

Subcelularna klasifikacija proteinskih obrazaca u ljudskim ćelijama

1st Nikola Zubić

Departman za informatiku
Fakultet tehničkih nauka
Novi Sad, Srbija
nikola.zubic@uns.ac.rs

2nd Vasilije Pantić

Departman za informatiku
Fakultet tehničkih nauka
Novi Sad, Srbija
vasilije.pantic@uns.ac.rs

Sažetak—Lokalizacija proteina, odnosno njihovih paterna je veoma značajan problem u oblasti biologije ćelije. Utvrđivanje kojim organelama pripada koji proteinski patern unutar ćelije produbljuje naše znanje o samoj lokaciji proteina u organizmu, te omogućuje i detaljniju analizu biohemiskih reakcija u organizmu, te istraživanja patoloških stanja koja se javljaju kod ljudi. Na početku, izvršena je eksplorativna analiza podataka koja podrazumijeva izvođenje zaključaka o pojedinim klasama (labelama), njihovoj zastupljenosti i ostalim karakteristikama. Potom su korišćene razne tehnike procesiranja slike i model Bernulijevih miješavina za prečišćavanje i modifikaciju skupa podataka, te dodavanje novih obilježja. Tako pretprocesirane slike sa dodatnim obilježjima su korišćene kao ulaz u tri različite arhitekture neuronske mreže (od kojih je trenutno najbolja korišćena sa pretreniranim težinama). Za evaluaciju rješenja korišćena je makro F1 mjera zato što se ona najbolje ponaša kada radimo sa neizbalansiranim skupom podataka, što je ovdje slučaj. Najbolje rješenje ostvaruje pretrenirana GapNet-PL arhitektura neuronske mreže koje iznosi 0.785, dok naša rješenja iznose 0.568 (za ResNet) i 0.477 (za InceptionNet). Ljudski ekspert dostiže rezultat oko 0.710, a problem se može smatrati riješenim ukoliko se ostvari rezultat preko 0.410.

Ključne riječi—lokalizacija proteina, konvolutivne neuronske mreže, makro F1 mjera, klasifikacija, biologija ćelije

I. UVOD

Problem koji ovdje rješavamo se tiče lokalizacije proteina odnosno njihovih paterna. Ovaj problem je veoma značajan u oblasti biologije ćelije. Klasifikacija proteinskih paterna u pogledu pripadnosti određenim organelama unutar ćelije dovodi i do uvida u samu lokaciju proteina u čovjekovom tijelu. Cilj projekta jest da se pravilno pretprocesiraju slike i prouči skup podataka, te naprave modeli mašinskog učenja koji mogu da identifikuju gdje se proteini nalaze u sklopu ćelije na osnovu fluoroscentne mikroskopije, odnosno kojoj organeli ili djelu ćelije pripadaju.

Proteine možemo nazvati “nanomašinama” unutar živih organizama. Oni izvršavaju mnoge funkcije bez kojih ne bi bilo života na Zemlji. Kako bi u potpunosti razumjeli kompleksnost ljudske ćelije, bilo bi poželjno formirati model koji je u stanju da odredi kojim organelama pripadaju neki proteinski paterni u različitim ćelijama. Korišćenje metoda mašinskog učenja i istraživanja i analize podataka u problemu biomedicinske analize slika je veoma korisno zato što poboljšava razumijevanje ljudske ćelije, što može dovesti i do boljeg razumijevanja različitih oboljenja.

Nakon eksplorativne analize skupa podataka i pretprocesiranja slika koristeći različite tehnike, te izdvanjanja dodatnih obilježja, koriste se tri arhitekture neuronskih mreža. Jedna arhitektura je pretrenirana i ona ostvaruje trenutno najbolje rezultate, tj. predstavlja *state-of-the-art* arhitekturu za ovaj problem. Ova arhitektura je korišćena na početku kako bi se uvidjelo kakve smo rezultate dobili nakon pretprocesiranja, a potom su implementirane dvije arhitekture koje su zahtijevale dodatni rad sa skupom podataka i dobijeni su rezultati koji prelaze prag koji je potreban da bi se rješenje smatralo uspješnim. Nakon toga, vršeno je poređenje rezultata između različitih arhitektura i analiza grešaka modela kako bi utvrdili gdje je model pogriješio i interpretirali grešku.

Ono po čemu se naš rad razlikuje od postojećih jeste što smo sami izvršili detaljnu eksplorativnu analizu podataka i što koristimo različite tehnike pretprocesiranja slike i dodavanja novih obilježja kako bi dobili što pogodnije ulaze za same arhitekture neuronske mreže. U drugom poglavlju je izvršen pregled drugih rješenja na zadatu temu, dok je treće poglavlje najobimnije i sastavljeno od više potpoglavlja koja prate kompletну metodologiju, odnosno plan pomoću kojeg smo riješili problem. Četvrto poglavlje bavi se mjerom evaluacije metoda, poređenjem metoda (arhitektura neuronskih mreža), eksperimentalnom procedurom (podijela na trening/validacioni/test skup), hiperparametrima, rezultatima i na kraju analizom grešaka modela. Nakon ove analize, u petom poglavlju se sumarizuje naš rad, dok se nakon zaključka nalazi korišćena literatura.

II. PREGLED POSTOJEĆE RELEVANTNE LITERATURE

A. Definicija problema i arhitektura InceptionNet v3

U radu [1] upoznali smo se sa samim problemom subcelularne klasifikacije proteinskih obrazaca u ljudskim ćelijama što je ekvivalentno lokalizaciji proteina. Takođe, ovdje je ponuđena i analiza mogućih pristupa na visokom nivou i njihovo poređenje sa ljudskim ekspertima koristeći *InceptionNet* arhitekturu neuronske mreže. Ovo je bitno kako bi bolje analizirali greške modela, tako što ćemo izdvojiti podskup primjera na kojima model grijesi i pokušati da razumijemo razlog (na primjer, da li postoji neka veza između greške modela i greške ljudskog eksperta).

Definisana je i metrika za ovaj problem sa više labela (engl. *multi-label problem*). Pošto je glavna prepreka u treniranju ovog modela činjenica da je skup podataka neizbalansiran, dati su prijedlozi o razriješavanju tog problema. Nije bio fokus na objašnjavanju tehnika procesiranja slike (engl. *image processing*), što je naš doprinos. Posmatran je uticaj korišćenja i poređenja različitih arhitektura konvolutivnih neuronskih mreža (engl. *Convolutional Neural Networks*) - *CNN* [2] i odabir odgovarajućih aktivacionih funkcija. Ujedno je u radu [1] objavljen i skup podataka.

