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 računalnoj infrastrukturi za potrebe izrade ovog rada.

**Sadržaj**

[1. Uvod 5](#_Toc9413587)

[2. Pregled dosadašnjih istraživanja 6](#_Toc9413588)

[3. Klasifikacija histopatoloških slika limfnih čvorova 8](#_Toc9413589)

[3.1 Klasifikacija slika 8](#_Toc9413590)

[3.2 Skup podataka 9](#_Toc9413591)

[3.3 Odabir pristupa rješavanja problema klasifikacije 11](#_Toc9413592)

[4. Duboko učenje u analizi histopatoloških slika 12](#_Toc9413593)

[4.1 Korišteni modeli 13](#_Toc9413594)

[4.1.1 AlexNet 13](#_Toc9413595)

[4.1.2 ResNet 14](#_Toc9413596)

[4.1.3 Inception 14](#_Toc9413597)

[4.1.4 DenseNet 14](#_Toc9413598)

[4.2 Učenje prijenosom značajki 14](#_Toc9413599)

[4.3 Proširivanje skupa podataka 15](#_Toc9413600)

[5. Rezultati 16](#_Toc9413601)

[6. Zaključak 17](#_Toc9413602)

[7. Literatura 18](#_Toc9413603)

[Sažetak 20](#_Toc9413604)

[Summary 21](#_Toc9413605)

[Prilog – programski kod 22](#_Toc9413606)

**Popis slika**

[Slika 1. Primjer histopatoloških slika iz skupa podataka PCam [5] koji sadrže tumor 10](#_Toc9413607)

[Slika 2. Primjeri histopatoloških slika iz skupa podataka PCam [5] koji ne sadrže tumor 11](#_Toc9413608)

# 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 omogućili su kvantitativni opis medicinskih slika što je bitno kako bi smanjili pogreške liječnika te kako bi ubrzali proces označavanja slika za potrebe dijagnoze. 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 se mogu između ostalog koristiti i za 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 slikovnih elemenata na najvećoj rezoluciji s 3 bajta memorije za svaki slikovni element što predstavlja ukupno 55.88 GB 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].

U sljedećoj tablici X istaknute su razlike između radioloških i histopatoloških snimaka na konkretnim primjerima snimki te na temelju podataka preuzetih iz [11].

Tablica 1. Primjer razlika između radioloških i histopatoloških snimaka temeljena na podacima iz [11]

|  |  |  |
| --- | --- | --- |
|  | Radiološka snimka | Histopatološka snimka |
| Opis snimke | CT snimka visoke rezolucije prsnog koša | WSI biopsija prostate |
| Dimenzije snimke |  |  |
| Vrsta informacije | vrijednosti predstavljaju nijanse sive boje | višespektralni podaci (RGB ili za svaku mjerenu valnu duljinu) |
| Svrha korištenja | određivanje prisutnosti raka i prostorne rasprostranjenosti | određivanje prisutnosti raka i stupanj diferenciranosti raka (engl. *cancer grade*) |

Što su ljudi još radili s njima, kakvi su pristupi bili na camelyonu?

# 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 histopatološke slike 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 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 iz skupa primjera za učenje . Skup primjera za učenje sastoji se od parova primjera i pripadnih oznaka, gdje je ukupni broj primjera za učenje. predstavlja skup svih mogućih primjera, a pojedini primjer iz možemo definirati kao vektor značajki pri čemu je broj značajki, odnosno dimenzija vektora. 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 pri čemu možemo definirati skup oznaka .

