Algoritmi klasifikacije za detekciju primarnog tumora na osnovu mikroskopskih slika metastaza u kostima

Slađan Kantar, Aleksandar Pluškoski i Igor Ciganović, Jelena Vasiljević

***Sadržaj* - U ovom radu predstavljene su tehnike analize mikroskopskih slika u cilju nalaženja primarnog tumora na osnovu metastaza u kostima. Rađena je algoritamska klasifikacija u tri grupe, bubreg, pluca i dojka. Sa ciljem da se ubrza lečenje pacijenta i olakša posao lekarima i time smanji prostor za ljudsku grešku. Analizirane su digitalne mikroskopske slike metastaza u kostima, za koje je poznato da je primarni tumor u jednom od tri organa, bubregu, plućima ili dojci. Testirali smo više rešenja za klasifikaciju. Testirana su dva metoda analize slike. Multifraktalna analiza i konvolucione neuralne mreže. Oba metoda su testirana sa i bez predprocesiranja slika. Rezultati multifraktalne analize su zatim klasifikovani pomoću različitih algoritama. Slike su obrađene pomoću CLAHE i k-means algoritama. Na kraju su prikazani rezultati dobijeni upotrebom različitih tehnika.**

***Ključne reči* –Klasifikacija kancera, mikroskopske slike, preprocesiranje slika, multifraktalna analiza, algoritmi klasifikacije**

I. UVOD

Kancer je vodeći uzrok smrti u svetu. Svake godine sve više ljudi oboleva, a napredak medicine nažalost nije dovoljno brz, iako je u poslednjih nekoliko godina dosta učinjeno. U oblasti računarskih nauka mnogi naučni radovi se bave pitanjima automatizacije postavljanja dijagnoze i određivanja načina lečenja karcinoma. Naš rad se bavi analizom različitih tehnika za algoritamsku kalsifikaciju (određivanje) primarnog tumora na osnovu mikroskopiskih slika metastaza u kostima.

Poslednjih nekoliko godina algoritmi mašinskog učenja doživeli su veliku popularizaciju i ekspanziju. Sve se više primenjuju u različitim oblastima industrije i nauke. Postali su i deo proizvoda koje svakodnevno koristimo. Takođe, koriste se u medicini sve više. Kako su ti algoritmi u osnovi statistički, potrebna je jako velika količina podataka za njihovu primenu.

Poslednjih godina sve veće kolekcije podataka postaju dostupne što je glavni uzrok sve veće upotrebe ovog tipa algoritama. Osnovna ideja svakog od ovih algoritama je da na osnovu velike količine podataka "uoče" statističku sličnost unutar klasa trening podataka i zatim pokušaju da ekstrapoliraju "naučeno" na nove podatke i time ih kalsifikuju u jednu od zadatih klasa.

Kada je reč o analizi slike algoritmima mašinskog učenja, konvolucione neuralne mreže su nezaobilazne, a u medicini se dodatno primenjuje fraktalna i multifraktalna analiza, zbog uočenih fraktalnih karakteristika medicinskih slika. Konvoluciona neuralna mreža predstavlja kompletan klasifikacioni algoritam koji sliku koja je dovedena na ulaz kasifikuje u unapred zadate klase. Fraktalna i multifraktalna analiza daju kao rezultat nizove brojeva koji sadrže informacije o fraktalnim dimenzijama zadate slike. Tako da je potrebno koristiti neki od algoritama klasifikacije da bi se ti rezultati svrstali u neku od zadatih klasa. U oba slučaja slike koje se analiziraju mogu biti "sirove", dobijene direktno sa mikroskopa, ili prethodno obrađene. Mi smo testirali oba slučaja radi poređenja rezultata.

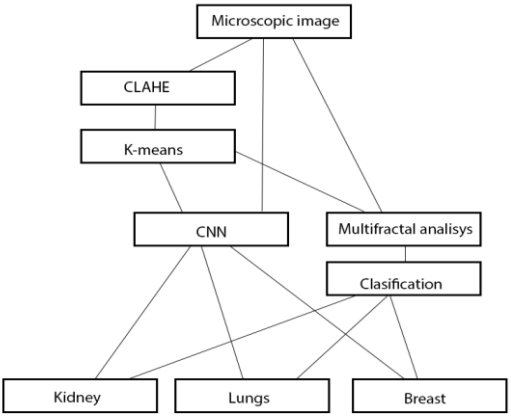
II. MODEL I METODE

Detekcija i klasifikacija kancera pomoću mikroskopskih slika sa biopsije jeste naporan i složen zadatak. Više različitih tehnika je ušlo u predloženu metodu, koja uključuje, menjanje kontraskta slike, segmentaciju ćelija, multifraktalnu analizu, i na kraju klasifikaciju. Na Sl. 1, su predložena dva puta za klasifikaciju mikroskopskih slika. Prvi predlaže, za poboljšanje mikroskopskih slika tkvia biopsije korišćenje CLAHE algoritma, koja se zatin propušta kroz k-means algoritam za segmentaciju ćelija. Ovom metodom smo istakli važne biološke i kliničke oblike, kao i morfološke osobine, koje uključuju nijasnu sive, segmentaciju boja i segmentaciji boja po teksturi. /13,14/ Na kraju se radi multifraktalna analiza i neki algoritam klasterizacije. Drugi predlog, jeste da se na dobijenim mikroskopsim slikama biopsije direktno radi multifraktalna analiza, a potom algoritam klasterizacije. Ovi algoritmi su tesitrani na nasumično izabranih 1050 (po 350 u svakoj grupi – bubreg, pluca, dojka) mikroskopsih slika biopsije. Na kraju je merena i analizirana uspešnost svakog od klasifikacionih algoritama. U nastavku je dat rad svakog od korišćenih algoritama u eksperimentu.

