

**Universitatea „Politehnica” din Timișoara
Facultatea de Automatică și Calculatoare
Departamentul Calculatoare**

REZUMAT

TEZĂ DE DOCTORAT

**REPREZENTARE FORMALĂ ȘI INFERENȚĂ ÎN
PROGNOZA BAZATĂ PE IMAGINILE MEDICALE
MICROSCOPICE. APLICAȚIE ÎN DOMENIUL
GRADĂRII CANCERULUI MAMAR.**

*FORMAL REPRESENTATION AND REASONING FOR MICROSCOPIC
MEDICAL IMAGE-BASED PROGNOSIS. APPLICATION TO BREAST CANCER
GRADING.*

CONDUCĂTORI ȘTIINȚIFICI:

**Prof. Dr.Ing. Vladimir Ioan CREȚU
UNIVERSITATEA “POLITEHNICA” DIN TIMIȘOARA**

**A/Prof. Daniel RACOCEANU HDR
UNIVERSITÉ DE FRANCHE-COMTÉ, FRANȚA**

**DOCTORAND:
Adina Eunice TUȚAC**

TIMIȘOARA, 2010

Teza de față constituie rezultatul original al cercetărilor efectuate în vederea găsirii unei abordări de reprezentare formală calitativă a cunoștințelor medicale, folosind ontologiile și mecanismul de inferență ca suport automatizat în prognoza cancerului mamar. Investigatiile au avut ca scop dezvoltarea unei ontologii de gradare a cancerului mamar numită *BCGO*, care să permită creșterea reproductibilității și eficienței acestei analize medicale în raport cu procedura manuală, să asigure o terminologie formală standardizată care să beneficieze de expresivitate ridicată și putere computațională, pentru a permite integrarea ontologiei într-o platformă de microscopie virtuală cognitivă (ca asistent virtual în prognozare) și cercetări ulterioare în domeniul reprezentării formale.

Lucrarea este structurată în 9 capitole, desfășurându-se pe un număr de 126 pagini și având ca suport științific 154 referințe bibliografice.

Capitolul 1 prezintă contextul și obiectivele tezei. Reprezentarea formală a cunoștințelor din diverse domenii constituie o componentă esențială a sistemelor perceptive și cognitive. O direcție de interes pentru această lucrare este reprezentarea spațială datorită strânsei legături cu aspectul perceptiv și cel cognitiv. Această reprezentare poate fi clasificată în: cantitativă (numerică) și calitativă (simbolică) funcție de contextul și domeniul de reprezentat. În aplicațiile medicale, o reprezentare calitativă corespunde „metaforei acvariului”, în care la momentul diagnozei sau prognozei informațiile nu sunt toate disponibile (de exemplu, radiologul nu dispune de date de la patolog), sau diagnoza/prognoza depind de interpretarea oferită de medicul specialist, de percepția și experiența personală a acestuia. Nu în ultimul rând, imaginile care sunt strâns legate de aspectul spațial, joacă un rol important în procesul de diagnostic și prognoză.

Gradarea cancerului mamar (BCG) este un exemplu ilustrativ în acest context. Gradarea cancerului mamar este considerată la ora actuală analiza esențială de prognoză în practica patologiei moderne. Sistemul standard folosit de medicii patologi este sistemul de gradare Nottingham (NGS). Acest sistem se bazează pe analiza manuală a țesutului sub forma slide-ului/frame-ului microscopic digitizat, analiză care constă în aplicarea a trei criterii: identificarea formațiunilor tubulare, pleomorfismul nuclear și numărul de mitoze.

Cercetările și experiențele medicale au arătat ca această analiză este puternic influențată de experiența fiecărui patolog în parte. De asemenea, aspectele de reproductibilitate a intra-observabilității precum și cele de eficiență a inter-observabilității duc la inconsistențe în acordul asupra gradării, ceea ce constituie o problemă majoră pentru prognoza cancerului mamar. Pe de altă parte, lipsa unei reprezentări formale semantice standardizate, care să ajute la indexarea și clasificarea terminologiei și să folosească un mecanism de inferență care să asiste în prognoză, constituie o altă problemă esențială.

Prin urmare, obiectivele acestei teze sunt următoarele:

- un studiu a doua abordări diferite în indexarea și obținerea caracteristicilor imaginilor și a celor de la nivel semantic: Content-Based Image Retrieval (CBIR) și Case-Based Reasoning (CBR).
- o metodă nouă pentru reducerea inconsistențelor în acordul medicilor asupra BCG care constă în dezvoltarea unui model formal calitativ numit Breast Cancer Grading Ontology (BCGO) sub forma unei ontologii de aplicație.
- o teorie de reprezentare spațială pentru relațiile/conceptele spațiale aferente BCG, integrând concepte mereo-topologice, metrice, geometrice și de scală.
- o metodă de evaluare calitativă a BCGO însoțită de validarea medicală.

