

**SISTEM INTELIGENT DE ASISTARE AUTOMATA
ÎN DIAGNOSTICUL CANCERULUI CERVICAL (PAPDIA)**

- faza a 2-a -

Rezumat

Am dezvoltat primul sistem de diagnostic Papanicolau care funcționează în mod complet automat și care abordează provocările ce sunt întâlnite într-o imagine completă. Pentru abordarea acestor provocări sunt necesare un proces rezonabil de rapid de detecție a nucleelor și un proces de clasificare care funcționează fără ajutorul unui citopatolog. Pentru a demonstra că sistemul nostru de clasificare este competitiv cu alți clasificatori, cele trei proceduri de detecție a nucleelor, descriere și clasificare au fost testate pe baza de date Herlev. Acesta a aratat o performanță relativ bună - o eroare de 0,11. Alte studii au raportat o eroare mai mică, dar se bazează pe rezultatul segmentării efectuate de către un citopatolog [9, 13, 14].

Rezultatul clusterizării așa cum este prezentat în Fig. 10 este extrem de util pentru un citopatolog. Citopatologul poate identifica cu ușurință acele clustere (rânduri), care conțin potențiale nuclee afectate. Această situație este exploatată acum pentru a crea un "set de date de referință". Pentru aceste potențiale clustere, se extrag apoi zonele corespondente mai mari - incluzând întreaga frontieră a celulei - și se salvează ca imagini separate. Citopatologul poate apoi să extragă ușor prin clicuri de mouse aceste zone de imagine, pentru a crea un "set de referință", evitându-se astfel adnotarea manuală pe imagini de mari dimensiuni. Suntem în curs de realizare a unei astfel de "baze de date de referință", întrucât baza de date Herlev reprezintă un set destul de "curat" în care nucleele afectate sunt majoritare. Adevărata provocare va fi de a găsi câteva nuclee afectate într-o multime uriasă de nuclee sănătoase.

Raport de cercetare UPB

Clasificarea automata a cancerului de col uterin

1. Introducere

Sistemul de clasificare a cancerului de col uterin constă din două părți. Prima parte utilizează tehnici de prelucrare a imaginilor pentru a detecta celulele și nucleele lor (secțiunea 2). A doua parte constă în caracterizarea și clasificarea nucleelor detectate folosind tehnici de tip "machine learning" (învățare automatizată); în acest scop, am creat, de asemenea, propria noastră bază de date de celule (secțiunea 3).

Partea de detecție a fost încununată de un succes special, întrucât că am dezvoltat o tehnică nouă de prelucrare a imaginilor cu care putem localiza relativ repede în special nucleele celulelor (punctul 2.1). Suntem acum capabili să procesăm o întreaga imagine microscopica, de cinci miliarde de pixeli, într-o singură zi - alte metode existente necesitând mai multe zile, dacă nu chiar săptămâni. Partea de clasificare s-a dovedit, de asemenea, a fi încununată de succes - alte metode nedemonstrand aplicabilitatea lor la astfel de imagini de mari dimensiuni. Prezentăm astfel primul sistem de lucru care analizează în mod complet automat imaginile microscopice în întregime, în termen rezonabil și fără asistența unui medic. (Studiile existente privind diagnosticul frotiului Papanicolaou s-au concentrat în mod obișnuit pe sarcinile individuale ale

diagnosticului, și anume detectarea ori clasificarea, în mod separat, prin urmare, nu pe demonstrarea folosirii lor într-un sistem complet automat.)

2. Detectia celulelor si a nucleelor

Partea de detectie constă din două faze. O fază este identificarea rapidă a nucleelor potențiale utilizand o metodă de segmentare rapidă (secțiunea 2.1). A doua parte constă aplicarea unei metode de segmentare lente pentru a obține rezultate imbunatatite (secțiunea 2.2).

2.1. Detectia rapida a nucleelor

Majoritatea metodelor de segmentare utilizate în prezent pe imagini ce reprezinta frotiuri Papanicolaou sunt foarte consumatoare de timp, intrucat acestea folosesc adeseori metode de propagare ce presupun un volum ridicat de calcule. Un exemplu obisnuit consta din metodele de tip "level-set", care încep cu un izo-contur si alte cateva trasaturi din imagine și care apoi evolueaza treptat catre o stare optimă [7, 8]. În afară de problema lor legata de durata mare de calcul, nu este clar modul în care aceste metode se comporta în cazul nucleelor afectate, mari, luminoase - cele mai multe studii concentrandu-se asupra nucleelor sănătoase, mici, întunecate. Am dezvoltat, prin urmare, algoritmi noi, și anume analiza de izo-contururi. Pentru a obține puncte de start ("seed points"), vom aplica mai intai o metoda de tip "blob-detection" (secțiunea 2.1.1). Ulterior este efectuată o analiză locală de izo-contururi în jurul fiecarui punct de start, care servește pentru atingerea a două scopuri: unul este eliminarea ulterioara a unor puncte de start și obtinerea unei multimi finale de nuclee candidate (punctul 2.1.2); celalalt scop este de a extrage attributele pentru reprezentarea acestora (secțiunea 3).

2.1.1. Puncte de start

Procesul de detecție a punctelor de start este împărțit în următoarele trei faze, care sunt ilustrate în Fig. 1:

Filtrarea trece-banda si praguirea. Imaginea originală I_{org} este trecuta printr-un filtru trece-banda de tip "difference-of-Gaussian" (DoG), apoi ieșirea I_{DoG} este praguata pentru a obține o imagine alb-negru B_{sink} cu regiunile care corespund pozitiiilor potențiale ale nucleelor, numita acum "image puț" ("sink image"). În scop ilustrativ, în imaginea din partea dreaptă sus în fig. 1, pixelii de interes ("the on-pixels") din B_{sink} se înlocuiesc cu nivelurile de gri corespunzătoare I_{org} . Utilizarea filtrării trece bandă este oarecum similara cu metoda prezentata în [10], dar din motive de complexitate, în acest studiu este utilizat doar un singur nivel în loc de un întreg spațiu.

Transformata de distanță și determinarea axelor simetrice. Transformata de distanță este aplicată imaginii puț B_{sink} [11], urmată de determinarea axelor sale simetrice (sym-axe), folosind un simplu algoritm de detecție de creastă (a se vedea ilustratia din stânga jos în Fig. 1). Rezultatele acestei faze constau într-o listă S_i de segmente de sym-axe ($i = 1, ..n_{sax}$), fiecare dintre ele conținând o listă de coordonate x și y și o valoare pentru distanța simetrică.

Detecția minimelor de luminanță de-a lungul axelor simetrice. Sunt determinate maximele în axele simetrice S_i . Cu toate acestea, valoarea nivelurilor lor de gri corespunzătoare în I_{org} ar putea sa nu corespunda minimului local în I_{org} . Astfel, vom determina intensitatea minimă într-o vecinătate locală în jurul acestor maxime, care apoi sunt luate drept puncte de start P_{seed} . Punctele de start sunt prezentate în graficul din dreapta jos în Fig. 1.

2.1.2. Iso-contururi și selecția candidaților

Analiza izo-contururilor în jurul punctelor de start este identică cu analiza noastră precedentă introdusă în [12]. Mai explicit, din jurul fiecărui punct de start P_{seed} este extrasă o mică zonă asupra careia este aplicată analiza de iso-contururi, candidații mai îndepărtați sunt eliminați, având ca rezultat o mulțime finală de mici zone candidate. Acei candidați - și izo-contururile locale asociate lor - sunt prezentate în Fig. 2. Pentru fiecare astfel de "grup" de izo-contururi, cel cu cea mai mică alungire este selectat ca silueta nucleului. Siluetele nucleelor selectate și interiorul lor sunt apoi descrise la punctul 3.

2.2. Segmentarea automată a nucleelor folosind filtrarea mean-shift

În raportul de cercetare anterior au fost prezentate câteva rezultate preliminare obținute prin aplicarea metodei de filtrare mean-shift (a lui Comaniciu și Meer [16]) asupra unor imagini de frotiuri Papanicolaou.

În această etapă am continuat cercetarea privind aplicarea filtrării *mean-shift* și integrarea acesteia ca un pas preliminar unei metode de segmentare a nucleelor mai eficiente. Scopul segmentării este acela de a extrage din imagine doar micile regiuni reprezentând nucleele celulelor, cu o precizie cât mai ridicată. Extragerea nucleelor este o etapă esențială în lanțul de prelucrare a imaginilor, deoarece pe această etapă se bazează analiza anormalităților celulare (întotdeauna o celulă anormală va prezenta modificări la nivelul nucleului, deoarece procesul de cancerogeneză implică mutații în ADN-ul nuclear). Nucleele anormale sunt de obicei mai închise la culoare, mai voluminoase, uneori cromatina este dispusă neuniform, iar membrana nucleară este neregulată.