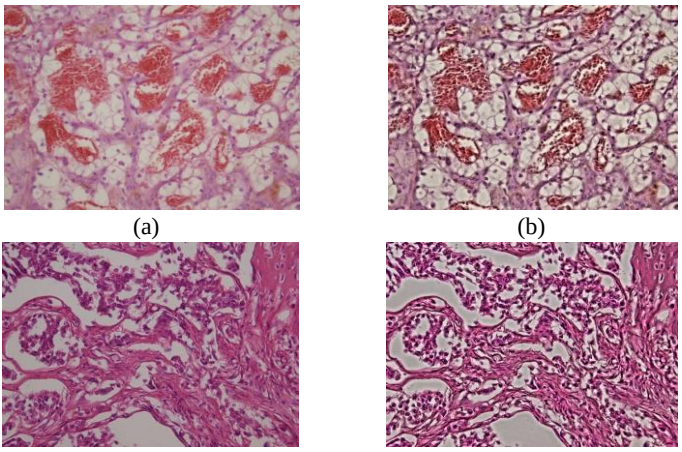
A. Poboljašavanje slike

Glavni cilj predprocesiranja bio nam je da umanjimo šum i da napravimo veći kontrast. Tako bi došli do izražaja nama važni regioni. Prilikom

fotografisanja mikroskopskih slika tkiva biopsije dešava se deformacija slike, usled male žižine daljine (riblje oko) i loše osvetljenosti. Samim tim dešava se da su slike previše ili premalo eksponirane i lošeg kontrasta, pa i nejasne ivice. Stoga je korišćen CLAHE algiratam da bi poboljšali mikroskopske slike. Sl. 2 pokazuje nam sliku pre i posle upotrebe CLAHE algoritma.



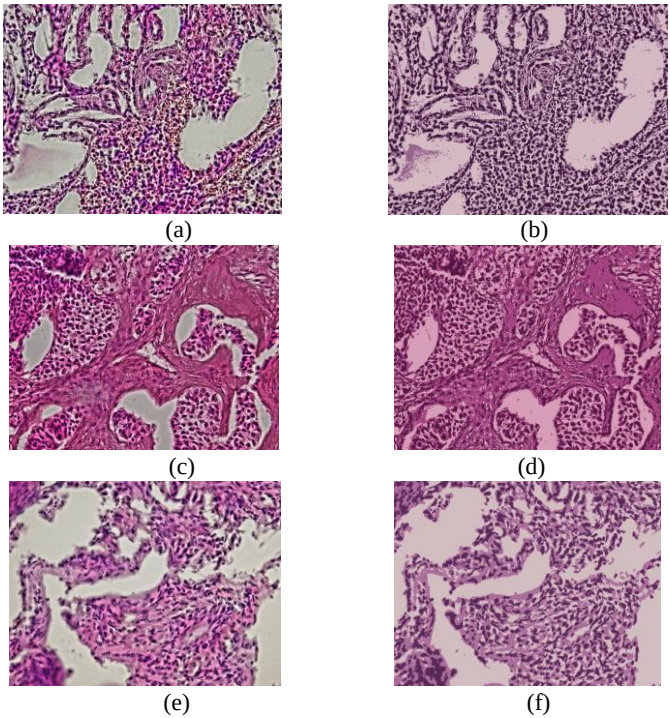
Slika 1.Šematski prikaz rada predstavljenon u ovom radu



Slike 2. Slike (a) i (c) su uzete direktno sa mikroskopa, dok su (b) i (d) je posle CLAHE algoritma

B. Segmentacija

Postoji nekoliko metoda segmentacije slika, na osnovu ćelija, citoplazme ili jedra. Postoje dva načina da se ovo postigne. Prvi je da čovek radi segmentaciju regiona od interesa, ali taj način bi bio vremenski zahtevan i dozvolio bi ljudsku grešku. Drugi način je, automatski, pomoću nekog od algoritama. Testirano je više algoritama, ali najbolji, i jedini koristan, u slučaju mikroskopskih slika biopsije bio je k-menas algoritam. K-menas je vektorska kvantizacija, koja je preuzeta iz oblasti obrade signala. K-menas algoritam, radi tako što particioniše n posmatranih objekata u k klastera. /15/ Pošto naše mikroskopske slike imaju 256 nijansi boja, mi smo grupisali u 3 klastera, i onda se svaka slika predstavlja sa samo tri boje. Ovom metodom se neki od podataka gube, ali zato uprošćujemo šablone kojima se klasifikuje vrsta carcinoma na mikroskopskim slikama. Na Sl. 3 ispod je prikazana, k-means klasterizacija, tkiva bubrega, pluca i dojke.



