată, analizăm cinci tehnici de extragere a caracteristicilor, atât bazate pe puncte de interes, cât și pe textură, și evaluăm patru metode de selecție a caracteristicilor. Pentru a facilita interpretarea, utilizăm algoritmi clasici de învățare automată pentru clasificarea benign/malign și testăm opt metode diferite. Sistemul propus atinge o acuratețe între 95% și 97% atunci când este testat pe imagini din seturile de date mini-MIAS, mini-DDSM și RDBMC, oferind utilizatorilor posibilitatea de a analiza fiecare etapă a procesului. În plus, facem primii pași către un sistem multi-modal, explorând posibilitatea diagnosticării cancerului de sân pe baza bio-fluidelor și a rapoartelor mamografice.

Lista publicațiilor

Standardele de evaluare folosite sunt cele valide din 1 octombrie 2018. Lista conferințelor, precum și categoria lor, este bazată pe clasificarea internațională CORE ¹. Lista jurnalelor este cea folosită de UEFSCDI pentru premierea articolelor publicate în jurnale științifice internaționale ².

1. ***Comparison of Gradient-Based Edge Detectors Applied on Mammograms***, **Cristiana Moroz-Dubenco [1]**, publicat în *Studia Universitatis Babeș-Bolyai Informatica* (2021): pp. 5-18, rang D: 1 punct;
2. ***Mammography Lesion Detection Using an Improved GrowCut Algorithm***, **Cristiana Moroz-Dubenco, Laura Dioșan, Anca Andreica [2]**, prezentat la *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (2021), rang B: 4 puncte;
3. ***SERS liquid biopsy in breast cancer. What can we learn from SERS on serum and urine?***, **Ștefania D. Iancu, Ramona G. Cozan, Andrei Ștefancu, Maria David, Tudor Moişoiu, Cristiana Moroz-Dubenco, Adél Bajcsi, Camelia Chira, Anca Andreica, Loredana F. Leopold, Daniela Eniu, Adelina Staicu, Iulian Goidescu, Carmen Socaciu, Dan T. Eniu, Laura Dioșan, Nicolae Leopold [3]**, publicat în *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* (2022): 120992, rang B: 0.267 puncte;
4. ***An Unsupervised Threshold-based GrowCut Algorithm for Mammography Lesion Detection***, **Cristiana Moroz-Dubenco, Adél Bajcsi, Anca Andreica, Camelia Chira [4]**, prezentat la *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (2022), rang B: 2 puncte;
5. ***Towards an Unsupervised GrowCut Algorithm for Mammography Segmentation***, **Cristiana Moroz-Dubenco, Laura Dioșan, Anca Andreica [5]**, prezentat la *International Conference on Computer Vision Systems* (2023) și publicat în *Lecture Notes in Computer Science* (2023): pp. 102-111, rang C: 2 puncte;
6. ***Generalizing an Improved GrowCut Algorithm for Mammography Lesion Detection***, **Cristiana Moroz-Dubenco, Laura Dioșan, Anca Andreica [6]**, prezentat la *International Conference on Hybrid Artificial Intelligence Systems* (2023) și publicat în *Lecture Notes in Computer Science* (2023): pp. 709-720, rang C: 2 puncte;

¹<http://portal.core.edu.au/conf-ranks/>

²<https://uefscdi.ro/premierea-rezultatelor-cercetarii-articole>

7. *Linear Discriminant Analysis Tumour Classification for Unsupervised Segmented Mammographies*, Cristiana Moroz-Dubenco, Anca Andreica [7], prezentat la *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (2023), rang B: 4 puncte;
8. *Machine-Learning-Based Approaches for Multi-Level Sentiment Analysis of Romanian Reviews*, Anamaria Briciu, Alina-Delia Călin, Diana-Lucia Miholca, Cristiana Moroz-Dubenco, Vladia Petrascu, George Dascălu [8], publicat în *Mathematics* 12, no. 3 (2024): p.456, rang C: 0.5 puncte;
9. *Towards an Interpretable Breast Cancer Detection and Diagnosis System*, Cristiana Moroz-Dubenco, Adél Bajcsi, Anca Andreica, Camelia Chira [9], publicat în *Computers in Biology and Medicine* 185 (2025): 109520, rang A: 4 puncte.

Scorul total al publicațiilor: 19.767

Capitolul 1

Introducere

1.1 Motivație

Conform Organizației Mondiale a Sănătății, 2.3 milioane de femei au fost diagnosticate cu cancer de sân în 2020, iar alte 7.8 milioane de femei au fost diagnosticate între 2015 și 2020. Această boală apare în țesutul glandular al sânelor, dar poate invada țesuturile învecinate și se poate răspândi către organe apropiate sau îndepărtate. Femeile care mor din cauza cancerului de sân, mor din cauza metastazelor. Prin urmare, dacă boala este descoperită din timp și tratată corespunzător, creșterea și răspândirea sa pot fi prevenite, salvând viața pacientului.

Cea mai eficientă metodă de detectare a cancerului de sân în stadii incipiente este realizarea regulată a examenelor de screening. Mamografia este una dintre cele mai utilizate metode de screening, constând într-o imagine cu raze X a sânelor, analizată de medici pentru identificarea semnelor timpurii ale cancerului de sân. În Europa, protocoalele de screening includ, de obicei, interpretarea dublu-orb a imaginilor mamografice. Din acest motiv, tot mai mulți radiologi utilizează sisteme de detecție și diagnoză asistate de calculator (CAdE/CAdx) ca a doua opinie, pentru a reduce volumul de muncă și a îmbunătăți acuratețea diagnosticării.

Mai mult decât atât, sistemele automate pot ajuta la analiza rapoartelor mamografice anterioare pentru a evalua schimbările apărute într-o leziune între mamografii consecutive. De asemenea, acestea pot integra imagistica medicală cu informații suplimentare, precum bio-markeri din bio-fluide, pentru a îmbunătăți precizia diagnosticului.

Inteligența artificială este utilizată pe scară largă în astfel de sisteme, având potențialul de a asista radiologii, crescând astfel acuratețea diagnosticului, reducând volumul de muncă și minimizând riscul de supra-diagnosticare. Totuși, în domenii critice, cum este medicina, unde viețile pacienților sunt în joc, interpretabilitatea acestor sisteme este crucială. Așadar, ne concentrăm pe modelele interpretabile.

1.2 Obiective

Obiectivele acestei teze pot fi rezumate astfel:

1. Construirea unui sistem interpretabil automat pentru detecția și diagnoza cancerului de sân pe baza mamografiilor
 - (a) identificarea unei tehnici de pre-procesare robuste, capabile să elimine informațiile nedorite și să ofere explicații vizuale;
 - (b) îmbunătățirea unei tehnici existente de segmentare interpretabilă, eliminând necesitatea intervenției umane, crescând performanța și reducând timpul de calcul;
 - (c) determinarea celei mai potrivite metode de extragere a caracteristicilor, care să funcționeze în tandem cu noua metodă de segmentare și să genereze caracteristici ușor de înțeles;
 - (d) alegerea unei metode de selecție a caracteristicilor pentru reducerea dimensionalității într-un mod transparent;
 - (e) selectarea unei tehnici de clasificare care să echilibreze cât mai bine raportul dintre interpretabilitate și performanță predictivă.
2. Analiza altor tipuri de date medicale care pot fi utilizate împreună cu mamografiile pentru creșterea confidenței diagnosticului
 - (a) identificarea celui mai potrivit clasificator pentru analiza SERS a bio-fluidelor;
 - (b) alegerea metodelor de reprezentare și clasificare potrivite pentru extragerea unui diagnostic din rapoarte medicale în două limbi (română și engleză).

În plus, această teză își propune și următoarele obiective secundare:

- analiza impactului caracteristicilor setului de date asupra procesului de detecție și diagnoză,
- studiul influenței numărului de puncte de pornire asupra unei metode de segmentare bazată pe regiuni,
- examinarea efectului densității sânului asupra performanței sistemului,
- elaborarea unei metodologii de evaluare a interpretabilității unui sistem,
- crearea unui set de date care să conțină și descrierile textuale ale mamografiilor împreună cu imaginile,

- investigarea influenței limbii asupra rezultatelor clasificării în cazul rapoartelor medicale,
- evaluarea relației dintre vârsta pacientului și diagnostic.

1.3 Contribuții originale

Pentru atingerea obiectivelor propuse, contribuțiile originale ale acestei teze sunt următoarele:

1. *Algoritmul Threshold-based GrowCut (TbGC)*, o versiune îmbunătățită a metodei de segmentare semi-supervizată GrowCut. Acest algoritm introduce o constrângere privind numărul de iterații, reducând astfel semnificativ timpul de convergență. De asemenea, include un mecanism de prag care rafinează procesul de etichetare, îmbunătățind acuratețea segmentării și conservând o detecție robustă, în ciuda reducerii timpului computațional.
2. *Trei metode pentru generarea automată a punctelor de pornire pentru fundal*, utilizabile atât cu TbGC, cât și cu alte tehnici de segmentare bazate pe regiuni. Aceste metode au scopul de a reduce dependența de utilizator, identificând automat regiuni fiabile pentru fundal, îmbunătățind consistența segmentării și acuratețea, în special în cazurile în care selecția manuală a punctelor este dificilă sau predispusă la variabilitate.
3. *Două metode pentru generarea automată a punctelor de pornire pentru prim-plan*, care pot fi, de asemenea, utilizate cu orice tehnică bazată pe regiuni. Aceste metode elimină complet necesitatea intervenției umane, permițând identificarea automată a leziunilor suspecte. Asigurând o inițializare mai consistentă și mai obiectivă, aceste metode îmbunătățesc acuratețea segmentării și reduc variabilitatea introdusă prin intervenția umană.
4. *Integrarea TbGC (cu puncte de pornire generate automat) într-un sistem CAD complet*. Prin combinarea capabilităților crescute de segmentare ale algoritmului TbGC cu metodele propuse pentru generarea automată a punctelor de pornire, sistemul atinge robustețe, eficiență și consistență crescute în detecția și delimitarea leziunilor, permițând o identificare complet automată a regiunii de interes. Încorporarea acestei abordări de segmentare într-un sistem pentru detecția și diagnoza cancerului de sân oferă un canal simplificat pentru analiza mamografică, în care fiecare pas poate fi vizualizat și examinat.
5. *Romanian Dense Breast Mammography Collection*, un nou set de date menit să sprijine cercetarea asupra detecției cancerului de sân în țesut mamar dens. Acesta conține imagini mamografice 2D și 3D, împreună cu rapoarte histopatologice și evaluări radiologice, devenind astfel o resursă

de valoare pentru dezvoltarea și evaluarea unor sisteme CAD avansate. Includerea adnotărilor clinice cuprinzătoare și a meta-datelor pacientului permite studii aprofundate asupra relației dintre caracteristicile imagistice și rezultatele diagnozei. Din cunoștințele noastre, nu există nicio altă colecție disponibilă public care să se concentreze exclusiv pe pacienți cu țesut mamar dens, astfel că RDBMC abordează o lacună critică în cercetarea imagistică a cancerului de sân.

