Predikcija B-cell epitopa u antigenima

Ante Sorić
0036539765
ante.soric@fer.hr

Domagoj Sviličić
0036540224
domagoj.svilicic@fer.hr

Lana Barić
0036538219
lana.baric@fer.hr

Martin Bugarin
0036539403
martin.bugarin@fer.hr

Anđelko Prskalo
0036544169
anđelko.prskalo@fer.hr

Diego Mišetić
0036543343
diego.misetic@fer.hr

I. Uvod

B-cell epitopi predstavljaju specifične dijelove antigena (proteina ili peptida) koji aktiviraju imunološki odgovor tijela vezanjem na B-cell, te tako omogućujući stvaranje kompleksa antigen-antitijelo. Ova svojstva B-cell epitopa ključna su za imunološka istraživanja poput razvoja peptidnih cjepiva, dijagnosticiranje bolesti te istraživanja povezana s alergijama. Mogu se podijeliti u dvije osnovne kategorije: linearne i konformacijske. U ovom projektu fokus je na linearnim epitopima zbog njihove lakše obrade i mogućnosti sinteze umjetnim peptidima. Unatoč znanju, eksperimentalne metode za identifikaciju epitopa zahtijevaju mnogo vremena i resursa.

Linearni B-cell epitopi, koji su sačinjeni od uzastopnih aminokiselinskih ostataka, osobito su korisni u dizajniranju cjepiva i istraživanju imunoloških odgovora. U novije vrijeme, algoritmi temeljeni na umjetnim neuronskim mrežama pokazali su potencijal u preciznijem predviđanju ovih epitopa. Posebno se ističe primjena rekurentnih neuronskih mreža (RNN), koje zbog svoje sposobnosti obrade sekvencijskih podatka nude značajne prednosti u odnosu na klasične mreže.

Antigeni, proteini koji pokreću imunološke reakcije, sastoje se od peptida izgrađenih od 20 standardnih aminokiselina. B-cell epitopi definirani su kao podnizovi unutar ovih sekvenci na koje se vežu B-cell. Glavni izazov u predikciji epitopa je varijabilna duljina sekvenci te kompleksne interakcije između aminokiselina koje utječu na njihova imunološka svojstva. Rekurentne neuronske mreže idealne su za obradu takvih podataka jer mogu uzeti u obzir sekvencijalne odnose pamtiti prethodne elemente unutar niza.

U ovom projektu koristimo bazu IEDB za stvaranje skupa podataka. Pozitivni uzorci obuhvatit će poznate B-cell epitope, dok će negativni uzorci biti generirani iz nasumičnih proteinskih sekvenci. Ovakav pristup osigurava biološku relevantnost i pouzdanost modela.

Primjenom RNN-a, projektom će se istražiti kako duljina epitopa i okruženje unutar proteinske sekvence utječu na imunološka svojstva.

II. Pregled postojećih pristipa

Jedno od postojećih rješenja za predikciju kontinuiranih B-cell epitopa u antigenima je sustav ABCpred, temeljen na rekurentnim neuronskim mrežama (RNN). Razvio ga je tim istraživača kako bi unaprijedili točnost predikcije epitopa, koji su ključni za razvoj peptidnih cjepiva, dijagnostiku bolesti i istraživanja alergija, a svi detalji opisani su u [1].

