

Detekcija abnormalnih stanica iz slika dobivenih PAPA-testom

1. Nikolina Đuranec

Matematika i informatika, smjer nastavnički
Prirodoslovno-matematički fakultet
Zagreb, Hrvatska

2. Ana Nedić

Primijenjena matematika
Prirodoslovno-matematički fakultet
Zagreb, Hrvatska

3. Borna Radoš

Matematika i informatika, smjer nastavnički
Prirodoslovno-matematički fakultet
Zagreb, Hrvatska

4. Fran Špigel

Primijenjena matematika
Prirodoslovno-matematički fakultet
Zagreb, Hrvatska

Abstract—Uspoređujemo dvije različite metode za binarnu klasifikaciju slika stanica iz vrata maternice - konvolucijsku neuronsku mrežu te segmentaciju putem k-means algoritma. Ulazne podatke klasificiramo na 'normalne' (negativne) te 'abnormalne' (pozitivne). U rezultatima želimo maksimizirati metriku 'recall'.

I. UVOD

Nakon dobivanja slika putem PAPA-testa, posao je citoloških tehničara da procjene da li uzorci stanica predstavljaju normalne, zdrave stanice, ili abnormalne stanice koje ukazuju na rak vrata maternice ili neki sličan budući problem. Kada je riječ o stotinama ili tisućama stanica u jednom uzorku, taj proces može potrajati, pa bi bilo izravno korisno automatizirati ga, ili barem razviti mehanizam koji tehničaru ubrzava posao tako da mu ukazuje na stanice koje bi mogle biti abnormalne. Takav problem spada u područje računalnog vida, i kao takvom mu pristupamo s tri različite metode, te uspoređujemo njihove rezultate pri klasificiranju na 'normalne' i 'abnormalne' stanice.

Jedna metoda koju koristimo je konvolucijska neuronska mreža, a druga je nenadgledana metoda za segmentaciju bazirana na algoritmu k-means. Ulazni podaci su slike u boji pojedinačnih stanica, a izlaz je binarna klasifikacija - algoritam mora odlučiti da li se na slici radi o normalnoj ili abnormalnoj stanici. Pri tome je iznimno važno minimizirati lažne negativne rezultate. Ako algoritam za poneku normalnu stanicu zaključi da je abnormalna, u najgorem slučaju to stvara više posla tehničaru koji kasnije pregledava rezultate. Ali ako se često događa da abnormalne stanice prolaze neopaženo, to predstavlja rizik od toga da ozbiljna bolest ne bude dijagnosticirana. S time na umu ćemo se fokusirati na metriku 'recall', koja je zapravo vjerojatnost da abnormalna stanica bude klasificirana kao takva, čak i na štetu preciznosti (vjerojatnost da je stanica klasificirana kao abnormalna, doista abnormalna).

Obje metode ćemo obraditi u zasebnom poglavlju.

II. ULAZNI PODACI

Podaci na našem raspolaganju sastoje se od slika pojedinačnih stanica iz dva različita izvora - jedan je bolnica Herlev u Danskoj, a drugi je otvorena baza Sipakmed [CITAT].

Među sobom sadrže otprilike 4,000 slika stanica, podijeljenih ravnomjerno između normalnih i abnormalnih. Na slici 1 možemo vidjeti primjer normalne i abnormalne stanice. Takve pojedinačne slike će biti ulazni podaci naših klasifikatora. Tamna mrlja u sredini predstavlja jezgru (nukleus) stanice, obojana površina oko nje je citoplazma, a izvan toga je pozadina koju ćemo se truditi zanemarivati kroz našu analizu. U našem su iskustvu često normalne stanice vrlo jasne - iz prve je očigledno kako je stanica podijeljena na jezgru, citoplazmu i pozadinu, dok abnormalne stanice često sadržavaju mrlje i druge anomalije oko sebe koje otežavaju taj posao.