Ovaj rad nam je omogućio da se na adekvatan način upoznamo sa domenom ovog problema, pošto je za njegovo riješavanje ono zaista potrebno (riječ je o oblastima računarske biologije i biologije ćelije). Pored toga, upoznali smo se i sa korisnim modelima neuronskih mreža i aktivacionih funkcija za ovaj domen.

B. ResNet arhitektura neuronske mreže

U radu [3] korišćena je *ResNet* arhitektura neuronske mreže za riješavanje problema subcelularne klasifikacije proteinskih obrazaca. Rezidualna arhitektura neuronske mreže je vještačka neuronska mreža (engl. *Artificial Neural Network*) - *ANN* koja je zasnovana na ideji piridalnih ćelija u cerebralnom korteksu. One su zasnovane na tome da se propagacija unaprijed (engl. *forward propagation*) ne radi iz sloja u sloj, nego se može vršiti preskakanje određenog sloja, što podsjeća i na ljudsko učenje. Ne moramo da naučimo sve da bi shvatili nešto, možemo neke dijelove i preskočiti, ali je bitno da naučimo suštinu, što su u ovom kontekstu značajni *feature-i*. Tipične *ResNet* mreže preskaču dva do tri sloja koji sadrže *ReLU* i *batch normalization* elemente.

C. GapNet-PL arhitektura neuronske mreže

U radu [4] predstavljena je *GapNet-PL* arhitektura neuronske mreže koja je specijalno konstruisana za riješavanje ovog problema. Trenutno predstavlja najbolju arhitekturu za ovaj problem. Dodatno, ovaj rad uključuje i eksperimente vezane za podešavanje hiperparametara (engl. *hyperparameters*) modela. Važno je napomenuti da ovaj model postiže bolje rezultate od ljudskog eksperta. Ovaj model je iskorišćen sa pretreniranim težinama nakon eksplorativne analize, prečišćavanja skupa podataka i tehnika procesiranja slike (engl. *image processing*) koji predstavljaju naš doprinos. Za razliku od ove arhitekture, *ResNet* i *InceptionNet* smo implementirali od nule.

III. METOD

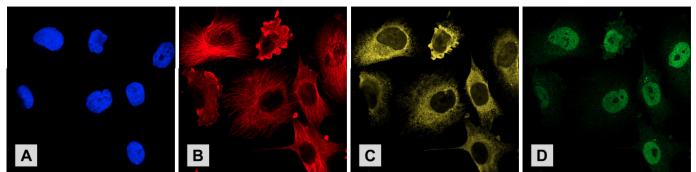
Metodologija (plan) ovog rješenja se može podijeliti u tri cjeline: eksplorativna analiza skupa podataka, korišćenje tehnika procesiranja slike i pregled arhitektura neuronskih mreža.

A. Eksplorativna analiza skupa podataka

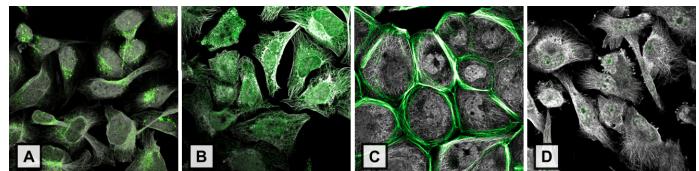
1) *Opis skupa podataka:* U radu [1] je objavljen skup podataka. Ulaz našeg sistema čine PNG slike dobijene konfokalnom mikroskopijom, čije su dimenzije $512 \times 512 \times 4$. Riječ je o *multi-label* klasifikacionom problemu, što znači

da za svaku sliku dobijamo izlaz koji čini više labela (engl. *multi-label classification problem*) - ukupno 28 klase (koje su numerisane brojevima od 0 do 27). Na primjer, za neku sliku, izlaz može biti: 0, 4 i 10, gdje su 0, 4 i 10 klase koje odgovaraju imenima konkretnih dijelova ćelije (organeli i ostali ćelijski dijelovi). Potrebno je napomenuti da u skupu podataka imamo i 27 različitih tipova ćelija, koje su morfološki znatno drugačije, što utiče na proteinske paterne. Dakle, neće za istu organelu biti isti proteinski paterni kod različitih ćelija. Ovaj skup je znatno neizbalansiran po pitanju klasa (engl. *class imbalance problem*), te su i sami proteinski paterni često slični, što čini ovaj problem znatno težim od standardnih klasifikacionih problema (kao što su npr: *ImageNet*¹ i ostali).

Navedimo primjer: Data nam je jedna crno-bijela slika (imamo 4 filtera: plavi - nukleus, crveni - mikrotubule, žuti - endoplazmatični retikulum, zeleni - protein od interesa). Na osnovu toga mi zaključujemo da je taj protein od interesa prisutan, na primjer, u nekim organelama, kao što su Goldžijev aparat, citosoli itd. Primjer je prikazan na slici 1.



Slika 1. Četiri fluorescencntna kanala jednog primjera ćelije iz skupa podataka. (A) nukleus, (B) mikrotubule, (C) endoplazmatični retikulum, (D) protein od interesa.



Slika 2. (A) Primjer kada je proteinski patern od interesa lociran u Goldžijevom aparu i vezikulama, (B) citosoli, nukleus i plazmatična membrana, (C) aktinski filamenti, (D) nukleoliji i centrozomi.

Ostala tri kanala (osim zelenog) formiraju strukturalni dio ćelije. Ovdje kanali nemaju standardno značenje, ne posmatraju se direktno u kontekstu kao sa *RGB* kanalima za sliku. Ukoliko bi imali crvenu i zelenu boju na istom pikselu (regionu piksela) to znači samo da imamo mikrotubule i ciljni protein na istom mjestu u ćeliji. Mikrotubule i endoplazmatični retikulum obično se nalaze u istom dijelu ćelije, tako da su određeni autori izbacili jedan kanal od ta dva, jer na neki način smatraju da već sa jednim imaju dovoljne informacije.

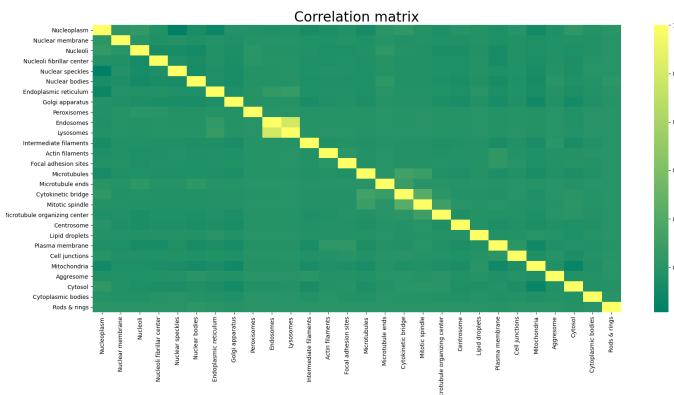
Na slici 2 se uočava da zeleni filter predstavlja proteinske paterne, jer se samo oni fluorescentnom (konfokalnom) mikroskopijom registriraju sa zelenim filterom. Na osnovu njihovog paterna, te ostala tri kanala (zadužena za strukturu) treba da

¹<http://www.image-net.org/>

dobijemo izlaz koji čine neke organele tj. dijelovi ćelije, a to ukazuje na to u kojim dijelovima ćelije se nalaze koji proteini.