Slika 3. Slike (a), (c) i (e) su pre, a (b), (d) i (f) posle obrade sa k-means algoritmom. Gledano odozgo da dole, prikazani su bubreg, dojka i pluća.

C. Multifraktalna analiza

Za svaku od ove tri grupe, izabrano je 350 slika (ukupno 1050). Multifraktalna analiza digitalnih medicinskih slika je rađena korišćenjem slobodnog programa ImageJ, i plugin-a FracLac /1/. Parametri koji su bili prosleđeni program su: Q u opsegu od -5 do +5 sa korakom Qs od 0.25, broj skeniranja po slici bio je 1.

D. Algoritmi klasifikacije

Klasifikacija je jedan od najčešćih zadataka mašinskog učenja Ona predstavlja problem indentifikovanja kom unapred definisanom skupu pripada nepoznata posmatrana pojava. Klasifikacija nekog objekta zasniva se na pronalaženju sličnosti sa unapred određenim objektima, koji pripadaju različitim klasama. Pri klasifikaciji svaki objekat se dodeljuje nekoj klasi, sa određenom verovatnoćom tačnosti. Klasifikacioni algoritam ima zadatak da na osnovu trening podataka napravi model pomoću koga bi vršio klasifikaciju novih nepoznatih objekata. U problemu klasifikacije, broj klasa je unapred poznat i nije promenljiv.

U ovom radu je korišćene python otvorena biblioteka scikit-learn 0.18. Pisana je u Python, C i C++ programskim jezicima. Sadrži razne algoritme za klasifikaciju, od kojih su su ovde korišćeni: SVM, Nearest Neighbors, Naive Bayes, Decision Trees, Ensemble methods; algoritme za regresiju i klasterizaciju, a korišćeni su: K-Means i Spectral clustering /3,4/.

*D.*1. *Linear Discriminant Analysis-LDA*

LDA je motod koji se koristi u statictici i mašinskom učenju da se pronađe vektor (span, linearna kombinacija) posmatranih objekata, koji razdvaja dve ili više klasa. Rezultat se može koristii kao linearni klasifikator ili za smanjenje broja dimenzija, pa da se potom primeni neki drugi dobro poznati klasifikatori.

LDA klasifikator za više klasa, je samo generalizacija LDA za dve klase koji može da barata sa proizvoljnim brojem klasa. LDA za više klasa se zasniva na dve kovarijaciono-disperzione n-dimenzionalne matrice: kovarijaciono-disperzione n-dimenzionalne matrice unutar klase i izlmeđu klasa. Pretpostavimo da su trening podaci $x_1, \dots, x_n$, a njihove labele kojim klasama pripadaju $y_1, \dots, y_n$, onda je kovarijaciono-disperzione n-dimenzionalne matrice unutar klase definisana kao:

$$\mathbf{S}_w = \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu}_{y_i})(\mathbf{x}_i - \boldsymbol{\mu}_{y_i})^T$$

Gde je μ_k matematičko očekivanje za *k*-tu klasu.

Kovarijaciono-disperziona n-dimenzionalna matrica između klase je definisana kao:

$$\mathbf{S}_b = \sum_{k=1}^{m} n_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T$$

Ovde je m, broj klasa, μ je matematičko očekivanje, a n_k je broj uzoraka u k-toj klasi.

Tada, LDA za više klasa možemo formulisati, kao problem koji traži vektor w koji maksimizuje razliku među klasama koje su unutar klase.

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

U oopštem slučaju, potrebno je najviše m-1 vektor, da bi dobili disktretne vrednosti za m klasa . Rešenje u opštem slučaju može biti zapisano kao:

$$\mathbf{S}_b \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w}$$

Ovaj algoritam nije preporučljivo koristiti kada imamo jako veliki broj klasa, zbog njegove kompleksnosti za izračunavanje. /16/

D.2. Gaussian naive Bayes

Bayes klasifikatori su skup algoritama koji koriste Bajesovu teoremu, pod pretpostavkom da su podaci koje uzimaju medjusobno nezavisni. Često koriti kao osnovni klasifikator, zato što ga je jednostavno i brzo implemntirati. Ovaj algoritam je degradiran ozbog loših rezultata, međutim, dobrim predprocesiranjem podataka i dobro podešenim parametrima, može da bude konkurentan sa SVM-om. /10, 11/.