- integrarea ontologiei BCGO în platforma de microscopie virtuală MICO, ca suport pentru adnotarea, explorarea vizuală și extragerea semantică a conceptelor aferente caracteristicilor de imagine și reprezentând astfel ridicarea microscopiei virtuale la nivel cognitiv.

Capitolul 2 își propune să realizeze o incursiune în abordările conceptuale privitoare la reprezentarea formală și mecanismul de inferență, de la reprezentările existente la *nivel de imagine* până la *nivel semantic*. Au fost identificate două strategii de reprezentare la nivel de imagine: *Content-Based Image Retrieval* și respectiv *Case-Based Reasoning* și au fost cercetate în detaliu 3 tehnici pe care ambele le folosesc: *indexing*, *retrieval* și *refining*, evidențiindu-se avantajele și dezavantajele acestora la fiecare nivel. Pe baza analizei lor din punct de vedere tehnologic și metodologic, se argumentează că ambele paradigme pot fi văzute ca și metodologii și se propune o combinație CBIR-CBR sub forma unei noi metodologii, care să beneficieze de avantajele fiecărei abordări în parte, pentru aplicarea acestora în domeniul medical. La nivel semantic, au fost analizate limbajele formale *Description-Logics* (DL), *Ontology-Web Language* (OWL) și *Semantic Web Rule Language* (SWRL) și de asemenea tipurile de ontologie, cu scopul identificării unei reprezentări avantajoase pentru gradarea cancerului mamar. Ultima secțiune a acestui capitol prezintă o parcurgere a literaturii de specialitate în reprezentarea spațială, cu accent pe reprezentarea calitativă și diversele teorii care au fost propuse: mereologie, topologie, mereo-topologie, și teorii pentru reprezentarea distanței și a orientării.

Urmând structura capitolul 2, reprezentare la nivel de imagine și reprezentare semantică, **Capitolul 3** este dedicat analizei aplicării CBIR și CBR în domeniul medical, la nivel de imagine, urmate de o trecere în revistă a ontologiilor bio-medicale de la cele construite pe baza de frames și rețele semantice până la cele care au formalismul logic ca suport. În expunere se pun bazele pentru propunerea de dezvoltare a unui nou model ontologic pentru BCG. Sunt așadar prezentate motivele pentru care o reprezentare ontologică formală rezolvă problema *gap-ului semantic* și a *gap-ului de context*, generic definite *gap-ul de conținut*. Analiza CBIR-CBR a condus la enunțarea unuia dintre principiile abordării propuse în lucrare și anume: raționamentul este comun atât în CBIR cât și în CBR, la CBIR este un raționament bazat pe imagine pe când la CBR este un raționament bazat pe informația structurată pe cazuri, de obicei sub formă de text. În termeni de logică formală, raționamentul este mecanismul de inferență. Ca atare, *mechanismul de inferență este esențial în CBIR și CBR, și reprezintă suportul pentru indexarea semantică și extragerea semantică a conceptelor aferente caracteristicilor obiectelor din imagine*. Un alt principiu care derivă din analiza ontologiilor bio-medicale este următorul: *ontologiile însoțite de mecanism de inferență DL reprezintă soluția pentru o reprezentare structurată și computabilă a cunoștințelor domeniilor de interes*. De asemenea, datorită faptului că imaginile histopatologice joacă un rol important în prognoză, este subliniată *necesitatea utilizării unei reprezentări formale spațiale precum și a unei teorii spațiale cu ajutorul căreia să poată fi eliminate inconsistențele și ambiguitățile în reprezentare*. Pe baza acestor principii, se va construi model formal prezentat în capitolul 4 și teoria spațială din capitolul 5.

