Milica Medić

Klasifikacija tumora dojke pomoću konvolucionih neuronskih mreža

Histopatološka analiza je i dalje zlatni standard u prepoznavanju i dijagnozi tumora. Ovaj rad predstavlja izradu softvera za klasifikaciju histopatoloških slika tumora dojke na podtipove malignih i benignih tumora. Klasifikacija je zasnovana na konvolucionim neuronskim mrežama specijalizovanim za rad na slikama, koje su istrenirane u 60 epoha, a čiji su atributi određeni emprijiski kroz veći broj treninga. Arhitektura mreže se zasniva na naizmeničnom smenjivanju konvolucionih slojeva i slojeva sažimanja. Za izvor podataka uzeta je velika, javno dostupna The BreakHis baza obeleženih tipova tumora dojke. Slike su prethodno obrađene posebnim programima za isecanje belina i podeljene na manje delove, te na taj način pripremljene za dalju analizu. Tačnost klasifikacije ovog softvera približno je jednaka tačnostima modela slične primene, čime se može zaključiti da program uspešno vrši klasifikaciju. Tačnost može povećati daljim prilagođavanjem atributa tipovima tumora. Program je uprošćen i lak za razumevanje, samim tim i pristupačan korisnicima kojima je ova oblast manje poznata.

Uvod

Mašinsko učenje je proces koji omogućava veštačkoj inteligenciji da oponaša ljudske sposobnosti učenja i rešavanja problema analizirajući podatke i obrasce unutar njih. U radovima fokusiranim na obradu slika koriste se konvolucione neuronske mreže (engl. convolutional neural networks – CNNs). Konvolucione neuronske mreže su ime dobile po konvolucionim filterima koji se primenjuju da bi se uočile specifične karakteristike i formirale mape tih karakteristika. Postoje tri osnovna tipa slojeva unutar svake konvolucione mreže: konvolucioni slojevi (engl. convolutional layers) koji su i osnovni gradivni blokovi, slojevi sažimanja (engl. pooling layers), kojim se smanjuje broj parametara i kontroliše pretreniranost modela, i potpuno povezani slojevi (engl. fully connected layers), u kojima je svaki neuron povezan sa svim neuronima prethodnog sloja (Nielsen 2015).

Konvolucione mreže su našle svoju primenu i u medicinskim istraživanjima, gde se koriste za obradu slika dobijenih magnetnom rezonancom (Pereira *et al*. 2016) i kompjuterizovanom tomografijom (Nikolov *et al*. 2018), kao i snimaka histopatoloških preparata (Saltz *et al*. 2018). Ovakva prmena konvolucionih neuronskih mreža je od velikog značaja prilikom postavljanja dijagnoze tumora. Primenom mašinskog učenja greška dijagnoze može da se redukuje, jer se odstranjuje subjektivni faktor, a u obzir se uzimaju obrasci koje čovek ne može da prepozna, i time se povećava preciznost dijagnsotike (Saria *et al*. 2018).

Tumor predstavlja skup ćelija koje odlikuje abnormalana i nekontrolisana deoba unutar tkiva. Razlikuju se dva tipa tumora: benigni i maligni. Benigni tumori uglavnom ostaju dobroćudni tokom celog života, dok maligni tumori poseduju posebno izmenjene ćelije, koje se karakterišu nekontrolisanom deobom, dospevaju u krv i limfu, čime ova bolest može metastazirati i dovesti do smrti. Maligni tumor koji potiče od ćelija vezivnog tkiva naziva se još i sarkom, a onaj koji potiče od epitelnih ćelija naziva se karcinom, kao što je karcinom dojke (Key *et al*. 2001). Karcinom je druga po redu najsmrtonosnija grupa bolesti, i prema podacima Svetske zdravstvene organizacije u 2018. godini je umrlo