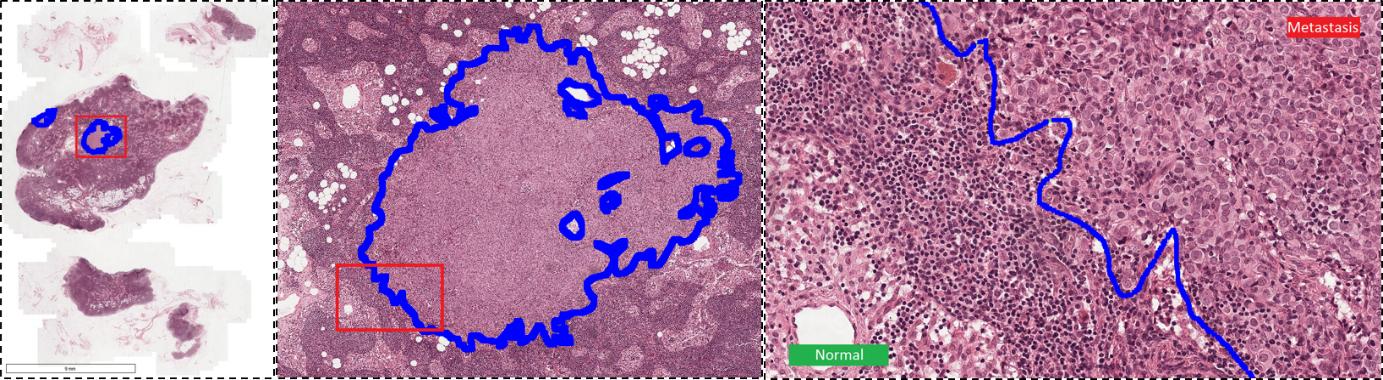
Značajke primjera za učenje prilikom klasifikacije slika mogu biti vrijednosti boja slikovnih elemenata, no moguće je i koristiti drugačije značajke. Problem s izravnim korištenjem vrijednosti slikovnih elemenata je velika dimenzionalnost prostora značajki, dok s korištenjem izvedenih značajki možemo postići manju dimenzionalnost s očuvanom informacijom potrebnom za klasifikaciju. Primjeri korištenih značajki su histogram orijentiranih gradijenata (engl. *histogram of oriented gradients*, HOG), histogram pojavljivanja iste vrijednosti sive boje na određenoj udaljenosti (engl. grey level co-ocurring matrix, GLCM), histogram lokalnih binarnih uzoraka (engl. local binary patterns, LBP) [1] te razne informacije o geometrijskim strukturama na slikama [Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning]. Navedene značajke su korištene u analizi medicinskih slika, međutim kako su modeli dubokog učenja postigli bolje performanse od prijašnjih modela strojnog učenja danas se često umjesto ručnog oblikovanja značajki prepušta modelu dubokog učenja da nauči izdvojiti značajke izravno iz vrijednosti slikovnih elemenata [1].

Funkcija pogreške

## 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 zapisanih u 8 bita po kanalu. Primjer slike preuzete iz Camelyon16 [6] skupa podataka s oznakom tumorskog tkiva na tri različite razine povećanja dan je na slici X.

**

Slika 1. Primjer slike histopatološkog uzorka slikanog WSI metodom preuzet iz Camelyon16 [6] skupa podataka s oznakom tumorskog tikva na tri razine uvećanja

Želja autora PCam skupa podataka [5] je bila napraviti skup podataka za klasifikaciju koji bi služio kao standardizirani skup podataka za ocjenjivanje modela (engl. *benchmark dataset*) 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: . Ukupno skup podataka sadrži slika. Skup podataka i podjele po podskupovima za treniranje, validaciju i testiranje su balansirani što znači da svaki od podskupova sadrži pozitivnih i 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 2. Primjer histopatoloških slika iz skupa podataka PCam [5] koji sadrže tumor



Slika 3. 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

Što su ljudi koristili, duboko učenje se sve više koristi, motivacija iza dubokog učenja

# Duboko učenje u analizi histopatoloških slika

Duboko učenje je grana strojnog učenja temeljena na razvoju višeslojnih neuronskih mreža koje omogućavaju izdvajanje značajki na visokom stupnju apstrakcije iz ulaznih podataka [3]. Algoritmi dubokog učenja, posebice konvolucijske neuronske mreže, su u zadnjih nekoliko godina postale najčešća metoda u analizi medicinskih slika [4].

Prednosti konvolucijskih modela za razliku od potpuno povezanih neuronskih mreža i nekih drugih modela strojnog učenja poput SVM-a (engl. *support vector machine*) je taj što oni uzimaju u obzir raspored slikovnih elemenata i njihovo susjedstvo. Duboki konvolucijski modeli su zbog toga izrazito pogodni za podatke s hijerarhijskom strukturom. Na primjeru prepoznavanja čovjeka na slici prvo možemo prepoznati rubove, pa vidjeti koji rubovi tvore nos, oči, rub lica zatim lice te u konačnici možemo klasificirati lice u određenu klasu. Pretpostavka hijerarhijske strukture podataka pokazala se kao dobra pretpostavka na problemima računalnog vida [dive to deep].

Modeli dubokog učenja pokazali su se odličnim izborom za analizu slika na natjecanjima kao što je ImageNet, PascalVoc te u analizi medicinskih slika na natjecanjima Camelyon16, Camelyon17, BraTS, BrakeHis. Osim uspješnosti koje su duboki modeli pokazali imaju i prednost što su omogućili automatsko oblikovanje značajki. Zbog toga više nije potrebno veliko domensko znanje specifičnog područja kako bi se mogao razviti prediktivni model.