Capitolul 4 introduce BCG, sistemul standard NGS de gradare folosit la momentul curent de către medicii patologi și evidențiază principalele probleme pe care procedura le ridică, cu precădere faptul că analiza necesită atenție și timp (aceasta se face pe imagini microscopice de mari dimensiuni, iar un slide conține peste 4000 de frame-uri care sunt digitizate și studiate la microscop). Apoi este prezentat modelul ontologic formal propus pentru reprezentarea conceptelor medicale aferente BCG. Sunt propuse două variante: *image-driven* și *semantic-driven*. Modelul *image-driven* este construit în 2 etape: *knowledge acquisition* și *knowledge translation*. În prima etapă, cunoștințele sunt preluate atât de la medicul patolog (*subjective knowledge*) cât și din NGS (*objective knowledge*). *Knowledge translation* se face pe baza unui translator MK-CV (Medical Knowledge-Computer Vision). Acest model prezintă însă câteva limitări: nu beneficiază de capacitatea computațională a logicii formale DL și este orientat pe nivel de imagine, ceea ce și algoritmi de procesare de imagini pot să facă în mod independent, nu neapărat având suport ontologic, adică de nivel semantic. Pe de altă parte, expresivitatea reprezentării nu este atât de ridicată. Prin urmare, s-a propus al doilea model, care constituie *ontologia de aplicație BCGO*. Aceasta vizează modelarea calitativă a conceptelor de tip *perdurant* (*obiecte*) având ca suport o teorie formală spațială. Spre deosebire de primul model, proiectarea ontologiei urmărește aici 3 pași:

- *knowledge acquisition* – cunoștințele medicale sunt preluate, pe de-o parte prin segmentarea ontologiei de referință National Cancer Institute (NCI) care conține concepte legate de cancerul mamar, și pe de altă parte, din NGS care conține, conceptele specifice pentru gradare. Se constată o abordare diferită, mai completă față de *knowledge acquisition* din primul model.
- *knowledge translation*- se bazează pe un modul structural care folosește logica OWL-DL și un modul de reguli, construit folosind limbajul SWRL. Avantajul acestei abordări constă în obținerea unei expresivități ridicate și în același timp, a puterii de decidabilitate, specifică logicii formale, spre deosebire de alte ontologii biomedicale care nu beneficiază de mecanism de inferență sau de reguli.
- *knowledge refining* – rafinarea reprezentării se face pe baza feedback-ului medical și pe baza reasoner-ului Pellet de tip DL, obținându-se astfel modelul inferat al ontologiei.

Teoria formală spațială și mecanismul de inferență sunt prezentate în **capitolul 5**. Teoria formală propusă tratează concepte mereo-topologice, metrice, geometrice și de scală în vederea contribuției la consistența reprezentării spațiale. Se definesc relații mereo-topologice precum *SurroundedBy* pornind de la definiția, axiomele și teoremele pentru relația *LocatedIn* propuse pentru ontologiile biomedicale FMA și GALEN. Se definesc de asemenea relații metrice precum *CloseTo*. În cazul acesta, se apelează la o altă metodă tot de reprezentare calitativă, și anume la Region Connection Calculus (RCC-8) și la tabela de compoziție. Cu ajutorul teoriei spațiale pot fi *eliminate ambiguitățile și inconsistențele în reprezentare*, aspect ilustrat de asemenea în acest capitol. Mecanismul de inferență se realizează în două moduri: manual și automat. Cel manual constă în aplicarea axiomelor și teoremelor teoriei în reprezentarea cunoștințelor din BCG. Se

analizează mai ales aplicarea acestora de la nivel de instanțe (*individuals*) la nivel de clase. Inferența automată propusă are drept suport algoritmul tablou implementat ca și mecanism de inferență DL în Pellet, unealta utilizată în această lucrare pentru procesul de clasificare automată și verificare a consistenței ontologiei. Este prezentată funcționarea acestui algoritm în BCGO și sunt analizate rezultatele raționamentului logic.

Capitolul 6 conține o descriere detaliată a implementării BCGO respectând modelul propus în capitolul 4, folosind Protégé-OWL ca mediu de programare (bazat pe limbajul de programare Java) și Pellet ca reasoner (bazat pe limbajul formal DL). Pe parcursul implementării sunt tratate principalele caracteristici ale OWL, DL și SWRL, în descrierea claselor, proprietăților, instanțelor și regulilor. Implementarea corespunde etapelor de *knowledge translation* și *knowledge refining*. Se pornește de la faptul că DL knowledge-base este împărțită în *TBox* și *ABox*. *TBox* conține axiomele, definițiile conceptelor și specificarea constrângerilor iar în termeni de OWL acestea reprezintă clasele, iar *ABox* conține aserțiunile conceptelor și proprietăților (relațiilor), adică instanțele în termeni OWL. Sunt definite 4 mari categorii: *AnatomicalEntity*, *ConceptualEntity*, *MicroscopicalEntity* și *SpatialEntity*. BCGO conține în total un număr de 129 de clase, 169 de instanțe și 86 de relații. Se arată apoi cum sunt construite clasele definite și clasele primitive în OWL-DL, clasele disjuncte, și este ilustrat principiul Open World Assumption (OWA) după care funcționează OWL, fapt ce implică folosirea unei *axiome de închidere* în reprezentarea clasei. Sunt discutate în continuare *object properties* și *datatype properties* și sunt prezentate modalitățile de implementare care oferă expresivitatea cea mai ridicată și decidabilitate funcție de tipul de proprietate ales. Modulul de reguli SWRL poate funcționa independent sau legat de OWL, unele reguli pot fi traduse în OWL fără ajutorul sintaxei SWRL (numite *syntactic sugar rules*). Pe de altă parte, deși regulile exprimate folosind limbajul SWRL completează reprezentarea, ridică în același timp problema indecidabilității. Soluția propusă este folosirea regulilor de tip *SWRL DL safe*.