Milica Medić (2001), Zrenjanin, učenica 3. razreda Zrenjaninske gimnazije

MENTOR: Luka Velimirov, student Biološkog fakulteta Univerzitet u Beogradu

9.6 miliona ljudi od ove bolesti u svetu (WHO 2018). Posle karcinoma pluća, karcinom dojke je najzastupljeniji tip karcinoma kod žena (Key *et al*. 2001). I pored napretka u digitalnoj analizi tumora dojke, najviše primenjivana metoda i dalje je pregled histoloških preparata od strane patologa. U klasifikaciji ovog tumora na maligni i benigni na osnovu manjeg ili većeg broja slika pomoću neuronskih mreža, u raznim istraživanjima dostignuta je tačnost preko 90% (Kowal *et al*. 2013; Filipczuk *et al*. 2013; George *et al*. 2013).

Dosadašnji radovi na ovu temu kao bazu podataka koristili su veoma mali broj slika, sve dok nije objavljena velika baza podataka *The BreakHis Database* koja sadrži 7909 snimaka histoloških preparata tumora dojke od 82 pacijenta (Spanhol *et al*. 2015). Na već treniranu arhitekturu AlexNet (Krizhevsky *et al*. 2012) za klasifikaciju obojenih predmeta, primenjena je ova baza i mreža je pokazala tačnost od 80% do 85%, u zavisnosti od uveličanja na mikroskopu (Spanhol *et al*. 2016). Rezultatima ovog istraživanja je pokazano da se arhitektura stvorena za rad na slikama niske rezolucije može primeniti i na slike visoke rezolucije, što prestavlja prednost u korišćenju već treniranih mreža za klasifikaciju histopatoloških slika. Najveću tačnost imali su treninzi vršeni na slikama uveličanja mikroskopa 40 i 100 puta, jer su najviše odgovarali veličini slike 64×64 prilikom smanjenja rezolucije.

Ovaj rad predstavlja izradu softvera za klasifikaciju tumora dojke, na osnovu histopatoloških slika, uz primenu klasifikatora koji omogućuju razlikovanje četiri podtipa malignih tumora: duktalni (*ductal carcinoma* – DC), lobularni (*lobular carcinoma* – LC), mucinozni (*mucinous carcinoma* – MC) i papilarni karcinom (*papillary carcinoma* – PC), kao četiri podtipa benignih tumora: fibroadenom (*fibroadenoma* – F), filoldes tumor (*phyllodes tumors* – PT), adenoza tumor (*adenosis* – A) i tubularni adenom (*tubular adenoma* – TA).

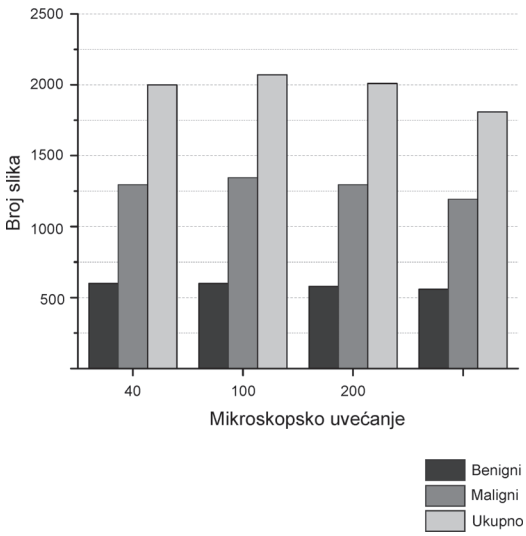
Metodologija

Da bi se istrenirala konvoluciona neuronska mreža, bilo je neophodno napraviti novu arhitekturu. Za trening nove arhitekture korišćena je The BreakHis baza podataka. Program je napisan u programskom jeziku Python 3.7.