Međutim bitno je i istaknuti neka od ograničenja koje duboki modeli imaju s naglaskom na analizu medicinskih slika. Prvi problem je što je za učenje dubokih modela potrebna velika količina podataka. Specifičnost prikupljanja medicinskih skupova podataka je da je potrebno da se dobije pristanak pacijenata za sudjelovanje u istraživanje te je potrebno anonimizirati podatke kako se ne bi na temelju nalaza mogao odrediti identitet pacijenta. Takvi problemi ujedno ograničavaju i javnu dostupnost medicinskih podataka. Kako je većina primjene dubokih modela nadzirano učenje uz podatke je potrebno imati i oznake podataka. Zbog navedenog ističe se potreba za označavnjem velikog broja kompleksnih podatka za čije je označavnje potrebno stručno znanje eksperata. Drugi problem je interpretabilnost dubokih modela koja je značajno manja u odnosu na ručno modelirane značajke. Zbog toga se danas sve više istražuju različiti načini kako omogućiti interpretabilnost dubokih modela. U istraživanju interpretabilnosti dubokih modela ističu se dva pravca, lokalna i globalna interpretabilnost. Cilj metoda lokalne interpretabilnosti je objasniti kako je model došao do predikcije za jedan konkretan ulazni primjer, dok je cilj metoda globalne interpetabilnosti

Treći problem je pronalaženje optimalnih hiperparametara.

U okviru ovog poglavlja opisane su prednosti i ograničenja dubokih modela prilikom primjene u analizi histopatološkim slikama. Opisani su korištene arhitekture dubokih modela (poglavlje 4.1.), učenje modela korištenjem metode prijenosa značajki (poglavlje 4.2) te metode proširivanja skupa podataka kako bi se poboljšala generalizacijska sposobnost modela (poglavlje 4.3).

## Korišteni modeli

U okviru ovog rada korišteno je nekoliko arhitektura dubokih modela koji su postali popularni u području dubokog učenja. Sažeti pregled korištenih modela i broja njihovih parametara dan je u tablici X.

|  |  |  |
| --- | --- | --- |
| Naziv modela | Broj parametara | Broj slojeva s aktivacijskom funkcijom |
| AlexNet | 57,877,824 | 8 |
| ResNet18 | 11,186,645 | 10 |
| Inception v3 |  |  |
| DenseNet |  |  |

### AlexNet

U okviru ovog rada korišten je AlexNet model koji je definiran u radu A. Krizhevsky, „One weird trick for parallelizing convolutional neural netowrks“ [13] te je modifikacija originalnog modela opisanog u A. Krizhevsky et al., „ImageNet Classification with Deep Convolutional Neual Networks“[12]. Korišteni model i parametri modela treniranog na ImageNet skupu podataka dostupni su u okviru radnog okvira PyTorch [14].

Arhitektura korištenog modela opisana je na slici X. Model se sastoji od osam slojeva s parametrima koji se mogu učiti pri čemu su pet slojeva konvolucijska, a tri potpuno povezana. Kao aktivacijska funkcija koristi se u svim slojevima osim u zadnjem u kojem se koristi sigmoidalna prijenosna funkcija .

To je različito od dotadašnjih modela konvolucijskih neuronskih mreža koje su uglavnom koristile i kao aktivacijske funkcije.

Model je različit od dotadašnjih modela i zbog veće dubine u usporedbi s primjerice LeNet5 modelom koji je korišten 90-tih godina 20. stoljeća na MNIST skupu podataka koji je imao 5 slojeva, je je AlexNet imao čak deset puta više konvolucijskih kanala. Kako bi omogućili treniranje modela toliko većeg kapaciteta autori su odlučili osim kažnjavanja težina L2 normom koristiti i dropout tehniku regularizacije. Kako bi omogućili dovoljnu količinu podataka za treniranje autori su u okviru svog rada koristili tehnike povećanja skupa podataka poput translatiranja slike, horizontalnog zrcaljenja i mijenjanje intenziteta boja na slikama.

### 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.

Novost koju model donosi je učenje rezidualnih funkcija.

### Inception

### DenseNet

## Učenje prijenosom značajki

Što je to, prijašnja istraživanja su proučavala prijenos s natural na medical, no ne na histopataološka, u okviru ovog rada ćemo to istražiti

## Proširivanje skupa podataka

Što je to, koje operacije imaju smisla u okviru histopatoloških slika

# Rješavanje problema klasifikacije histopatoloških slika

Koji su eksperimenti, što želimo proučiti , koje se metrike koriste

# 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>
12. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 1097-1105.
13. Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. <http://arxiv.org/abs/1404.5997>
14. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch. U NIPS-W. <https://openreview.net/pdf?id=BJJsrmfCZ>

# 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