6. *Evaluarea sistemului CAD propus pe noul set de date RDBMC*, care conține doar pacienți cu sâni denși. Deoarece țesutul mamar dens reprezintă o provocare semnificativă pentru detectarea leziunilor din cauză contrastului redus și probabilității crescute de fals pozitive, această evaluare oferă informații despre cât de bine se generalizează sistemul la scenarii complexe din lumea reală. Rezultatele sunt comparate cu interpretări ale radiologilor și cu metode automate existente, evidențiind punctele forte ale sistemului în gestionarea cazurilor dificile, îmbunătățirea ratelor de detectare și reducerea diagnosticelor greșite.
7. *Dezvoltarea unei metodologii pentru evaluarea interpretabilității unui sistem CAD*, bazată pe un cadru de evaluare pas cu pas, care cuantifică transparența deciziilor sistemului pentru experții umani. Metodologia evaluează fiecare etapă a sistemului CAD, de la pre-procesare până la clasificarea finală, atribuind scoruri de interpretabilitate pe baza feedback-ului experților. Această abordare structurată oferă un mijloc de a compara diferite sisteme CAD în ceea ce privește interpretabilitatea, facilitând dezvoltarea unor instrumente de diagnosticare mai transparente și de încredere, bazate pe IA.
8. *Diagnosticarea cancerului de sân și evaluarea BI-RADS pe baza rapoartelor mamografice textuale în limba română*, împreună cu o analiză asupra influenței limbajului, vârstei și reprezentării asupra rezultatelor clasificării. Abordarea propusă prezice malignitatea și detectează scorul BI-RADS, gestionând terminologia medicală și variațiile stilurilor de raportare. În plus, analiza efectuată pentru a determina modul în care demografia pacientului și nuanțele lingvistice afectează acuratețea modelului asigură robustețea sistemului. Această contribuție abordează sub-reprezentarea seturilor de date medicale în alte limbi decât engleză și evidențiază provocările în diagnosticarea automată bazată pe rapoarte cu text liber.
9. *Analiza diagnosticării automate a cancerului de sân din bio-fluide*. Prin acest studiu, ne propunem să evaluăm performanța serului și a urinei în biopsia lichidă a cancerului de sân folosind analiza SERS fără etichete. Metodologia propusă ar putea fi integrată în continuare într-un sistem CAD multi-modal pentru cancerul de sân, crescând acuratețea diagnosticului.

Capitolul 2

Detecția și diagnoza leziunilor din mamografii

Cancerul de sân este cea mai frecvent diagnosticată formă de cancer la femei și a doua cauză principală de deces prin cancer, după cancerul pulmonar [10]. În cele mai multe cazuri, cancerul de sân poate fi tratat cu succes prin chirurgie, radioterapie și chimioterapie, dacă este detectat într-un stadiu incipient. În schimb, dacă este diagnosticat târziu, probabilitatea de metastază crește semnificativ, ceea ce poate duce la un rezultat fatal. Prin urmare, screening-ul regulat este esențial pentru detectarea timpurie și creșterea șanselor de succes ale tratamentului. Una dintre principalele metode utilizate în acest scop este mamografia, o tehnică de imagistică medicală care permite examinarea detaliată a țesutului mamar.

Interpretarea mamografiilor are două obiective principale: detecția leziunilor și diagnoza. Aceste două procese distincte, dar interconectate, sunt esențiale pentru screening-ul și evaluarea cancerului de sân. Detecția leziunilor se referă la identificarea inițială a regiunilor anormale în cadrul unei imagini mamografice, având ca scop evidențierea zonelor care pot necesita o evaluare suplimentară. Diagnoza, pe de altă parte, implică determinarea naturii leziunii detectate – dacă este benignă sau malignă. În timp ce detecția se concentrează pe asigurarea faptului că nicio leziune potențial dăunătoare nu este trecută cu vederea, diagnosticul este crucial pentru minimizarea biopsiilor inutile și optimizarea managementului pacientului.

Mamografiile sunt obținute prin utilizarea unui sistem de raze X cu doze mici [11], caracterizat printr-o serie de caracteristici speciale, care pot ridica provocări în detectarea leziunilor mamare [12]:

- sunt imagini în tonuri de gri;
- pot conține contururi neclare;

- prezintă zgomot gaussian;
- pot conține un alt tip de zgomot de fundal: artefacte, precum etichetele medicale;
- pot avea calitate scăzută, contrast scăzut și iluminare slabă, în funcție de aparatura folosită;
- țesutul dens, ca și țesutul conjunctiv sau glandular, apare mai luminos, ceea ce face dificilă diferențierea de tumori, care sunt, de asemenea, alcătuite din țesut dens;
- o masă poate fi nu numai o tumoră, ci și un chist sau un fibroadenom;
- conturul unei mase, mai ales a uneia maligne, nu este întotdeauna bine definit.

În plus, chiar dacă detecția este efectuată cu succes, pot apărea provocări suplimentare atunci când vine vorba de diagnostic, din cauza leziunilor care pot varia semnificativ în dimensiune, formă, densitate și textură [13]. Leziunile maligne pot părea subtile, în timp ce anumite anomalii benigne imită cancerul [14]. Ratele ridicate de fals pozitiv și fals negativ complică și mai mult procesul, ducând la biopsii inutile sau la întârzierea tratamentului [15].

Sistemele de detecție și diagnoză asistate de calculator sunt menite să ajute medicii în interpretarea imaginilor medicale, putând fi aplicate, deci, și pentru mamografii, oferind o opinie secundară judecății lor [11]. Aceste sisteme procesează imaginile căutând secțiuni și structuri remarcabile, bazându-se pe recunoașterea modelelor foarte complexe. Imaginile medicale și, în particular, mamografiile, sunt furnizate sistemului și analizate în mai mulți pași:

1. *Pre-procesare* – îmbunătățirea calității imaginii prin reducerea zgomotului, normalizarea contrastului și eliminarea artefactelor pentru a îmbunătăți vizibilitatea leziunii;
2. *Segmentare* – identificarea și izolarea regiunilor de interes (ROI) care pot conține anomalii, cum ar fi mase sau microcalcificări;
3. *Extracție de caracteristici* – analiza regiunilor de interes pentru a extrage caracteristici relevante, inclusiv forma, textura și densitatea, care ajută la diferențierea țesutului normal de cel anormal;
4. *Selecție de caracteristici* – selectarea celor mai semnificative caracteristici;
5. *Clasificare* – atribuirea leziunilor detectate unei categorii specifice – benigne sau maligne – pe baza caracteristicilor extrase și a tiparelor învățate.

2.1 Pre-procesare

Pre-procesarea este o etapă critică în sistemele de detecție și diagnoză asistate de computer, concepută pentru a îmbunătăți calitatea imaginilor mamografice pentru o analiză mai precisă și mai fiabilă. Această etapă implică o serie de tehnici de procesare de imagini menite să îmbunătățească consistența, să reducă zgomotului și să îmbunătățească vizibilitatea potențialelor anomalii.

Procesul începe cu normalizarea, care compensează variațiile parametrilor de achiziție a imaginii, cum ar fi nivelurile de expunere, diferențele de contrast și setările scannerului. Acest pas este esențial pentru standardizarea caracteristicilor imaginii, asigurând uniformitatea în diferite mamografii și facilitând o analiză consecventă.

După normalizare, se aplică tehnici de reducere a zgomotului, pentru a elimina artefactele care ar putea ascunde detalii importante sau ar putea interfera cu extragerea caracteristicilor. Pot fi utilizate diferite tehnici de eliminare a zgomotului, cum ar fi filtrarea mediană, netezirea gaussiană sau transformarea wavelet, în funcție de tipul și intensitatea zgomotului prezent în imagine.

Odată ce zgomotul este redus la minimum, tehnicile de îmbunătățire a imaginii sunt folosite pentru a mări claritatea și contrastul imaginilor mamografice, făcând mai anomaliile subtile mai ușor de distins. Metode precum întinderea contrastului, egalizarea histogramei și egalizarea histogramei adaptive sunt utilizate în mod obișnuit pentru a amplifica detalii vizuale importante, în special în țesutul mamar dens, unde leziunile pot fi mai dificil de detectat.

Prin aplicarea acestor tehnici de pre-procesare, sistemele CAD pot optimiza calitatea imaginii, asigurându-se că pașii ulteriori – cum ar fi segmentarea, extragerea caracteristicilor și clasificarea – sunt efectuați cu acuratețe și fiabilitate mai ridicate, îmbunătățind în cele din urmă detecția și diagnoza cancerului de sân. În funcție de caracteristicile imaginilor și de rezultatul dorit, ordinea în care aceste tehnici sunt aplicate poate fi ajustată sau anumiți pași pot fi omiși.