Na primjer, za neku sliku ukoliko dobijemo izlaz koji predstavlja 'Endoplazmatični retikulum' i 'Lizozomi', to znači da se taj proteinski patern (vidljiv zelenim filterom) nalazi u lizozomu i EPR-u. Znači da je ovo zapravo problem lokalizacije proteina. Riješavanje ovog problema odgovara na pitanje: Na osnovu proteina (njegovog paterna) u ćeliji, definijiši zapravo u kojim dijelovima ćelije je on lociran, jer sam patern određuje i gdje će se nalaziti u ćeliji. Često se može naći na više mesta zato što je baš transport proteina recimo omogućen preko endoplazmatičnog retikuluma koji preko transportne vezikule prenosi protein.

2) *Analiza skupa podataka:* U skupu podataka čiji je naziv *Human Protein Atlas*² koji je predstavljen u radu [1] nalaze se 31072 slike od kojih se svaka crno-bijela slika propušta kroz 4 kanala, tako da imamo ukupno 124288 slika. Zbog veličine skupa podataka, obezbijeden je poseban .csv fajl koji omogućuje pribavljanje slike po *id*-ju i za svaku sliku iz trening skupa imamo njen *id* i odgovarajuće izlaze numerisane sa brojevima od 0 do 27.

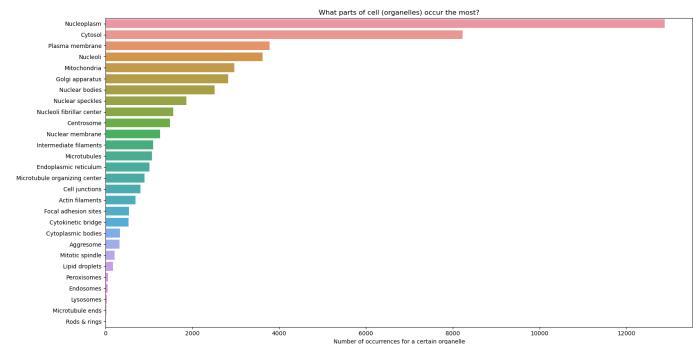


Slika 3. Matrica korelacije gdje se na horizontalnoj i vertikalnoj osi nalazi svih 28 labela i posmatramo korelacije svake labele sa svakom. Mnoge organele i dijelovi ćelije su u slaboj korelaciji (zelena boja na grafiku). Neke od njih imaju znatno jače korelacije (žuta boja). Na primjer, endozomi i lizozomi koji su locirani unutar endoplazmatičnog retikuluma imaju znatno jače korelacije.

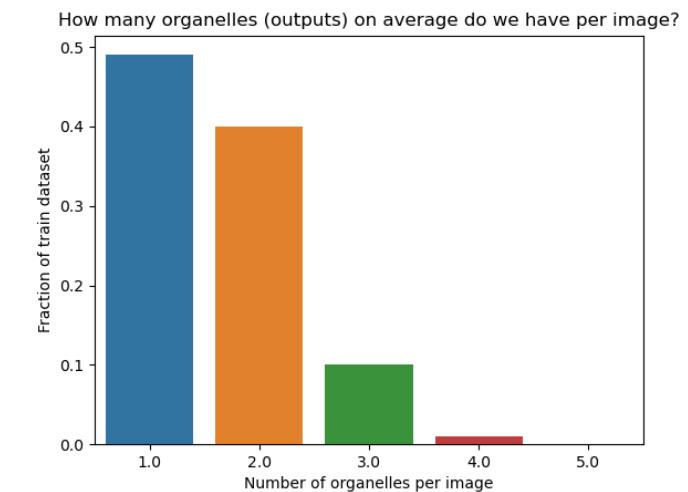
Može se primjetiti da se veći i značajniji dijelovi ćelije više pojavljuju u trening skupu podataka. Većinom su to nukleoplazma i citoplazma (fluidi koji ispunjavaju nukleus i sadržaj ćelije, respektivno), potom ćelijski nukleoliji, citosoli i plazmatične membrane. Takođe, na osnovu Slike 4 se primijeti da je skup podataka značajno neizbalansiran.

B. Tehnike procesiranja slike

Nakon eksplorativne analize skupa podataka, odradene su tehnike procesiranja slike kako bi se izdvojila dodatna obilježja koja pomažu prilikom treniranja modela. Korišćene tehnike su: *Yen-ov* i *Otsu-ov thresholding*, redukcija i kompresija slike, binarno otvaranje i zatvaranje (engl. *binary opening/closing*), segmentacija i spektralno klasterovanje, optička granulometrija



Slika 4. Frekvencija pojavljivanja konkretnih ćelijskih organeli (dijelova ćelije). Na horizontalnoj osi se nalazi broj pojavljivanja, a na vertikalnoj osi naziv organele ili nekog dijela ćelije.



Slika 5. Na horizontalnoj osi se nalazi broj organeli ili dijelova ćelije po slici, dok se na vertikalnoj osi nalazi procenat zastupljenosti u trening skupu podataka. Uočavamo da se većinom na slikama nalaze 1 do 2 organeli.

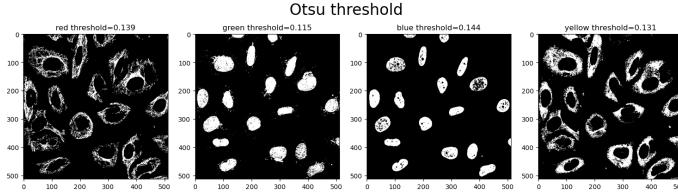
i selekcija kontura, *Canny*-jeva detekcija ivica i *blob* detekcija, segmentacija bazirana na regionima i binarno popunjavanje praznina, te eksperimentisanje sa šumom i korišćenje Sobel-lovog operatora. Neke od najreprezentativnijih tehnika ćemo opisati u nastavku.