Baza podataka

Slike korišćene za trening pripadaju The BreaKHis bazi podataka koja sadrži 7909 slika prikupljenih od 82 pacijenta (slika 1). Slike su prikupljane tokom kliničkog istraživanja u Brazilu 2014. godine (Spanhol *et al*. 2016). Svrstane su u dve kategorije: maligne i benigne. Slike malignih tumora podeljene su u četiri potkategorije: duktalni, lobularni, mucinozni i papilarni karcinom. Slike benignih tumora su razvrstane isto u četiri potkategorije: fibroadenomi, filodes tumore, adenozu i tubularni adenom. Slike su napravljene pri uvećanjima 40, 100, 200 i 400 puta, i sačuvane u RGB obliku.

Za obradu slika su napisane tri skripte za sečenje slika na manje delove, menjanje veličine na 54×54 i otklanjanje belih delova slika, jer na njima nema smisla učiti mrežu, a doprinose po-



Slika 1. Broj histopatoloških slika tumora dojke po uveličanjima i klasama u The BreakHis bazi podataka koje su korišćene za treniranje modela u ovom radu

Figure 1. Number of histopathological images of breast cancer by magnification and class from The BrekHis database used for training the model described in this paper

goršanju performansi. Obrađene slike sačuvane su u .pkl formatu.

Arhitektura modela

Program je pisan u programskom jeziku Python 3.7, u okruženju Spyder. Biblioteka za duboko učenje je TFLearn. Arhitektura za model koji klasifikuje tumore na njihove podtipove formirana je po principu Conv-Pool-Conv-Pool, gde se smenjuju konvolucioni slojevi i slojevi sažimanja (šema u prilogu ovog članka). Formirana je po uzoru na programe slične namene koji su obrađivali slike tumora mozga i dojke (Pereira *et al*. 2016; Spanhol *et al*. 2016). Model se sastoji od četiri konvoluciona sloja (kerneli veličine 3×3, 3×3, 5×5 i 5×5; broj filtera 16, 32, 64 i 128, respektivno), četiri Max Pooling sloja (veličine 2×2) za sažimanje, tri potpuno povezana sloja (512 neurona i 4 izlazna neurona kao broj izlaznih podataka) i jedan Dropout sloj (sloj izbacivanja koji sprečava pretreniranost nuliranjem neurona koji se nalaze u skrivenim slojevima), kao i jednog sloja za poravnavanje. Broj epoha je 60. Za poboljšanje performansi korišćen je ReLu sloj, kao aktivaciona funkcija nakon svakog konvolucionog sloja.

Unakrsna validacija predstavlja podelu originalne baze podataka na određeni broj jednakih delova – *k*, što je ujedno i broj ponavljanja procesa. Svaki put, mreža izdvaja jedan deo i koristi ga za validaciju, dok se na ostalima vrši trening. U ovom radu je korišćena unakrsna validacija u kojoj je broj *k* iznosio 6. Na ovaj način se obezbeđuje da se u svakom delu nalazi odgovarajuća proporcija podataka.

Trening i prikaz rezultata

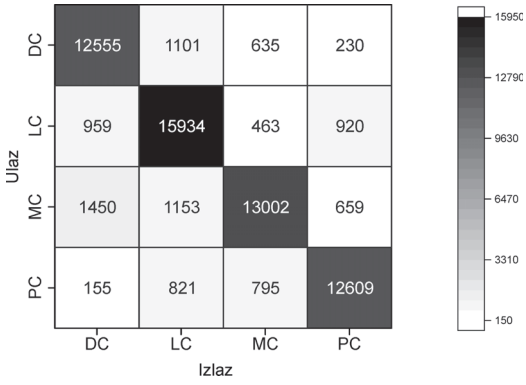
Trening mreža rađen je na virtuelnom cloud računaru operativnog sistema Windows 10 (7 Gb RAM memorije; 4 jezgra), kojem je pristupljeno preko usluge Microsoft Azure.