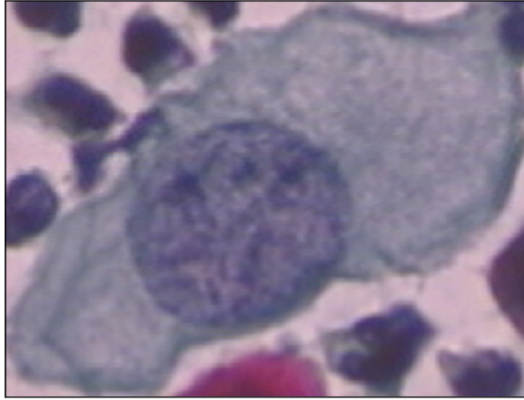
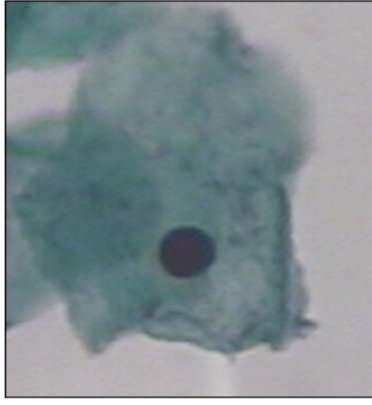
III. PRVA METODA: SEGMENTACIJA PUTEM K-MEANS ALGORITMA

Prvo što primjećujemo kada sagledamo više primjera različitih stanica je to da se jezgra abnormalnih stanica čini puno većom nego kod normalnih stanica. Preciznije, promatramo razliku između omjera površine koju zauzima jezgra, te površine cijele stanice. Stoga, kada bismo mogli segmentirati sliku na tri ključne komponente - jezgru, citoplazmu i pozadinu, mogli bismo trivijalno izračunati taj omjer i na taj način formirati svoju procjenu. Taj problem nastojimo riješiti metodom k-means za segmentiranje slike. Pri tome je metoda preslikavanja u crno-bijeli format te odšumiranja uvelike inspirirana radom [1].

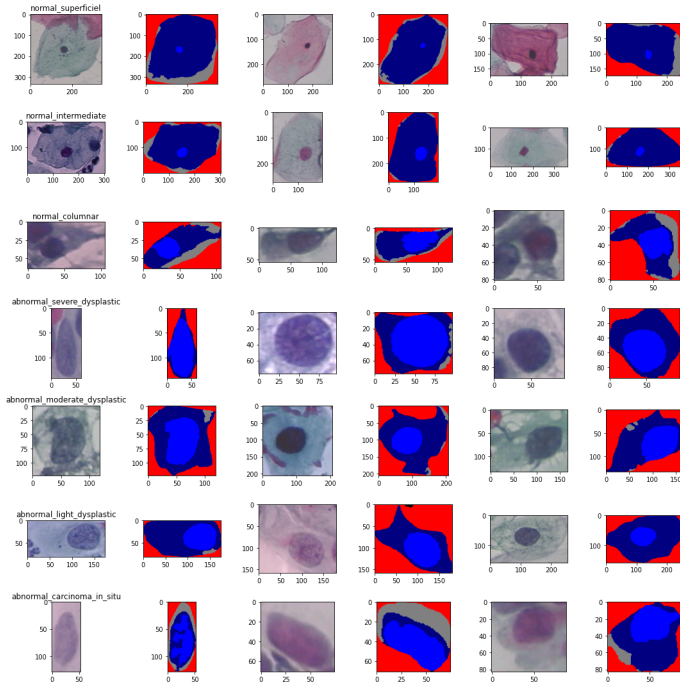
A. Pred-procesiranje

Da bismo olakšali posao k-means (dalje KM) algoritmu, slike je potrebno pred-procesirati. Za početak svaku sliku pretvaramo u crno-bijeli format koristeći prvo transformaciju u CIELAB format, koji sliku prikazuje u tro-dimenzionalnom prostoru (L, a, b), pri čemu dimenzija L predstavlja svjetlinu pojedinih piksela. Stoga, ako umjesto svakog piksela prikažemo njegovu L-vrijednost, dobivamo crno-bijeli prikaz slike.

Nadalje, potrebno je odstraniti što je više moguće šuma sa slike. Pri tome koristimo tzv. ne-lokalne sredine, metodu koja uspostavlja funkciju sličnosti pojedinih piksela, te piksele



Sl. 1: Normalna stanica (gore) i abnormalna (dolje)



Sl. 2: Usporedba omjera veličine jezgre kod normalnih i abnormalnih stanica. Svaka druga slika prikazuje prethodnu sliku ručno segmentiranu na jezgru (svjetlo plavo), citoplazmu (tamno plavo) i pozadinu (crveno)

zamjenjuje prosjekom između njih i njima sličnih piksela sa drugih dijelova slike. Suštinski, to je funkcija