2.2 Segmentare

Segmentarea este procesul de împărțire a unei imagini în seturi de pixeli, denumite regiuni de interes, conform unor criterii specifice. Aceasta este utilizată pentru a recunoaște, extrage sau identifica obiecte în imagini. Scopul segmentării este de a simplifica reprezentarea imaginii astfel încât să devină mai ușor de analizat sau procesat. Cu alte cuvinte, segmentarea imaginii presupune atribuirea de etichete pixelilor astfel încât toți pixelii care aparțin unei anumite clase să împărtășească caracteristici comune.

Segmentarea imaginilor este utilizată într-o varietate de domenii, inclusiv în viziunea computerizată, planificarea intervențiilor chirurgicale, sistemele de control al traficului, recunoașterea fețelor și

detectarea luminilor de frână. În această teză, ne concentrăm pe segmentarea imaginilor medicale, în special pe detectarea tumorilor în mamografii. Identificarea regiunii de interes este esențială pentru a putea analiza și clasifica ulterior leziunile ca fiind benigne sau maligne.

Segmentarea joacă un rol esențial în interpretarea mamografiilor, deoarece forma și dimensiunea unei mase sunt factori importanți în diferențierea dintre masele benigne și maligne [16]. Totodată, dimensiunea unei tumori influențează decizia privind posibilitatea efectuării unei intervenții chirurgicale. Cu toate acestea, datorită naturii țesutului mamar, segmentarea mamografiilor poate fi o sarcină dificilă. Tumorile pot avea forme și dimensiuni variabile, pot avea densități și localizări diferite, iar sânul poate conține și alte anomalii, cum ar fi chisturi, care pot fi confundate cu tumorile. În plus, conturul unei mase, mai ales dacă este malignă, nu este întotdeauna bine definit. Dacă luăm în considerare posibilitatea de a avea imagini cu contrast scăzut, calitate scăzută a imaginii, niveluri ridicate de zgomot sau iluminare slabă [17], putem afirma că segmentarea mamografiei s-ar putea dovedi o provocare.

Printre tehnicile utilizate pentru segmentarea mamografiilor, enumerăm:

1. *Thresholding* – convertește o imagine în tonuri de gri într-una binară;
2. *Clustering* – împarte pixelii în grupuri astfel încât cei dintr-un grup să fie mai asemănători între ei decât cu cei din alte grupuri;
3. *Histogram based* – utilizează histograma imaginii pentru a detecta zonele de interes;
4. *Region growing* – compară un pixel cu vecinii săi și, dacă îndeplinește un criteriu de similitudine, îl adaugă la regiunea din care face parte vecinul său.

Dintre aceste metode, segmentarea bazată pe regiuni este considerată mai interpretabilă decât celelalte, datorită capacității sale de a păstra coerența spațială și de a oferi o reprezentare semnificativă a obiectelor. Spre deosebire de abordările de tip prag și histogramă, care se bazează exclusiv pe valorile intensității pixelilor și pot întâmpina probleme din cauza iluminării sau a zgomotului neuniform, metodele bazate pe regiune agregă pixelii în baza caracteristicilor comune, asigurându-se că regiunile segmentate corespund structurilor coerente din imagine [18]. Această consistență spațială este deosebit de valoroasă în aplicații precum imagistica medicală, unde structurile anatomice trebuie delimitate cu precizie pentru diagnostic și planificare a tratamentului. În plus, tehnicile de creștere a regiunii grupează pixelii într-un mod care se aliniază direct cu obiectele din lumea reală, făcând rezultatele segmentării mai intuitive și mai ușor de interpretat. În comparație cu metodele de clustering, care segmentează imaginile pe baza similitudinii statistice, dar pot ignora relațiile spațiale, segmentarea

bazată pe regiuni produce limite mai robuste, luând în considerare interacțiunile pixelilor locali [19]. Aceste avantaje fac din segmentarea bazată pe regiuni o alegere preferată în aplicațiile care necesită interpretabilitate și precizie ridicate, cum ar fi imagistica biomedicală și teledetecția. Prin urmare, în conformitate cu obiectivul nostru de interpretabilitate, ne concentrăm pe segmentarea bazată pe regiuni.

Tehnicile de segmentare bazate pe creșterea regiunilor necesită selectarea unor puncte inițiale – un număr de pixeli etichetați înainte de începutul algoritmului. Selectate pe baza criteriilor utilizatorului, locațiile acestor puncte sunt considerate regiunile inițiale. Apoi, regiunile sunt crescute la puncte adiacente, în funcție de similitudinea dintre un pixel și pixelii dintr-o anumită regiune.

2.3 Extracția caracteristicilor

Extracția caracteristicilor implică identificarea și extragerea caracteristicilor relevante din imaginile mamografice segmentate, care sunt ulterior utilizate pentru clasificarea și identificarea potențialelor tumori maligne. Extragerea eficientă a caracteristicilor poate îmbunătăți acuratețea și fiabilitatea diagnosticului cancerului de sân, conducând în cele din urmă la rezultate mai bune pentru pacient.

În sistemele CAD pentru cancerul de sân, extracția caracteristicilor este de obicei efectuată folosind tehnici de analiză a imaginilor care pot identifica diferite proprietăți ale potențialelor leziuni. Aceste caracteristici pot include, printre altele, textura, forma și intensitatea. Alegerea tehnicii de extracție a caracteristicilor depinde de cerințele specifice ale sistemului CAD și de caracteristicile leziunilor cancerului de sân analizate.

Analiza texturală este o tehnică comună de extracție a caracteristicilor utilizată în sistemele CAD pentru cancerul de sân. Caracteristicile texturale se bazează pe distribuția spațială a valorilor de intensitate din imaginile mamografice și pot fi extrase folosind tehnici precum analiza Gray-Level Co-occurrence Matrix (GLCM) și analiza Local Binary Patterns (LBP).

Extracția caracteristicilor bazate pe puncte cheie joacă un rol important în diagnosticarea cancerului de sân din mamografii, permițând identificarea modelelor distinctive în imagini. Acest proces implică detectarea și descrierea caracteristicilor localizate, cum ar fi marginile, texturile sau variațiile de intensitate, care sunt esențiale pentru caracterizarea regiunilor suspecte. Aceste tehnici, printre care se numără Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF) și Histogram of Oriented Gradients (HOG), îmbunătățesc capacitatea sistemului de a diferenția între structurile normale și anormale ale țesuturilor. Prin captarea caracteristicilor invariante și robuste, extracția punctelor cheie ajută la îmbunătățirea detecției leziunilor, clasificării și acurateței diagnosti-

cului general.

Analiza formei este o altă tehnică importantă de extracție a caracteristicilor în sistemele CAD pentru cancerul de sân. Caracteristicile formei se bazează pe caracteristicile geometrice ale potențialelor leziuni maligne și pot fi extrase folosind tehnici precum analiza contururilor sau analiza morfologică. Analiza contururilor implică măsurarea contururilor leziunilor potențiale, în timp ce analiza morfologică implică analiza mărimii, formei și locației leziunilor în raport cu țesutul din jur.

Măsurile de intensitate sunt, de asemenea, utilizate în mod obișnuit drept caracteristici în diagnosticul cancerului de sân. Aceste caracteristici se bazează pe luminozitatea și contrastul imaginii mamografice și pot fi extrase folosind tehnici precum analiza histogramei și analiza wavelet. Analiza histogramei implică măsurarea distribuției valorilor intensității pixelilor în imaginea mamografică, în timp ce analiza wavelet implică descompunerea imaginii în mai multe benzi de frecvență.

Metodele de extracție a caracteristicilor bazate pe textură și pe puncte cheie sunt preferate față de abordările bazate pe formă și intensitate în sistemele CAD pentru detecția și diagnoza cancerului de sân, deoarece captează modele complexe care ar putea să nu fie ușor de perceput de ochiul uman. În timp ce radiologii pot evalua direct forma și intensitatea din imaginile mamografice, textura și punctele cheie oferă informații suplimentare, cuantificabile, care îmbunătățesc analiza automată și sprijină luarea deciziilor într-o manieră mai informată.

2.4 Selecția caracteristicilor

Selecția caracteristicilor are ca scop identificarea celor mai relevante caracteristici din setul de caracteristici extrase din imaginile mamografice segmentate. Caracteristicile selectate sunt utilizate ulterior pentru clasificarea și identificarea potențialelor afecțiuni maligne, care pot ajuta radiologii în diagnosticarea cu acuratețe a cancerului de sân. Tehnicile de selecție a caracteristicilor pot fi clasificate pe scară largă în metode de filtrare, metode wrapper și metode încorporate.

Metodele de filtrare implică evaluarea fiecărei caracteristici în mod independent și clasarea pe baza puterii discriminatorii folosind măsuri statistice, cum ar fi teste t , ANOVA sau informații reciproce. Caracteristicile de top sunt apoi selectate pentru a fi utilizate în etapa de clasificare ulterioară. Metodele de filtrare sunt eficiente din punct de vedere computațional, dar nu iau în considerare interacțiunea dintre caracteristici.

Metodele wrapper evaluează subseturi de caracteristici prin antrenarea și testarea unui clasificator folosind diferite subseturi de caracteristici. Aceste metode evaluează performanța clasificatorului pentru fiecare subset de caracteristici și selectează subsetul care produce cea mai bună acuratețe de

clasificare. Metodele wrapper pot lua în considerare mai bine interacțiunea dintre caracteristici, dar pot fi costisitoare din punct de vedere computațional.

Metodele încorporate încorporează selecția caracteristicilor ca parte a procesului de formare a clasificatorului. Cea mai comună metodă încorporată este regularizarea, care implică adăugarea unui termen de penalizare la funcția obiectiv a clasificatorului pentru a descuraja utilizarea caracteristicilor irelevante.