Rezultati su prikazani u obliku matrica konfuzije (engl. confusion/error matrix) koje pokazuju broj slika koje je program tačno odnosno netačno predvideo u svakoj klasi. Broj slika poredi se sa skalom koja se primenjuje na slikovnom prikazu matrice, te se na ovaj način može odrediti tačnost programa. Na vertikalnoj osi prikazane su tačne vrednosti (ulaz), a na horizontalnoj osi predviđene vrednosti (izlaz). Na osovu podataka iz matrica računa se tačnost i gubitak modela, koji se prikazuju pomoću grafika. Tačnost predstavlja sposobnost programa da predvidi klasu slike, izražava se u procentima i očekuje se njen rast tokom treninga. Gubitak predstavlja skup svih grešaka tokom treninga, odnosno relativni udeo tačno ili pogrešno klasifikovanih pozitiva, i očekuje se njegov pad tokom treninga.

Kodovi za obradu slika, kao i obrađena baza podataka dostupni su na GitHub-u zajedno sa arhitekturom za klasifikaciju tumora koja je korišćena za trening modela u ovom radu može se naći na repozitorijumu (Medić 2019).

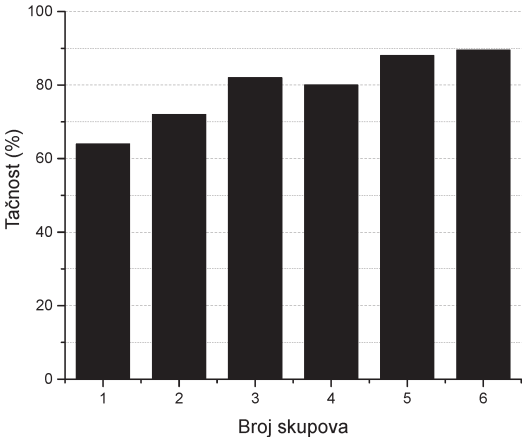
Rezultati testiranja

Prvobitna ideja je bila korišćenje ovog modela za klasifikaciju podataka iz The BreakHis baze na maligne i benigne tumore, što bi podrazumevalo podelu na osam klasa. Nakon 13



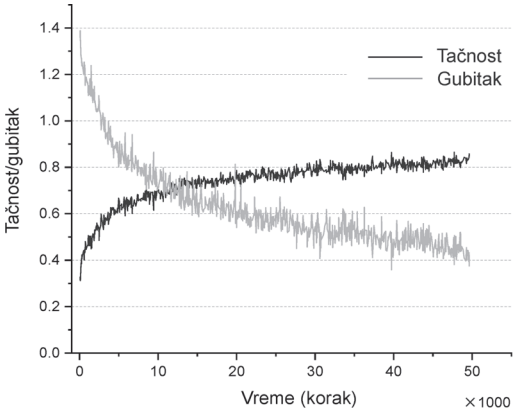
Slika 2. Matrica konfuzije, rezultat nakon treninga na slikama malignog tumora dojke; izlazni podaci su vrednosti predviđene od strane programa. Oznake: DC – *ductal carcinoma*, LC – *lobular carcinoma*, MC – *mucinous carcinoma* i PC – *papillary carcinoma*.

Figure 2. Confusion matrix, the result of training on malignant breast tumor images; the input data are the values provided by the model. Tags: DC – *ductal carcinoma*, LC – *lobular carcinoma*, MC – *mucinous carcinoma* and PC – *papillary carcinoma*.



Slika 3. Unakrsna validacija prilikom treninga programa na slikama malignog tumora dojke

Figure 3. Cross-validation through the training on malignant breast tumor images

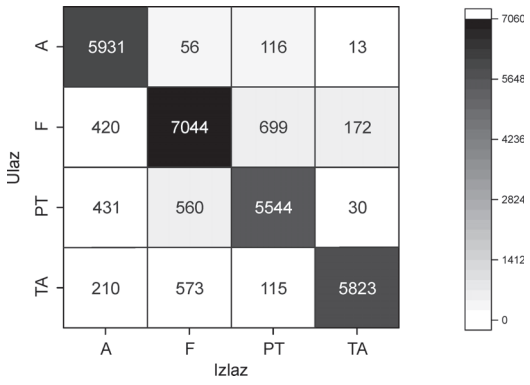


Slika 4. Prikaz tačnosti i gubitka tokom treninga na slikama malignog tumora dojke