ABCpred sustav koristi podatke iz baze Bcipep, koja sadrži 2479 kontinuiranih B-staničnih epitopa. U tom modelu su primijenjene RNN i unaprijedne neuronske mreže (FNN), pri čemu su RNN pokazale značajno bolje rezultate. Kako bi se prilagodio zahtjevima neuronskih mreža za fiksnom duljinom ulaznih uzoraka, maksimalna duljina epitopa ograničena je na 20 aminokiselina zbog zahtjeva neuronskih mreža za fiksnom duljinom ulaznih uzoraka. Kraći epitopi nadopunjuju se susjednim aminokiselinama iz njihovih sekvenci, čime se stvaraju uzorci jedinstvene duljine. Osim toga, koristi se metoda petostruke unakrsne validacije kako bi se osigurala točnost predikcija i izbjegla prekomjerna prilagodba podacima. Najveća točnost ABCpred sustava postignuta je s duljinom prozora od 16 aminokiselina i jednim slojem od 35 skrivenih jedinica. Rezultati su pokazali osjetljivost od 67,14%, specifičnost od 64,71% i ukupnu točnost od 65,93%. U usporedbi s postojećim metodama, ABCpred pokazuje bolje rezultate u pristupima temeljenim na fizikalno-kemijskim svojstvima aminokiselina.

ABCpred omogućuje predikciju epitopa putem online sučelja gdje korisnici mogu odabrati parametre kao što su duljina prozora i prag osjetljivosti, što povećava fleksibilnost njegove primjene. Ipak, sustav ima određene izazove poput nemogućnosti preciznog definiranja granica epitopa i ograničenu dostupnost podataka o ne-epitopima. ABCpred sustav predstavlja značajan napredak u području bioinformatike, omogućujući precizniju identifikaciju B-cell epitopa. Međutim, zbog određenih ograničenja, preporučuje se korištenje u kombinaciji s drugim postojećim metodama za optimalne rezultate.

III. Opis rješenja problema

Prvi korak svakog istraživačkog rada je pretprocesiranje podataka. Iz prethodno navedenih baza podataka izvučeno je 4097 proteinskih sekvenci koje sadrže B-cell epitope te isti broj sekvenci koji ih ne sadrže. Također, iskorišteno je i 1400 redaka podataka koji su korišteni u Kako je u našem radu korišten pristup u kojem u mrežu šaljemo podatke fiksne duljine aminokiselina, a epitopi mogu biti varijabilne duljine, korišten je pristup u kojem epitope nadopunjavamo aminokiselinama iz njihove izvorišne proteinske sekvence. To se radi kako bi mreža mogla uzeti u obzir i kontekst u kojem se epitop nalazi u proteinu, što može biti važno za točnost predikcije. Dopunjavamo ga do fiksne veličine recimo (16,18,20). Npr. za epitop AAALPGKCGV, za definiranu veličinu ulaza 16 dopunjavamo na niz: PNNAAALPGKCGVHIP (samo ga okružimo sa aminokiselinama iz izvorne sekvence). Nadopunjavanje s obje strane (simetrično) osigurava da mreža dobije informacije o sekvencama koje okružuju epitop, jer bi sekvence prije i poslije epitopa mogle imati jednaku važnost u određivanju njegovih svojstava. Na primjer, neka funkcionalna svojstva epitopa mogu biti bolje prepoznata kroz interakcije s okolinom epitopa, a ne samo kroz sam epitop. Ako je epitop smješten blizu kraja proteinske sekvence, ne možemo ga ravnomjerno nadopuniti s obje strane jer možda nema

dovoljno aminokiselina na jednoj strani. U tom slučaju, isključujemo ga iz skupa za treniranje.

Neuronska mreža ne može direktno raditi s tekstualnim podacima (aminokiselinama), stoga moramo pretvoriti te sekvence u numeričke podatke. Najčešće korištena metoda u bioinformatici za kodiranje aminokiselina je one-hot enkodiranje. One-hot enkodiranje pretvara svaku aminokiselinu u binarni vektor duljine 21, gdje svaka pozicija u vektoru odgovara jednoj aminokiselini. U podacima također postoji oznaka *X*, koja označava nepoznatu aminokiselinu na toj poziciji, te zbog toga imamo vektore duljine 21 (20 aminokiselina te oznaka *X*).