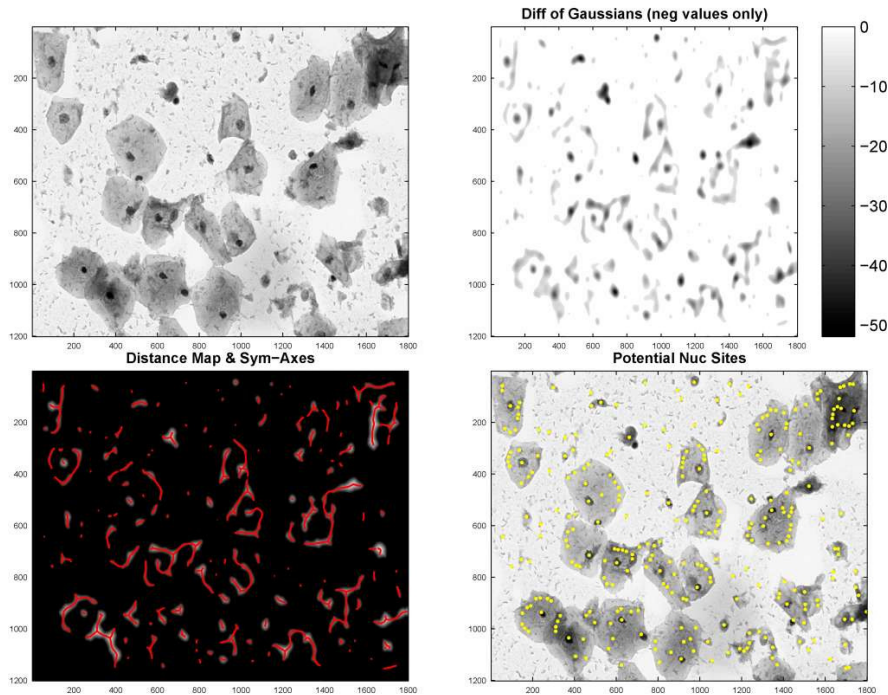


Figura 1: Detecția pozițiilor potențiale ale nucleelor. **Stânga sus:** o imagine tipică de frotiu Papanicolaou (imaginea originală I_{org}). **Dreapta sus:** imaginea puț B_{sink} cu valorile regiunilor luate din I_{DoG} (valorile negative ale I_{org} filtrată printr-un filtru trece-bandă de tip "difference-of-Gaussian" (DoG); valorile pozitive stabilite la 0). **Stânga jos:** Harta de distanțe și axele simetrice (roșii pe imagine) ale B_{sink} . **Dreapta jos:** punctele de start P_{seed} , minimele de luminanță de-a lungul axelor simetrice (galbene pe imagine).

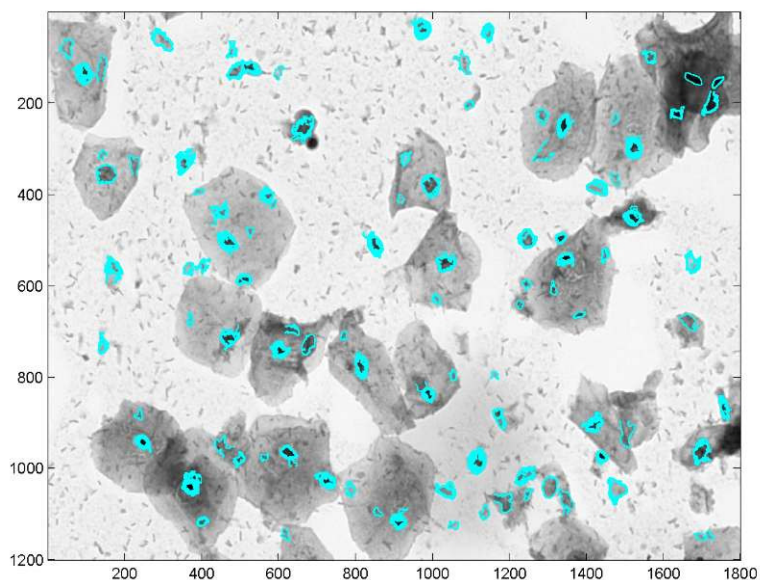


Figura 2: Izo-contuururile din jurul pozitiei nucleelor candidate. Pentru fiecare dintre aceste "clustere" de izo-contururi, cel cu cea mai mica alungire este luat ca silueta nucleului. Pentru aceasta silueta a nucleului vor fi analizate forma și interiorul acesteia, după cum se explică în secțiunea 3.

În Figura 3 a) este prezentată o mică regiune dintr-o imagine, regiune conținând două nuclee (pentru o mai bună vizualizare 'suprafața' imaginii este inversată iar nucleei apar ca niște dealuri). Forma rotunjită a nucleelor (tranzițiile lente ale nivelurilor de gri) reprezintă o piedică majoră în extragerea corectă a frontierei nucleare. În Figura 3 b) este prezentată aceeași regiune filtrată cu mean-shift. Se poate observa imediat că după filtrare regiunea nucleelor apare ca un platou plan (același nivel de gri), iar frontierele sunt foarte abrupte și precis delimitate. Astfel, segmentarea nucleelor va fi mult mai precisă.

Tranzițiile abrupte din zonele de frontieră ale nucleelor ne-a condus al ideea de a aplica un algoritm de extragere a contururilor, pentru a evidenția aceste frontiere. Aceasta deoarece cei mai simpli și rapizi algoritmi de extragere a contururilor sunt de tip derivativ, iar răspunsul lor pe o regiune de tranziție abruptă este foarte puternic. Am aplicat clasicul operator Sobel cu o fereastră de 3x3 pixeli pentru a extrage contururile.

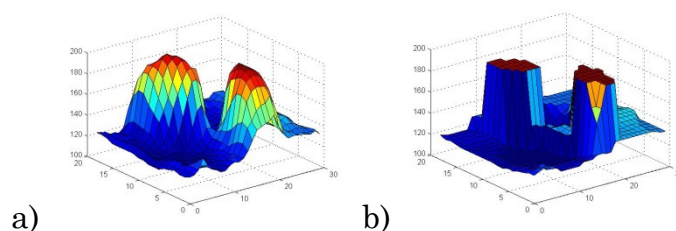


Figura 3. O mică regiune din imagine a) din imaginea originală
b) din imaginea filtrată cu mean-shift

În Figura 4 a) se poate observa o regiune dintr-o imagine de frotiu Papanicolaou; în Figura 4 b) este prezentată imaginea trecută la nivele de gri și filtrată cu mean-shift (se poate observa un contrast mai ridicat), iar în Figura 4c) este prezentat rezultatul obținut după aplicarea operatorului Sobel pe imaginea filtrată cu mean-shift și păstrarea doar a contururilor importante prin prăguire (thresholding). Setarea unui prag pe imaginea de contururi are rostul de a elimina contururile slabe (de exemplu

zgomot, diferențe ale nivelurilor de gri din interiorul citoplasmei sau frontierele între citoplasmă și fundal) și a păstra doar frontierele nucleelor.

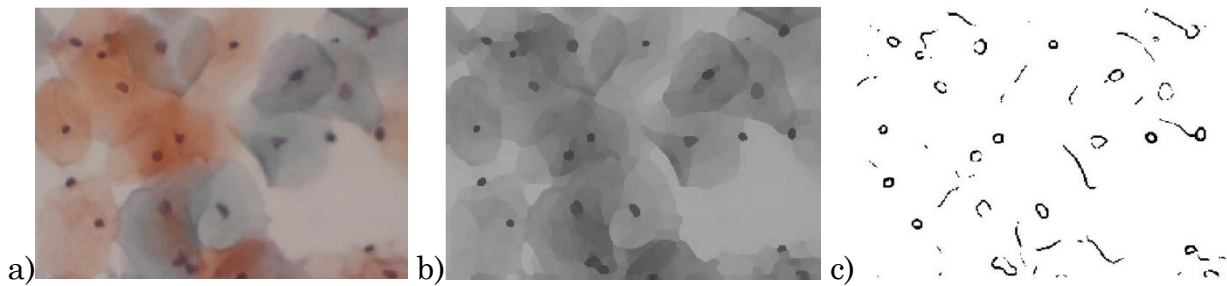


Figura 4. a) imaginea originală b) imaginea filtrată cu mean-shift
c) imaginea celor mai importante contururi

Totuși, după cum se poate observa în Figura 4c) imaginea obținută mai conține regiuni de citoplasmă, iar aceasta impune aplicarea unei metode de segmentare suplimentare.

Am ales pentru acest scop metoda de creștere a regiunilor (region growing), deoarece aceasta funcționează foarte bine atunci când regiunile sunt uniforme (nu există mari diferențe ale nivelurilor de gri în interiorul unei regiuni), iar prin filtrarea mean-shift s-au obținut aceste regiuni uniforme (platourile evidențiate în Figura 3b)). Aceste platouri, reprezentând nucleele sunt de fapt, cum am menționat, regiuni de minimă intensitate (nucleele sunt cele mai întunecate obiecte din imagine).

Segmentarea bazată pe creșterea regiunilor funcționează după următorul mecanism: se aleg niște puncte de start (semințe sau "seeds") în imagine și apoi se aglomerează pixeli la fiecare din aceste puncte, uniform în toate direcțiile, până se atinge un prag stabilit. Așadar un punct de start bine plasat în cea mai joasă (închisă la culoare) regiune a nucleului va duce la o segmentare completă a aceluia nucleu, pragul fiind stabilit ca o anumită fracțiune din zona de frontieră a nucleului. În algoritmul nostru, ne-am folosit de imaginea de maxime de contururi (Figura 4c)) pentru a stabili punctele de start pentru creșterea regiunilor. Mai precis, câte un punct de start a fost ales pentru fiecare fracțiune continuă de contur, căutând întâi minimul din acea fracțiune, apoi minimul dintr-o vecinătate de 3x3 pixeli în jurul aceluia minim. Astfel algoritmul setează punctele de start în zonele cele mai întunecate ale nucleelor.

