SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 2016

KLASIFIKACIJA HISTOPATOLOŠKIH SNIMAKA DIJELOVA LIMFNIH ČVOROVA POMOĆU STROJNOG UČENJA

Domagoj Pluščec

Zagreb, lipanj 2019.

Zahvaljujem mentoru doc. dr. sc. Marku Čupiću i suradniku dr. sc. Tomislavu Lipiću s Instituta Ruđer Bošković koji su mi pomogli prilikom izrade ovog rada. Ujedno zahvaljujem Institutu Ruđer Bošković i Istraživačkom centru mladih na ustupljenoj infrastrukturi za potrebe izrade ovog rada.

**Sadržaj**

[1. Uvod 5](#_Toc8476381)

[2. Pregled dosadašnjih istraživanja 6](#_Toc8476382)

[3. Klasifikacija histopatoloških slika limfnih čvorova 7](#_Toc8476383)

[3.1 Skup podataka 7](#_Toc8476384)

[3.2 Proširivanje skupa podataka 7](#_Toc8476385)

[3.3 Odabir pristupa rješavanja problema klasifikacije 7](#_Toc8476386)

[4. Duboko učenje u analizi histopatoloških slika 8](#_Toc8476387)

[4.1 Korišteni modeli 8](#_Toc8476388)

[4.1.1 AlexNet 8](#_Toc8476389)

[4.1.2 Resnet 8](#_Toc8476390)

[4.1.3 Inception 8](#_Toc8476391)

[4.1.4 DenseNet 8](#_Toc8476392)

[4.2 Učenje prijenosom značajki 8](#_Toc8476393)

[5. Rezultati 9](#_Toc8476394)

[6. Zaključak 10](#_Toc8476395)

[7. Literatura 11](#_Toc8476396)

[Sažetak 12](#_Toc8476397)

[Summary 13](#_Toc8476398)

[Prilog – programski kod 14](#_Toc8476399)

**Popis slika**

[Slika 1. Primjer histopatoloških slika iz skupa podataka PCam [5] koji sadrže tumor 9](#_Toc8657562)

[Slika 2. Primjeri histopatoloških slika iz skupa podataka PCam [5] koji ne sadrže tumor 9](#_Toc8657563)

# Uvod

Računarstvo je promijenilo zdravstvene sustave iz temelja. Računala su se počela koristiti u administraciji u bolnicama, promijenila su način na koji liječnici bilježe pacijentove podatke sustavima kao što je e-karton u hrvatskom zdravstvenom sustavu, počeli su se koristiti za naručivanje lijekova, izdavanje recepata, a između ostalog i u analizi medicinskih slika i u potpomognutoj dijagnostici korištenjem algoritama strojnog učenja.

Računalni sustavi za analizu medicinskih slika omogućava kvantitativni opis medicinskih slika što je bitno kako bi smanjili pogreške liječnika te kako bi ubrzali proces označavanja slika. Također, računalne sustave u obradi medicinskih slika moguće je koristiti u istraživanju, primjerice za proučavanje bioloških mehanizama razvoja bolesti [11].

Cilj ovog rada je dati pregled područja analize histopatoloških slika i izrada programske implementacije za klasifikaciju histopatoloških slika limfnih čvorova pomoću strojnog učenja.

Ovaj diplomski rad strukturiran je kroz šest poglavlja. U ovom (uvodnom) poglavlju predstavljena je motivacija za istraživanje, proučavanje i primjenu računalnih metoda za analizu podataka iz medicinskih snimaka. Nakon motivacije predstavljen je cilj ovog diplomskog rada.

U drugom poglavlju dan je pregled područja analize medicinskih slika korištenjem strojnog učenja s naglaskom na analizu histopatoloških slika. U trećem poglavlju opisan je konkretan problem koji će se rješavati, predstavljen je korišteni skup podataka te odabir pristupa rješavanju problema. U četvrtom poglavlju opisane su korištene metode dubokog učenja, podskupa strojnog učenja. Rezultati provedenih eksperimenata dani su u petom poglavlju.

# Pregled dosadašnjih istraživanja

Analizu medicinskih slika prvenstveno su radili doktori na otisnutim radiološkim nalazima, odnosno na mikroskopima u analizi histopatoloških uzoraka[1]. Napretkom računala i omogućavanjem spremanja i učitavanja medicinskih slika u računalima omogućena je i njihova analiza korištenjem algoritama. Tako su se i algoritmi strojnog učenja počeli koristiti u analizi medicinskih slika već od sredine 60 tih godina prošlog stoljeća [2][3].

Sustavi potpomognute dijagnostike (engl. *computer aided diagnosis*, CAD) predstavljaju cjelovite sustave koji mogu između ostalog koristiti i analizu medicinskih slika kako bi pomogli doktoru u donošenju zaključaka temeljem nalaza pretraga i u konačnici u donošenju dijagnoze pacijentove bolesti. Primjenom analize medicinskih slika sustavi potpomognute dijagnostike su 80tih godina prošlog stoljeća počeli pomagati doktorima u detekciji raka i dijagnostici na radiološkim slikama prsa [3]. Od tada analiza medicinskih slika se proširila i na druge vrste radioloških slika kao što su računalna tomografija (engl. *computed tomography*, CT), magnetska rezonancija (engl. *magnetic resonance imaging*, MRI) i ultrazvuk [3].