Metodele de reducere a dimensionalității, cum ar fi Principal Component Analysis (PCA), urmăresc să reducă numărul de caracteristici dintr-un set de date, păstrând în același timp cât mai multe informații relevante posibil. Deși aceste tehnici nu efectuează selecția caracteristicilor în sensul tradițional, ele reduc eficient dimensionalitatea prin selectarea celor mai informative componente, îmbunătățind astfel generalizarea modelului. Ele sunt de obicei preferate în sistemele CAD, deoarece transformă spațiul original al caracteristicilor într-o reprezentare dimensională inferioară care păstrează cele mai relevante informații, în timp ce metodele clasice de selecție a caracteristicilor îndepărtează doar caracteristicile mai puțin importante, pierzând informații potențial valoroase pentru diagnostic [20].

Pe lângă tehnicile deja enumerate, selecția caracteristicilor poate fi ghidată și de cunoștințe despre domeniu, cum ar fi caracteristicile leziunilor cancerului de sân și caracteristicile despre care se știe că au fost relevante în studii anterioare.

2.5 Clasificare

Etapa de clasificare este o componentă crucială a oricărui sistem automat de diagnoză asistată de computer. Aceasta implică utilizarea caracteristicilor selectate din mamografiile segmentate pentru a clasifica anomaliile ca fiind benigne sau maligne.

Diferiți algoritmi de învățare automată, atât supervizați, cât și nesupervizați, pot fi utilizați pentru clasificarea în sistemele CAD pentru cancerul de sân. Metodele supravegheate necesită date de antrenament etichetate pentru a antrena modelul de clasificare, în timp ce metodele nesupravegheate încearcă să descopere tipare în spațiul caracteristicilor.

Metodele de clasificare supravegheate includ Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN) și Naive Bayes (NB), printre altele. Metodele de clasificare nesupravegheate includ tehnici de clustering, cum ar fi K-Means, clustering ierarhic și modele de amestec Gaussian. Aceste metode încearcă să grupeze puncte de date similare împreună pe baza reprezentărilor caracteristicilor lor. Metodele nesupravegheate pot fi utilizate pentru a identifica subtipuri de cancer

de sân și pot ajuta la descoperirea unor modele necunoscute anterior în date.

Alegerea algoritmului de clasificare depinde de cerințele specifice ale sistemului CAD și de caracteristicile leziunilor de cancer de sân analizate. Evaluarea performanței modelului de clasificare se face de obicei folosind metrici precum acuratețea, sensibilitatea, specificitatea și aria sub curba caracteristică de funcționare a receptorului (AUC-ROC).

Pe lângă clasificarea binară (benign vs. malign), sistemele CAD pot îndeplini și sarcini de clasificare mai nuanțate, cum ar fi identificarea subtipului specific de cancer de sân sau prezicerea probabilității de malignitate. Aceste sarcini necesită modele de clasificare mai complexe și pot necesita caracteristici sau modalități de imagistică suplimentare.

Capitolul 3

Un nou sistem de detecție și diagnoză a leziunilor din mamografii

Propunem un sistem complet, pe deplin automat, de detecție și diagnoză asistate de computer pentru analiza mamografică. Sistemul primește o mamografie anormală (o mamografie care conține o leziune) la intrare și emite o valoare binară, însemnând că leziunea este malignă sau benignă. Pentru aceasta, mamografia este mai întâi pre-procesată pentru a elimina orice informație redundantă. Imaginea rezultată este apoi segmentată și sunt extrase caracteristici, din care sunt selectate cele mai relevante. În cele din urmă, aceste caracteristici servesc ca intrare pentru un clasificator, care produce diagnosticul.

3.1 Pre-procesare

Mamografiile sunt imagini cu raze X ale țesutului mamar. Aceste imagini au de obicei o calitate scăzută, iar intensitatea lor poate varia în funcție de mașinile utilizate. Prin urmare, obiectivul pre-procesării este de a îmbunătăți mamografiile și de a le pregăti pentru segmentare.

Folosim metoda de preprocesare propusă de Bajcsi et al. [21] și generalizată în [22]. Aceasta constă în îmbunătățirea imaginii și îndepărtarea artefactelor externe și interne. Prin artefacte externe ne referim la informații din afara sânelui, iar prin artefacte interne ne referim la informații care sunt incluse în masca sânelui. Pentru a îmbunătăți imaginea, sunt folosite deschiderea morfologică (pentru reducerea zgomotului) și egalizarea histogrammei (pentru accentuarea conturului).

3.2 Segmentare

3.2.1 Threshold-based GrowCut

Abordarea noastră, algoritmul Threshold-based GrowCut, este o îmbunătățire a algoritmului GrowCut [23], o tehnică de segmentare semi-supervizată bazată pe regiuni. Aceasta este menită să realizeze segmentarea cu mai multe etichete folosind un automat celular, imaginea fiind spațiul celulelor, iar pixelii fiind celulele. Un pixel p din imaginea P este caracterizat de un triplet format din eticheta l_p , tăria θ_p – certitudinea că pixelul aparține clasei l_p – și vectorul caracteristic $\vec{C}_p$. Începe cu un număr de pixeli etichetați de utilizator, cărora li se atribuie o tărie de 1, și iterează în spațiul celulelor, actualizându-și etichetele și tăriile până când nicio celulă nu mai este actualizată în timpul unei iterații, ceea ce înseamnă că automatul a converș la o stare stabilă.

Având în vedere că algoritmul se oprește numai atunci când automatul converge, timpul de calcul poate crește rapid. Prin urmare, prin metoda TbGC, ne propunem să reducem timpul de calcul, menținând în același timp un nivel ridicat de acuratețe. Ca mijloc de atingere a acestui obiectiv, metodei originale i-au fost aduse următoarele modificări:

- modificarea regula de evoluție a celulei pentru a actualiza eticheta unui pixel doar dacă noua tărie este mai mare decât o valoare prag, aleasă experimental;
- limitarea algoritmului la un număr fix de iterații, ales experimental, obținându-se astfel un rezultat fie când automatul converge, fie când este atins numărul maxim de iterații – oricare se întâmplă mai întâi.

Cu aceste modificări, complexitatea generală de timp a algoritmului TbGC rămâne aceeași cu cea a algoritmului originalului, deoarece introducerea unei condiții suplimentare la regula evoluției celulei nu modifică complexitatea computațională fundamentală. Cu toate acestea, prin aplicarea unei limite superioare pentru numărul de iterații, versiunea noastră îmbunătățită previne iterațiile excesive în cazurile în care modificările continuă să fie detectate, dar se diminuează în semnificație, ceea ce duce la o reducere a timpului de execuție fără a afecta complexitatea de bază a metodei.

3.2.2 Generarea automată a punctelor inițiale

Pe lângă algoritmul TbGC, propunem trei metode pentru generarea automată a punctelor inițiale de fundal și două metode pentru punctele inițiale din prim-plan. Aceste metode, deși evaluate numai împreună cu TbGC, pot fi adaptate și utilizate cu orice tehnică de segmentare bazată pe regiuni.

3.2.2.1 Generarea punctelor inițiale de fundal

Mamografiile neprelucrate conțin nu numai sânul, ci și un fundal, ridicând astfel întrebarea: punctele inițiale de fundal trebuie selectate în interiorul sânului sau în afara acestuia? Dacă punctele inițiale de fundal sunt selectate în interiorul sânului, ele s-ar putea suprapune cu țesutul conjunctiv sau glandular care, după cum a fost menționat, au aceleași proprietăți ca și tumorile. De asemenea, un lucru demn de menționat este că masele pot fi localizate foarte aproape de marginea sânului, ceea ce face mai dificilă indicarea manuală a punctelor de fundal în interiorul sânului și în afara masei. Pe de altă parte, dacă punctele inițiale de fundal sunt selectate în afara sânului, pentru sânii denși, întregul sân ar putea fi segmentat ca regiune de interes, algoritmul nefiind capabil să diferențieze între mase și țesut normal, sănătos, dens.

Acestea fiind spuse, pentru a reduce nevoia de intervenție umană și pentru a rezolva problema alegerii punctelor inițiale de fundal, propunem trei soluții posibile:

1. Generarea punctelor inițiale de fundal în interiorul sânului;
2. Folosirea punctelor inițiale de fundal în afara sânului;
3. Folosirea doar a punctelor inițiale de prim-plan.

Pentru prima variantă, folosim un pătrat care cuprinde leziunea, în timp ce pentru a doua variantă, folosim cei mai întunecați pixeli din imagine ca puncte inițiale de fundal.

3.2.2.2 Generarea punctelor inițiale de prim-plan

Plecând de la ideea că centrul anomaliei are cea mai mare șansă de a fi corect etichetat de către utilizator, ne propunem să construim un set optim de puncte inițiale de prim-plan în mod automat, pornind de la centrul tumorii. Pentru ca acest lucru să se întâmple, alegem punctele de prim plan ca un cerc cu centrul corespunzător centrului tumorii și raza aleasă experimental.

După ce am decis cu privire la valoarea cea mai potrivită a razei, care va fi denumită în continuare r , sugerăm o metodă pentru generarea automată a punctelor inițiale de prim-plan ca un cerc cu o rază de r pixeli. Ținând cont de faptul că, într-o mamografie, țesutul care compune o masă apare mai luminos decât restul sânului, ne propunem să găsim cel mai luminos cerc cu o rază de r pixeli în interiorul sânului.

Pentru a atinge acest obiectiv, iterăm peste imagine, considerând fiecare pixel la un moment dat ca centrul unui cerc de rază r pixeli și calculăm luminozitatea cercului, prin însumarea intensităților tuturor pixelilor care se află în interiorul cercului. Ulterior, comparăm valoarea obținută cu cele

obținute anterior și reținem coordonatele pixelului care reprezintă centrul cercului care dă cea mai mare sumă. După ce întreaga imagine este analizată, rămânem cu coordonatele centrului celui mai luminos cerc din imagine. Folosim aceste coordonate pentru a construi două tipuri de cerc cu o rază r -pixel: (1) cerc și (2) disc, pentru a analiza impactul numărului de puncte inițiale asupra rezultatelor segmentării și a alege varianta care duce la o segmentare mai bună.