Postoji više vrsta bajesovih klasifikatora i razlika je samo u tome kako prave raspodelu P(xi | y). Ali, pošto se u prirodi iza naizgled nasumučnih vrednosti krije gausova raspodela, koja se još naziva i zvonasta kriva, stoga je jedan on najkorišćenijih klasifikatora u ovoj grupi GNB klasifikator. /3, 4, 12/ Pretpostavimo da su trening podaci X, dok je Y lista svih klasa, kojima Xi može pripadati. Mi prvo podelimo trening podatke na klase, i za svaku klasu nadjemo matematičko očekivanje y i varijansu 2y. Onda je verovatnoća raspodele

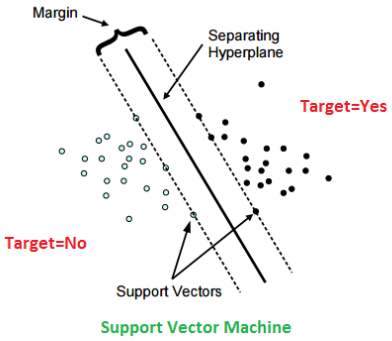
$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

D.3. SVM (SVC)

Support Vector Machine (SVM) je popularna metoda mašinskom učenju

sa nadgledanjem za klasifikaciju, regresiju i druge zadatke. SVM mogu dati vrlo tačne predikcije, imaju malo parametara koji zahtevaju podešavanje i nisu preterano osetljivi na veličinu podataka.

SVM prvo mora da nadje granicu razdvajanja. Za to ne uzima sve podatke, već traži iz skupa najprbližnije vrednosti da da pripadaju različitim klasama. Tačke koje se nalaze na samim granicama nazivamo ih support vectors odnosno support planes. Kao što smo već pomenuli, nama su potrebne samo granične vrednosti odnosno support vectors za računanje optimalne (granične) hiperravni.



Slika 4. Šema SVM algoritma

Algoram traži tu granicu na sledeći način? Prvo se nalaze dve hiperravni koje najbolje predstavljaju podatke. Neka su svetle tačkice (Target=No) predstavljene kao $w^T x + b_0 > 1$, dok su crne $w^T x + b_0 < -1$, gde nam je **w** težinski vektor, **x** je skup trening podataka, odnosno vektora, a $\mathbf{b_0}$ je disperzija. Težinski vektor određuje orijentaciju granice, dok disperzija odlučuje tačku razdvajanja. Pomoću ove dve vrednosti algooritam pravi granicu odlučivanja, koja se nalazi na pola puta. /17/

SVC algoritam jeste samo dodatni skup klasa koji se sposoban da kada napravi hiperravni, klasifikuje podatke unutar njih. SVC uzima kao ulaz niz trening podataka koje klasterizuje u n klasa, a njih je dobio kao drugi parametar. /3, 4/

*D.*4. *K neighbors*

K-NN je neparametarska metoda za klasifikaciju i regresiju, koja je uspešna u slučajevima kada je granica odlučivanja nepravilna. /2/. Ulaz nam predstavljaju trening podaci koji su dobijeni upotrebom multifraktalne

analize. Izlaz jeste klasa kojoj pripada. Objekat pripada onoj klasi, kojoj pripadaju najviše njegovih k suseda (k > 0). Broj k je definisan pre početka izvršavanja algoritma. Ukoliko je k = 1 onda se rezultat dodeljuje klasi tog jednog objekta. Rastojanje između tačaka može biti bilo koja merna jedinica, ali uglavnom jeste Euklidsko rastojanje. Optimalni izbor broja k je složen process. On zavisi on vrste podataka. Ukoliko je k veliko, onda ono suzbija efekat šula, ali zato čini da se granice odlučivanja manje razlikuju /3/. U ovom predlogu rešenja, mi imamo 3 klase pripadanja, na osnovu tkiva (bubreg, dojka, pluća).

*D.*5. *Decision Tree*

DT je neparametarsko učenje sa nadzorom koje se koristi za klasifikaciju i regresiju. Cilj je da se napravi model koji predviđa vrednost, tako što uči jednostvna pravila odlučivanja, zaključena iz trening podataka /3, 4/. DT lako barata za velikom količinom ulaznih podataka i to ga čini dobrim za hiperspektralne podatke /8/. DT sastoji se od grana i čvorova. Svki čvor je povezan sa skupom mogućih odgovora i na ovaj način se od ulaznog skupa podataka pravi manji podskup, koji odgovara različitim test slučajevima. (Slika). Stablo odlučuvanja kao i drugi algoritmi mašinskog učenja koristi rekurzivnu podelu podataka da bi napravio (izveo) pravila za klasifikaciju iz trening podataka /9/.

*D.*6. *Random Forests*

RF je metoda grupnog učenja (ensemble learning method) za klasifikaciju i regresiju. RF radi tako što u toku treniranja pravi više stabala odlučivanja, a daje izlaz onu klasu koja se najviše puta pojavila kao izlaz svakog pojedinačnog stabla /4, 5/. Ovo znači da se prave raznoliki skupovi klasifikatora, tako što se uvodi slučajnost, u samu konstrkticju RF klasifikatora. RF uzima slučajan vector (X,Y) iz trening skupa. Vektor $X=\{X_1 , \dots , X_n\}$ i neka $\forall X \in R^+$ predststavlja predispozicije, dok $Y \in \bar{Y}$ gde je $\bar{Y}$ ili labela klase ili numeričku odgovor. Klasifikator t mapira