Otsu thresholding je metoda koja vraća *threshold* jedinstvenog intenziteta koji razdvaja piksele u dvije klase: unutrašnja oblast i pozadina. Ovaj *threshold* je određen minimizacijom varijanse intenziteta za unutrašnju klasu, ili ekvivalentno maksimizacijom varijanse vanjske klase.

Za razdvajanje unutrašnje oblasti od pozadine, potrebno je koristiti prag koji predstavlja numeričku vrijednost između 0 i 255, tako da vrijednosti piksela veće od te vrijednosti ulaze u unutrašnju oblast, dok vrijednosti koje su manje pripadaju pozadini. Postoji i mogućnost da se postavi i odgovarajući interval povjerenja. Osim ovog pristupa, postoji i *Yen-ov*, *Liu-ov* i *IsoData* pristup. Na slici 6 dat je primjer rada *Otsu* metoda koji prag nalazi automatski na osnovu histograma nad

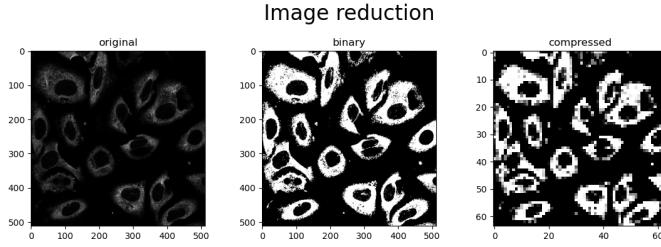
²<https://www.kaggle.com/c/human-protein-atlas-image-classification/data>

intenzitetima vrijednosti piksela slike.



Slika 6. Primjena Otsu tehnike nad sva četiri kanala slike za konkretan primjer iz trening skupa podataka.

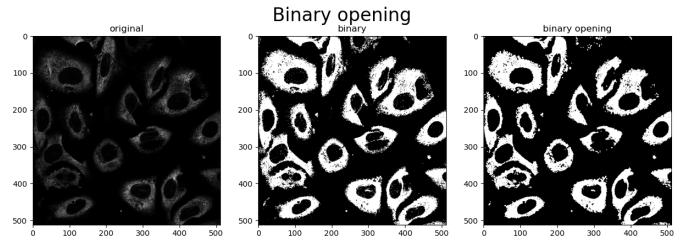
Reducija slike (engl. *Image Reduction*) je tehniku čiji je cilj smanjenje broja piksela slike po dužini i širini tako da se značajne informacije koje se tiču sadržaja slike ne izgube, što omogućuje da se rezolucija vratí na prvobitnu. Svrha ove tehniku jeste da smanji memorijsko zauzeće skupa podataka sa jedne strane, a sa druge strane i da omogući brže treniranje modela neuronskih mreža. Kao ulaz, ova tehniku prima sliku čiji pikseli imaju vrijednosti 0 ili 1 (binarna slika). Na slici 7 dat je primjer rezultata tehniku redukcije slike, koji funkcioniše na način da smanji sliku 8 puta po širini i visini (ukupno 64 puta manje piksela). To postiže tako što piksele u grupi 8×8 konatenira u jedan veliki broj u kojem su enkapsulirana sva 64 piksela.



Slika 7. Rezultat tehniku redukcije slike za konkretan primjer iz trening skupa podataka.

Prije nego što objasnimo tehniku binarnog otvaranja i zatvaranja, potrebni su nam pojmovi erozije i dilacije. Erozija predstavlja tehniku koja od binarne slike generiše novu sliku u kojoj se pikseli vezani za unutrašnju oblast pretvaraju u pozadinske ukoliko u svojoj okolini (susjedni pikseli) nemaju sve piksele tog tipa. Sa druge strane, dilacija predstavlja tehniku koja sve piksele koji su pozadinski, a imaju bar jednog iz unutrašnje oblasti u okolini, pretvaraju u unutrašnju oblast. Binarno otvaranje koristi tehniku erozije nakon koje se primjenjuje dilacija, dok kod binarnog zatvaranja nakon tehniku dilacije primjenjujemo eroziju. Kod binarnog otvaranja dolazi do uklanjanja šuma sa slike, dok kod binarnog zatvaranja popunjavamo praznine koje se nalaze u segmentu unutrašnje oblasti slike. Na slici 8 dat je primjer primjene tehniku binarnog otvaranja.

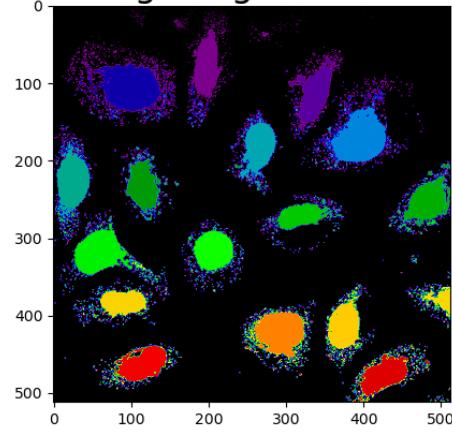
Tehniku segmentacije slike (engl. *Image Segmentation*) je proces partionisanja slike u višestruke segmente (skupove piksela, odnosno dijelove slike). Cilj segmentacije je da se pojednostavi i/ili izmjeni reprezentacija slika u nešto što je od



Slika 8. Rezultat tehniku binarnog otvaranja za konkretan primjer iz trening skupa podataka.

većeg značaja i lakše za analizu nekog konkretnog problema. Najčešće se koristi da se detektuju objekti i granice (linije, krive, itd) na slikama. Još preciznije, segmentacija slike je proces koji dodjeljuje labelu svakom pikselu slike tako da pikseli sa istom labelom (bojom) dijele slične karakteristike. Na slici 9 dat je primjer primjene tehniku segmentacije.

Image segmentation



Slika 9. Rezultat tehniku segmentacije za konkretan primjer iz trening skupa podataka.

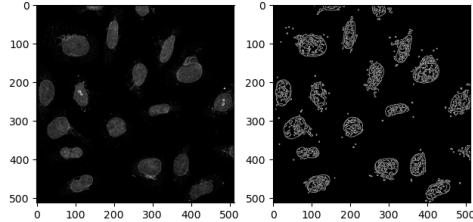
Canny-jeva detekcija ivica je tehniku za izdvajanje ivica sa ulazne slike tako što znatno smanjuje količinu podataka na slici, te prikazuje segmente od interesa preko tih ivica. Ova tehniku koristi više koraka obrade nad slikom kako bi na adekvatan način došla do svih ivica na slici. Na početku smanjuje šum na slici pomoću Gausovog filtera, a potom koristi Sobelov operator. Osnovna ideja je zasnovana na činjenici da ivicu na slici definiše velika razlika u intenzitetu između susjednih piksela. Nakon primjene Sobelovog operatorka koristi se segmentacija pomoću praga kako bi se dobole ivice. Na slici 10 dat je primjer primjene detekcije ivica.