3.3 Extracția caracteristicilor

Extracția caracteristicilor joacă un rol cheie în sistemele automate de detecție a cancerului de sân. Experimentăm cu două tipuri de caracteristici: bazate pe puncte cheie și pe textură. Pentru prima categorie, folosim două metode: Orientat FAST și Rotated BRIEF (ORB) [24] și Scale Invariant Feature Transform (SIFT) [25], în timp ce pentru a doua, comparăm trei tehnici: Local Binary Pattern (LBP) [26], LeGLRthLMtrix-LengLMtrix [27] și Gray Level Co-occurrence Matrix (GLCM) [28]. Metodele au fost alese în funcție de obiectivul nostru de interpretabilitate: caracteristicile rezultate pot fi fie vizualizate, fie ușor de calculat.

3.4 Selecția caracteristicilor

Selecția caracteristicilor are ca scop identificarea celor mai relevante caracteristici din setul de caracteristici extrase din imaginile mamografice segmentate. Caracteristicile selectate sunt utilizate ulterior pentru clasificarea și identificarea potențialelor afecțiuni maligne, care pot ajuta radiologii în diagnosticarea cu acuratețe a cancerului de sân. Analizăm și experimentăm cu patru tehnici, și anume: Principal Component Analysis (PCA) [29, 30], Kernel Principal Component Analysis (KPCA) [31], Incremental Principal Component Analysis (ICPA) [32], și Singular Value Singular Value Decomposition (SVD).

3.5 Clasificare

Scopul acestei teze este de a construi un sistem CAD interpretabil. Prin urmare, selectăm metodele de clasificare în funcție de acest obiectiv: folosim algoritmi transparenți, ușor de înțeles, care se caracterizează prin comprehensibilitatea lor de către oameni. În acest scop, experimentăm cu următorii algoritmi: (1) Decision Tree (DT) [33], (2) Random Forest (RF) [34], (3) Gaussian Naive Bayes (GNB) [35], (4) K-Means [36], (5) K-Nearest Neighbors (KNN) Fix [37], (6) Linear Discriminant Analysis (LDA) [38], (7) Quadratic Discriminant Analysis (QDA) [39], și (8) Logistic Regression (LR) [40].

Capitolul 4

Evaluarea sistemului de detecție și diagnoză a leziunilor din mamografii propus

4.1 Detecție

Pentru a evalua clar metoda de segmentare descrisă în această teză, evaluăm algoritmul TbGC cu puncte inițiale automate pe mini-MIAS [41] și mini-DDSM [42]. Ca mijloc de a evalua corect îmbunătățirile noastre, comparăm abordarea noastră cu metoda GrowCut originală.

4.1.1 Mini-MIAS

Pentru mini-MIAS, în ceea ce privește generarea automată a punctelor inițiale de fundal, TbGC dă rezultate foarte similare pentru toate cele trei alternative, în timp ce rezultatele produse de algoritmul original se înrăutățesc atunci când se folosesc puncte de fundal în afara sânelor sau nu se folosesc deloc. Pentru GrowCut clasic, zona segmentată se extinde până la fundalul mamografiei, pe măsură ce setăm punctele de fundal, în timp ce, în lipsa punctelor de fundal, se extinde și la fundalul mamografiei. Deși odată cu creșterea numărului de iterații, centrul zonei segmentate, obținut prin abordarea inițială, se apropie de centrul leziunii, aria sa se mărește, rezultând astfel fals-pozitive. Pentru abordarea propusă de noi – TbGC –, aria zonei segmentate crește și ea, dar diferența este aproape inobservabilă. Prin urmare, deși rezultatele obținute cu abordarea clasică sunt mai bune decât rezultatele noastre pentru punctele inițiale de fundal bine definite, în interiorul sânelor, se constată o îmbunătățire incontestabilă a rezultatelor obținute cu TbGC pentru celelalte două cazuri.

Atunci când se generează și punctele inițiale de prim-plan într-o manieră automată, cea mai bună variantă pentru un TbGC nesupervizat este combinația dintre punctele inițiale de fundal din afara sânului și punctele inițiale de prim-plan generate ca un cerc cu centrul corespunzând centrului masei și o rază egală cu 25. Cu această configurație, abordarea noastră atinge 98.52% acuratețe, 67.76% precizie și 57.36% senzitivitate.

4.1.2 Mini-DDSM

Pentru mini-DDSM, când numai punctele de fundal sunt generate automat, TbGC obține rezultate mai bune pentru toate cele trei variante, cu excepția senzitivității. Subliniem faptul că o valoare mai mare pentru numărul maxim de iterații ar duce la segmentarea întregii mase de către algoritmul TbGC și, astfel, la o senzitivitate crescută. De asemenea, dorim să remarcăm faptul că acuratețea obținută de TbGC este de peste 0.9 pentru toate cele trei cazuri, în timp ce, pentru GrowCut, putem observa o diferență de aproape 0.39 între primul și al treilea caz. Deși valorile obținute pentru celelalte metrice nu sunt satisfăcătoare, ele sunt în mod evident mai bune pentru algoritmul TbGC (cu excepția senzitivității, așa cum sa evidențiat deja). Putem concluziona că existența și localizarea (în interiorul sau în afara sânului) punctelor de fundal influențează rezultatul segmentării, dar, deoarece diferențele dintre rezultatele obținute cu TbGC pentru cele trei variante sunt mult mai mici decât cele dintre rezultatele obținute cu algoritmul original GrowCut, putem afirma că TbGC este un algoritm flexibil în comparație cu GrowCut.

Pentru versiunea nesupravegheată a abordării noastre se obține un scor de acuratețe de 97,12%, puțin mai mic decât cel obținut pe mini-MIAS. Din scorurile de precizie de 66.43% și de senzitivitate de 51.01%, putem afirma că regiunea de interes segmentată nu coincide perfect cu leziunea, fiind fie mai mică (ducând la o senzitivitate scăzută), fie mai mare (ducând la o precizie scăzută), conținând totuși cel puțin jumătate din pixelii de interes. Cu toate acestea, reiterăm faptul că abordarea propusă de noi este complet automatizată, nefiind nevoie de nicio intervenție umană și, astfel, având în vedere că TbGC este o tehnică bazată pe regiuni și punctele inițiale sunt generate automat, considerăm că rezultatele sunt satisfăcătoare.

4.2 Diagnoză

După obținerea regiunilor de interes din imaginile mamografice pre-procesate, următorul pas este extracția caracteristicilor, selecția celor mai relevante și efectuarea clasificării. Deoarece unele dintre metodele utilizate sunt parametrizate, trebuie să găsim valorile care conduc la cele mai bune rezultate

de clasificare. Pentru ca acest lucru să se întâmple, efectuăm o căutare exhaustivă, analizând toate combinațiile posibile între valorile parametrilor utilizați pentru extracția și selecția caracteristicilor și clasificare. Numărul total de configurații diferite utilizate în experimentele noastre poate fi calculat după cum urmează: 5800 configurații pentru extracția caracteristicilor $\cdot$ 280 configurații pentru selecția caracteristicilor $\cdot$ 267 configurații pentru clasificare = 433608000 combinații de parametri care sunt comparate pentru a obține cea mai bună clasificare posibilă.

4.2.1 Mini-MIAS

Din rezultatele obținute pe setul de date mini-MIAS rezultă câteva observații. În primul rând, toți clasificatorii folosiți în experimentele noastre obțin cele mai bune performanțe atunci când sunt aplicați pe caracteristicile LBP. Performanțe similare sunt obținute cu GLCM pentru DT, GNB, KNN și LDA. Acestea duc la primele concluzii:

- caracteristicile texturale sunt mai potrivite pentru sistemul nostru decât cele bazate pe puncte cheie,
- extracția caracteristicii LBP funcționează cel mai bine pentru segmentările realizate folosind algoritmul TbGC nesupervizat.

În continuare, examinând rezultatele din perspectiva selecției caracteristicilor, PCA cu variantele sale (KPCA și IPCA) atinge o performanță generală mai bună decât SVD. Cu toate acestea, atunci când vine vorba de alegerea unei anumite variante, trebuie luate în considerare scorurile individuale: utilizarea KPCA este singura care duce la un scor de acuratețe de 95%, în timp ce la PCA și IPCA, cea mai mare acuratețe atinsă este de 90%. Prin urmare, considerăm că KPCA este cea mai eficientă metodă de selecție a caracteristicilor folosită în experimentele noastre.

În ceea ce privește clasificarea, patru metode ating o acuratețe și un F1-score de 95%: DT, RF, QDA și LR, dovedind astfel mai mult potențial atunci când vine vorba de a fi integrate în sistemul propus. Cu toate acestea, unele metode obțin scoruri de precizie și specificitate de 100% (DT cu GLCM și KPCA, RF cu LBP și KPCA, QDA cu LBP și KPCA și LR cu LBP și KPCA), în timp ce altele, senzitivitate perfectă (DT cu LBP și KPCA și QDA cu LBP și SVD). Prin urmare, înainte de a decide asupra unui clasificator, trebuie luată o altă decizie: precizie sau senzitivitate? În timp ce o senzitivitate mai mare înseamnă că nicio malignitate nu este omisă, o precizie mai mare înseamnă că pacienții ar putea începe tratamentul direct, fără investigații suplimentare. Dacă nu se creează un sistem perfect, va exista întotdeauna un compromis între senzitivitate și precizie și considerăm că utilizatorii finali ar trebui să decidă care dintre ele este mai importantă.

În continuare, ne bazăm alegerea pe încă un alt compromis: acuratețea diagnosticului versus transparentă (obținută fie prin interpretabilitate, fie prin explicabilitate). Considerăm că acesta este un aspect important atunci când alegem un model, în special pentru scopul nostru, care presupune că sistemul va fi folosit de persoane fără cunoștințe tehnice legate de IA. În timp ce toate metodele folosite în experimentele noastre oferă un anumit nivel de interpretabilitate, decidem să folosim Random Forest ca mijloc de a echilibra compromisul dintre transparentă și performanță. În concluzie, avansăm folosind LBP pentru extracția caracteristicilor, KPCA pentru selecția caracteristicilor și RF pentru clasificare. Pentru această configurație, am calculat și scorul AUC-ROC, obținând o valoare de 0.93.