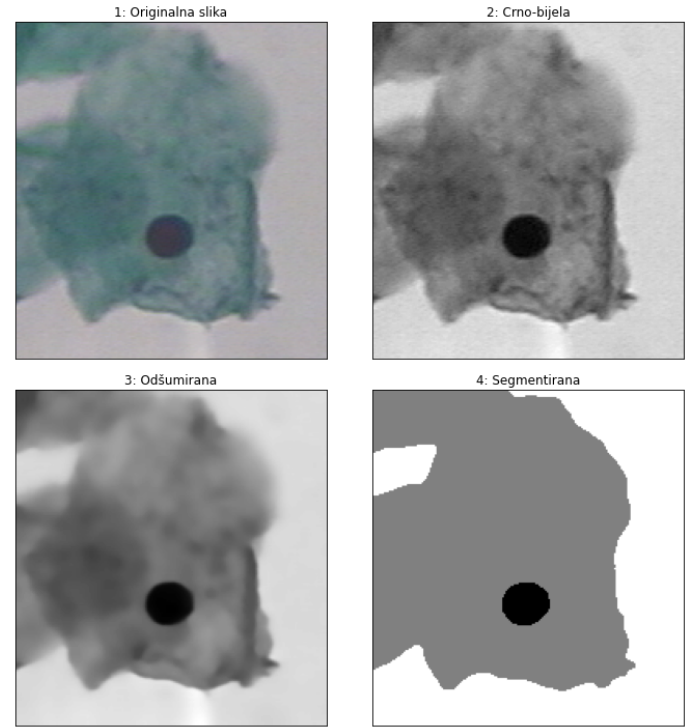
$$v(x) = \frac{1}{C(x)} \int_{\Omega} u(x) f(x, y) dy,$$

pri čemu su x i y pikseli, $C(x)$ je normalizirajući faktor za slijedeći integral, $v(x)$ nova vrijednost piksela, $u(x)$ stara vrijednost piksela, a $f(x, y)$ neka mjera sličnosti dva piksela, na primjer

$$f(x, y) = e^{-\frac{|B(q) - B(p)|^2}{h^2}}$$

U radu [1] je ta metoda odabrana jer dobro utječe na šum na slici, a ipak čuva strukture koje bi druge metode, poput zamućivanja (blur), oštetile. Obje opisane operacije možemo vidjeti na slici 3

Zadnji alat za pred-procesiranje kojeg smo uzimali u obzir bio je povećavanje kontrasta na slici u svrhu naglašavanja razlike između pojedinih regija, ali to nije urodilo zadovoljavajućim rezultatima, pa je ideja odbačena.

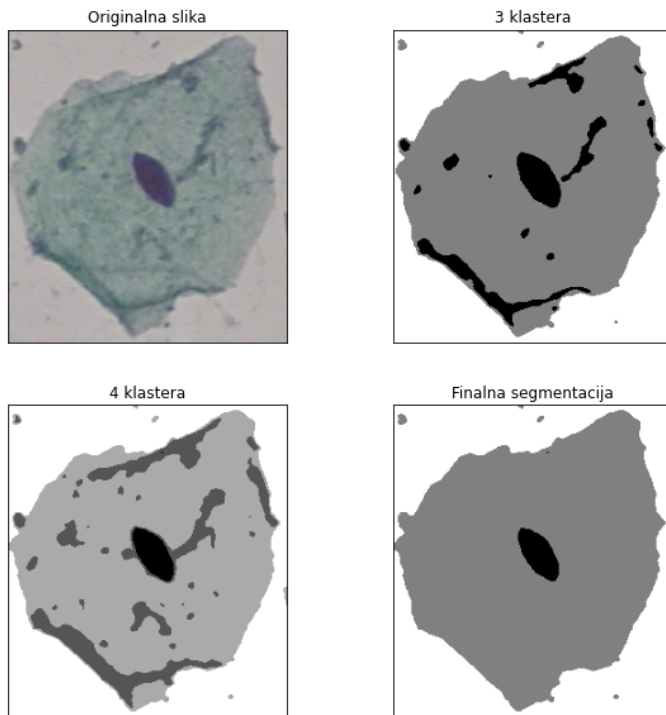


Sl. 3: Proces pred-procesiranja i segmentiranja slike

B. K-means

KM je dobro poznati algoritam, pa ovdje nećemo ulaziti u previše detalja. Ukratko, skup elemenata (u ovom slučaju, piksela), u n -dimenzionalnom prostoru dijelimo na grupe, tj. klastere, ovisno o njihovom položaju u prostoru. U našem slučaju, svaki piksel opisujemo sa tri karakteristike: njegova svjetlina, prosjek svjetlina njegovih susjeda, te medijan svjetlina njegovih susjeda. Time ga smještamo u trodimenzionalni prostor, a zatim na taj prostor primjenjujemo k-means.