C. Arhitekture neuronskih mreža

Korišćene su tri arhitekture neuronskih mreža, od kojih je jedna pretrenirana (*GapNet-PL* [4]), a druge dvije su implementirane od nule (*InceptionNet v3* [1] i *ResNet* [3]).

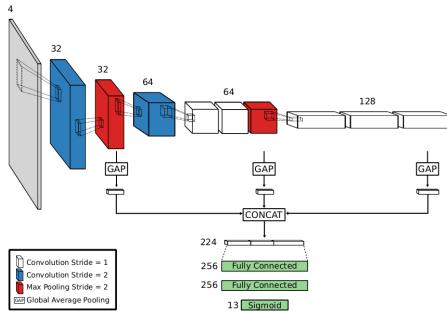
1) *GapNet-PL*: Ova arhitektura (predstavljena u radu [4]) je specijalno dizajnirana za potrebe procesiranja mikroskopskih

Canny edge detector



Slika 10. Rezultat primjene Canny-jeve detekcije ivica za konkretni primjer iz trening skupa podataka.

slike. Radi sa visoko i nisko-rezolucionim slikama i samim tim ima mogućnost da uči finije (detaljnije) strukture unutar slike, te ne zahtijeva *downscaling*. Ovo se ostvaruje kroz dva koraka. U prvom koraku koristi se enkoder koji je sastavljen od par konvolutivnih slojeva (neki imaju korak 2) zajedno sa *max-pooling* slojevima kako bi naučili apstraktna obilježja za različite prostorne rezolucije.

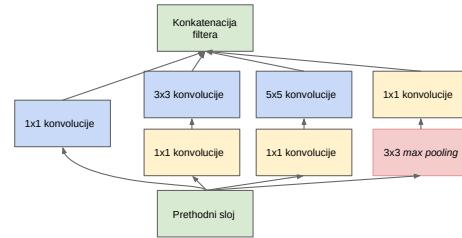


Slika 11. *GapNet-PL* arhitektura neuronske mreže

U drugom koraku, redukujemo mape obilježja (engl. *feature maps*) iz tri različita sloja koristeći *global average pooling* veličine jednog piksela i konateniramo rezultujuće vektore obilježja. Ova *pooling* operacija omogućuje smanjivanje uticaja labela koje se rijetko pojavljuju. Rezultujuća obilježja, koja predstavljaju različite prostorne rezolucije, se potom prosljeđuju do potpuno-povezane mreže sa dva skrivena sloja kako bi se izvršila konačna predikcija. Ova arhitektura ima relativno mali broj parametara (oko 600000) i može se prilično lako iskoristiti i za neki drugi specifičan zadatak, tako što se na kraj doda još konvolutivnih blokova u enkoderu i koristi više obilježja u drugom koraku.

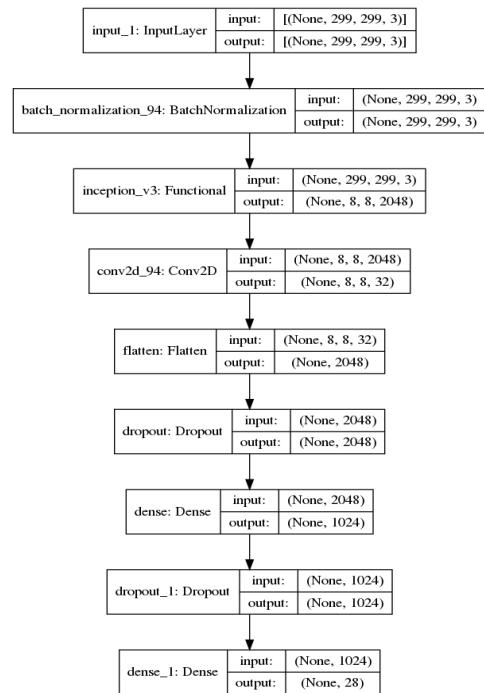
2) *InceptionNet v3*: Glavna ideja *Inception* arhitektura riješava problem optimalnog prekrivanja lokalnih rijetkih struktura konvolutivne mreže sa gustim komponentama uz pomoć aproksimacije [5]. Do izuma *InceptionNet-a*, najpopularniji CNN-ovi su formirani redanjem (engl. *stacking*) konvolutivnih slojeva jednog za drugim, kreirajući tako sve dublje i dublje strukture, u nadi da ćemo dobiti bolje performanse. Za razliku od ovog pristupa, *InceptionNet* je iskoristio niz 'trikova' koji su znatno poboljšali performanse u pogledu

brzine i preciznosti, ali bez prevelikog povećavanja dubine mreže. Suštinski, zasnovana je na ideji da imamo filtere različitih dimenzija ($\text{širina} \times \text{visina}$) koji operišu na istom sloju neuronske mreže. Na taj način naša mreža postaje šira, a ne dublja. Na primjer, data je operacija konvolucije nad ulazom sa 6 filtera različitih dimenzija uz *max pooling* i nakon toga se ti izlazi konateniraju i šalju u naredni sloj, kao što se vidi na Slici 12.



Slika 12. Prikaz *Inception* modula

InceptionNet v3 je CNN arhitektura iz *Inception* familije koja uvodi par unaprijeđenja u odnosu na prethodne verzije. Jedna od novina jeste *Label Smoothing* [6] koji je tip regularizirajuće komponente koja se dodaje na funkciju gubitka kako bi se preveniralo da mreža postane previše sigurna u neku klasu (spriječavanje *overfitting-a*). Takođe, uvodi korišćenje *RMSProp* optimizatora, te faktorizovanih 7×7 konvolucija. U svrhe ovog problema, radi bolje klasifikacije, uklonjeni su originalni potpuno-povezani slojevi sa *InceptionNet v3* arhitekture, a dodat je konvolutivni sloj, a potom dva potpuno-povezana sloja, što je prikazano na Slici 13.



Slika 13. *Inception v3* arhitektura iskorisćena za dati problem.

3) *ResNet*: Prije nego što je iskorišćen *ResNet* model, kako bi se formirala dodatna obilježja, iskoristili smo model Bernulijevih miješavina (engl. *Bernoulli Mixture Model*) i *oversampling*.