După mai multe teste, un pas suplimentar a fost inclus, pentru a elimina posibilele regiuni de citoplasmă mai întunecată: când se obțin două regiuni conexe (lipite), se alege regiunea care are valoarea medie a nivelelor de gri cea mai mică (regiunea nucleului care e mai întunecat).

În Figura 5b) este prezentat rezultatul obținut după aplicarea algoritmului de creștere a regiunilor. Se observă că în imagine apar și obiecte (fals-pozitive) care nu reprezintă nuclee de celule, ci regiuni mai întunecate de citoplasmă. Aceste regiuni au fost eliminate printr-un pas de analiză morfologică a imaginii, mai precis: eliminarea obiectelor care sunt prea mici sau prea mari pentru a fi considerate nuclee, și eliminarea obiectelor prea alungite pentru a fi nuclee.

Putem stabili un prag pe dimensiunea nucleelor deoarece cunoaștem rezoluția de scanare (40x). Pentru eliminarea obiectelor alungite am folosit raportul de excentricitate λ_1 / λ_2 , unde λ_i sunt valorile proprii ale matricei de covarianță a coordonatelor pixelilor aparținând obiectului analizat. Am ales un prag (de exemplu

0.2) pentru excentricitate, eliminând astfel obiectele prea alungite (știm că nucleele sunt rotunzi iar excentricitatea unui disc ideal este 1).



Figura 5. a) imaginea originală b) imaginea după creșterea regiunilor
c) rezultatul final - imaginea segmentată

În Figura 5c) este prezentat rezultatul final al segmentării; nucleele segmentate sunt evidențiate cu o bordură albă. Pentru validarea experimentală, am testat algoritmul pe două baze de date de imagini: una publică disponibilă pe Internet (cele 38 de imagini asociate articolului [17] - Plissiti ș.a, fiecare imagine având o rezoluție de 1536x2048, numărul total de nuclee fiind 5617), și o bază de date de imagini din cele achiziționate în cadrul proiectului (10 imagini cu o rezoluție de aprox. 500x500 și un număr total de aprox. 400 nuclee). Toate imaginile utilizate au avut o versiune "de referinta" pentru a putea verifica acuratețea segmentării.

În Figura 6 este prezentată curba ROC (Receiver operator characteristic curve) pentru baza de date de imagini "Plissiti", în funcție de pragul ales pentru binarizarea imaginii de gradient (cea obținută cu operatorul Sobel). Pragul optim a fost 50, pentru care am obținut o sensibilitate (true positive rate, TPR) de 90.83% și o specificitate (true negative rate, TNR) de 98.62%. TPR-ul maxim obținut de Plissiti (în articolul [17]) a fost de aproximativ 91%, dar cu un TNR de 75%, care este mult mai mic decât cel obținut de noi.

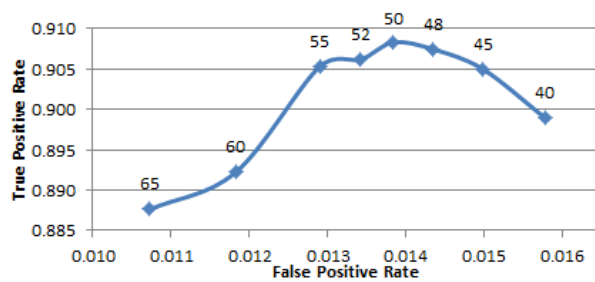


Figura 6. Curba ROC pentru baza de date de imagini Plissiti
(pragurile de gradient sunt marcate pe grafic)

Tabel 1. Rezultatele detecției de nuclee

<i>Baza de date</i>	<i>Sensibilitate</i>	<i>Specificitate</i>
Plissiti	91.21%	98.65%
Proprie	98.23%	95.53%

În tabelul 1 sunt prezentate rezultatele detecției de nuclee pentru cele două baze de date de imagini de test. Aceste rezultate au fost obținute după o îmbunătățire a metodei, și anume adăugarea unei etape de alegere automată a pragului de gradient în funcție de

contrastul imaginii (a fost utilizată o măsură a frecvențelor înalte din spectrul Fourier al imaginii). Astfel, dacă imaginea are un contrast scăzut, se folosește un prag mai scăzut, și vice-versa.

3. Clasificarea

În alte studii privind clasificarea frotiurilor Papanicolaou, celulele de col uterin și nucleeele lor au fost redată prin 20 de atribute care descriu cele mai simple caracteristici geometrice și de aspect extrase atât din nucle, cât și din forma și regiunile celulei (tabelul III de la pagina 73 din [9], [13], [14]). Dar acestea au fost testate doar pe o multime mică de date, cu mai puțin de o mie de celule. Realizările lor nu reflectă realitatea de fapt - și anume prezența a zeci de mii de celule. În continuare vom efectua clasificarea pe baza izo-contururilor, așa cum este utilizată pentru detecție în secțiunea 2. Astfel, se economisește durata de calcul și se obțin rezultate excelente.

3.1. Extragerea trasaturilor

Atributele nucleelor care au fost extrase în alte studii privesc forma lor geometrică - cel mai scurt diametru, cel mai lung diametru, alungirea și rotunjimea -, precum și aspectul său - stralucirea, numărul de maxime și minime [9]. Aici am ales în plus aspecte texturale pe care le luăm în mod convenabil de la izo-contururile pe care le-am folosit pentru a detecta nucleeele.

Pentru a descrie frontiera regiunii - silueta nucleului - folosim transformata Fourier aplicată unidimensional pe semnatura radială a frontierei. Studiile privind recuperarea formei au arătat că aceasta este descrierea cea mai eficientă a formei în ceea ce privește complexitatea spațială și temporală [15]: prin comparație, descrieri ale formei care prezintă o mai mare performanță de recuperare folosesc o complexitate spațială și temporală excesivă. Mai concret, pentru o frontieră $B(s)$ cu lungime arc s variabilă, se determină semnătura ei radială $R(s)$, din care sunt extrasi primii patru descriptori (fast) Fourier, f_1 , f_2 , f_3 și f_4 . Mai mulți descriptori Fourier nu îmbunătățesc performanța în mod semnificativ.

Sunt determinate următoarele trasaturi de aspect pentru regiunea nucleului: intensitatea sa medie i_{mean} , deviația standard a valorilor intensității pixelilor săi i_{std} , precum și contrastul, și intervalul de valori ale intensității i_{rng} .

Pentru a capta aspectele texturale, sunt determinate în primul rând izo-contururile din interiorul unei siluete a nucleului (Fig. 7). Acest număr al izo-contururilor din interior, n_{iso} , poate varia între câteva - tipic pentru nucleeele mici sanatoase - și câteva zeci - tipic pentru nucleeele umflate.

Pentru a capta unele aspecte structurale, sunt analizate două distribuții. Una dintre ele este distribuția razelor $R(i)$, care este multimea ordonată a razelor medii ale izo-contururilor ($i = 1, \dots, n_{\text{iso}}$).

Distribuția $R(i)$ este ilustrată în graficul din dreapta jos al figurilor ce arată exemple de nucleee (fig. 7, 8 și 9). Prima valoare a distribuției razelor este raza (medie) a siluetei nucleului, ultima valoare corespunde celui mai mic izo-contur.

Cealaltă distribuție este nivelul de luminanță al izo-contururilor (graficul din dreapta sus în figuri), în care ordonarea este efectuată în concordanță cu distribuția razelor, de unde și "neregulile" accidentale de pe partea dreaptă a distribuției.

Aceste două distribuții pot fi caracteristice pentru diferite tipuri de nuclee, lucru care este demonstrat pe trei celule din baza de date Herlev, un nucleu normal (sanatos) (Fig. 7), un nucleu afectat ușor (Fig. 8) și nucleu cancerigen (Fig. 9).

Scopul este, prin urmare, de a caracteriza aceste distribuții prin cativa parametri. Acest lucru se face cu ajutorul declinului liniar între primul și ultimul punct al distribuției, prezentat prin linia punctată pe grafice, pentru care se măsoară marimea "abaterii" în domeniul pozitiv și negativ (graficul din stânga jos în figuri). Pentru distribuția razelor, se determina atât abaterea pozitivă, cât și abaterea negativă, normalizate prin intervalul de valori ale razelor, δ_R^+ și respectiv, δ_R^- . Pentru nivelul de luminanță se determina numai abaterea negativă δ_L^- , intrucat o abatere pozitivă este aproape imposibila datorită valorilor mai mici ale intensității în regiunea nucleului.

Din toate măsurile atributelor specificate mai sus, se formeaza vectorul atributelor $\mathbf{n}$:

$$\mathbf{n} = [a, i_{\text{mean}}, i_{\text{std}}, i_{\text{rng}}, f_1, f_2, f_3, f_4, r_{\text{min}}, r_{\text{mean}}, \delta_R^+, \delta_R^-, \delta_L^-, n_{\text{iso}}] \quad (1)$$

unde a este aria nucleului și r_{min} este raza minimă, $\min_i R(i)$, iar r_{mean} este raza medie, $\sum_i R(i) \cdot 1/n_{\text{iso}}$.