Nakon svega ovoga, imamo 8640 redaka koji su korišteni u modelima. Skup podataka podijeljen je na skup za treniranje, skup za validaciju te skup za provjeru, u omjeru 60:20:20. Redci su promiješani tijekom prije svake epohe kako bi se spriječilo da model nauči redoslijed podataka, čime se povećava učinkovitost treniranja. Pri treniranju modela korištena je tehnika ranog zaustavljanja (engl. *early stopping)* koja se koristi za sprječavanje pretreniranosti modela i poboljšanje njegove generalizacije. Ako se učinak modela prestane poboljšavati tijekom određenog broja epoha (u našem modelu odabrano 20 epoha), treniranje se prekida prije nego što dosegne maksimalni broj epoha.

Za rad sa sekvencama odabrana su dva modela RNN-a, Elmanovov RNN –„obični“ RNN te LSTM (Long-short Term Memory). Elmanova mreža te LSTM tipovi su rekurentnih neuronskih mreža dizajniranih za obradu sekvencijalnih podataka. Elmanova mreža sastoji se od ulaznog sloja, skrivenog sloja i izlaznog sloja. Mreža prima ulazni vektor, x_t, koji se prosljeđuje skrivenom sloju. Skriveni sloj je povezan sam sa sobom, na način da se skriveno stanje prethodnog vremenskog koraka prosljeđuje nazad u trenutni skriveni sloj. Ovakav mehanizam stvara kratkoročnu memoriju, omogućujući mreži da uhvati određene ovisnosti između dijelova sekvenci.

Nakon obrade cijele sekvence, završno skriveno stanje h_t mora se linearnom transformacijom dovesti do oblika iskoristivog za procjenu točnosti modela, tj. trebamo dobiti samo jedan broj. Za to nam je dovoljan jedan sloj, koji će izvršiti linearnu transformaciju završnog skrivenog stanja u željeni izlaz. U našem modelu korištena su 2 potpuno povezana sloja. Na izlaz primjenjujemo sigmoidu koja izlaz pretvara u broj između 0 i 1, te je krajnja klasifikacija modela 1 ako je taj broj veći od 0.5, inače je 0.
Ovaj te ostali jednostavniji modeli poznati su po problemu nestajanja ili eksplodiranja gradijenata. Nestajanje gradijenata znači da kako se gradijenti propagiraju unazad, oni postaju sve manji, što otežava učenje dugoročnih ovisnosti, osobito u zadnjim slojevima. Ista stvar vrijedi i za eksplodiranje gradijenata, samo što tamo postaju sve veći. Ove, te ostale probleme s RNN-ovima riješilo se LSTM-ovima.

LSTM mreže podvrsta su RNN-ova, dizajnirana da efektivno uhvate dugoročne ovisnosti između sekvencijalnih podataka. Sastoji se od nekoliko glavnih komponenata. Stanje ćelije (engl. *cell state*) c_t ponaša se kao memorija mreže, čuvajući informacije tijekom vremena. Za razliku od običnih RNN-ova, gdje informacije mogu nestati tijekom dugih sekvenci (prethodno naveden problem nestajanja gradijenata), LSTM koristi stanje ćelije kako bi sačuvao informacije tijekom dužeg vremena. Glavna funkcija stanja ćelije je pohranjivanje dugoročnih informacija koje su relevantne za zadatak. Skriveno stanje h_t predstavlja izlaz iz LSTM-a tijekom svakog vremenskog koraka obrade sekvence. Skriveno stanje prenosi informaciju o tome što je LSTM dosad naučio u obradi sekvence i prenosi se u sljedeći vremenski korak obrade sekvence. Nakon svakog vremenskog koraka, LSTM ažurira stanje ćelije te skriveno stanje.