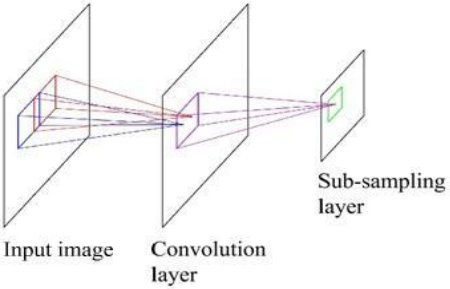
$$t: R^+ \in \bar{Y}$$

U RF algoritmu, svako stablo u grupi (šumi) zavisi od jednog nasumično izabranog vektora iz trening skupa i jednaka je verovatnoća raspodele za sva stabla u šimi. /6, 7/

*D.*7. *Konvoluciona neuralna mreža*

Konvoluciona neuralna mreža (CNN) predstavlja neuralnu mrežu čiji je

primarni zadatak prepoznavanje šablona. Najčešće se koristi za probleme prepoznavanja slika. U ovom radu govorićemo o dvo-dimenzionalnoj konvolucionoj neuralnoj mreži, međutim broj dimenzija koje možemo koristiti nije ograničen. Konvolucione neuralne mreže koriste princip kreiranja varijanti neuralnih mreža za nekoliko transformacija ulaza. Upravo ovo predstavlja problem sa potpuno povezanim neuralnim mrežama, jer potpuno povezani slojevi parcijalno uklanjaju deo informacija ulaznih podataka.



Slika 5. Arhitektura standardne konvolucione mreže. Prikazuje konvolucio sloj za kojim sledi pooling sloj.

Konvolucione neuralne mreže uključuju specijalne konvolucione i slojeve za objedinjavanje (pooling). Izlaz predstavlja spljošteni vektor (matrica transformisana u vektor) koji se kasnije koristi kao ulaz u potpuno povezanu mrežu za klasifikaciju. Arhitektura konvolucione neuralne mreže prikazana je na Sl. 5. Konvolucija i pod odabiranje se može vršiti više puta. Postoje tri osnovna principa u osnovi konvolucionih neuralnih mreža: lokalna receptivna polja, deljenje težina i pododabiranje.

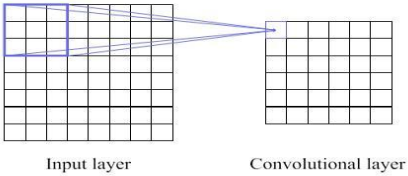
D.7.A. Local receptive fields

U potpuno povezanoj neuralnoj mreži svaki neuron određenog sloja povezan je sa svakim neuronom narednog. U konvolucionim neuralnim mrežama ovo nije slučaj. Svaki ulaz konvolucionog sloja povezan je sa određenom grupom prethodnog sloja. Ovo se naziva lokalno receptivno polje. Upravo ovim principom čuvamo prostorne informacije sa ulaza. Primer ove ideje prikazan je na Sl. 6. Vidimo da je prvi neuron konvolucionog sloja povezan samo sa devet pikslela ulaznog sloja. /17/

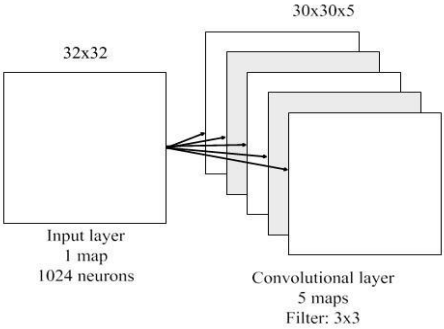
D.7.B. Weight sharing

Svaki konvolucioni sloj organizovan je kao mnoštvo paralelnih skrivenih

slojeva. Ovakvi slojevi se još nazivaju i mape funkcija. Neuroni istih indeka iz različitih mapa su povezani sa istom ulaznom površinom. Cilj deljanja težina je upravo da neuroni iz iste mape funkcija dele istu težinu. Ovaj princip prikazan je na Sl. 7.



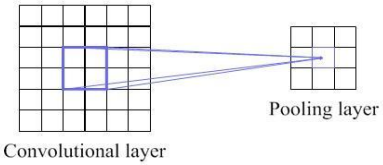
Slika 6. Primer lokalnog polja u convolucionom sloju.



Slika 7. Primer konvolucionog sloja. Filter je veličine 3x3 i 5 feature mapa.

D.7.C. Sub-sampling

Sloj pod odabiranja, se najčešće koristi odmah nakon konvolucionog sloja. Slika ovog sloja je da kreira varijante osobenosti. Ovo se radi korišćenjem statističkih informacija dobijenih iz receptivnog polja u fokusu konvolucionog sloja. Videti sliku Sl. 8.



Slika 8. Primer pooling sloja u feature mapi. Ovaj primer koristi pooling veličine 2x2.