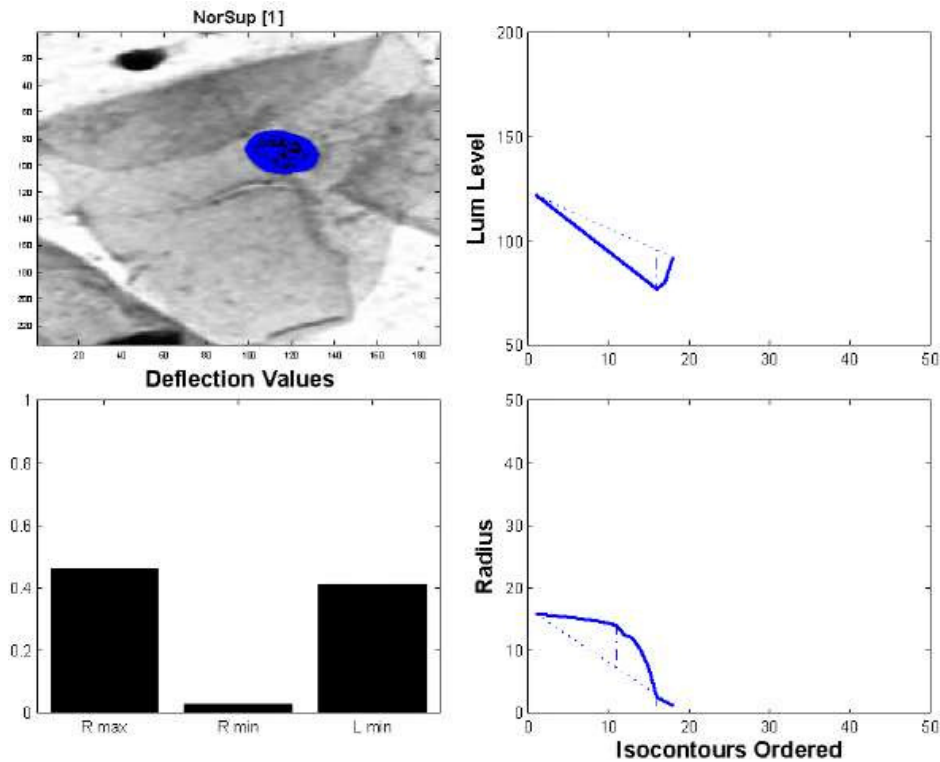


Figura 7: Iso-contururi pentru un nucleu normal (din baza de date Herlev). **Dreapta sus:** nivelul de luminanta (intensitate) raportat la izo-contururile ordonate in functie de $R(i)$. Linia punctata oblică conectează liniar prima si ultima valoare. Linia intrerupta verticală indică declinul liniar; abaterea negativa este marcata printr-o linie punctat-intreruptă. **Dreapta jos:** razele izo-contururilor, sortate în ordine descrescătoare. Numai abaterea pozitivă este vizibila in mod clar. **Stânga jos:** valorile abaterilor δ_R^+ , δ_R^- și δ_L^- .

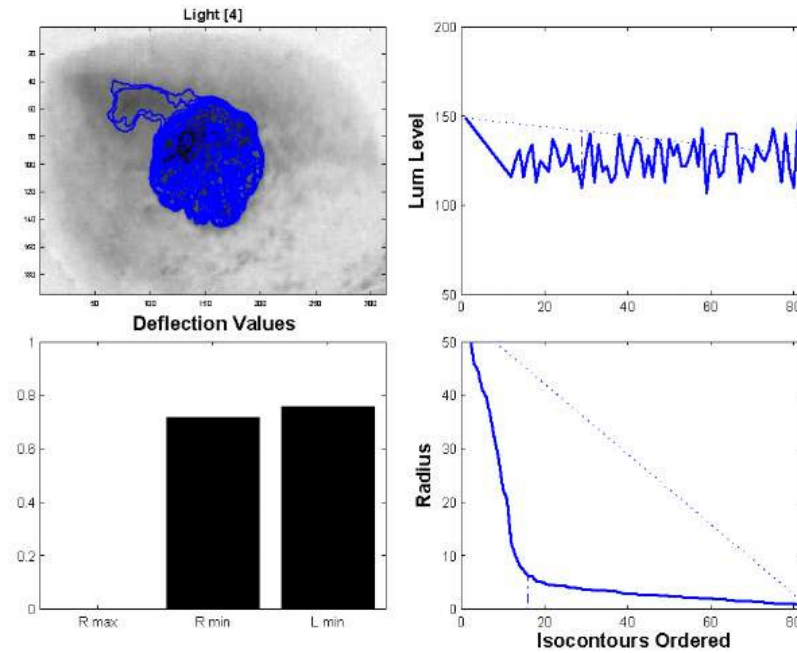


Figura 8: Izo-contururile unui nucleu usor afectat (baza de date Herlev). Numarul de izo-contururi din interior n_{iso} este mult mai mare.

3.2. Evaluarea

Se folosesc două seturi de date, setul Herlev și propriul nostru set. Baza de date Herlev este utilă pentru compararea metodelor: vom explica ce performanță de clasificare am obținut pe ea (secțiunea 3.2.1). Apoi, vom ilustra performanta de clusterizare pe nuclee detectate pe imagini complete, unde ne confruntăm si cu alte structuri celulare (secțiunea 3.2.2).

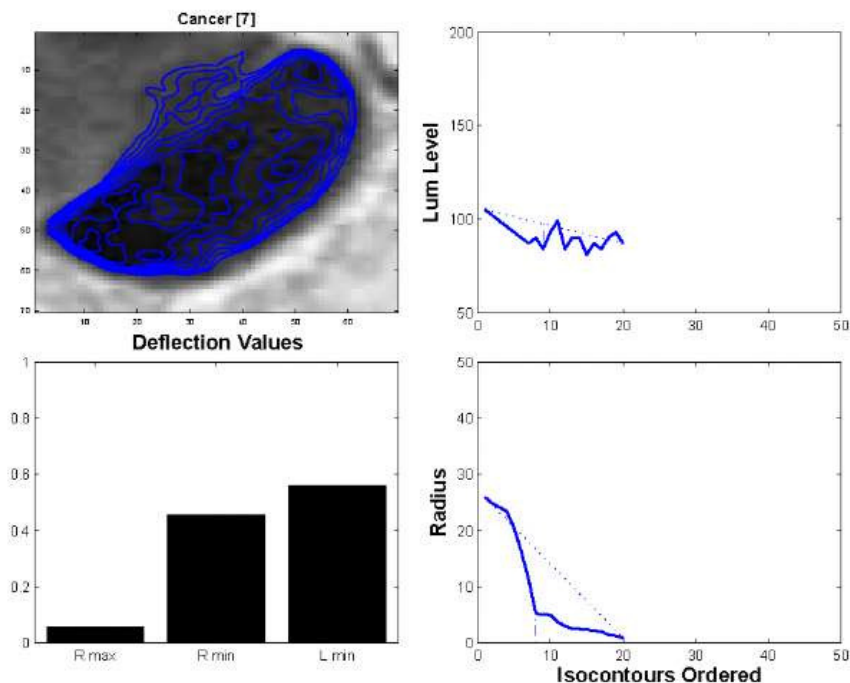


Figura 9: Izo-contururile unui nucleu cancerigen (baza de date Herlev). Notă: scala nu este aceeași ca în figura anterioară - celulele cancerigene sunt doar de 3 ori mai mare decât un nucleu normal.

3.2.1. Performanța de clasificare pe baza de date Herlev

Baza de date Herlev conține 917 de imagini de celule [9] din 7 clase de celule (în funcție de sistemul de Bethesda): trei clase de celule sanatoase și patru clase de celule afectate, cu un număr inegal de celule în fiecare clasă. Clasele de celule sănătoase sunt de tip "normal-superficial", "normal-intermediar", "normal-columnar"; clasele de celule afectate sunt de tip "usor", "moderat", "sever" și "cancerigen" (a se vedea secțiunea 4.2.2 din [9]).

Fiecare imagine este însoțită de o imagine segmentată în care în care nucleeele și citoplasma celulelor au fost etichetate de un medic citopatolog. Aceste hărți de segmentare au fost exploatate prin alte studii referitoare la clasificare pentru a extrage caracteristicile din regiunile corespondente [9, 13, 14], prin urmare, acele studii se bazează pe segmentarea umană. În schimb, în experimentele noastre am folosit silueta nucleului - așa cum a fost obținută în urma aplicării metodei noastre de detecție (secțiunea 2) - și, în acest fel, nu am făcut deloc uz de aceste hărți de segmentare realizate prin intervenție umană.

Ca și în alte studii, am testat o clasificare a tuturor celor 7 clase și o altă clasificare ce discriminează numai între 2 clase: sănătoase și afectate. Pentru cazul cu 7 clase, am utilizat un clasificator SVM, iar ca eticheta a clasei a fost ales maximul posterior rezultat dintr-o analiză "1-contra-toti".

Am testat un număr diferit de echipartiții ale bazei de date (trei, cinci, opt), dar performanțele abia au diferit, așa ca raportăm doar pe cel pentru clasificarea recomandată ce utilizează cinci echipartiții. Pentru 7 clase, precizia medie este de 50,1 la sută. Pentru două clase, valorile preciziei, sensibilității (recall) și erorii sunt 0.95, 0.95 și, respectiv, 0.11.