Capitolul 7 prezintă evaluarea BCGO din punct de vedere calitativ și a constrângerilor sintactice DL. Metricile calitative folosite au vizat granularitatea, scopul reprezentării și gradul de integrare a ontologiei în alte ontologii. Granularitatea reprezentării în BCGO este analizată funcție de *relationship richness* și *attribute richness*. Scopul sau gradul de reflecție a domeniului medical BCG reprezentat în BCGO este evaluat în termeni de *class richness* și *average population*. Iar gradul de integrare a ontologiei în alte ontologii analizează nivelul de conexiune cu alte concepte sau relații din alte ontologii cu care BCGO are elemente comune. Sunt discutate aspecte de constrângeri sintactice ale DL, precum restricții cardinale pentru proprietățile tranzitive, sau instanțe în enumerări. Validarea ontologiei se bazează pe *semantic retrieval* constând în query-uri semantice de tip RDF, SQWRL și de asemenea pe feedback-ul medical și integrarea ontologiei în platforma Open Biomedical Ontologies. Se concluzionează că ontologia implementată reflectă în mare măsură domeniul medical și care poate fi integrată ca pas următor într-o ontologie de referință.

Capitolul 8 prezintă un studiu de caz: integrarea ontologiei BCGO implementate, în platforma MICO de microscopie virtuală (construită de echipa IPAL din care face parte și doctoranda), unde ontologia joacă un rol esențial pentru aspectul cognitiv. Demersul capitolului pornește de la definirea microscopiei virtuale și evidențierea importanței în

contextul gradării cancerului mamar, continuând cu prezentarea principiilor și caracteristicilor microscopiei cognitive. Se arată apoi ca platforma MICO poate fi considerată o platformă microscopică cognitivă și datorită ontologiei. Platforma MICO este împărțită în 4 module: *image acquisition*, *Region of Interest (ROI) detection*, *ROI grading si ontology-based retrieval and validation support*. În primul modul dezvoltat de echipa IPAL, se efectuează achiziționarea imaginilor (frame-urilor) care sunt apoi convertite în formatul Digital Imaging and Communications in medicine (DICOM), standardul folosit în clinici, și se aplică apoi un algoritm de *stitch and blending* a frame-urilor pentru a se forma slide-ul complet numit WSI (Whole Slide Imaging). În modulul de *ROI detection*, sunt aplicați algoritmi de procesări de imagini pentru detecția pleomorfismului nucleului, la nivel multi-scală. Pentru detecția formațiunilor tubulare și a mitozelor, se argumentează faptul că algoritmi de procesări de imagini folosiți independent nu sunt eficienți și este necesară intervenția ontologiei pentru a asigura segmentarea corectă a acestor formațiuni din imagine. Este propusă de asemenea o metodă prin care reprezentarea semantică descrisă în limbajul formal OWL poate fi mapată cu reprezentarea informațiilor în alte limbaje de programare, cu scopul de a conecta nivelul semantic cu cel de imagine. Modulul *ROI grading* se ocupă de gradarea regiunilor de interes, pe baza criteriilor de scor și gradare reprezentate semantic în BCGO. Ultimul modul asigură suportul de validare a întregului sistem. Astfel, cu ajutorul BCGO se realizează adnotarea semantică a imaginilor (slide-urilor) histopatologice, explorarea acestor slide-uri de mari dimensiuni precum și semantic retrieval. BCGO contribuie de asemenea la procesul de gradare automată a slide-urilor, urmând metodologia CBIR-CBR propusă în capitolul 2, și putând opera ca și consultant virtual în gradare.