Prisustvo proteina od interesa je dato kao binarna vrijednost: 1 ukoliko je prisutan, 0 ukoliko nije. Pošto želimo formirati klastere za diskrete vrijednosti, zato koristimo algoritam za diskretno klasterovanje, što je u ovom slučaju model Bernulijevih miješavina. Za svaki primjerak u skupu podataka x_n od naših N primjera, postoji latentna varijabla z_n koja je jednaka 1 za komponentu k koja je generisala x_n i jednaka je 0 za sve ostale. Ukoliko pretpostavimo da ih već znamo, onda se gustina vjerovatnoće može predstaviti na sledeći način:

$$P(X) = \sum_Z P(X, Z | \theta) = \sum_Z P(Z | \theta) \cdot P(X | Z, \theta) \quad (1)$$

Svaki ciljni protein x_d je binaran. Ukoliko želimo da opišemo distribuciju specifične organele (dijela ćelije), to simuliramo 'bacanjem novčića'. Za svaku od $D = 28$ cilnjih labela 'bacamo' novčić, sa 0 na jednoj strani i 1 na drugoj strani. Ukoliko su ove ciljne varijable nezavisne unutar jedne komponente k , onda možemo uočiti trenutne ciljne proteinske obrasce i reći da pripadaju jednoj grupi. Tako da, ovi protinski paterni (organele) koje zajedno imaju 1 su one koje predstavljaju jednu ciljnu grupu (jedan klaster).

Na primjer, ukoliko imamo lizozome i endozome, te znamo da oni pripadaju grupi čija je komponenta $k = 2$, onda je vjerovatnoća da ih uočimo unutar te grupe jednaka: $\mu_{2,lizozom} = \mu_{2,endozom} = 0.99$. Sa druge strane, vjerovatnoća za neku drugu organelu biće 0.30, dok će za sve ostale biti 0.02. Ovo znači da ukoliko imamo primjerak (sliku) koja nam daje takve vjerovatnoće, da se na toj slici vrlo vjerovatno nalaze te organele (lizozom i endozom), jer je za njih najveća vjerovatnoća.

Želimo da opišemo podatke preko gustine vjerovatnoće $P(X)$. Prepostavljamo da su svi primjeri (engl. *sample*) nezavisno izvučeni iz ove distribucije. Sa pretpostavkom možemo da podijelimo sve primjerke u proizvod po N primjera.

$$\begin{aligned} P(X) &= \sum_Z P(X, Z | \theta) = \sum_Z P(Z | \theta) \cdot P(X | Z, \theta) = \\ &\prod_n \sum_z p(z | \pi) \cdot p(x | z, \mu) \end{aligned} \quad (2)$$

Pošto postoji jedna stvarna komponenta $z_{n,k}$ za svaki primjerak, možemo reći da je z_n *one-hot* enkodovano, te da se može opisati sa multinomijalnom distribucijom:

$$p(z | \pi) = \prod_{k=1}^K \pi_k^{z_k} \quad (3)$$

Isto važi i za uslovnu vjerovatnoću $p(x | z, \mu)$:

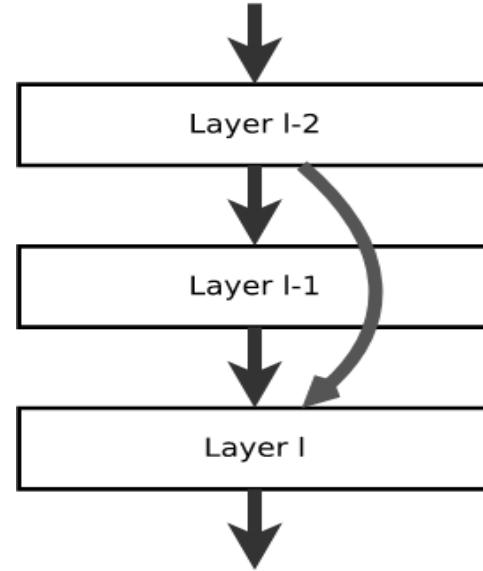
$$p(x | z, \mu) = \prod_{k=1}^K p(x | \mu_k)^{z_k} \quad (4)$$

Na kraju, konačno dobijamo:

$$\begin{aligned} P(X) &= \prod_n \sum_z p(z | \pi) \cdot p(x | z, \mu) = \prod_n \sum_k^K \pi_k \cdot p(x_n | \mu_k) \\ &= \prod_n \sum_{k=1}^K \pi_k \cdot \prod_{d=1}^D \mu_{k,d}^{x_{n,d}} (1 - \mu_{k,d})^{(1-x_{n,d})} \end{aligned} \quad (5)$$

Sa ovim modelom, želimo da opišemo ciljne podatke. Takođe, potrebno je fitovati ovaj model tako da smo u stanju da ova obilježja iskoristimo kao dodatna prilikom analize šta vidimo na slici (ukoliko, na primjer, vidimo lizozome, vrlo vjerovatno da ćemo vidjeti i endozome). To radimo maksimizacijom gustine vjerovatnoće $P(X)$ u funkciji od parametara modela π i μ . Ova maksimizacija se radi koristeći *Expectation maximization* pristup.

Kao što je opisano u Podsekciji II-B, te se vidi na Slici 14, rezidualna neuronska mreža (engl. *ResNet*) preskače određene slojeve prilikom određenih propagacija unaprijed.



Slika 14. Kanonska forma rezidualne neuronske mreže. Sloj $l-1$ se preskače preko aktivacije iz sloja $l-2$.

Postoje dva osnovna razloga zbog kojeg se uvode konekcije za preskakanje (engl. *skip connections*). Jedan od njih jeste kako bi izbjegli problem isčezavajućih gradijenata (engl. *vanishing gradients*) ili kako bi izbjegli problem saturacije (dodavanje novih slojeva u mreži dovodi do veće trening greške).

Efektivno preskakanje dovodi i do pojednostavljivanja mreže jer koristimo manje slojeva u inicijalnim fazama treiranja. Ovo ubrzava učenje redukcijom uticaja isčezavajućih gradijenata, jer postoji manje slojeva kroz koje se izvršava propagacija unaprijed.

IV. REZULTATI I DISKUSIJA

Algoritmi učenja većinom daju dobre rezultate na čestim klasama, ali imaju probleme sa klasama koje se rijetko poja-

vljuju. Kako bi podstakli klasifikaciju zasnovanu na jednakoj distribuciji po svih 28 klasa, koristićemo makro $F1$ mjeru. Ova mjera daje značaj i preciznosti i senzitivnosti/odzivu (engl. *recall*). Računa se za svaku klasu prije nego što se uprosiječe za svih 28 klasa. Ovo znači da bi model ostvario zadovoljavajuće ili dobre rezultate, on mora funkcionisati dobro i za rijetke klase, a to nam je ovdje bitno jer je riječ o neizbalansiranom skupu podataka.