3.2.2. Clusterizarea pe imagini complete

Imaginile noastre sunt obținute cu un scanner VENTANA iScan Coreo, cu un zoom optic de 40. Detectia nucleelor durează aproximativ o zi pe o imagine completă, folosind un procesor de 2.66 GHz Intel cu un cod scris în Matlab. Sunt găsiți până la 10^5 candidați. Vectorii corespunzători nucleelor sunt grupați cu un algoritm k-Means, cu k ales în mod arbitrar 30 - mai mare decât numărul de clase, întrucât ne așteptăm, de asemenea, să găsim și alte structuri celulare. Pentru fiecare cluster, se determină un centru, prin medierea membrilor săi și distanțele sunt ordonate crescător. Fig. 10 arată primele 10 exemplare ale fiecărui cluster (pe fiecare rand) pentru o imagine completă. Clusterizarea grupează și structuri celulare - acestea sunt prezente, deoarece detectia nucleelor nu este perfectă și, de asemenea, identifică alte structuri mici, cum ar fi granulocitele. Cu toate acestea, această sortare s-a dovedit extrem de utilă în crearea unui "set de date de referință", care urmează să fie discutat în secțiunea finală.

3.2.4. Setul de date de referință

Nucleeele au fost clasificate în zece tipuri, dintre care trei sunt patologice, iar șapte sunt normale. Fig. 13 arată distribuția nucleelor găsite pentru fiecare tip, precum și distribuția totală. Au fost colectate în total 7269 de nucleee, din care 1,074 sunt cazuri patologice. Aceasta este de aproximativ trei ori mai mult ca populara bază de date Herlev.

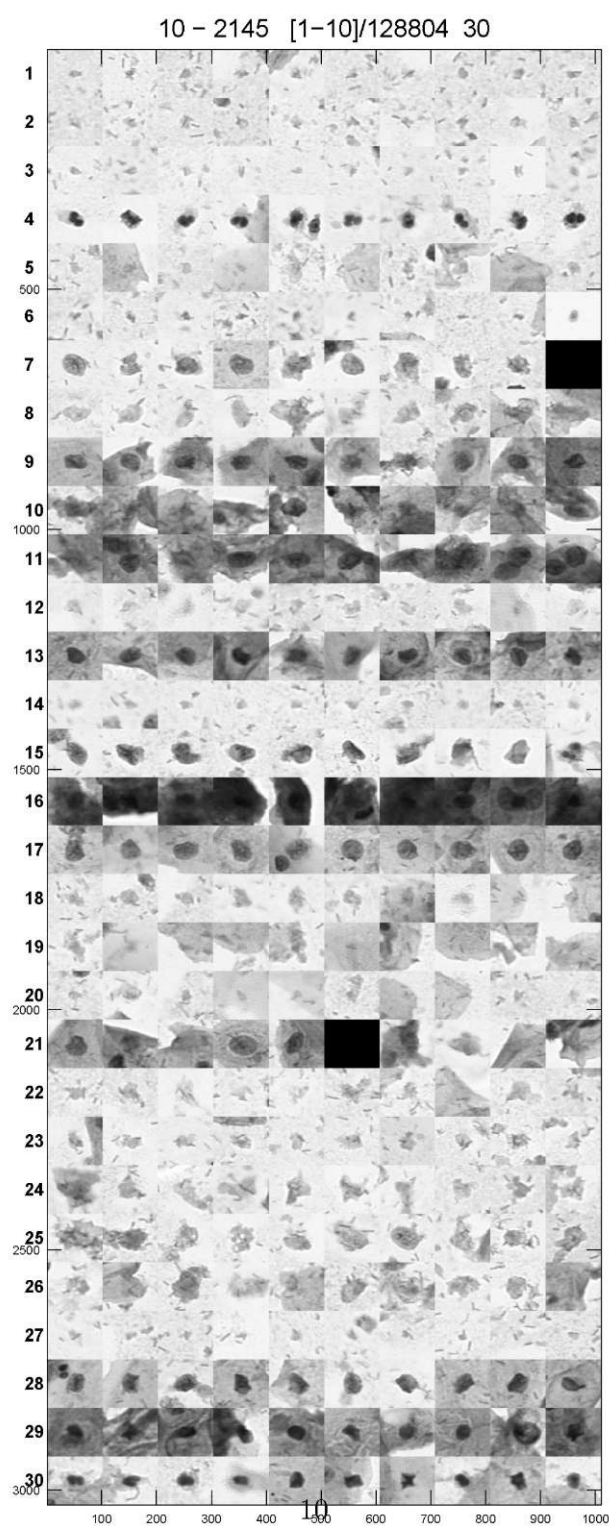


Figura 10: Exemple de clustere create cu
algoritmul k-Means cu $k=30$, luate dintr-o
imagine intreaga. Fiecare rand arata cele mai
apropiate 10 nuclee candidate. Imaginile goale
(negre) sunt candidati ignorati fiind situati
langa o frontiera

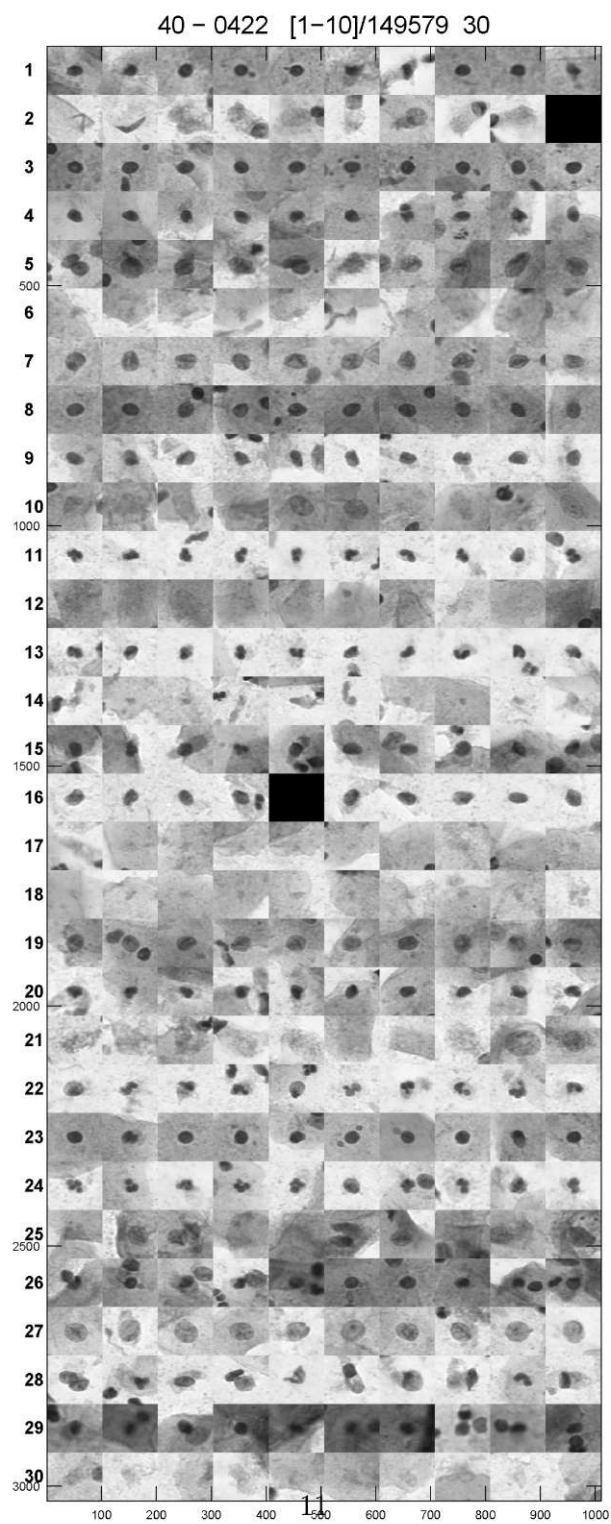


Figura 11: Exemple de clustere (ca in figura precedenta).