LSTM sadrži nekoliko naprednijih mehanizama u odnosu na Elmanov model. Vrata zaborava (engl. *forget gate*) odlučuju koliko informacija iz ćelije stanja treba odbaciti, na temelju trenutnog ulaza i prethodnog skrivenog stanja. Vrata ulaza (engl. *input gate*) kontroliraju koliko novih informacija treba biti dodano ćeliji stanja. Vrata izlaza (engl. *output gate*) odlučuju o tome kako će izgledati novo skriveno stanje, kombinirajući trenutno stanje ćelije te ulazne vektore podataka. Nakon obrade cijele sekvence, kao i kod Elmanovog modela, korištenjem 2 potpuno povezana sloja završno skriveno stanje LSTM-a dovodimo u oblik pogodan za predikciju.

Za oba modela koristi se postupak propagacije pogreške unazad kroz vrijeme (BPTT), što efektivno znači *razmotavanje* RNN-a kroz vremenske korake, tretirajući svaki korak kao zaseban sloj. Tijekom treniranja, BPTT izračunava gradijente funkcije pogreške te ažurira težine mreže kako bi smanjio pogrešku.

Za evaluaciju točnosti modela korišteno je nekoliko mjera: osjetljivost – udio stvarno pozitivnih slučajeva koji su ispravno identificirani od strane testa, specifičnost – udio stvarno negativnih slučajeva koji su ispravno identificirani od strane testa te pozitivna prediktivna vrijednost(PPV) – udio pozitivnih rezultata testa koji su zaista točni, te točnost na testnom skupu podataka.

IV. Opis eksperimentalnih rezultata

Tijekom treniranja mreže korištene su različite fiksne veličine epitopa: 16,18,20,22,24 i 26. Za svaku mrežu isprobane su različite kombinacije hiperparametara, te su u tablicama ispisani najbolji rezultati. Pokazalo se da najbolji modeli koji su koristili Elmanovu arhitekturu postižu točnost na testnom skupu od oko 70%, dok najbolji LSTM modeli postižu točnost od oko 80%. Međutim, ovo su rezultati nad fiksnim duljinama sekvenci, te će kasnije biti isprobani i nad cijelim sekvencama proteina kako bismo vidjeli koliko dobro predviđaju mjesta epitopa unutar proteina.

TABLICA I - PERFORMANSE ELMANOVOG MODELA RNN-A

Veličina prozora	Osjetljivost (%)	Specifičnost (%)	PPV(%)	Točnost(%)
16	42.50	91.52	79.38	70.22
18	45.38	88.10	73.77	69.97
20	60.53	77.19	68.99	69.60
22	45.51	89.77	78.27	69.96
24	61.26	81.27	72.20	72.41

26	70.70	75.47	70.60	73.30

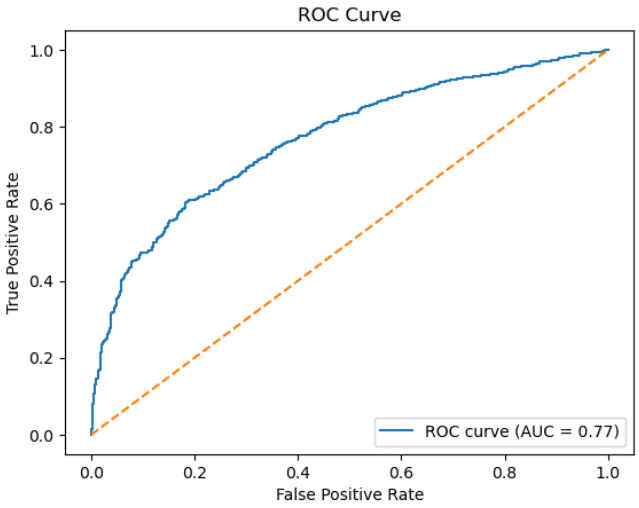
TABLICA II - PERFORMANSE LSTM MODELA

Veličina prozora	Osjetljivost	Specifičnost	PPV	Točnost
16	70.18	76.20	69.38	73.59
18	71.76	76.46	69.21	74.47
20	75.64	77.07	73.44	76.41
22	81.74	75.83	73.26	78.47
24	72.59	82.73	76.95	78.24
26	76.68	83.69	79.66	80.51