Makro $F1$ mjera računa se na sledeći način:

$$F_1 = 2 * \frac{\text{preciznost} * \text{odziv}}{\text{preciznost} + \text{odziv} + \varepsilon} \quad (6)$$

gdje je ε veoma mali broj koji spriječava dijeljenje sa nulom.

Preciznost se računa kao:

$$\text{preciznost} = \frac{TP}{TP + FP + \varepsilon} \quad (7)$$

Odziv je:

$$\text{odziv} = \frac{TP}{TP + FN + \varepsilon} \quad (8)$$

gdje su TP stvarno pozitivna (engl. *True Positive*), FP lažno pozitivna (engl. *False Positive*) i FN lažno negativna (engl. *False Negative*) klasa.

Poređenje metoda koje smo mi koristili (*GapNet-PL* [4]) i implementirali (*InceptionNet v3* [1] i *ResNet* [3]) sa ostalim metodama na osnovu makro $F1$ mjere su prikazani u donjoj tabeli. Uočava se da je *GapNet-PL* metoda najbolja, dok je najlošija metoda *Convolutional MIL* [7]. Svi rezultati preko 0.410 se smatraju zadovoljavajućim, a ljudski eksperti ostvaruju rezultat oko 0.710.

Tabela I

POREDENJE PERFORMANSI RAZLIČITIH ARHITEKTURA NEURONSKIH MREŽA ZA PROBLEM LOKALIZACIJE PROTEINA. ZA POREDENJE SE KORISTI MAKRO $F1$ MJERA.

Metoda	Makro $F1$ mjera
<i>GapNet-PL</i> [4]	0.785
<i>ResNet</i> [3]	0.568
<i>InceptionNet v3</i> [1]	0.477
<i>M-CNN</i> [8]	0.729
<i>DeepLoc</i> [9]	0.520
<i>Convolutional MIL</i> [7]	0.411

Što se tiče eksperimentalne procedure, korišćenje unakrsne validacije nema smisla za probleme dubokog učenja, nego samo za male ili manje skupove podataka. Što se tiče *GapNet-PL*-a [4], preprocesirani primjerici su se nasumično podijelili u trening (87.5%) i validacioni (12.5%) skup. Za *InceptionNet v3* [1] podjela na trening i validacioni skup iznosila je 95% : 5%, dok je kod *ResNet* [3] modela podjela bila 90% : 10%.

Veličina *batch*-eva za modele zavisi od memoriske iskorušenosti i bira se da bude što veća moguća. Za treniranje *ResNet* modela korišćen je Google Colab Pro³, dok je za

treniranje *InceptionNet v3* korišćen *Intel UHD Graphics 620* sa 3 GB memorije i 16 GB RAM-a. Kao optimizator za *GapNet-PL* korišćen je *Stochastic Gradient Descent* (SGD) sa momentumom od 0.9. Kako bi se trening stabilizovao korišćen je i *gradient clipping* gdje su gradijenti normalizovani ukoliko je njihova globalna $L2$ norma bila preko praga (koji je podešen na 5). Za optimalne rezultate, koristio se promjenljivi *learning rate*, tako da se mijenja sa vremenom. Inicijalni *learning rate* je određen pretragom hiperparametara na validacionom skupu. Kako bi izbjegli *overfitting* korišćene su sledeće regularizacione tehnike: $L1$ norma sa vrijednošću 10^{-7} i $L2$ norma sa vrijednošću 10^{-5} .

Korišćeni hiperparametri za *InceptionNet v3* model su: veličina slike (299×299), veličina *batch*-a (16), *dropout* (50%), broj epoha za *warm-up* novododatih slojeva (2), broj epoha za regularne slojeve (20), *Adam*-ov optimizator, *learning rate* za *warm-up* (10^{-3}), *learning rate* za regularne slojeve (10^{-4}).

Kod *ResNet* modela, korišćeni su hiperparametri sa sledećim vrijednostima: veličina slike (512×512), veličina *batch*-a (64), *gradient clipping* (0.99), *dropout* (50%), *learning rate* (0.005), *Adam*-ov optimizator, *Focal Loss* prilikom treniranja.

Povodom rezultata dobijenih sa *InceptionNet v3* arhitekturom, model za neke klase uspešno klasificuje i uspijeva prepoznati organelu (dio ćelije), dok je kod drugih klasa uspešnost manja. Labele koje odlično klasificuje su: *Nucleoplasm*, *Nucleoli*, *Microtubules* i *Cytosol*. Labele sa kojima je na test skupu imao problem prilikom klasifikacije (većinom nije uspio da klasificuje) su: *Rods & Rings*, *Cytoplasmic bodies*, *Lipid droplets* i *Microtubule ends*.

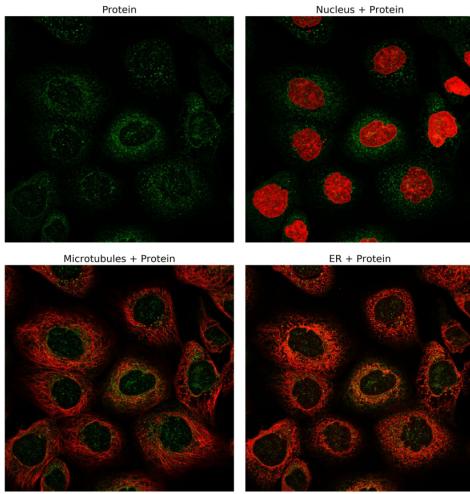
Za *ResNet* model, za razliku od *InceptionNet v3* modela nema ekstrema u pogledu klasifikacije labela. Dok *InceptionNet v3* za neke ulaze radi vrhunski, a za druge radi skroz loše, *ResNet* nema toliko velike ekstreme. Većinom radi dobro za sve, što ga u prosjeku čini boljim. Za visoko zastupljene labele, on radi dobro, ali su mu lošije performanse nego kod *InceptionNet v3* modela, dok za rijede i rijetko zastupljene labele radi takođe dobro, a *InceptionNet v3* tu ima ozbiljne probleme, tj. loše performanse.