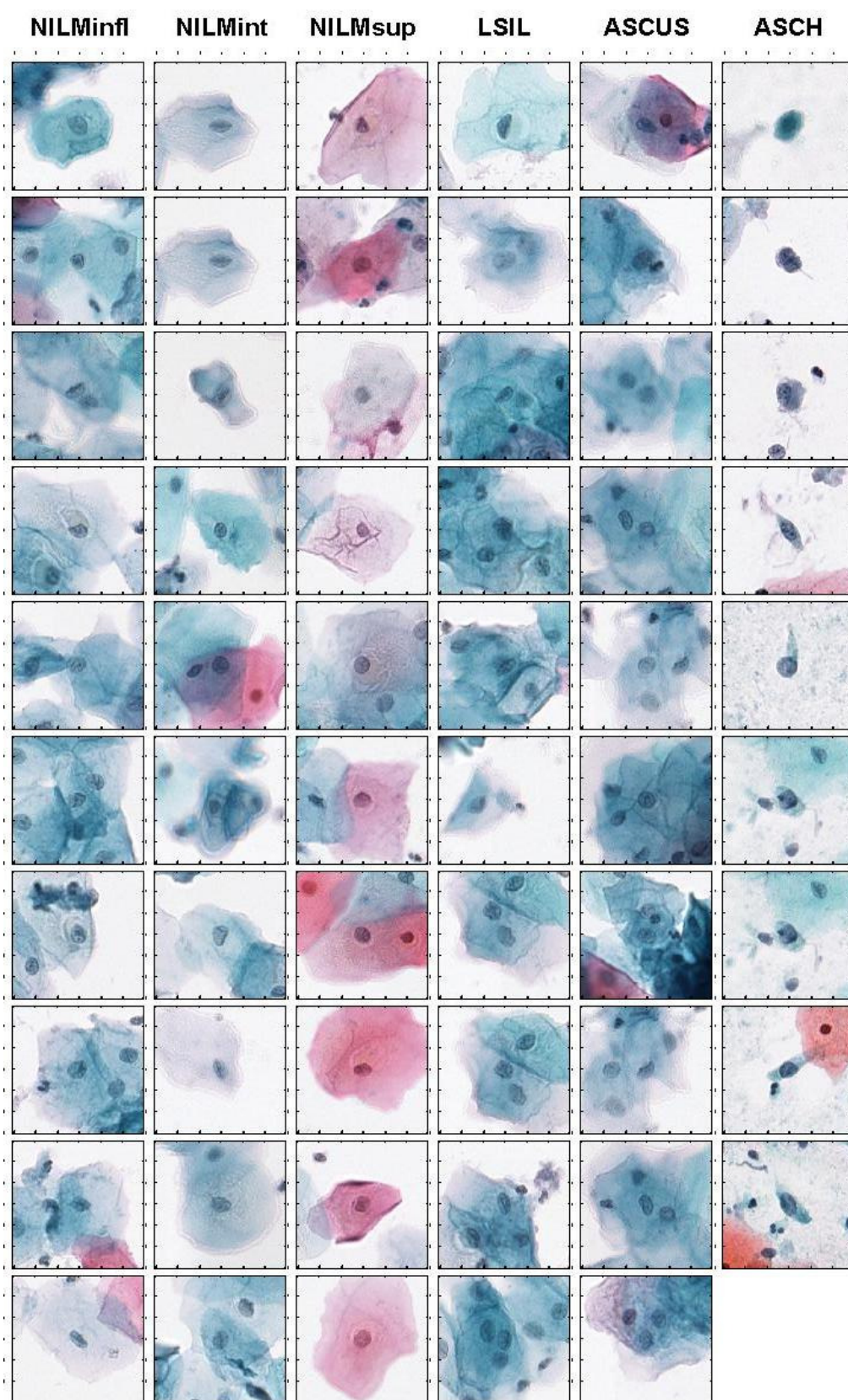


Fig. 12: Exemple de nuclee clasificate de catre un histopatolog

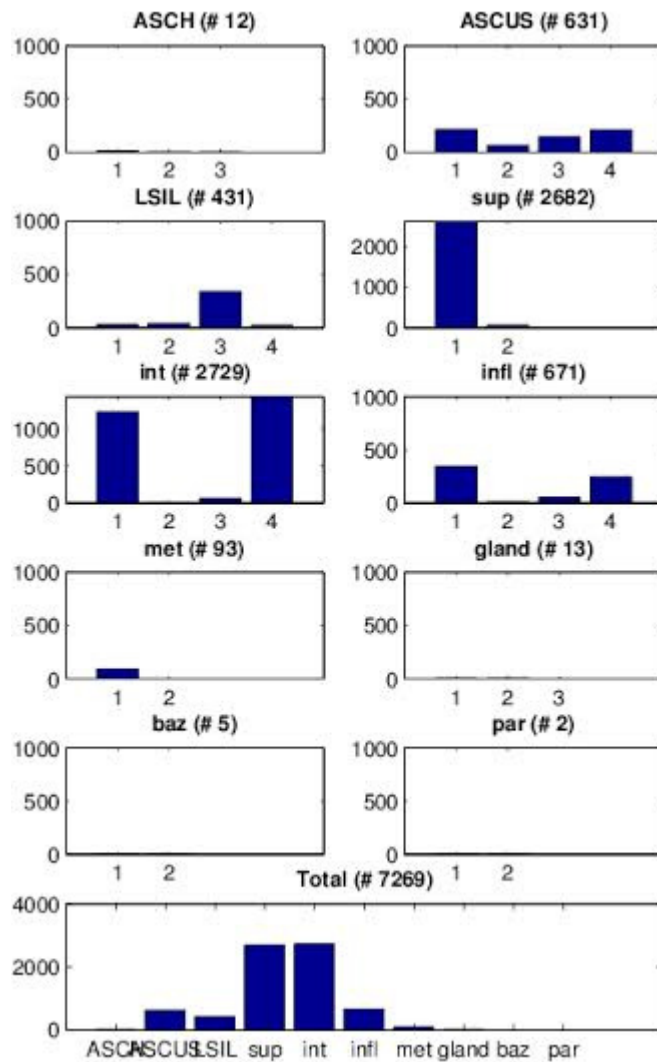


Fig. 12 - distribuția nucleelor pe fiecare tip și distribuția totală.

Referinte

- [1] C. Duanggate, B. Uyyanonvara, and T. Koanantakul, "A review of image analysis and pattern classification techniques for automatic pap smear screening process," in *The 2008 International Conference on Embedded Systems and Intelligent Technology February*, February 27-29, 2008.
- [2] S. N. Sulaiman, N. A. Mat Isa, and N. H. Othman, "Semi-automated pseudo colour features extraction technique for cervical cancer's pap smear images," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, vol. 15, no. 3, pp. 131-143, 2011.
- [3] M. E. Plissiti, C. Nikou, and A. Charchanti, "Combining shape, texture and intensity features for cell nucleee extraction in pap smear images," *Pattern Recognition Letters*, vol. 32, no. 6, pp. 838-853, 2011.
- [4] R. Moshavegh, B. E. Bejnordi, A. Mehnert, K. Sujathan, P. Malm, and E. Bengtsson, "Automated segmentation of free-lying cell nucleee in pap smears for malignancy-associated change analysis," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 5372-5375.

- [5] K. Li, Z. Lu, W. Liu, and J. Yin, "Cytoplasm and nucleus segmentation in cervical smear images using radiating GVF snake," *Pattern Recognition*, vol. 45, no. 4, pp. 1255-1264, 2012.
- [6] M. E. Plissiti, C. Nikou, and A. Charchanti, "Automated detection of cell nucleee in pap smear images using morphological reconstruction and clustering," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 15, no. 2, pp. 233-241, 2011.
- [7] J. Cheng and J. C. Rajapakse, "Segmentation of clustered nucleee with shape markers and marking function," *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 3, pp. 741-748, 2009.
- [8] Z. Lu, G. Carneiro, and A. P. Bradley, "Automated nucleus and cytoplasm segmentation of overlapping cervical cells," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2013*, ser. Lecture Notes in Computer Science, K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds. Springer Berlin Heidelberg, 2013, vol. 8149, pp. 452-460.
- [9] Y. Marinakis, G. Dounias, and J. Jantzen, "Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification," *Computers in Biology and Medicine*, vol. 39, no. 1, pp. 69-78, 2009.
- [10] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nucleee in histopathology images," *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 4, pp. 841-852, 2010.
- [11] C. R. Maurer Jr, R. Qi, and V. Raghavan, "A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 2, pp. 265-270, 2003.
- [12] C. Rasche, S. Oprisescu, A. Sultana, and T. Radulescu, "Analysis of pap smear images with iso- and edge-contours," in *IEEE 11th International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, RO*, 2015.
- [13] A. Genctav, S. Aksoy, and S. Onder, "Unsupervised segmentation and classification of cervical cell images," *Pattern Recognition*, vol. 45, no. 12, pp. 4151-4168, 2012.
- [14] M. E. Plissiti and C. Nikou, "Cervical cell classification based exclusively on nucleus features," in *Image Analysis and Recognition*. Springer, 2012, pp. 483-490.
- [15] D. Zhang and G. Lu, "Study and evaluation of different fourier methods for image retrieval," *Image and Vision Computing*, vol. 23, pp. 33-49, 2005.
- [16] Comaniciu, Dorin, and Peter Meer. "Mean shift: A robust approach toward feature space analysis." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24.5 (2002): 603-619.
- [17] M. E. Plissiti, C. Nikou, and A. Charchanti, "Automated detection of cell nucleee in pap smear images using morphological reconstruction and clustering," *Inf. Techn. in Biomedicine, IEEE Trans. on*, vol. 15, no. 2, pp. 233-241, 2011.

Raport de cercetare GeneticLab

Activitatea Actualizarea bazei de date cu imagini

În această etapă a proiectului se urmărește mărirea bazei de date cu noi imagini de citologie lichida în vederea acoperirii unei palete cât mai largi de probe și clasificate conform Sistemului de raportare a citologiei colului uterin Bethesda 2001.

În acest scop s-au recoltat și scanat peste 450 de probe în cadrul clinicii, urmărind protocoalele raportate în prima fază:

1. Prelevare
2. Colorare
3. Analiza anatomo-patolog
4. Eliberare buletin
5. Scanare lame citologice
6. Arhivare

Metodă - citologie în mediu lichid, colorație Papanicolaou.

Toate analizele și experimentele au fost făcute într-o configurație de laborator de citologie clasică. Probele au fost prelevate de la pacienți care au fost îndrumate de către medici ginecologi pentru controale de rutină periodice. La cele mai multe s-a făcut în paralel și controlul infecției cu virusul HPV, HCV și HBV. Controlul virusurilor hepatice este realizat în vederea protejării operatorilor citologi care vin în contact cu probele prelevate. Lamele au fost analizate de către anatomo-patologi, după care au fost preluate imagini pe un microscop Olympus C21 cu camera Olympus C7070 la o rezoluție de 7 Mpixeli. Pentru fiecare pacientă s-au achiziționat cel puțin 30 de câmpuri de microscopie la o mărime de 20X. Formatul imaginilor a fost JPEG pentru a micșora mărimea fișierelor.