4.2.2 Mini-DDSM

Pentru a-i demonstra robustețea, aplicăm abordarea propusă și asupra imaginilor din setul de date mini-DDSM. În experimente, optăm pentru metodele care obțin cele mai bune rezultate pe mini-MIAS. Cu toate acestea, procesul de achiziție a imaginii diferă între seturile de date, datorită tipului de mașină utilizată. Pentru a găsi cele mai potrivite valori ale parametrilor pentru fiecare metodă, se execută o altă căutare în grilă. Prin urmare, metodele folosite pentru fiecare pas rămân aceleași și doar valorile parametrilor sunt modificate.

Rezultatele testelor obținute pe imaginile din mini-DDSM, după o împărțire de 70%-30% antrenare-testare, sunt următoarele: *acuratețe* – 0.97, *precizie* – 0.95, *senzitivitate* – 1.00, *specificitate* – 0.95, *F1-score* – 1.00 *AUC-ROC* – 0.98.

Aceste rezultate dovedesc robustețea abordării propuse. Sistemul antrenat pe mini-DDSM depășește sistemul antrenat pe mini-MIAS din punct de vedere al acurateții. Scorul de sensibilitate perfectă indică faptul că toate leziunile maligne sunt corect etichetate, în timp ce scorurile de precizie și specificitate de 95% arată că o singură anomalie benignă este clasificată greșit drept malignă. La aplicare sistemul propus pe setul de date mini-MIAS, o anomalie malignă a fost clasificată drept benignă – deci, invers. Cu toate acestea, așa cum s-a subliniat anterior, compromisul dintre sensibilitate și precizie va fi întotdeauna prezent, iar decizia cu privire la care dintre aceste metrice cântărește mai greu ar trebui să aparțină utilizatorilor finali.

Vrem să subliniem faptul că, deși experimentăm cu un set de date aproape triplu ca dimensiune decât cel pe care abordarea noastră a fost validată inițial, aceasta încă clasifică greșit doar o singură imagine. Prin urmare, putem concluziona că sistemul propus poate fi adaptat cu ușurință pentru diferite seturi de date prin modificarea doar a valorilor unor parametri, menținând în același timp o performanță ridicată.

4.2.3 RDBMC

Ca o validare finală a abordării noastre propuse, o evaluăm pe setul de date RDBMC, care conține doar mamografii mamare dense. Potrivit Colegiului American de Radiologie [43], femeile cu sân dens nu numai că au un risc crescut de a dezvolta cancer de sân, dar este și mai greu de identificat, din cauza leziunilor care imită țesutul dens pe mamografii.

Din acest motiv, un sistem CAD cu performanțe satisfăcătoare pe țesutul mamar dens este foarte de dorit. Ca și pentru mini-DDSM, folosim aceleași metode care au avut ca rezultat cea mai bună performanță pe mini-MIAS și reglăm parametrii. Deși RDBMC nu oferă nicio indicație pentru segmentare, considerăm acest pas ca fiind esențial pentru o bună clasificare și, astfel, îl aplicăm în continuare, chiar dacă nu putem evalua individual acuratețea acestuia.

Această combinație de metode oferă 96% acuratețe, 100% precizie și specificitate, 88% sensibilitate și 93% F1-score. Precizia și specificitatea perfecte, identice cu rezultatele obținute pe mini-MIAS, arată că abordarea propusă de noi este una prudentă, valorificând deciziile sigure peste îndoieli. Scăderea sensibilității, cauzată de etichetarea greșită a două afecțiuni maligne, poate fi atribuită datelor destul de dezechilibrate utilizate atât pentru antrenament, cât și pentru testare, ceea ce duce la o ușoară părtinire în favoarea clasificărilor benigne.

Considerăm că aceste rezultate dovedesc că sistemul nostru este capabil de generalizare și poate fi folosit indiferent de aparatul folosit pentru examenele mamografice, obținând performanțe satisfăcătoare pe imagini cu diferite caracteristici tehnice. Mai mult, evaluarea sistemului pe un set de date care conține doar sâni denși și obținerea unui scor de acuratețe de 96% arată potențialul său pentru utilizare într-un mediu clinic real.

4.3 Un punct de vedere clinic

Pentru a evalua alinierea sistemului propus de noi cu utilizatorul final, rezultatele obținute pentru 25 de imagini din setul de date RDBMC, alese aleatoriu din subsetul de date de testare, au fost revizuite manual de un radiolog cu experiență, atât în ceea ce privește detecția, cât și diagnoza. Din cele 25 de imagini incluse în acest studiu, a fost identificată o suprapunere între rezultatul segmentării automate și tumora reală pentru 19 imagini. Din cele 6 imagini rămase, expertul nu a putut identifica leziunea în 2 dintre ele, astfel că nu putem spune dacă sistemul a identificat corect leziunea sau nu. Excluzând aceste 2 imagini, procentul de imagini pentru care rezultatul segmentării se suprapune cu leziunea este de 82.61%. Din cele 19 imagini pentru care regiunea segmentată intersectează tumora, 2 arată o aliniere perfectă, 3 leziuni sunt supra-segmentate, iar restul sunt subestimate.

Folosind aceleași 25 de imagini utilizate pentru validarea detecției, radiologul a oferit diagnostic pentru 22 de imagini, omițând cele 2 mamografii în care nu a putut fi identificată tumora. O mamografie suplimentară a fost ignorată, pentru care tumora se suprapunea mușchiului pectoral, cauzând erori la segmentarea automată. Pentru cazurile rămase, au existat 16 potriviri între diagnosticul radiologului și rezultatul sistemului nostru CAD propus. Una dintre aceste potriviri a fost un diagnostic greșit. Din imaginile pentru care radiologul a oferit un alt diagnostic decât sistemul, 5 conțin leziuni benigne etichetate drept maligne, iar una conține o leziune malignă etichetată drept benignă. Toate aceste șase imagini au fost clasificate corect cu abordarea propusă de noi.

Ca mijloc de a ne asigura că radiologii înțeleg cum funcționează abordarea propusă de noi și, de asemenea, că se vor simți încrezători să o folosească, am cerut expertului în radiologie să evalueze sistemul și din punct de vedere al interpretabilității.

Metodologia de pre-procesare este absolut clară, deoarece radiologul nu înțelege doar funcționarea internă, dar este și încrezător să modifice manual rezultatul, dacă este necesar. Pentru etapa de segmentare, apar unele dificultăți atunci când vine vorba de procesul efectiv de creștere a regiunii, cu privire la modul exact în care eticheta unui pixel este modificată în funcție de vecinii săi. Pentru următorii doi pași, care se referă la extracția și selecția caracteristicilor, apar unele probleme de înțelegere și când vine vorba de rezultatul așteptat. Cu toate acestea, clasificarea este mai bine înțeleasă, deși expertul nu înțelege pe deplin datele de intrare (numerice) ale acestui pas. În timp ce abordarea propusă ar beneficia cu siguranță de o interpretare vizuală mai clară a proceselor de extracție și selecție a caracteristicilor, considerăm că această evaluare este încurajatoare pentru un sistem CAD interpretabil.

Capitolul 5

Către un sistem multi-modal de detecție și diagnoză a cancerului de sân

În procesul de detecție și diagnoză a cancerului de sân, diverse informații pot juca un rol important. În timp ce imagistica medicală oferă de obicei o vedere suficient de clară asupra sânilor unui pacient, unele anomalii se pot „ascunde” în țesutul dens, făcând detecția aproape imposibilă, chiar și pentru cei mai experimentați profesioniști. Prin urmare, mai ales pentru pacienții cu sâni denși, combinarea mamografiilor cu informații suplimentare s-ar putea dovedi de mare ajutor. Din acest motiv, analizăm două tipuri de date care pot fi integrate în continuare într-un sistem multi-modal de detecție și diagnoză a cancerului de sân, împreună cu sistemul propus pentru mamografii.

5.1 Bio-fluide

Există o mare nevoie de noi abordări în screening-ul sau diagnosticul medical care nu necesită operații de specialitate și au un nivel ridicat de accesibilitate, senzitivitate și specificitate. Atenția este îndreptată către tehnici rapide, minim invazive și ușor de utilizat, capabile de identificare timpurie a bolilor, dintre care analiza spectrelor Raman (SERS) a câștigat un mare interes. Biopsia lichidă SERS – o abordare în care algoritmi de învățare automată permit clasificarea probelor de la pacienți și a celor de control pe baza semnalului SERS al bio-fluidelor –, a demonstrat progrese intense în ceea ce privește transpunerea în mediul clinic în ultimul deceniu.

În ciuda numărului tot mai mare de articole de cercetare în domeniul biopsiei lichide bazate pe SERS pentru diagnosticul medical, există încă multe îmbunătățiri de făcut și întrebări de răspuns. Scopul nostru este de a evalua performanța bio-fluidelor serice și urinare în detecția cancerului de sân.

Ca atare, comparăm patru modele de clasificare – și anume, DT, RF, GNB și LDA –, pe un set

de date privat compus din 60 de probe de ser și urină obținute de la aceleași pacienți, dintre care 39 de la pacienți cu cancer de sân (confirmat prin biopsie) și 21 de probe de la subiecții de control. Informațiile sunt apoi împărțite în două seturi, care sunt utilizate în continuare pentru antrenarea și validarea clasificatorilor.

5.1.1 Rezultate experimentale

LDA este cel mai des folosit clasificator pentru diagnosticul bazat pe SERS și într-adevăr, pentru spectrele SERS serice, LDA a dat cele mai bune rezultate de clasificare - 0.83 acuratețe. Cu toate acestea, cea mai mare acuratețe de clasificare a fost obținută de DT (0.89), pe baza spectrelor SERS ale urinei, în timp ce LDA a dat o acuratețe generală de 0.78, precum și RF și GNB.