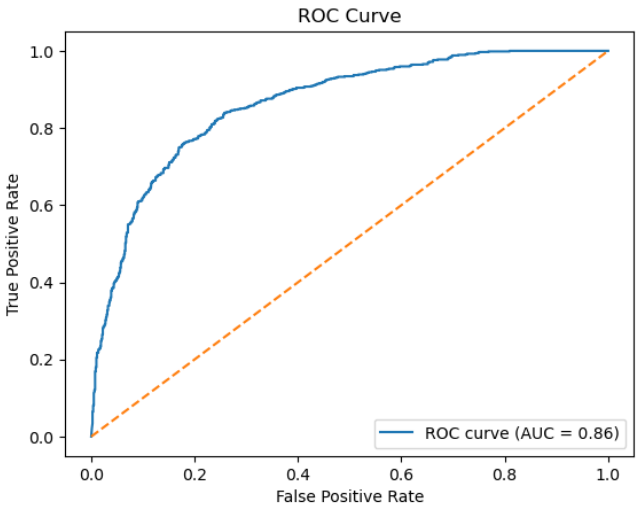
Na sljedećim slikama prikazane su ROC krivulje za oba modela, za veličinu prozora 26. Hiperparametri korišteni u ovim modelima su stopa učenja od 0.005, veličina skrivenog sloja od 30, *batch* size 64 te je korišten *Adam* optimizator za Elman model, odnosno stopa ulenja 0.01, veličina skrivenog sloja 40, *batch* veličina 128 te Adam optimizator za LSTM. Adam optimizator je algoritam za optimizaciju koji koristi adaptivne stope učenja za svaku težinu u mreži, kombinirajući prednosti trenutnih i prethodnih gradijenata. Time omogućuje bržu konvergenciju i bolje performanse u treniranju dubokih neuronskih mreža, te se pokazao boljim od SGD optimizatora. (stohastički gradijentni spust). ROC krivulja prikazuje performanse modela pri različitim pragovima klasifikacije. Na y-osi je prikazan *True Positive Rate* (TPR), koji mjeri postotak točno identificiranih pozitivnih primjera, dok je na x-osi *False Positive Rate* (FPR), koji mjeri postotak negativnih primjera pogrešno klasificiranih kao pozitivni. Površina ispod krivulje (AUC – *Area Under the Curve*) predstavlja ukupan kapacitet modela da razlikuje pozitivne od negativnih primjera.

Za Elmanov RNN model, AUC iznosi 0.77, što ukazuje na umjerene performanse. Vrijednost od 0.5 značila bi nasumično pogađanje, dok bi 1.0 predstavljala savršenu klasifikaciju. Ovo sugerira da Elmanov RNN model ima ograničene mogućnosti u prepoznavanju obrazaca u podacima. Za LSTM model, AUC iznosi 0.86, što je znatno bolje u usporedbi s RNN-om. LSTM model bolje razlikuje pozitivne i negativne primjere te pokazuje bolju ukupnu točnost. Krivulja za LSTM bliže je gornjem lijevom kutu, što ukazuje na superiorne performanse u klasifikaciji.