Lamele analizate au fost arhivate prin digitizarea acestora cu un scanner histologic achiziționat în prima etapă a proiectului, iScan Coreo de la Roche Diagnostic:

Caracteristicile tehnice ale acestui aparat sunt următoarele:

- Capacitate de scanare continuă până la 160 de lame
- Viteză mare de scanare
- Magnificare optică selecționabilă 10 X, 20 X, 40 X
- Identificare automată a regiunilor tisulare
- Posibilitatea scanării în volum cu Z plan
- Formatul fișierelor este TIFF
- Software de vizualizare a imaginilor.

Mărimea fișierelor rezultate este variabilă și depinde de rezoluția de scanare optată. Aceasta poate varia pentru lamele de citologie lichida între 1 Gb și 50 Gb. Mărimea foarte mare a fișierelor rezultate ridică și probleme de stocare.

S-a urmărit cu precădere selecționarea cazurilor cu clase ridicate de patologie. Aceste cazuri sunt descoperite accidental în urma unor controale de rutină și sunt extrem de rare. Marea majoritate a cazurilor sunt normale sau cu modificări minore. Pentru actualizarea bazei de date cu cazuri mai grave s-au inițiat colaborări cu spitale din București.

Activitatea Actualizare platformei web a proiectului

S-a discutat și stabilit cu partenerul CO modalitatea de realizare și structura bazei de date aferentă proiectului. S-au elaborat modalitățile de upload a imaginilor, formatul acestora, dimensiunea maximă posibilă și modalitatea de adnotare. S-a actualizat baza de date aferentă proiectului cu toate imaginile scanate. S-a proiectat o a doua platformă web cu titlul www.pathonet.ro. Aceasta platformă va permite cercetătorilor implicați în anatomie patologică să scaneze și să primească imaginile digitizate direct de pe site. Termenul de finalizare al platformei este de 31 Martie 2016.

Activitatea Proiectare și implementare metode de segmentare

In cadrul acestei activitati s-a elaborat impreuna cu partenerul CO si descrise de acesta in raport, a modalitatilor de segmentare a imaginilor.

Activitatea Selectia trasaturilor relevante specifice imaginilor citologice

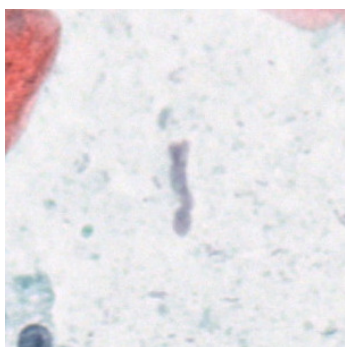
Din cauza dimensiunilor foarte mari a imaginilor scanate s-a stabilit numarul si tipul trasaturilor care trebuie reprezinte fiecare imagine astfel incat acestea sa poata fi utilizate in procesul de training si de recunoastere de catre software-ul propus.

Sunt necesare mai multe conditii obligatorii:

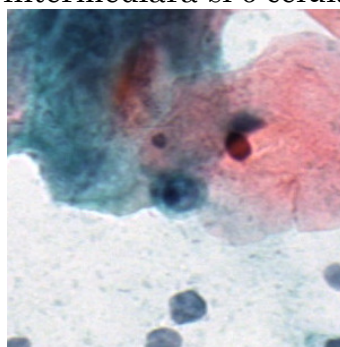
- dimensiunea acestor imagini sa fie cat mai mica
- fiecare imagine sa contina cat mai putine celule in camp si acestea sa fie uniforme
- celule existente in camp sa fie clasificate si adnotate conform sistemului Bethesda 2001. O prezentare sumara a acestui tip de clasificare interna este prezentat mai jos:

Tip celula/imagine exemplu

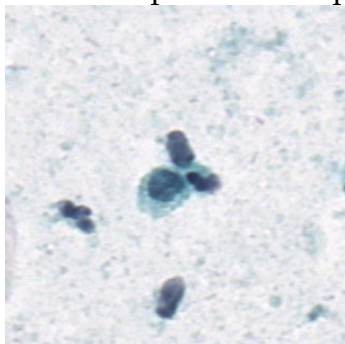
1. Artefact.



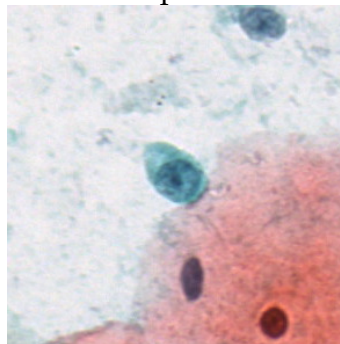
2. Celula bazala suprapusa cu o celula intermediara si o celula superficiala.



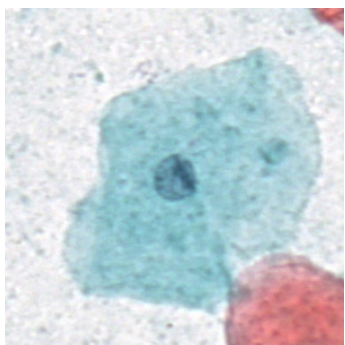
3. Celula bazala. In fundal se observa citoliza si prezenta de polimorfonucleate.



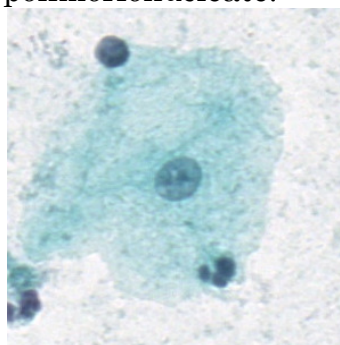
4. Celula bazala. In vecinatate se observa 2 celule superficiale suprapuse.



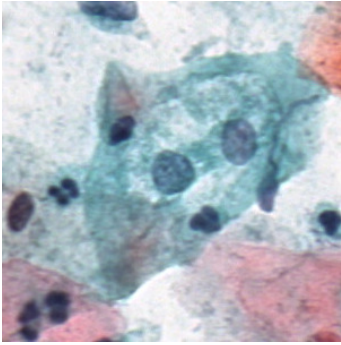
5. Celula intermediara.



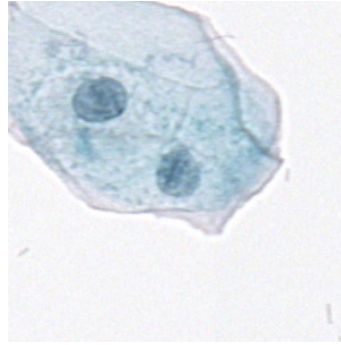
6. Celula intermediara acoperita de polimorfonucleate.



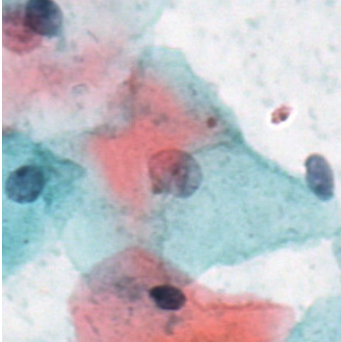
7. Celula intermediara binucleata, cu nucleii usor crescuti de volum, ce schiteaza halou perinuclear.



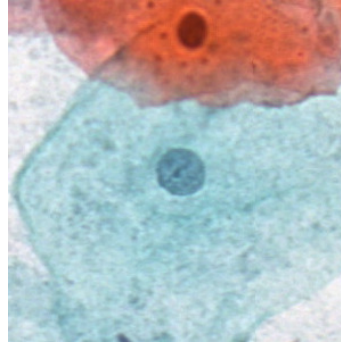
8. Celula intermediara binucleata, cu nucleii usor crescuti de volum.



9. Celula intermediara cu modificari celulare reactive, falsa eozinofilie.



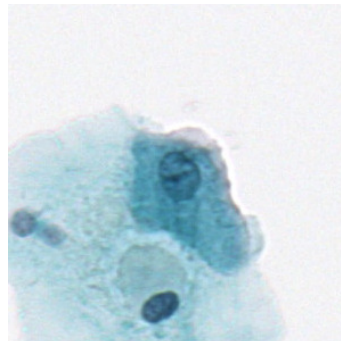
10. Celula intermediara suprapusa cu o celula superficiala.



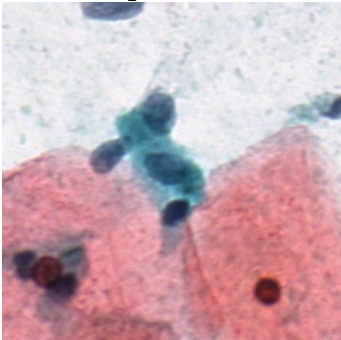
11. Celula metaplaziata.



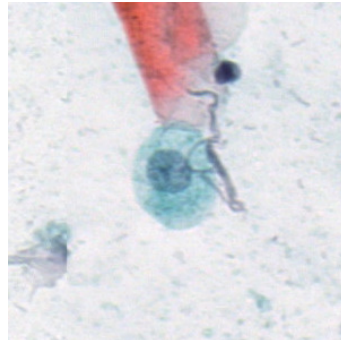
12. Celula metaplaziata suprapusa cu celule intermediare.