Budući da nastojimo sliku segmentirati na tri ključna područja (jezgra, citoplazma, pozadina), prirodno bi bilo tražiti od KM algoritma svrstavanje u tri klastera. Međutim, ovaj pristup se pokazao problematičnim. Naime, zbog jakih varijacija u svjetlini unutar citoplazme, često dio citoplazme upada u klaster namijenjen za jezgru ili pozadinu. Primjer ovog problema možemo vidjeti na slici 4. Vidimo da su dijelovi citoplazme nešto tamniji, pa ih KM smješta u područje jezgre. Ako mu dopustimo da različite djelove citoplazme smjesti u različite klustere, te zatim ručno spojimo te klustere u jedan, dobivamo puno bolji rezultat.



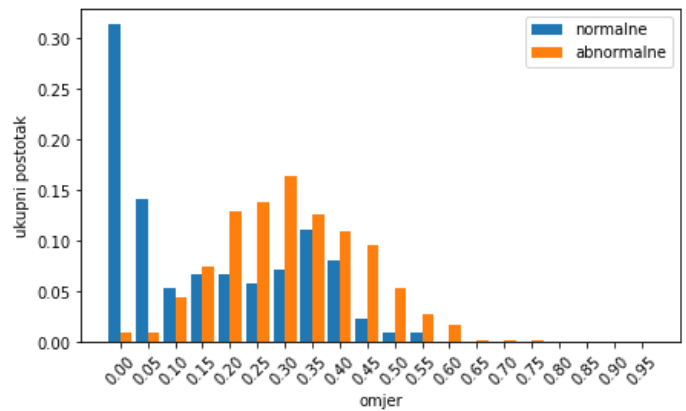
Sl. 4: Segmentacija na 3 i 4 klastera

C. Klasifikacija

Nakon što je KM segmentirao sliku, trivijalan je posao izračunati površinu jezgre odnosno citoplazme. Za svaku sliku računamo omjer r površine jezgre i čitave stanice. Sada možemo proučiti da li procjenjujemo značajno različit omjer kod slika normalnih i abnormalnih stanica. Iako su u sredini omjeri prilično ispremešani, odmah vidimo da je omjer kod normalnih stanica većinom ispod 0.1, a kod abnormalnih iznad. Stoga, možemo uspostaviti jednostavno pravilo za klasifikaciju: ako je procjena omjera površine jezgre u odnosu na čitavu stanicu veća od $x = 0.1$, smatramo da je stanica abnormalna. Inače, smatramo da je normalna.

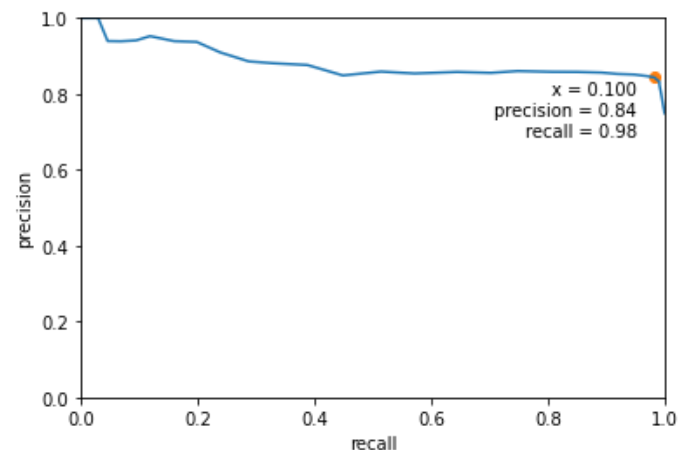
Sada možemo dobiveni mehanizam testirati na čitavom skupu podataka, te izračunati nama bitne metrike (precision i recall). Za podatke iz bolnice Herlev, one iznose:

- Accuracy: 85%
- Precision: 84%
- Recall: 98%



Sl. 5: Histogrami omjera površine jezgre i stanice za normalne i abnormalne stanice