SLIKA I – ROC KRIVULJA ZA ELMANOV RNN MODEL S VELIČINOM PROZORA 26



SLIKA II – ROC KRIVULJA ZA LSTM MODEL S VELIČINOM PROZORA 26



Također, nakon treniranja i testiranja modela na podacima fiksnih duljina, model je korišten i za predikciju epitopa na proteinskim sekvencama varijabilnih duljina. Očekivano, ovakav pristup dao je nešto nižu točnost predikcija, tj. bio je poveći broj lažno pozitivnih slučajeva. Jedan od razloga tomu je što korištenje fiksne duljine sekvenci proteina za treniranje onemogućava modelu da točno predvidi poziciju epitopa, tj. ukoliko koristimo npr. prozor duljine 20, a epitop je duljine 8, model označava cijeli prozor od 20 kao pozitivan. Na temelju prikazanih rezultata, može se zaključiti da LSTM model pokazuje značajno bolje performanse u svim važnim metrima u odnosu na Elmanov RNN model. LSTM ima veću sposobnost prepoznavanja pozitivnih slučajeva (osjetljivost), što znači da bolje prepoznaje epitopse, dok RNN ima slabiju sposobnost u tom području. Također, LSTM pokazuje veću specifičnost, što znači da bolje prepoznaje kada nema epitopa, dok RNN, iako ima solidnu specifičnost, nije tako dosljedan. U smislu prediktivne vrijednosti (PPV), LSTM je učinkovitiji u davanju točnih predikcija, a točnost modela također favorizira LSTM, s time da on postiže bolje rezultate u svim veličinama prozora. Ovi rezultati sugeriraju da LSTM bolje upravlja složenostima u podacima i omogućuje bolje

predikcije, osobito kada se koristi veći broj aminokiselina u ulaznim sekvencama. Općenito, dok veće veličine prozora poboljšavaju performanse oba modela, LSTM je jasni pobjednik, jer u svim veličinama prozora postiže bolje rezultate nego RNN, koji je skloniji nižim vrijednostima u gotovo svim metrikama.

Ovaj rezultat također može i zavarati, jer povećanjem prozora također taj prozor obuhvaća veći broj aminokiselina, a kako je većina epitopa duljine 20 i manje aminokiselina, važno je modele isprobati i na cijelim proteinskim sekvencama, koristeći posmični prozor. Zanimljivo, pokazalo se da pojedini Elmanovi model RNN-a imaju veću točnost na ovakvim proteinima od LSTM-a, međutim ovaj rezultat lako nas može zavarati. Naime, izvorni proteini mogu biti varijabilne duljine, često i preko 100 aminokiselina, a proučavanjem metrika može se vidjeti da je osjetljivost Elmanovog RNN modela poprilično niska, tj. davanjem oznake „0“ većini podataka, model je „točniji“, međutim broj negativnih sekvenci puno je veći od pozitivnih.

TABLICA III – PERFORMANSE RNN MODELA

Veličina prozora	Osjetljivost	Specifičnost	PPV	Točnost
16	35.93	81.53	37.45	70.80
18	28.77	87.16	40.81	73.42
20	47.74	72.55	34.87	66.71
22	69.09	41.55	26.68	48.03
24	66.82	54.24	31.01	57.20
26	60.41	57.77	30.57	58.39

TABLICA IV – PERFORMANSE LSTM MODELA

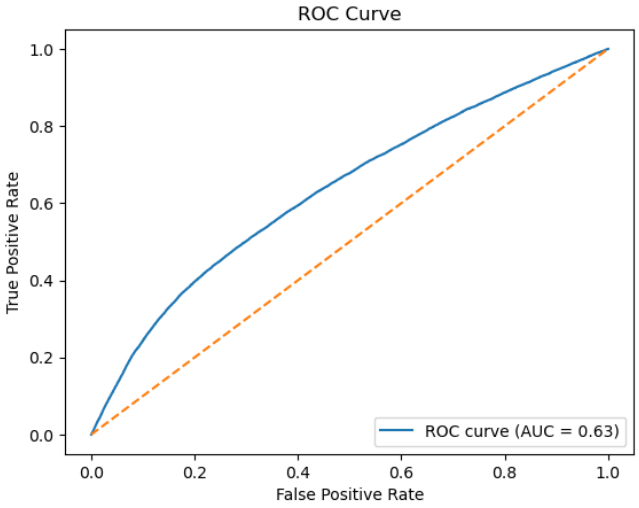
Veličina prozora	Osjetljivost	Specifičnost	PPV	Točnost
16	60.02	66.52	35.56	64.99
18	68.64	62.26	35.89	63.76
20	69.16	61.57	35.65	63.36
22	72.50	58.51	34.97	61.80
24	71.88	60.04	35.64	62.83
26	77.14	50.86	32.58	57.04