Kako bi demonstrirali praktične sposobnosti pretreiranog modela, izvršeno je poređenje performansi *GapNet-PL*-a sa 3 ljudska eksperta u oblasti patobiologije koji često rade sa fluorescentnim mikroskopskim slikama i 25 postdiplomskih studenata koji se bave prirodnim naukama i radili su sa fluorescentnom mikroskopijom, a rezultati su prikazani u Podsekciji IV-A. Za ovo poređenje, uzimamo u obzir samo primjerke gdje se proteinski patern veže za jedinstvenu lokaciju u ćeliji kako bi pojednostavili zadatak i smanjili količinu posla potrebnu ljudskim učesnicima eksperimenta. U prvoj interaktivnoj sesiji sa ekspertima, za svaki primjerak oni su dobili četiri slike, gdje je svaki od tri kanala prekriven sa proteinskim paternom, a kanal proteina od interesa je izdvojen kao zaseban. Svi eksperti su dobili mogućnost da samostalno pregledaju dijelove trening skupa i upoznaju se sa njim. Nakon toga, svi su potvrdili da će biti sposobni da lokalizuju proteine na osnovu ovih slika. Svi učesnici su dobili oko 200 primjeraka (odnosno slika) iz test skupa podataka, te su imali približno oko jednu sedmicu

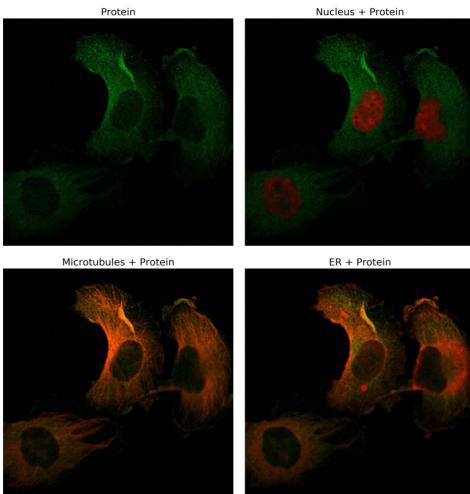
³<https://colab.research.google.com/>

da ih klasificuju.

A. Analiza grešaka modela kroz poređenje sa ljudskim eksperima i studentima



Slika 15. Primjeri koje je pogrešno klasifikovala i mreža *GapNet-PL*, a i svi ljudski ekserti i studenti. Stvarna klasa: Endoplazmatični retikulum. *GapNet-PL*: Vezikule. Ljudski ekserti i studenti: Vezikule, Citosoli. Sve tri klase su sličnih struktura van nukleusa i to je dovelo do pogrešne klasifikacije.



Slika 16. Primjerak koji su pogrešno klasifikovali ekserti i studenti, ali je *GapNet-PL* tačno klasifikovala. Stvarna klasa: Citosoli. *GapNet-PL*: Citosoli. Ljudski ekserti i studenti: Mikrotubule, Plazmatična membrana, Vezikule.

V. ZAKLJUČAK

U ovom radu, rješavali smo problem lokalizacije proteina gdje se na osnovu jedne ulazne crno-bijele slike, koja se propušta kroz 4 filtera (čime dobijamo boju), određuje unutar kojih dijelova ćelije (organela) se taj proteinski patern nalazi. Kada znamo u kojim dijelovima ćelije se on nalazi, ujedno znamo i lokaciju proteina. Motivacija za rješavanje ovog problema ogleda se u činjenici da proteini igraju značajnu

Tabela II

POREDENJE PREDIKTIVNIH PERFORMANSI MREŽE *GapNet-PL* SA LJUDSKIM EKSPERTIMA KORIŠĆENJEM MAKRO *F1* MJERE. DATI SU REZULTATI POJEDINAČNIH EKSPERATA I ANSAMBLA EKSPERATA. KOD ANSAMBLA EKSPERATA KONAČNI IZLAZ ODREĐEN JE VEĆINSKIM GLASANJEM. U SLUČAJU DA IMAMO IZJEDNAČENJE, JEDNA OD IZJEDNAČENIH KLASA SE BIRA NASUMIČNO.

Metoda	Makro <i>F1</i> mjera
<i>GapNet-PL</i> [4]	0.800
<i>Ekspert 1</i>	0.570
<i>Ekspert 2</i>	0.581
<i>Ansambl eksperata</i>	0.601

ulogu u fiziologiji svakog živog organizma, te disbalans u njihovoj proizvodnji dovodi i do mnogih patoloških stanja organizma. Nastojali smo da se izborimo sa neizbalansiranim skupom podataka koristeći pažljive funkcije gubitka tokom treniranja, preprocesiranje slika i dodavanje novih obilježja koja su nam bila od pomoći kako bi dobili što bolje rješenje. Tako preprocesirane slike, uz dodata nova obilježja, propuštane su u numeričkoj reprezentaciji tokom faze treniranja kroz tri različite arhitekture neuronskih mreža. Na kraju, dobili smo istrenirane modele koji su u stanju da uspješno prediktuju koji se proteinski patern nalaze u kojim dijelovima ćelije. Kroz eksperimente i poređenje sa ekspertima, potvrđeno je da je naš model primjenljiv.

LITERATURA

- [1] W. Ouyang, C. Winsnes, M. Hjelmare, A. Cesnik, L. Åkesson, H. Xu, D. Sullivan, S. Dai, J. Lan, P. Jinmo, S. M. Galib, C. Henkel, K. Hwang, D. Poplavskiy, B. Tunguz, R. Wolfinger, Y. Gu, C. Li, J. Xie, and E. Lundberg, “Analysis of the human protein atlas image classification competition,” *Nature Methods*, vol. 16, pp. 1254–1261, 12 2019.
- [2] Y. LeCun, Y. Bengio *et al.*, “Convolutional networks for images, speech, and time series,” *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.
- [3] W. J. Godinez, I. Hossain, S. E. Lazic, J. W. Davies, and X. Zhang, “A multi-scale convolutional neural network for phenotyping high-content cellular images,” *Bioinformatics*, vol. 33, no. 13, pp. 2010–2019, 02 2017. [Online]. Available: <https://doi.org/10.1093/bioinformatics/btx069>
- [4] E. Rumetschofer, M. Hofmarcher, C. Röhrl, S. Hochreiter, and G. Klambauer, “Human-level protein localization with convolutional neural networks,” in *International conference on learning representations*, 2018.
- [5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” 2015.
- [7] O. Z. Kraus, J. L. Ba, and B. J. Frey, “Classifying and segmenting microscopy images with deep multiple instance learning,” *Bioinformatics*, vol. 32, no. 12, p. i52–i59, Jun 2016. [Online]. Available: <http://dx.doi.org/10.1093/bioinformatics/btw252>
- [8] X.-S. Wei, C.-W. Xie, and J. Wu, “Mask-cnn: Localizing parts and selecting descriptors for fine-grained image recognition,” *arXiv preprint arXiv:1605.06878*, 2016.
- [9] J. J. Almagro Armenteros, C. K. Sønderby, S. K. Sønderby, H. Nielsen, and O. Winther, “Deeploc: prediction of protein subcellular localization using deep learning,” *Bioinformatics*, vol. 33, no. 21, pp. 3387–3395, 2017.