III. REZULTATI I ANALIZA

A. CNN

Koristili smo konvolucionu mrežu sa relativno jednostavnom aritehturom izrađenu pomoću TensorFlow framework-a. Sastojala se od 6 kovolucionih slojeva, gde su svi bili upareni sa pooling slojem. Kao aktivaciona funkcija korišćen je "ReLU". Parametri u konvolucionom sloju su: strides = [1, 1, 1, 1] i padding = 'SAME', dok je za pooling sloj bilo: strides = [1, 1, 1, 1], ksize = [1, 2, 2, 1] i padding = 'SAME'. Broj feature mapa je računat prema sledećoj formuli:

$$n = 2^{4+i},$$

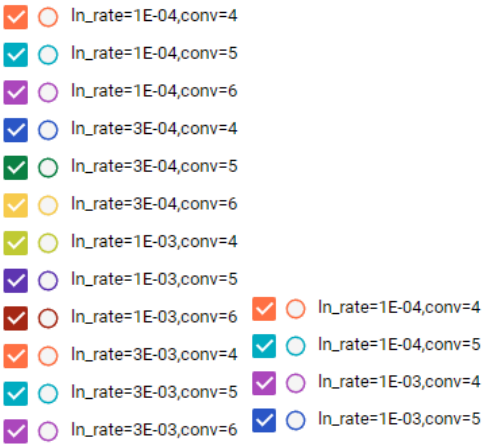
gde je "i" indeks trenutnog konvolucionog sloja. Izlaz poslednjeg pooling sloja je korišćen kao ulaz u potpuno povezanu mrežu koja se sastojala iz dva skrivena sloja i izlaznim slojem. Prvi potpuno povezan sloj imao je stopu odbacivanja od 30% da bi smanjio overfitting. Aktivaciona funkcija koja se koristila za sakriveni sloj je "sigmoid", a "softmax" se koristila za izlaz. Za treniranje mreže korišćen je "AdamOptimizer". Mreža je trenirana korišćenjem različitih hiperparametara i koristeći dva skupa trening podataka. Prvi skup je bio neprocesirane slike, a drugi procesirane. Slike u oba trening skupa su smanjene na 300x226 piksela, zbog smanjenja kompleksnosti mreže. Rezultati na Sl. 9 i Sl. 10 pokazuju poboljsanje preciznosti u funkciji vremena za neprocesirane i procesirane slike, respektivno. Legenda je prikazana odvojeno radi bolje preglednosti. Različite boje predstavljaju različitu brzinu učenja i različiti broj konvolucionih slojeva.

Kao što vidimo sa slike, jednostavna konvoluciona mreža nije dobro rešenje za ovu vrstu problema. Takođe, vidimo da se neki hiperparametri pokazuju boljim od drugih. Ali sa bilo koji hiperparametrima mreža kovergira vrlo brzo i ne nastavlja da dalje uči. Iz rezultata vidimo da nema velike razlike između neprocesiranih i procesiranih slika.

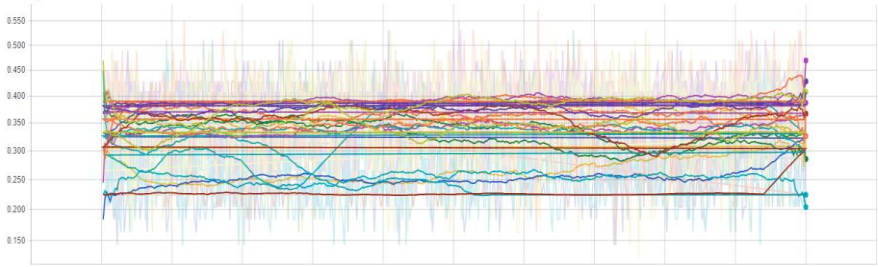
Rezultati pokazuju da daju bolje rešenje nego nasumično pogađanje. Naše mišljenje je da bi unifomniji podaci i kompleksija neuralna mreža dali bolje rešenje. Ali vršenje eksperimenata sa drugim vrstama arhitekture je izvan ovog projekta.

B. Algoritmi klasifikacije

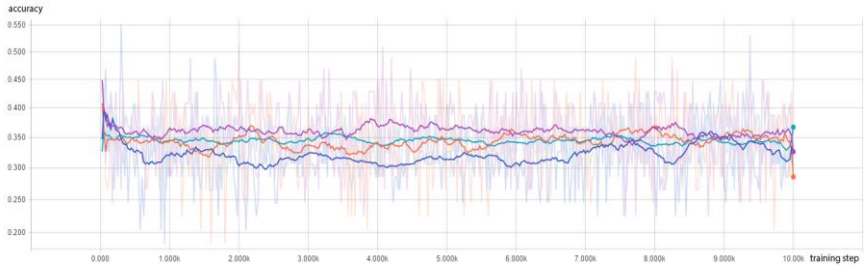
Pošto je ovo klasičan problem klasifikacije, testirali smo neke od dobro poznatih algoritama mašinskog učenja. Rezultati su prikazani na Sl. 12.