LSTM model postiže veću osjetljivost za sve veličine prozora u usporedbi s Elmanovim RNN modelom. Za veće veličine prozora LSTM model postiže znatno bolju osjetljivost, što ističe sposobnost LSTM-a da bolje uhvati kontekst. Elmanov RNN općenito ima veću specifičnost, osobito za manje veličine prozora. Međutim, za veće veličine prozora razlika u specifičnosti se smanjuje, pri čemu Elman RNN postiže 57,77% za veličinu 26 u usporedbi s 50,86% kod LSTM-a. PPV je relativno slična za oba modela kroz sve veličine prozora. LSTM model općenito pokazuje bolju točnost za većinu veličina prozora, s posebno dobrom izvedbom za veće veličine. Povećanje veličine prozora značajno poboljšava osjetljivost za LSTM model, ali smanjuje specifičnost. Za Elmanov RNN povećanje veličine prozora ne pokazuje konzistentan trend, što sugerira da model teže koristi dodatni kontekst duljih sekvenci.

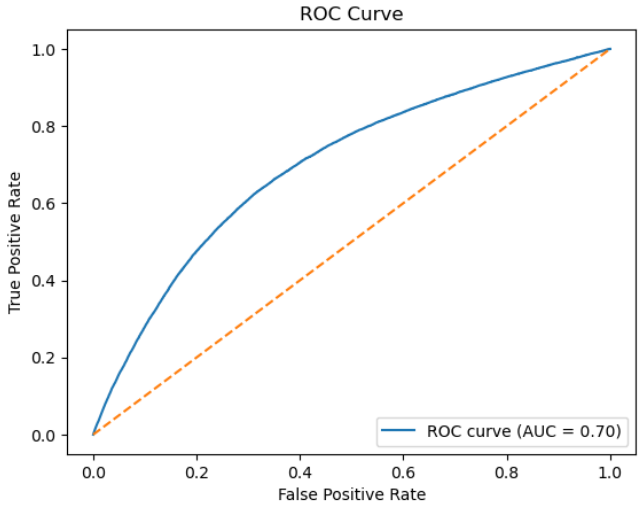
Elmanov RNN model ima veću specifičnost i pozitivnu prediktivnu vrijednost (PPV) u usporedbi s LSTM-om, što ukazuje da generira manje lažno pozitivnih rezultata. Međutim, zbog toga ima nižu osjetljivost, odnosno propušta identificirati dio pravih pozitivnih uzoraka. S druge strane, LSTM model postiže veću osjetljivost, ali na račun veće stope lažno pozitivnih rezultata, osobito kod većih veličina prozora. To znači da je LSTM bolji u hvatanju svih potencijalno važnih uzoraka, ali ponekad pogrešno označi nešto kao pozitivno.

Najbolji model među ovim nije lako za pronaći. Veličina prozora od 20 sekvenci se čini dobrom jer za RNN pruža dobar balans između prepoznavanja pozitivnih uzoraka i minimiziranja lažno pozitivnih, dok zadržava solidnu točnost. Za LSTM omogućava hvatanje šireg konteksta epitopa uz visoku osjetljivost, a istovremeno održava prihvatljiv broj lažno pozitivnih rezultata. Također, modeli s veličinom prozora 24 čine se približno dobrima. Povećanjem prozora, iako u Tablici 1 te Tablici 2 možemo vidjeti kako točnost raste, korištenjem tih modela na duljim sekvencama ne dobivamo najbolje rezultate. Na sljedećim slikama možemo vidjeti ROC krivulje za modele s veličinom prozora od 20. Lako je uočljivo kako je krivulja lošija u odnosu na krivulje sa slika III i IVV, što je očekivano, s obzirom da sada radimo s pomičnim prozorom te sekvencama varijabilnih duljina, što znači da će broj sekvenci koje treba označiti s NE biti dosta veći od onih pozitivnih.