- F1 score: 91%



Sl. 6: Ovisnost metrika recall i precision

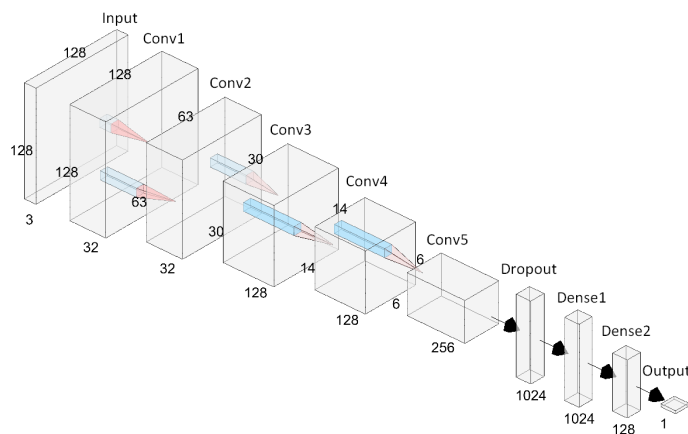
IV. DRUGA METODA: KONVOLUCIJSKA NEURONSKA MREŽA

KNN je vrlo popularan pristup problemima računalnog vida, te opet izbjegavamo ulaziti u općenite detalje toga kako funkcioniraju. Mreža koju smo konstruirali sastoji se od 5 konvolucijskih slojeva koji se nastavljaju na neuronsku mrežu sa 3 potpuno povezana sloja, kojima prethodi jedan dropout sloj, sa namjenom suzbijanja pre-treniranja, od kojih zadnji ima jedan jedini output - broj koji označuje vjerojatnost da ulazni podatak predstavlja abnormalnu stanicu. Sažetak njene arhitekture vidljiv je na slici 7, a detaljna arhitektura na slici 10

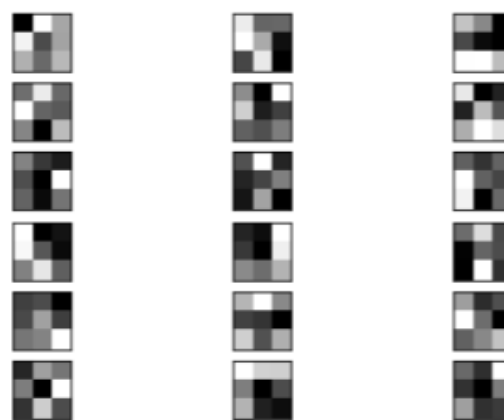
Pri testiranju smo došli do zaključka da je mreži dovoljno 20 epoha kako bi postigla optimalne rezultate. Treniranje nakon toga rezultira pre-treniranjem (engl. overfitting).

Nakon 20 epoha, mreža postiže vrlo dobre rezultate:

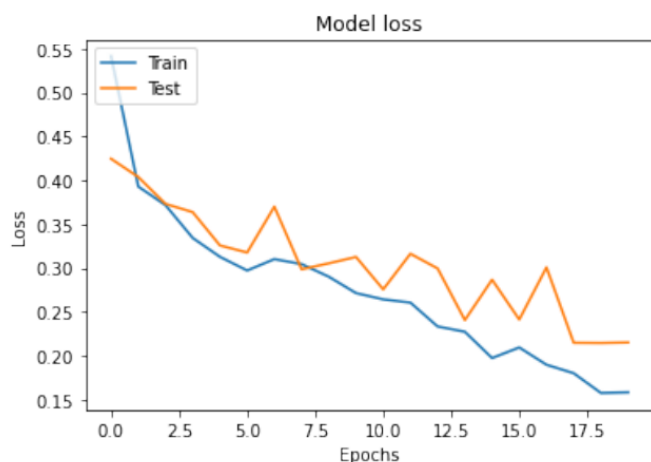
- Accuracy: 94%
- Precision: 94%
- Recall: 97%



Sl. 7: Arhitektura konvolucijske neuronske mreže



Sl. 9: Neki konvolucijski filteri iz mreže



Sl. 8: Proces treniranja

- F1 score: 94%

Iako je recall nešto lošiji, svi drugi rezultati su značajno bolji nego kod pristupa segmentacijom. Pri tome je još potrebno napomenuti da je naša implementacija segmentacije često neprecizna i na mnogim stanicama vrlo ne radi dobro. Nadalje, segmentacija je testirana samo na podacima iz bolnice Herlev, koji su općenito lakši zbog većeg kontrasta između citoplazme i pozadine na slikama. Stoga zaključujemo da je ova KNN značajno bolji alat za klasifikaciju od našeg prvog pristupa.