Figure 4. Accuracy (dark gray) and loss (light gray) through training on malignant breast tumor images

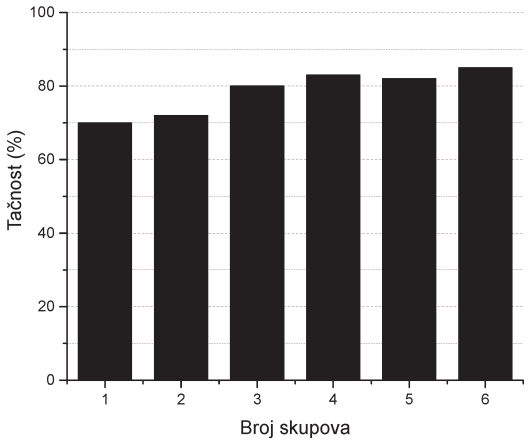
epoha tačnost je iznosila samo 8%, pa je trening prekinut.

Trening modela za klasifikaciju malignih tumora na podgrupe kao najbolju tačnost je imao 87.2% (za celu bazu podataka) i 80.9% (za test



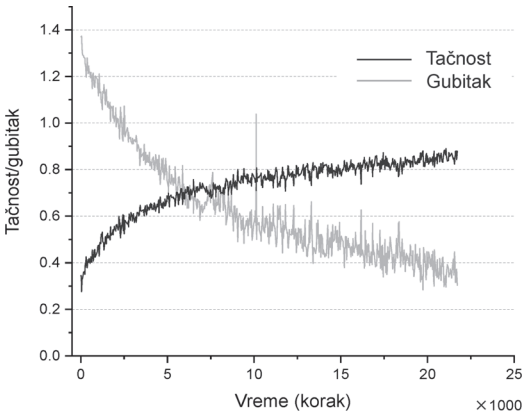
Slika 5. Matrica konfuzije, rezultat nakon treninga na slikama benignog tumora dojke; ulazni podaci su vrednosti predviđene od strane programa; izlazni podaci su vrednosti predviđene od strane programa. Oznake: F – *fibroadenoma*, PT – *phyllodes tumors*, A – *adenosis* i TA – *tubular adenoma*.

Figure 5. Confusion matrix, the result of training on benign breast tumor images; the input data are the values provided by the model.
Tags: F – *fibroadenoma*, PT – *phyllodes tumors*, A – *adenosis* and TA – *tubular adenoma*.



Slika 6. Unakrsna validacija prilikom treninga programa na slikama benignog tumora dojke

Figure 6. Cross-validation through the training on benign breast tumor images



Slika 7. Prikaz tačnosti i gubitka tokom treninga na slikama benignog tumora dojke

Figure 7. Accuracy and loss through training on benign breast tumor images

bazu podataka) i trajao je 13.8 sati. Najmanji gubitak je iznosio 0.359 (slike 3–5).

Trening konvolucione mreže za klasifikaciju benignih tumora kao najbolju tačnost je imao 89.1% (za celu bazu podataka) i 79.9% (za test bazu podataka) i trajao je 5.8 sati. Najmanji gubitak je iznosio 0.283 (slike 6–8).

Diskusija

Problem kod ovakvih programa jeste baza podataka, koja često nije dostupna, nije dovoljno velika, slike nemaju isti kvalitet i nije sigurno koliko su tačno obeležene, jer je to nešto što je potrebno uraditi ručno. Veliki broj podataka je bez provere postavljen na internetu gde je slobodno dostupan, ali ne sadrži potrebne podatke za klasifikaciju, odnosno slike nisu podeljene u kategorije (labels). Ručno obeležavanje slika je zahtevan proces, s obzirom da se baza često sastoji iz više hiljada, pa i desetina hiljada slika. Podatke je teško sakupiti, jer pacijenti često ne žele da daju svoje slike u ovakve biblioteke. Zbog svega ovoga je usporen dalji napredak ovakve primene mašinskog učenja u medicinske svrhe (Xu *et al*. 2017).