În concluzie, am testat performanța relativă a probelor SERS de ser și urină pentru clasificarea cancerului de sân și a probelor de control folosind patru algoritmi diferiți de învățare automată. Performanțe generale similare, în intervalul 61–89%, indică atât serul, cât și urina, ca potențiali candidați pentru biopsia lichidă SERS pentru detectarea cancerului de sân. Scoruri de acuratețe ușor mai mari au fost obținute utilizând probe de urină în comparație cu probele de ser, în ciuda variabilității mai mari a spectrelor SERS de urină, pe care le asociem cu variațiile pH-ului. Prin identificarea semnăturilor metabolice SERS ale probelor de ser și urină, probele clasificate greșit au fost corelate cu dezechilibre specifice în profilul metabolic al probelor respective.

Considerăm că rezultatele acestei analize SERS pot fi combinate cu examenul mamografic pentru o predicție mai precisă a cancerului de sân.

LDA is the most common classifier for SERS-based diagnosis and indeed, for serum SERS spectra, LDA yielded the best classification results – 0.83. However, the highest classification accuracy was obtained by DT (0.89), based on the SERS spectra of urine, while LDA yielded an overall accuracy of 0.78, as well as RF and GNB.

5.2 Rapoarte medicale

Rapoartele mamografice, care conțin descrieri textuale detaliate ale rezultatelor imagistice, pot fi procesate folosind tehnici de analiză la nivel de document (DLA) pentru a le clasifica automat. Cu toate acestea, un dezavantaj principal al DLA se referă la presupunerea că un document, indiferent de lungimea sa, aparține unei singure clase. În cazul nostru mai ales, un raport mamografic descrie de obicei toate anomaliile prezente într-un sân. Aceasta înseamnă că, dacă un sân conține atât leziuni benigne, cât și leziuni maligne, toate vor fi descrise în același raport. Mai mult, dacă examenul a fost

făcut pentru ambii sâni simultan, raportul va conține informații despre ambii sâni.

Propunem două tipuri de clasificare. Prima se referă la evaluarea generală a stării unui pacient în termeni de *sănătos*, *benign* sau *malign*. Al doilea este referitor la gravitatea leziunilor, conform clasificării BI-RADS.

Pentru experimente, folosim atât raportul original, în limba română, cât și traducerea acestuia în engleză, furnizată împreună cu setul de date. Mai mult, pentru a evalua legătura dintre vârstă pacientului și diagnosticul său, concatenăm vârsta (furnizată cu setul de date) la sfârșitul raportului.

În continuare, efectuăm un pas de pre-procesare, care presupune transformarea textului în litere mici. În ceea ce privește stopwords, experimentele sunt efectuate atât cu cât și fără eliminarea acestora, pentru a evalua impactul acestora asupra performanței modelului. După ce datele sunt pre-procesate, textul poate fi înglobat. În acest scop, experimentăm două metode: *Term Frequency-Inverse Document Frequency (TF-IDF)* și *Latent Semantic Indexing (LSI)*.

Pentru a evalua relevanța înglobărilor bazate pe TF-IDF și LSI atunci când vine vorba de clasificarea automată a rapoartelor medicale scrise în limba română și traduse în engleză, antrenăm și evaluăm patru modele standard de clasificare a învățării automate: DT, RF, NB și LR.

5.2.1 Rezultate experimentale

Pentru primul tip de clasificare, cea mai potrivită combinație de algoritmi (pentru înglobare și clasificare) pentru datele noastre este clasificatorul Random Forest și înglobările TF-IDF. Pentru limba română, cele mai bune rezultate se obțin numai pe rapoarte, excluzând stopwords: 80% acuratețe, precizie, senzitivitate și F1-score. Aceeași configurație oferă cea mai bună performanță și pentru limba engleză, cu o creștere de 2% în acuratețe și senzitivitate (82%), 4% în precizie (84%) și 1% în F1-score (81%).

Pentru evaluarea scorului BI-RADS al unui pacient, cele mai bune metrici de performanță pentru limba română se obțin cu aceeași combinație ca și pentru clasificarea cu trei clase: RF aplicat pe reprezentările TF-IDF ale rapoartelor (fără vârstă), excluzând stopwords: 73% acuratețe, 77% precizie, 73% senzitivitate și 67% F1-score. Datorită faptului că clasa corespunzătoare scorului BI-RADS 6 este sub-reprezentată, se constată o scădere a performanței față de clasificarea cu 3 clase. Pentru limba engleză, totuși, cea mai bună combinație pare să fie între NB și LSI pentru raport și vârstă, fără a exclude stopwords (75% acuratețe și senzitivitate, 74% precizie, 73% F1-score).

În ansamblu, analizăm și comparăm patru clasificatori și două metode de înglobare pe 24 de tipuri de date – rapoarte și rapoarte concatenate cu vârsta, în română și engleză, excluzând și păstrând stopwords – folosind drept clasă diagnosticul sau scorul BI-RADS, obținând o clasificare cu 3 clase și

o clasificare cu 5 clase. Rezultatele experimentale au arătat performanțe mai bune pentru clasificarea cu 3 clase, precum și robustețea algoritmului RF, obținând cea mai mare acuratețe pe ambele tipuri de clasificare pentru limba română.

Considerăm aceste combinații selectate pot fi utilizate pentru clasificarea rapoartelor mamografice și pot fi integrate în continuare într-un sistem multi-modal care ține cont de datele de imagistică actuale și anterioare.

Capitolul 6

Concluzii și extinderi

În această teză, am propus un sistem interpretabil automat de detecție și diagnoză a cancerului de sân, cu scopul de a servi drept a doua opinie medicilor care analizează și interpretează imaginile mamografice. Sistemul poate fi împărțit în cinci pași ușor de înțeles, dar robusteți, care pot fi vizualizați în așa fel încât utilizatorii să poată înțelege deciziile sistemului fără a avea nevoie de cunoștințe tehnice.

Sistemul propus începe cu eliminarea informațiilor redundante din mamografia brută și îmbunătățirea calității acesteia. Masele suspecte sunt apoi segmentate cu un algoritm nesupervizat, algoritmul Threshold-based GrowCut, obținându-se astfel regiunile de interes. Din acestea, sunt extrase caracteristicile texturale Local Binary Pattern, iar cele mai relevante sunt selectate cu ajutorul Kernel Principal Component Analysis. Aceste caracteristici servesc ca intrare la un clasificator Random Forest, care dă rezultatul final: leziune benignă sau malignă.

Abordarea noastră se distinge prin interpretabilitatea sa, realizată prin explicarea vizuală a fiecărei etape. În comparație cu rețelele neuronale artificiale, este mai ușor și mai rapid de antrenat, ceea ce este crucial în asistența medicală, unde seturile de date sunt adesea limitate ca dimensiune. În plus, utilizând algoritmul Threshold-based GrowCut, o metodă de segmentare iterativă, utilizatorul poate verifica fiecare pas al detecției.

Pentru a obține CAD-ul prezentat, am dezvoltat o metodă de segmentare nesupervizată, am analizat cinci metode de extracție a caracteristicilor, patru metode de selecție a caracteristicilor și opt metode de clasificare, am testat toate variantele pe 57 de imagini din setul de date mini-MIAS [41] și am ales-o pe cea care a obținut cele mai bune rezultate – 95% acuratețe, 100% precizie și 100% specificitate. Mai mult, am validat abordarea propusă pe un set de date diferit, și anume mini-DDSM [44], obținând o acuratețe de 97% și o sensibilitate de 100%. De asemenea, am propus un nou set de date – Romanian Dense Breast Mammography Collection – care conține doar mamografii ale sânilor cu țesut dens și am testat sistemul nostru și pe acest set de date, obținând o acuratețe de 96% și o

precizie și specificitate de 100%.

În plus, am făcut deja primii pași către un sistem multi-modal de diagnoză a cancerului de sân, prin efectuarea unei analize preliminare privind potențiala utilizare a bio-fluidelor și a rapoartelor medicale împreună cu imaginile mamografice pentru un nivel sporit de certitudine.

În timp ce sistemul propus în această teză demonstrează performanță și interpretabilitate puternice, există mai multe direcții pentru cercetări viitoare care ar putea îmbunătăți acuratețea, aplicabilitatea și integrarea în fluxurile de lucru clinice. Pentru început, intenționăm să continuăm eforturile noastre de integrare a mai multor tipuri de date (textuale și numerice) într-un sistem multi-modal pentru detecția și diagnoza cancerului de sân. Această integrare ar putea îmbunătăți acuratețea diagnosticului prin valorificarea informațiilor complementare din imaginile mamografice, rapoartele radiologice și analiza SERS pe bază de bio-fluide. În plus, ne propunem să explorăm includerea datelor genomice și a istoricului pacientului, permițând un cadru mai cuprinzător de evaluare a riscurilor.

De asemenea, ca mijloc de îmbunătățire a fiabilității diagnosticului, intenționăm să dezvoltăm metode care să cumuleze rezultatele obținute din mai multe imagini mamografice ale aceluiași pacient. Aceasta include combinarea constatărilor din vederile cranio-caudale (CC) și oblice medio-laterale (MLO), precum și din diferite modalități de imagistică, cum ar fi tomosinteza. Acest lucru ar permite sistemului CAD să reducă rezultatele fals-pozitive și fals-negative prin identificarea modelelor consistente în mai multe imagini.

Deoarece această teză a introdus un nou set de date concentrat exclusiv pe sânii denși, abordând o lacună critică în cercetarea imagistică a cancerului de sân, ca pas următor, intenționăm să introducem și un alt set de date care să cuprindă doar mamografii ale sânilor adipoși. Acest lucru va permite analize comparative pentru a evalua modul în care densitatea sânilor afectează performanța sistemului CAD, detectabilitatea leziunilor și interpretabilitatea. Efectuând evaluări încrucișate, ne propunem să ne perfecționăm modelul pentru a asigura generalizarea în diferite tipuri de țesut mamar.