Razvojem digitalne mikroskopije omogućeno je skeniranje većih uzoraka, s različitim uvećanjem i s velikom rezolucijom slika. Jedno od takvih dostignuća je tehnika skeniranja čitavog preparata (engl. *whole slide imaging*, WSI) [1]. Prvo natjecanje koje je ponudilo sudionicima podatke u obliku slika dobivenih WSI metodom je bio Camelyon16 [4]. Tipična veličina WSI primjera iz Camelyon17 skupa podataka je 200,000x100,000 slikovnih elemenata na najvećoj rezoluciji s 3 bajta memorije za svaki slikovni element što predstavlja ukupno 55.88GB podataka s jedne razine za jednog pacijenta [4]. Kako je slike takve rezolucije teško ručno analizirati došlo je do potrebe za razvojem automatiziranih sustava za analizu takvih slika. Zadatci takvih sustava su potpomognuta dijagnoza, dohvaćanje sličnih slika te pronalaženje anomalija [1].

# Klasifikacija histopatoloških slika limfnih čvorova

Histologija je grana biologije čiji je cilj proučavanje tkiva biljnog i životinjskog podrijetla od stanične i međustanične razine do organa. Patološka histologija (histopatologija) je grana histologije koja proučava građu bolesnih tkiva [7]. Predmet proučavanja ovog rada su slike histopatoloških slika limfnih čvorova. Limfni čvorovi su dio limfnog sustava, a služe za filtriranje limfe od mikroorganizama i stranih antigena [8].

Cilj ovog poglavlja je predstaviti zadatak klasifikacije slika, korišteni skup podataka skup podataka te odabir pristupa rješavanju problema.

## Klasifikacija slika

Pristupe strojnom učenju možemo podijeliti na nadzirano učenje, slabo nadzirano učenje, nenadzirano učenje i podržano učenje. [Understanding machine learning].

Zadatak ovog rada, klasifikacija, se svrstava u nadzirano učenje. Nadzirano učenje je oblik strojnog učenja u kojem su dostupni podaci i njihove oznake. Kada su oznake diskretne i konačne govorimo o klasifikaciji, a kada su oznake kontinuirane govorimo o regresiji. [Patter recognition and machine learning, bishop].

Formalno zadatak nadziranog učenja možemo definirati kao učenje preslikavanja f:X->Y iz skupa primjera za učenje D. Skup primjera za učenje D sastoji se od parova primjera i pripadnih oznaka, D={(x(i), y(i))}i=1,N gdje je N ukupni broj primjera za učenje. X predstavlja skup svih mogućih primjera, a pojedini primjer x iz X možemo definiratikao vektor značajki x=(x1, x2, ..., xn)T pri čemu je n broj značajki, odnosno dimenzija vektora. Y predstavlja prostor mogućih oznaka. [Šnajder, Strojno učenje; A brief introduction to weakly supervised learning, Zhi-Hua Zhoa]

Binarna klasifikacija, zadatak ovog rada, je vrsta klasifikacije u kojoj se ulazni primjeri svrstavaju u dva razreda.

\*što su značajke u klasifikaciji slika – to mogu biti sami slikovni elementi, no mogu biti i izvedene značajke kao što su HOG i sl.

## Skup podataka

U okviru ovog rada korišten je skup podataka PatchCamelyon opisan u radu Bastiaan S. Veeling et al., „Rotation Equivariant CNNs for Digital Pathology“ [5]. Skup podataka je nastao na temelju skupa podataka Camelyon16 [6].

Originalni skup podataka Camelyon16 [6] se sastoji od 399 slika limfnih čvorova snimanih WSI metodom s oznakama na razini slikovnih elemenata. Prilikom snimanja tkiva koristila se metoda bojanja hemalaun eozinom (engl. *hematoxylin and eosin*, H&E) kako bi se istaknuli dijelovi limfnih čvorova. Hemalaun ističe jezgre stanica plavom bojom, a eozin boji citoplazmu i spojno tkivo rozom bojom [11]. Jedan slikovni element predstavlja uzorak površine . Svaki slikovni element sadrži tri kanala, crveni, zeleni i plavi zapisanih u 8 bita po kanalu.

Želja autora PCam skupa podataka [5] je bila napraviti skup podataka za klasifikaciju koji bi služio kao standardizirani skup podataka za ocjenjivanje modela strojnog učenja na medicinskim slikama. Slični skupovi podataka su napravljeni za klasifikaciju prirodnih slika kao što su CIFAR-10[X], PASCAL VOC[X], ImageNet [X] te primjerice skup podataka za klasifikaciju rukom pisanih znamenki MNIST [X].