Prvobitni cilj da se ova baza podataka klasifikuje na maligne i benigne tumore je promenjena u klasifikaciju na podtipove za svaku od ove dve vrste odvojeno. To je uređeno zbog veoma male tačnosti programa konstatovane tokom treninga, pa je takva klasifikacija ubrzo prekinuta. Ovakav program, koji bi klasifikovao slike tumora dojke na maligne i benigne, zahtevao bi modifikaciju arhitekture, tako da bude primenljiva za rad sa osam klasa, po četiri podtipa za svaku od ove dve klase. To bi se moglo postići dodavanjem konvolucionih slojeva i menjanjem broja parametara prilikom podele baze slika.

Menjanjem parametara unutar arhitekture mogao bi da se poboljša rad i ishod treninga, ali samo treniranje traje dugo, i onemogućava lake varijacije modela. Dodavanjem još slojeva, menjanjem broja filtera i njihove veličine, povećanjem boja epoha i povećanjem broja podataka za unos, tačnost predikcije bi bila veća. Takođe, tačnost predviđanja se može uvećati i uočavanjem suvišnih atributa, kao i atributa koji nedostaju. Poznato je da povezivanje (fusion rules) više različitih mreža i njihova kombinacija dovodi do poboljšanja rada mreže (Spanhol *et al*. 2016), što bi se moglo primeniti i na ovu mrežu. Izbegavanje pretreniranosti bi se moglo izbeći i konkretnijom obradom slika pomoću opcije mirroring, gde bi se slika obrtala i služila kao poseban unos, kao i isecanjem okvira i pomeranjem slike za par piksela. Na ovaj način mreža ne bi napamet učila i bolje bi uviđala obrasce, te bi se pretreniranost svela na minimum.

Iz grafičkih prikaza unakrsne validacije (slika 4 i slika 7) kod ove mreže se može primetiti sličnost tokom svih šest iteracija, što potvrđuje da je procena uspešnosti modela pouzdana. Na osnovu podataka iz matrica konfuzije (slika 3 i slika 6) programom su generisani tačnost i gubitak modela, koji su prikazani na graficima (slika 5 i slika 8). Iz pomenutih grafika jasno se uviđa porast tačnosti i smanjenje gubitka, što je bio i očekivan rezultat. U poređenju sa modelim iz srodnih oblasti (Kowal *et al*. 2013; Filipczuk *et al*. 2013; George *et al*. 2013), u kojima tačnost programa iznosi od 90% do 100%, rad našeg programa može se smatrati uspešnim.

Zaključak

Ovaj rad predstavlja klasifikaciju malignih i benignih tumora dojke na njihove podtipove pomoću softvera zasnovanog na konvolucionoj neuralnoj mreži. Arhitektura konvolucione mreže je potpuno nova, i ovim radom pokazano je da ju je moguće koristiti za klasifikaciju histopatoloških preparata, posebno pri klasifikaciji slika tumora dojke na podtipove uz pomoć označenih slika unutar baze. Program ostavlja mesta za dalje usavršavanje i treniranje obe mreže pomoću novog materijala i uz pogodno menjanje parametara poput broja slika, slojeva, ili menjanjem biblioteka. Na ovaj način bi se empirisjki moglo doći do optimalnog rada ove mreže. Kodovi za obradu slika, kao i obrađena baza podataka dostupni su na GitHub-u zajedno sa kodovima za klasifikaciju tumora.

Zahvalnost. Zahvaljujem se Ognjenu Milićeviću, asistentu Medicinskog fakulteta Univerziteta u Beogradu i Ivi Veljković, studentkinji Elektrotehničkog fakulteta Univerziteta u Beogradu.

Literatura

Filipczuk P., Fevens T., Krzyżak A., Monczak R. 2013. Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. *IEEE Transactions on Medical Imaging*, **32** (12): 2169.