Capitolul 9. Teza se încheie cu o serie de concluzii reprezentând rezumatul contribuțiilor desprinse pe parcursul tratării subiectelor propuse în lucrare, și perspectivele de continuare a cercetărilor în domeniul ontologiilor biomedicale, precum și a tratării altor caracteristici ale logicilor formale.

În sinteză, aceste contribuții științifice sunt:

- un studiu comparativ din punct de vedere metodologic (spre deosebire de alte abordări care le privesc la nivel tehnologic) a două abordări aparținând unor domenii diferite, CBIR și CBR, extins spre domeniul medical și care construiește bazele pentru o nouă metodologie hibridă CBIR-CBR care să poată fi aplicată în BCG.
- o nouă abordare de a pune împreună *image representation* și *semantic representation*, întrucât ambele presupun reprezentarea de cunoștințe și ambele sunt în strânsă legătură cu percepția și cunoașterea. Acesta s-a concretizat prin dezvoltarea ontologiei de aplicație BCGO (o ontologie originală), bazată pe o reprezentare calitativă a conceptelor de tip perdurant. Metodologia de proiectare a ontologiei (cei trei pași) este o metodologie generică și independentă care poate fi aplicată oricărei alte reprezentări. Această abordare prezintă următoarele avantaje: rezolvă *gap-ul semantic* și *gap-ul de context* printr-o tehnică de indexare *semantică* a conceptelor aferente BCG, oferă expresivitate ridicată și decidabilitate, asigurate de cele două module: OWL-DL și regulile SWRL. Nu în ultimul rând prin reprezentarea formală standardizată însoțită de mecanismul

automat de inferență (*semantic reasoning*), BCGO contribuie la reducerea inconsistențelor în acordul medicilor asupra BCG (contribuție la nivel medical).

- evaluarea calitativă a BCGO folosind metricile de granularitate, gradul de reflecție a domeniului de reprezentat, și gradul de integrare în alte ontologii. Aceasta evaluare calitativă este însoțită de validarea medicală prin *semantic retrieval* și monitorizarea rezultatelor gradării automate. Ontologia este de asemenea integrată în OBO, cu scopul reutilizării și mapării ei în alte ontologii de referință cu care aceasta are legătură.
- Integrarea BCGO în platforma de microscopie virtuală, contribuind astfel la dezvoltarea unei noi direcții, cea de microscopie virtuală cognitivă – MICO, datorită aportului pe care îl aduce ontologia prin indexarea semantică, capacitatea de vizualizare, explorare a slide-urilor de mari dimensiuni și semantic retrieval.

Perspectivile de cercetare pe care le deschide această lucrare vizează în primul rând aspecte legate de *knowledge refining*, cum ar fi includerea conceptelor de scală, importante de altfel în procesul standard de gradare, extinderea teoriei spațiale cu relații geometrice precum și extinderea reprezentării cu concepte temporale. Apoi alte metode de evaluare ar putea fi cercetate, luând în considerare aspectele legate de ontologiile de aplicație versus ontologiile de referință și maparea lor, unde în ontologia din urmă ar putea fi reprezentați formal și alți factori importanți în prognostic, alături de BCG.

O altă direcție de interes o reprezintă complexitatea algoritmului de reasoning, problematica implementării ABox-urilor de mari dimensiuni, precum și analiza BCGO folosind diferite mecanisme de inferență, spre deosebire de Pellet, precum cel implementat în reasoner-ul FACT++ sau RacerPro.

Încapsularea informațiilor histopatologice într-un model ontologic poate fi aplicată altor boli precum ar fi cancerul de prostată. Pe de altă parte, poate contribui la descoperirea de noi cunoștințe folosind ontologia ca suport pentru explorarea imaginilor. Ontologia BCGO poate fi integrată într-un sistem complex care poate conține modele multiple din sisteme biologice și până la date de la nivel radiologic. Direcția aceasta este exploatată de comunitatea VPH (Virtual Physiologica Human) care pune mare accent pe dezvoltarea și simularea modelelor biologice și pe reprezentări formale heterogene integrate, care pot contribui la îmbunătățirea abilităților de prognozare, diagnosticare și de propunere a metodelor personalizate de tratament.

Întrucât ontologiile reprezintă esența web-ului semantic și fiind dezvoltate cu predilecție în domeniile medicale, la finalul lucrării se accentuează ideea ca strategia asistenței în prognoza bazată pe ontologii este o direcție de cercetare de mare relevanță pentru dezvoltarea viitoarelor sisteme de telepatologie sau a celor de Computer-Aided Diagnosis (CAD).