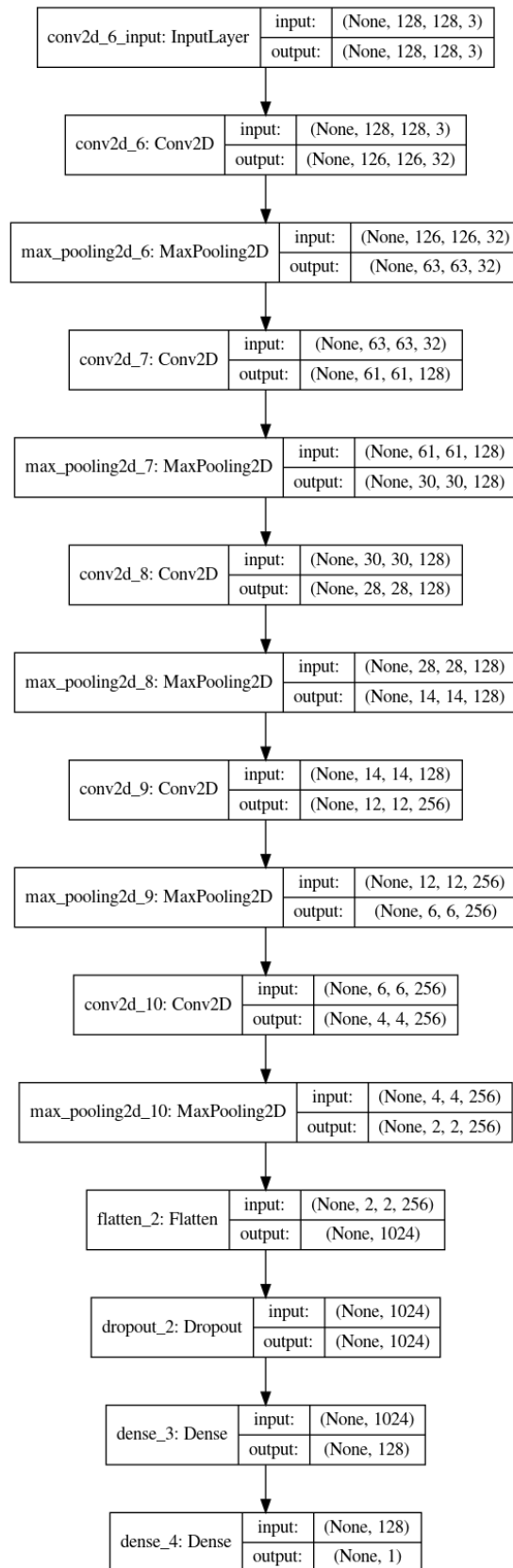
Slike 9 prikazuje neke od filtera koje je mreža izračunala, konkretno iz prvog konvolucijskog sloja.

V. DRUGI RADOVI NA ISTU TEMU

Budući da se radi o važnom problemu sa slobodno dostupnim podacima i izravno dostupnim podacima, za očekivati je da su se razni drugi timovi već pozabavili tim pitanjem. Već smo spomenuli rad [1] - grupu Kineskih znanstvenika koji su nakon segmentacije sa KM primijenili napredne metode poput RGVF-Snake kako bi vrlo precizno detektirali granice između jezgre, citoplazme i pozadine na svakoj slici. Nekoliko drugih radova se nalazi u bibliografiji ovog članka.

REFERENCES

- [1] Li, Kuan & Lu, Zhi & Liu, Wenyin & Yin, Jianping. (2012). Cytoplasm And Nucleus Segmentation In Cervical Smear Images Using Radiating Gvf Snake. Pattern Recognition. 45. 1255-1264. 10.1016/j.patcog.2011.09.018.
- [2] Bak E, Najarian K, Brockway JP (2004) Efficient segmentation framework of cell images in noise environments. Proc 26th Int Conf IEEE Eng Med Biol 1:1802-1805
- [3] Bamford P, Lovell B (1998) Unsupervised cell nucleus segmentation with active contours. Signal Process 71(2):203-213
- [4] Plissiti M.E., Nikou C. (2013) A Review of Automated Techniques for Cervical Cell Image Analysis and Classification. In: Andreus U., Iacoviello D. (eds) Biomedical Imaging and Computational Modeling in Biomechanics. Lecture Notes in Computational Vision and Biomechanics, vol 4. Springer, Dordrecht
- [5] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel, Non-Local Means Denoising, Image Processing On Line, 1 (2011), pp. 208-212. https://doi.org/10.5201/ipol.2011.bcm_nlm



Sl. 10: Detaljan prikaz arhitekture