George Y. M., Zayed, H. H., Roushdy, M. I., Elbagoury B. M. 2013. Remote computer-aided breast cancer detection and diagnosis system based on cytological images. *IEEE Systems Journal*, **8**: 949.

Key T. J., Verkasalo P. K., Banks E. 2001. Epidemiology of breast cancer. *The lancet oncology*, **2** (3): 133-140.

Kowal M., Filipczuk P., Obuchowicz A., Korbicz J., Monczak R. 2013. Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images. *Computers in biology and medicine*, **43** (10): 1563.

Krizhevsky A., Sutskever I., Hinton G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. U *Proceedings of the 25th International Conference on Neural Information Processing Systems* (ur. P. Bartlett). NIPS, str. 1097–1105.

Medić M. 2019. Petnica BMD2019: Klasifikacija histopatoloških slika raka dojke uz pomoć mašinskog učenja – Milica Medić. https://github.com/lxka/PetnicaBMD2019/tree/BreastCancerCNN

Nielsen M. A. 2015. *Neural networks and deep learning*. San Francisco: Determination press

Nikolov S., Blackwell S., Mendes R., De Fauw J., Meyer C., *et al*. 2018. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv preprint arXiv:1809.04430

Pereira S., Pinto A., Alves V., Silva C. A. 2016. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE transactions on medical imaging*, **35** (5): 1240.

Saltz J., Gupta R., Hou L., Kurc T., Singh P., *et al*. 2018. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, **23**: 181.

Saria S., Butte A., Sheikh A. 2018. Better medicine through machine learning: What's real, and what's artificial?. *PLoS Med*, **15** (12): e1002721. https://doi.org/10.1371/journal.pmed.1002721

Spanhol F. A., Oliveira L. S., Petitjean C., Heutte L. 2015. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, **63** (7): 1455.

Spanhol F. A., Oliveira L. S., Petitjean C., Heutte L. 2016. Breast cancer histopathological image classification using convolutional neural networks. U *2016 international joint conference on neural networks (IJCNN)*. IEEE, str. 2560-2567.

WHO (World health organization) 2018. Cancer – key facts. https://www.who.int/news-room/fact-sheets/detail/cancer

Xu Y., Jia Z., Wang L. B., Ai Y., Zhang F., *et al*. 2017. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC bioinformatics*, **18** (1): 281.