Skup podataka je podijeljen u skup za treniranje, validaciju i testiranje u sljedećim postotnim omjerima: 75%, 12.5% i 12.5%. Ukupno skup podataka sadrži 327,680 slika. Skup podataka i podjele po podskupovima za treniranje, validaciju i testiranje su balansirani što znači da svaki od podskupova sadrži 50% pozitivnih i 50% negativnih primjera. Slike u PCam skupu podataka se sastoje od slikovnih elemenata pri čemu su informacije za svaki slikovni element iste kao i kod originalnog skupa podataka.

Oznake primjera u skupu podataka su binarne. Pozitivna oznaka je dodijeljena slikama koje u regiji od slikovna elementa u sredini sadrže barem jedan slikovni element koji pripada tumorskom tkivu. Takva definicija pozitivnog primjera je napravljena kako bi se omogućilo korištenje dubokih konvolucijskih modela koji ne koriste ispunjavanje rubova slike nulama (engl. *zero padding*) kako bi osigurali konzistentno ponašanje prilikom primjene na cjelokupne slike dobivene WSI metodom.

Primjer pozitivnih primjera dan je na slici X., a negativnih na slici Y.



Slika 1. Primjer histopatoloških slika iz skupa podataka PCam [5] koji sadrže tumor



Slika 2. Primjeri histopatoloških slika iz skupa podataka PCam [5] koji ne sadrže tumor

Podaci u PCam skupu podataka su spremljeni u formatu HDF 5 (engl. *Hierarchical Data Format*) koji omogućava pohranu i organizaciju velikih količina podataka [10].

## Odabir pristupa rješavanja problema klasifikacije

# Duboko učenje u analizi histopatoloških slika

## Korišteni modeli

### AlexNet

### ResNet

Istraživači iz istraživačkog odjela Microsoft korporacije su 2015. godine osmisli arhitekturu neuronski mreža ResNet [9] koja je omogućila treniranje modela s više slojeva. Prethodna istraživanja su pokazala kako je dubina neuronske mreže važan parametar koji omogućava bolje rezultate. Problem koji se pojavljivao prilikom treniranja dubljih mreža je bio problem nestajućeg gradijenta. Gradijent se između svakog sloja značajno smanjivao jer se gradijent pogreške po težini u trenutnom sloju množi s umnoškom gradijenata pogreške svih idućih slojeva.

Model je prvotno testiran na skupovima podataka za klasifikaciju prirodnih slika Coco i ImageNet te je osvojio prvo mjesto na natjecanju ILSVRC 2015.

### Inception

### DenseNet

## Učenje prijenosom značajki

## Proširivanje skupa podataka

# Rezultati

# Zaključak

# Literatura

1. Komura, D., & Ishikawa, S. (2018). Machine Learning Methods for Histopathological Image Analysis. Computational and Structural Biotechnology Journal, 16, 34–42. <https://doi.org/10.1016/j.csbj.2018.01.001>
2. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42(1995), 60–88. <https://doi.org/10.1016/j.media.2017.07.005>
3. Sahiner, B., Pezeshk, A., Hadjiiski, L. M., Wang, X., Drukker, K., Cha, K. H., … Giger, M. L. (2019). Deep learning in medical imaging and radiation therapy. Medical Physics, 46(1), e1–e36. <https://doi.org/10.1002/mp.13264>
4. Litjens, G., Bandi, P., Bejnordi, B. E., Geessink, O., Balkenhol, M., Bult, P., … van der Laak, J. (2018). 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. GigaScience, 7(6), 1–8. <https://doi.org/10.1093/gigascience/giy065>
5. Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., & Welling, M. (2018). Rotation equivariant CNNs for digital pathology. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11071 LNCS, 210–218. <https://doi.org/10.1007/978-3-030-00934-2_24>
6. Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., … Venâncio, R. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA - Journal of the American Medical Association, 318(22), 2199–2210. <https://doi.org/10.1001/jama.2017.14585>
7. „histologija“. U: Hrvatska enciklopedija (on-line). Leksikografski zavod „Miroslav Krleža“. Pristup ostvaren 13.5.2019. <http://www.enciklopedija.hr/Natuknica.aspx?ID=25754>
8. „limfni čvor“. U: Hrvatska enciklopedija (on-line). Leksikografski zavod „Miroslav Krleža“. Pristup ostvaren 13.5.2019. [www.enciklopedija.hr/Natuknica.aspx?ID=36573](http://www.enciklopedija.hr/Natuknica.aspx?ID=36573)
9. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
10. The HDF Group. Hierarchical Data Format, version 5, 1997-2019. <http://www.hdfgroup.org/HDF5/>.
11. Can, A., Gurcan, M. N., Yener, B., Boucheron, L. E., Madabhushi, A., & Rajpoot, N. M. (2009). Histopathological Image Analysis: A Review. IEEE Reviews in Biomedical Engineering, 2, 147–171. <https://doi.org/10.1109/rbme.2009.2034865>

# Sažetak

Prilikom analize medicinskih slika ukazuje se potreba za automatiziranim sustavom koji bi

**Ključne riječi:**

# Summary

When analyzing medical images, there is a need for an automated system that can help

**Keywords:**

# Prilog – programski kod

Programski modul data.py s pomoćnim funkcijama za prikaz i obradu podataka