Pentru a valida în continuare sistemul, intenționăm să efectuăm studii asupra utilizatorilor pentru a-i evalua interpretabilitatea la o scară mai largă. În timp ce evaluările inițiale au demonstrat transparența sistemului, un studiu mai larg care implică mai mulți radiologi și clinicieni ar oferi informații mai solide. În acest scop, intenționăm să proiectăm un studiu controlat în care experții interacționează cu sistemul și oferă feedback cu privire la cât de bine se aliniază fiecare pas cu raționamentul lor clinic.

În cele din urmă, pentru a facilita adoptarea clinică, intenționăm să integrăm sistemul CAD în infrastructurile PACS (Picture Archiving and Communication System) existente. Această integrare le-ar permite radiologilor să acceseze fără probleme informații generate de CAD în fluxul lor de lucru.

În plus, intenționăm să implementăm generarea automată de rapoarte pentru a rezuma constatările cheie, a evidenția regiunile suspecte și a sugera recomandări ulterioare.

Pe lângă posibilele îmbunătățiri ale sistemului propus în această teză, intenționăm, de asemenea, să dezvoltăm modele predictive care evaluează riscul pacientului de a dezvolta cancer de sân pe baza datelor imagistice istorice. Analizând seturile de date mamografice longitudinale, astfel de modele ar putea identifica modificări subtile în timp care pot indica un risc crescut de malignitate. Folosirea tiparelor temporale în datele imagistice ar putea îmbunătăți eforturile de detecție timpurie, oferind clinicienilor informații valoroase asupra progresiei bolii și sprijinind luarea deciziilor mai informate cu privire la frecvența screening-ului și la măsurile preventive.

O altă direcție promițătoare pentru lucrările viitoare este extinderea metodologiei propuse dincolo de cancerul de sân, adaptând-o pentru detecția și diagnoza altor tipuri de cancer, cum ar fi cancerul pulmonar, de prostată sau cerebral. Aceasta ar implica reinstruirea sistemului pe scanări CT, RMN sau PET, ajustarea metodelor pentru a se adapta diferențelor de caracteristici imagistice dintre mamografie și alte modalități.

Mai mult, prin adaptarea metodologiei noastre la diferite modalități de imagistică, putem explora posibilitatea extinderii sistemului CAD propus la un cadru de analiză multi-organ, permițând evaluarea simultană a organului primar și a organelor în care cancerul este cel mai probabil să metastazeze. Această abordare ar îmbunătăți detectarea precoce a tumorilor secundare, ar îmbunătăți planificarea tratamentului și ar contribui la o înțelegere mai cuprinzătoare a progresiei cancerului în mai multe zone anatomice.

Bibliografie

- [1] C. Moroz-Dubenco. Comparison of gradient-based edge detectors applied on mammograms. *Studia Universitatis Babes-Bolyai Informatica*, 66(2):5–18, 2021.
- [2] Cristiana Moroz-Dubenco, Laura Dioşan, and Anca Andreica. Mammography lesion detection using an improved growcut algorithm. *Procedia Computer Science*, 192:308–317, 2021.
- [3] Stefania D Iancu, Ramona G Cozan, Andrei Stefancu, Maria David, Tudor Moisoiu, Cristiana Moroz-Dubenco, Adel Bajcsi, Camelia Chira, Anca Andreica, Loredana F Leopold, et al. Sers liquid biopsy in breast cancer. what can we learn from sers on serum and urine? *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 273:120992, 2022.
- [4] Cristiana Moroz-Dubenco, Adél Bajcsi, Anca Andreica, and Camelia Chira. An unsupervised threshold-based growcut algorithm for mammography lesion detection. *Procedia Computer Science*, 207:2096–2105, 2022.
- [5] Cristiana Moroz-Dubenco, Laura Dioşan, and Anca Andreica. Towards an unsupervised growcut algorithm for mammography segmentation. In *International Conference on Computer Vision Systems*, pages 102–111. Springer, 2023.
- [6] Cristiana Moroz-Dubenco, Laura Diosan, and Anca Andreica. Generalizing an improved growcut algorithm for mammography lesion detection. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 709–720. Springer, 2023.
- [7] Cristiana Moroz-Dubenco and Anca Andreica. Linear discriminant analysis tumour classification for unsupervised segmented mammographies. *Procedia Computer Science*, 225:2951–2960, 2023.
- [8] Anamaria Briciu, Alina-Delia Călin, Diana-Lucia Miholca, Cristiana Moroz-Dubenco, Vladieiela Petraşcu, and George Dascălu. Machine-learning-based approaches for multi-level sentiment analysis of romanian reviews. *Mathematics*, 12(3):456, 2024.

- [9] Cristiana Moroz-Dubenco, Adél Bajcsi, Anca Andreica, and Camelia Chira. Towards an interpretable breast cancer detection and diagnosis system. *Computers in Biology and Medicine*, 185: 109520, 2025.
- [10] ZiQi Tao, Aimin Shi, Cuntao Lu, Tao Song, Zhengguo Zhang, and Jing Zhao. Breast cancer: epidemiology and etiology. *Cell biochemistry and biophysics*, 72:333–338, 2015.
- [11] Ghada M El-Banby, Nourhan S Salem, Eman A Tafweek, and Essam N Abd El-Azziz. Automated abnormalities detection in mammography using deep learning. *Complex & Intelligent Systems*, 10(5):7279–7295, 2024.
- [12] Jelena Bozek, Mario Mustra, Kresimir Delac, and Mislav Grgic. A survey of image processing algorithms in digital mammography. *Recent advances in multimedia signal processing and communications*, pages 631–657, 2009.
- [13] Heang-Ping Chan, Ravi K Samala, and Lubomir M Hadjiiski. Cad and ai for breast cancer—recent development and challenges. *The British journal of radiology*, 93(1108):20190580, 2019.
- [14] Nicholas Konz, Mateusz Buda, Hanxue Gu, Ashirbani Saha, Jichen Yang, Jakub Chłędowski, Jungkyu Park, Jan Witowski, Krzysztof J Geras, Yoel Shoshan, et al. A competition, benchmark, code, and data for using artificial intelligence to detect lesions in digital breast tomosynthesis. *JAMA network open*, 6(2):e230524–e230524, 2023.
- [15] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):4165, 2018.
- [16] J Anitha and J Dinesh Peter. Mammogram segmentation using maximal cell strength updation in cellular automata. *Medical & biological engineering & computing*, 53(8):737–749, 2015.
- [17] Valliappan Raman, Putra Sumari, HH Then, and Saleh Ali K Al-Omari. Review on mammogram mass detection by machinelearning techniques. *International Journal of Computer and Electrical Engineering*, 3(6):873, 2011.
- [18] Asad Mohammed Khan and S Ravi. Image segmentation methods: A comparative study. *International Journal of Soft Computing and Engineering*, 3(4):84–92, 2013.
- [19] Yan Xu, Rixiang Quan, Weiting Xu, Yi Huang, Xiaolong Chen, and Fengyuan Liu. Advances in medical image segmentation: a comprehensive review of traditional, deep learning and hybrid approaches. *Bioengineering*, 11(10):1034, 2024.

- [20] Dilovan Asaad Zebari, Dheyaa Ahmed Ibrahim, Diyar Qader Zeebaree, Habibollah Haron, Merdin Shamal Salih, Robertas Damaševičius, and Mazin Abed Mohammed. Systematic review of computing approaches for breast cancer detection based computer aided diagnosis using mammogram images. *Applied Artificial Intelligence*, 35(15):2157–2203, 2021.
- [21] Adél Bajcsi, Anca Andreica, and Camelia Chira. Towards feature selection for digital mammogram classification. *Procedia Computer Science*, 192:632–641, 2021.
- [22] Adél Bajcsi, Camelia Chira, and Anca Andreica. Extended mammogram classification from textural features. *Studia Universitatis Babeş-Bolyai Informatica*, 67(2):5–20, 2023.
- [23] Vladimir Vezhnevets and Vadim Konouchine. Growcut: Interactive multi-label nd image segmentation by cellular automata. In *proc. of Graphicon*, volume 1, pages 150–156. Citeseer, 2005.
- [24] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. IEEE, 2011.
- [25] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.
- [26] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.
- [27] Mary M Galloway. Texture analysis using gray level run lengths. *Computer graphics and image processing*, 4(2):172–179, 1975.
- [28] Robert M Haralick, Karthikeyan Shanmugam, and Its’ Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, SMC-3(6):610–621, 1973.
- [29] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.
- [30] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.
- [31] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.
- [32] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *International journal of computer vision*, 77:125–141, 2008.

- [33] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.
- [34] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [35] Hajer Kamel, Dhahir Abdulah, and Jamal M Al-Tuwaijari. Cancer classification using gaussian naive bayes algorithm. In *2019 International Engineering Conference (IEC)*, pages 165–170. IEEE, 2019.
- [36] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.
- [37] Evelyn Fix. *Discriminatory analysis: nonparametric discrimination, consistency properties*, volume 1. USAF school of Aviation Medicine, 1985.
- [38] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.
- [39] Alaa Tharwat. Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, 3(2):145–180, 2016.
- [40] Jan Salomon Cramer. The origins of logistic regression. Technical report, Tinbergen Institute, 2002.
- [41] John Suckling. The mammographic images analysis society digital mammogram database. In *Excerpta Medica. International Congress Series, 1994*, volume 1069, pages 375–378, 1994.
- [42] M Heath, K Bowyer, D Kopans, R Moore, and P Kegelmeyer. The digital database for screening mammography, iwdm-2000. In *Fifth International Workshop on Digital Mammography., Medical Physics Publishing*, pages 212–218. Medical Physics Publishing, 2001.
- [43] American College of Radiology. Acr statement on reporting breast density in mammography reports and patient summaries, 2017. URL <https://www.acr.org/Advocacy/Position-Statements/Reporting-Breast-Density>. Accessed 12.11.2024.
- [44] Charitha Dissanayake Lekamlage, Fabia Afzal, Erik Westerberg, and Abbas Cheddad. Mini-ddsm: Mammography-based automatic age estimation. In *2020 3rd International Conference on Digital Medicine and Image Processing*, pages 1–6. Association for Computing Machinery, 2020.