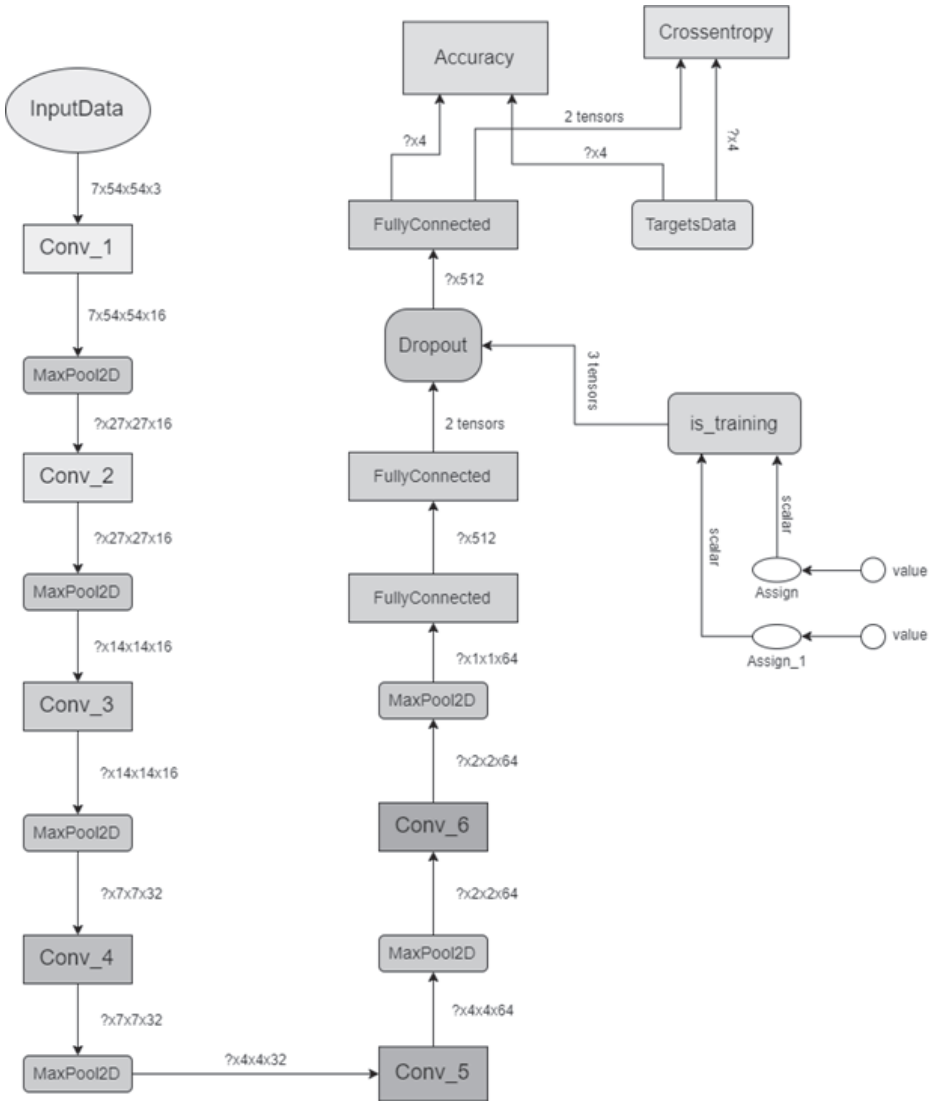
Milica Medić

Classification of Breast Cancer Histopathological Images Using Machine Learning

The usage of machine learning in scientific research has attracted significant attention recently, and has found its use in the medical field as well, especially in image and sound processing. Histopathological image analysis is still the most used method in cancer diagnosis, mostly done personally by a pathologist. In order to prevent misdiagnosis, this work focuses on using machine learning and artificial intelligence in programmed classification of tumor images. The architecture of the software is based on models that sort the images of different types of tumors. The database used in this work is The BreakHis database with labeled images of breast cancer, that can be easily found online. These images are already used in a model that classifies them into two categories: the malignant and the benign. This model is different from others because we have added the module that classifies malignant and benign tumors into their subtypes. Every program is made out of convolutional neural networks, which are specialized in image processing, and each of them was trained for 60 epochs. As expected, the accuracy during the training of every model is close to the accuracy of already existing programs, and this leaves room for further adjustment of features specific for this cancer type.

This software, simplified and easier to understand and use, is accessible to people less familiar with this matter. Programs used for the image processing and the whole data can be found, along with programs used for every classification, on GitHub (Medić 2019).

Prilog: arhitektura mreže



Arhitektura mreže korišćene prilikom treninga i klasifikacije malignih i benignih preparata tumora; shema algoritma je generisana u programskom jeziku Python pomoću biblioteke za duboko učenje TFLearn; Conv_n – konvolucioni sloj, MaxPool2D – sloj sažimanja dvodimenzionalnih prostornih podataka, FullyConnected – potpuno povezani sloj, Dropout – sloj izbacivanja, tensors – uopštavanje skalara, vektora i matrica, Accuracy – tačnost modela, Crossentropy – provera tačnosti predviđanja modela.

The architecture used for training and classification of malignant and benign tumor images; the scheme of the algorithm is generated in Python using the deep learning library TFLearn; Conv_n – convolutional layer, MaxPool2D – pooling layer of two dimensional spatial data, FullyConnected – fully connected layer, Dropout – dropout layer, tensors – generalisation of scalars, vectors and matrices, Accuracy – the accuracy of the model, Crossentropy – model accuracy check.