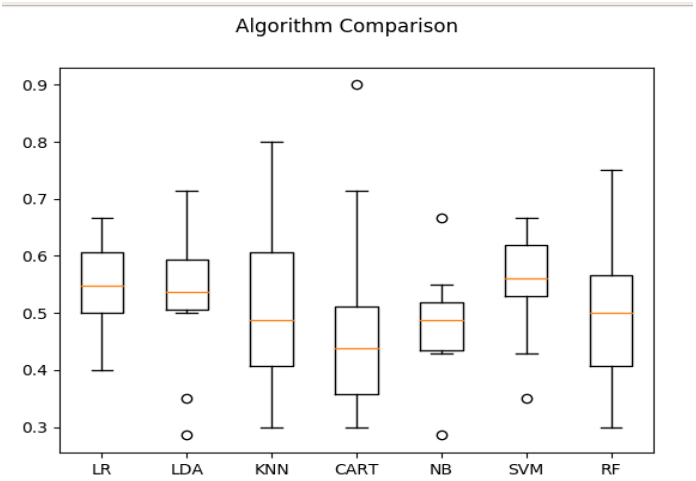
Slika 9. Legenda na levoj strani je za reprezentaciju neprocesoranih slika. Na desnoj je za procesirane slike.



Slika 10. Preciznost u funkciji vremena za neprocesirane slike.



Slika 11. Prezinost u funkciji vremena za procesirane slike.



Slika 12. Urađena ja više puta validacija. Narandžaste linije prikazuju nam srednju vrednost tačnosti, dok pravougaonici prikazuju oko njih prosečnu varijansu

*B.*1. *KNN analysis*

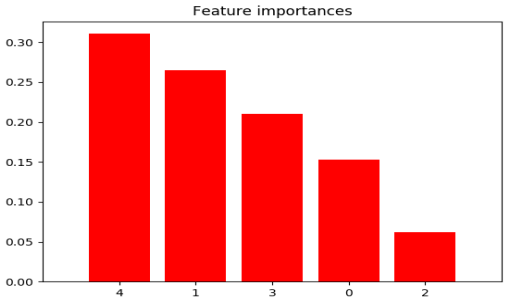
Videli smo da KNN ima najveću preciznost. Potom smo prišli proučavanju rezulta koje smo dobili od ovog algoritma. Prvi korak nam je bila izbor parametara.

Izbor parametara je proces u kome biramo one parametre koji najviše doprinose predviđanju izlaza koji nas interesuje.

Kada imamo previše nevažnih parametara u našim podacima mogu smanjiti preciznost modela. Tri glavne prednosti za dobar izbor parametara pre modelavanja podataka je:

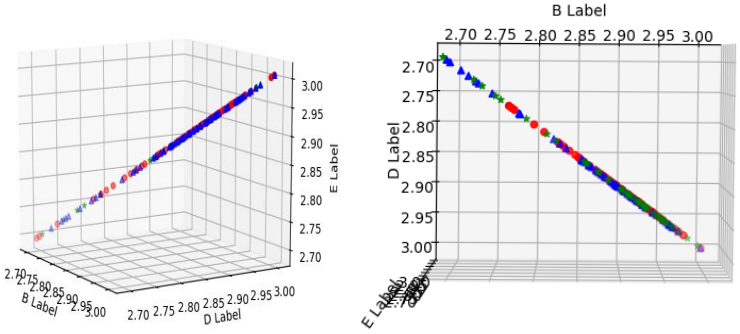
- Smanjenje overfitting-a: Manje redundantnih podatak znači manje šansi da se klasifikacija vrši na osnovu šuma.
- Povećanje preciznosti: Manje obmanjujućih podatata znači na se preciznost modela poboljšava.
- Smanjenje vremena potrebno za treniranje: Manje podataka znači brže treniranje algoritma.

Za naše parametre dobili smo podatke koji su prikazani na Sl. 13.



Slika 13. 1. parametar 4 (0.310558) 2. parametar 1 (0.264456) 3. parametar 3 (0.210032) 4. parametar 0 (0.152823) 5. parametar (0.062130)

Uzeli smo tri najvažnija parametra da bi napravili 3D grafikon koji predstavalj raspodelu za tri vrste organa (pluća, bubreg, dojka).



(a) (b)

Slika 14. Crvene tačke - Grudi , plave tačke - Bubreg, zelene tačke - Pluća. Slika (b) je pogled samo odozgo (iz ptičije perspektive).

Podešavanje algoritma je poslednji proces u mašinskom učenju pred prezentiranje dobijenih podataka.

Ovo se ponekad zove i **Hiperparametarska** optimizacija, gde se parametri algoritmi nazivaju **hiperparametri**, a koeficijenti koji su dobijeni sami po sebi od strane mašinskog učenja nazivaju se **parametri**. Optimizacija predlaže pretragu - prirodu problema.

Ako formulišemo kao problem pretraživanja, onda možemo koristiti različite strategije za pronalaženje dobrog i robusnog parametra ili skupa parametara za algoritam koji rešava naš problem.

Kao pristup za podešavanje parametara koristili smo slučajnu pretragu, koja uzima parametre algoritma na osnovu slučajne tj. uniformne raspodela. Ovo je rađeno fikstan broj puta. Za svaku iteraciju konstruisan je model, moje pro proučavan i testiran.

Najbolje parametre koje smo dobili su: **n_neighbors**=11 **metric**=euclidean. Rezultati sa tim parametrima su prikazani u Tabeli 1.

Tabela 1.

	Ukupno	Tačno	Procenat tačnosti
Pluća	14	7	50.00%
Bubreg	14	8	57.14%
Dojka	14	13	92.85%
Ukupno	42	28	64.29%

Tabela 2.