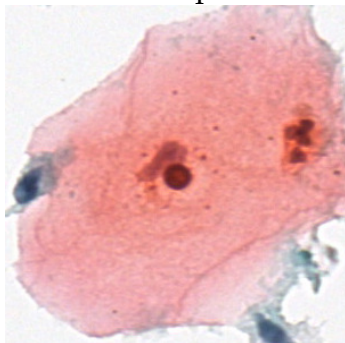
13. Celula metaplaziata suprapusa cu celule superficiale si nucleii denudati.



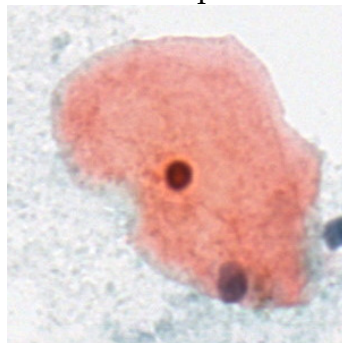
14. Celula parabazala.



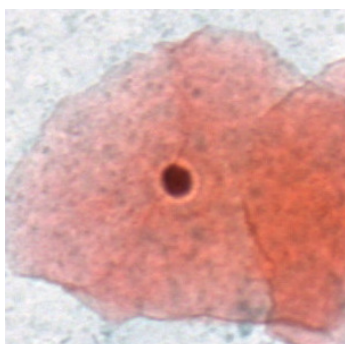
15. Celula superficiala.



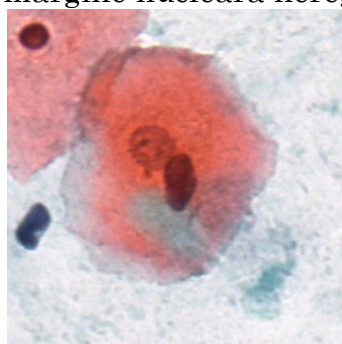
16. Celula superficiala.



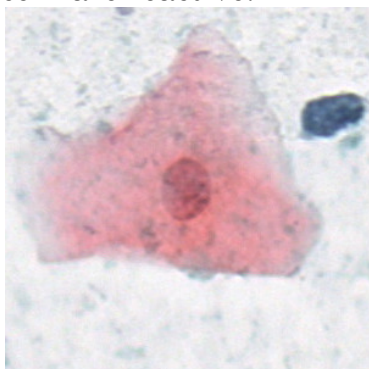
17. Celula superficiala.



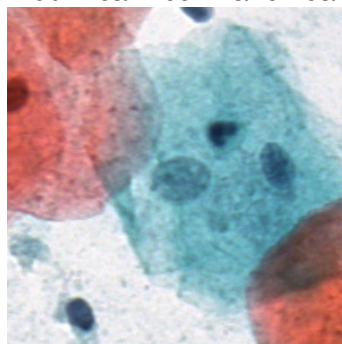
18. Celula superficiala binucleata, cu
nuclei crescuti de volum (x2,5-3), cu
margine nucleara neregulata, hiperchromi.



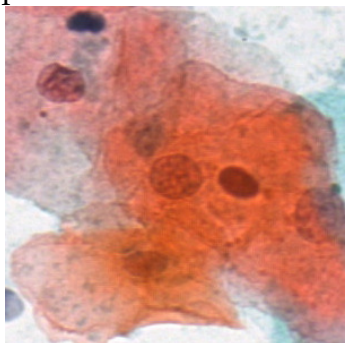
19. Celula superficiala cu modificari
celulare reactive.



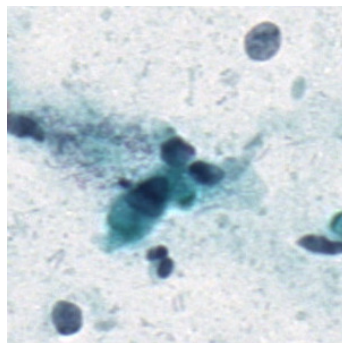
20. Celule intermediare suprapuse cu
modificari celulare reactive.



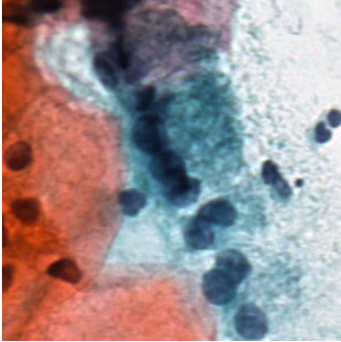
21. Celule superficiale suprapuse, unele cu
nuclei hipertrofi (x2,5-3), usor hiperchromi,
pleomorfism nuclear.



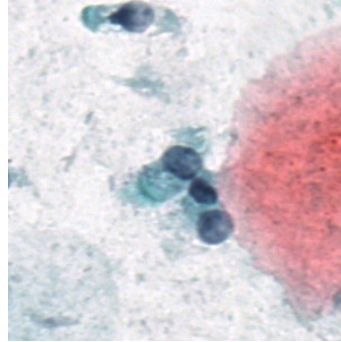
22. Grup de celule glandulare
endocervicale.



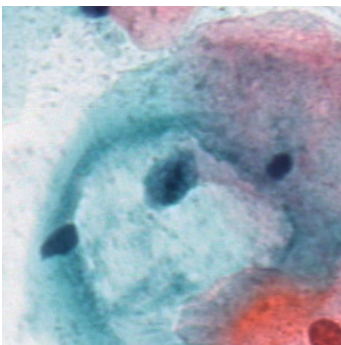
23. Grup de celule glandulare endocervicale.



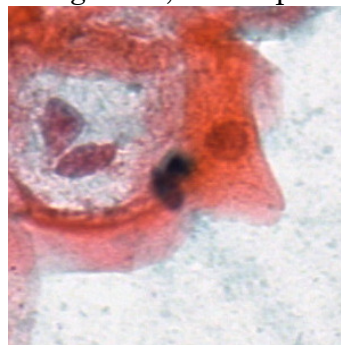
24. Grup de celule glandulare.



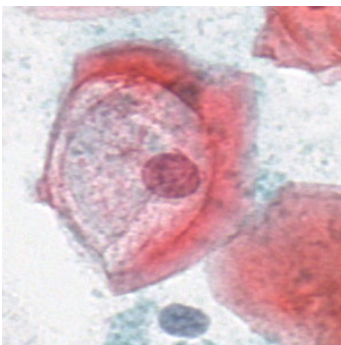
25. Koilocit. Celula intermediara cu nucleu crescut de volum (x3-4), hipercrom, cu margine nucleara neregulata.



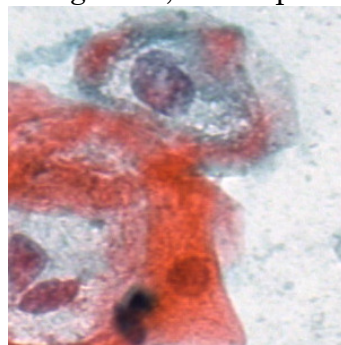
26. Koilocit. Celula superficiala binucleata, cu nucleu crescut de volum (x3-4), hipercromi, cu margine nucleara neregulata, halou perinuclear.



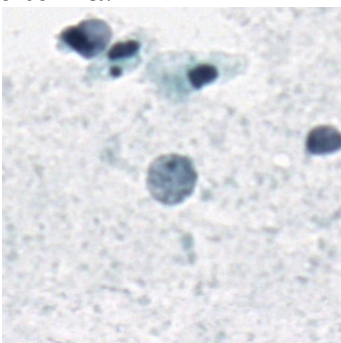
27. Koilocit. Celula superficiala cu nucleu crescut de volum si halou perinuclear.



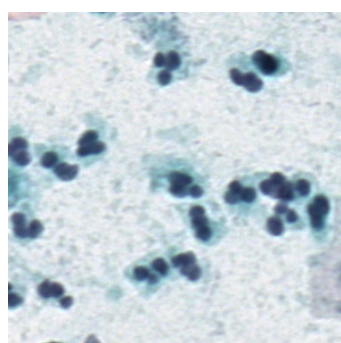
28. Koilocite. Celule intermediare si superficiale cu nucleu crescut de volum (x3-4), hipercromi, cu margine nucleara neregulata, halou perinuclear, binucleeri.



29. Nucleu denudat. In fundal se observa citoliza.



30. Polimorfonucleate partial degradate.



Concluzii

Toate obiective propuse in cadrul etapei au fost indeplinite.

Diseminare

Rezultatele obtinute au fost publicate in urmatoarele articole:

- Automatic Pap Smear Nuclei Detection Using Mean-Shift and Region Growing, S. Oprisescu, T. Radulescu, A. Sultana, C. Rasche, M. Ciuc, 12-th International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, 2015.
- Analysis of Pap Smear Images with Iso-, Edge-Contours, C. Rasche, S. Oprisescu, A. Sultana, T. Radulescu, International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj, Romania, 2015.
- Fast Probabilistic Pseudo-Morphology for Noise Reduction in Color Images, R.-M. Coliban, M. Ivanovici, I. Szekely, 9th International Conference Interdisciplinarity in Engineering, INTER-ENG, Tirgu-Mures, Romania, 2015.
- Cervical Nuclei Detection and Classification with Iso-Contours, C. Rasche, C. Tiganisteanu, M. Neghina, A.-E. Sultana, Proc. of IEEE International Symposium on Biomedical Imaging (ISBI): From Nano to Macro 2016 (*depus*).