SLIKA VI – ROC KRIVULJA ZA ELMANOV RNN MODEL S VELIČINOM PROZORA 20



SLIKA VII – ROC KRIVULJA ZA LSTM MODEL S VELIČINOM PROZORA 20



V. Diskusija

Razvoj i evaluacija novih modela za predikciju B-cell epitopa, uključujući LSTM i Elmanovog modela RNN arhitekture, omogućila je dobivanje detaljnijih uvida u sposobnosti suvremenih metoda u ovom području. ABCpred sustav, koji se temelji na fiksnim parametrima poput duljine sekvence (unutar istog eksperimenta) i jednostavnijim neuronskim mrežama, postigao je točnost od 65,93%. Za modele koje smo trenirali (LSTM i Elmanov RNN), najveću točnost je dao LSTM i to 80.51% (za duljinu sekvence 26). Za testiranje cijelih sekvenci proteina, pokazalo se da je LSTM bolji, pošto Elmanovi modeli imaju jako nisu osjetljivost.Ograničenja ABCpred sustava iz 2006. godine dijelom su proizlazila iz tadašnjih tehnoloških uvjeta, uključujući sporija računala i manje napredne algoritme dubokog učenja, što je utjecalo na sposobnost modela da postigne višu razinu preciznosti i fleksibilnosti. Naši rezultati dobiveni korištenjem LSTM i RNN modela, treniranih na sekvencama duljine 16, 18, 20, 22, 24 i 26 aminokiselina, predstavljaju korak prema adresiranju tih ograničenja. LSTM arhitektura omogućuje bolje razumijevanje dugoročnih ovisnosti unutar sekvenci, dok je Elmanov RNN ponudio jednostavniji pristup s nešto nižim performansama. Kombinacija različitih parametara poput veličine slojeva, brzine učenja i veličine *batch*-a osigurala je detaljnu evaluaciju utjecaja tih faktora na konačnu točnost.

Kao i što je navedeno u [1], modeli imaju velik broj lažno pozitivnih predviđanja i ne može predvidjeti granice B-staničnih epitopa. Jedan od razloga lošeg predviđanja jest činjenica da B-stanični epitopi nemaju fiksnu duljinu, a mi koristimo prozor s fiksnom duljinom. Također, korištenjem većeg fiksnog prozora u modelima, taj prozor obuhvati veći broj aminokiselina te time ostvaruje veću točnost, iako model neće moći točno zaključiti gdje točno počinje ili završava epitop, već cijeli prozor označava kao pozitivan.

Ovaj pristup mogao bi se poboljšati boljim upravljanjem epitopa s promjenjivim duljinama, možda kroz naprednije tehnike poput dinamičkog podešavanja veličine prozora ili korištenja dužih sekvenci podataka. Također, uključivanje više eksperimentalno potvrđenih ne-B-staničnih epitopa i bolja integracija s drugim prediktivnim modelima mogla bi poboljšati točnost i smanjiti lažno pozitivne rezultate.

VI. Zaključak

Modeli poput LSTM i Elmanovog RNN-a, omogućio je dublju analizu i usporedbu metoda za predikciju B-cell epitopa u odnosu na ABCpred sustav. Iako ABCpred sustav nudi solidne rezultate s točnošću od 65,93%, naši eksperimenti pokazali su napredak u točnosti. LSTM arhitektura posebno se istaknula prepoznavanjem složenih uzoraka u sekvencijama. Kombinacija naprednih arhitektura i optimizacije hiperparametara otvara nove mogućnosti za primjenu u dizajnu cjepiva i istraživanju imunoloških odgovora.

VII. Literatura

[1] Saha, Sudipto & Raghava, Gajendra. (2006). Prediction of Continuous B-cell Epitopes in an Antigen Using Recurrent Neural Network. Proteins. 65. 40-8. 10.1002/prot.2107