	Ukupno	Tačno	Procenat tačnosti
Pluća	11	8	72%
Bubreg	9	7	78%
Dojka	8	6	75%
Ukupno	28	21	75%

B.2. Analiza SVC algoritma

Ovaj algoritam je pokazao najbolje rezultate, što je prikazano u Tabeli 2. Koristili smo nasumičnu pretragu za podešavaje parametara algoritma. Najbolji rezultati koji su dobijeni da test podacima su: **C**=1000, **kernel**=linear.

ACKNOWLEDGEMENTS

This paper is translated and adapted for serbian language readers. It was published in 5th International Conference of Advanced Computer Science & Information Technology (2017) under the name *Classification algorithms for the detection of the primary tumor based on microscopic images of bone metastases.* Retrieved from http://airccj.org/CSCP/vol7/csit77004.pdf

LITERATURA

[1] G. Landini, "Fractals in microscopy," *Jurnal of Microscopy*, vol. 241, no. 1, pp. 1-8, 2011.

[2] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, vol. 46, no. 3, pp. 175-185, 2012.

[3] F. V. G. G. A. M. V. T. B. G. O. B. M. P. P. W. R. V. V. J. P. A. C. D. B. M. P. M. D. E. Pedregosa, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 12 10 2011.

[4] L. B. a. G. L. a. M. B. a. F. P. a. A. M. a. O. G. a. V. N. a. P. P. a. A. G. a. J. G. a. R. L. a. J. V. a. J. a. Bria, "API design for machine learning software: experiences from the scikit-learn project," *European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases*, pp. 108-122, 2013.

[5] R. T. J. F. Trevor Hastie, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.

[6] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[7] J.-M. P. C. T.-M. Robin Genuer, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225-2236, 2010.

[8] K. E. F. Glenn De'ath, "CLASSIFICATION AND REGRESSION TREES: A POWERFUL YET SIMPLE TECHNIQUE FOR ECOLOGICAL DATA ANALYSIS," *Ecology*, vol. 81, no. 11, pp. 3178-3192, 2000.

[9] S. Mannel, "Decision Tree Classification of a Forest Using AVIRIS and Multi-Seasonal TM Data," South Dakota School of Mines and Technology, Rapid City, South Dakota, 2012.

[10] L. S. J. T. D. R. K. Jason D. M. Rennie, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," in *ICML*, 2013.

[11] L. D.D., "Naive (Bayes) at forty: The independence assumption in information retrieval.," *Machine Learning: ECML-98.*, vol. 1398, pp. 4-15, 1998.

[12] B. G. K. M. S. Mohammad Ahsanullah, Normal and Student´s t Distributions and Their Applications, vol. 4, Atlantis Studies in Probability and Statistics, 2014, pp. 7-50.

[13] E. D. P. Z. M. H. D. E. J. M. P. B. M. Pizer, "Contrast Limited Adaptive Histogram Equalization image processing to improve the detection of simulated spiculations in dense mammograms," *Journal of Digital Imaging*, vol. 193, no. 11, 1998.

[14] R. J. R. P. G. a. H. P. Srivastava, "Enhancement and restoration of microscopic images corrupted with poisson's noise using a nonlinear partial differential equation-based filter," *Defence Science Journal,* vol. 61, no. 5, p. 452, 2011.

[15] S. Z. B. M. H. E. D. Pisano, "Contrast Limited Adaptive Histogram Equalization image processing to improve the detection of simulated spiculations in dense mammograms," *Journal of Digital Imaging,* vol. 11, no. 4, pp. 192-200, 1998.

[16] Y. Anzai, Pattern recognition and machine learning, Harcourt Brace Jovanovich, 2012.

[17] A. M. &. A. K, "OpenCV-Python Documentation", 2017. [Online]. Available: https://media.readthedocs.org/pdf/opencv-python-tutroals/latest/opencv-python-tutroals.pdf. [Accessed 23 05 2017].

[18] Dewi Suryani, S.KOM., M.ENG., "CONVOLUTIONAL NEURAL NETWORK", 2017. [Online]. Available: http://socs.binus.ac.id/2017/02/27/convolutional-neural-network [Accessed 27 05 2017].

ABSTRACT

This paper presents the analysis of techniques for microscopic images in order to find a primary tumor based on the of bone metastases. Was done alg orithmic classification into three groups, kidney, lung and breast. In order to speed up the treatment of the patient and easier for doctors and therefore reduce room for human error. Digital microscope images of bone metastases were analyzed, for which it is known that the primary tumor is in one of the three human organs: kidney, lung or breast. We tested several solutions for classification, were tested two methods of image analysis. Multifractal analysis and convolutional neural network. Both methods were tested with and without preprocessing image. Results of multifractal analysis were then classified using different algorithms. Images were processed using CLAHE and k-means algorithm. At the end, the results obtained using a variety of techniques are presented.

CLASSIFICATION ALGORITHMS FOR THE DETECTION OF THE PRIMARY TUMOR BASED ON MICROSCOPIC IMAGES OF BONE METASTASES

Slađan Kantar, Aleksandar Pluškoski i Igor Ciganović, Jelena Vasiljević