

Segmentacija slika magnetske rezonancije i dijagnostika pomoću virtualnog medicinskog asistenta

Jurdana, Mia

Master's thesis / Diplomski rad

2024

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Electrical Engineering and Computing / Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva**

Permanent link / Trajna poveznica: <https://urn.nsk.hr/urn:nbn:hr:168:889821>

Rights / Prava: [In copyright](#)/[Zaštićeno autorskim pravom.](#)

Download date / Datum preuzimanja: **2025-03-14**



Repository / Repozitorij:

[FER Repository - University of Zagreb Faculty of Electrical Engineering and Computing repository](#)



SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 495

**SEGMENTACIJA SLIKA MAGNETSKE REZONANCIJE I
DIJAGNOSTIKA POMOĆU VIRTUALNOG MEDICINSKOG
ASISTENTA**

Mia Jurdana

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 495

**SEGMENTACIJA SLIKA MAGNETSKE REZONANCIJE I
DIJAGNOSTIKA POMOĆU VIRTUALNOG MEDICINSKOG
ASISTENTA**

Mia Jurdana

Zagreb, lipanj 2024.

DIPLOMSKI ZADATAK br. 495

Pristupnica: **Mia Jurdana (0069069957)**

Studij: Računarstvo

Profil: Računarska znanost

Mentorica: doc. dr. sc. Jelena Božek

Zadatak: **Segmentacija slika magnetske rezonancije i dijagnostika pomoću virtualnog medicinskog asistenta**

Opis zadatka:

Algoritmi strojnog učenja omogućavaju učinkovitu i točnu segmentaciju iz medicinskih slika. Segmentacijom organa i promjena u tkivu iz slika magnetske rezonancije (MR) moguće je dobiti informacije o veličini abnormalnosti, točnoj lokaciji, odnosu zdravog i bolesnog tkiva. U radu je potrebno navesti pregled metoda strojnog učenja koje se koriste za segmentaciju tkiva iz MR slika. Cilj rada je napraviti aplikaciju koja korištenjem modela strojnog učenja iz MR slika automatski segmentira organe i promjene u tkivu te pruža informacije o volumenu, položaju i drugim karakteristikama segmentiranog područja. Model strojnog učenja potrebno je trenirati na nekoliko javno dostupnih skupova podataka. Aplikacija mora omogućiti korisniku unos nepoznatog skupa podataka te pomoću predtreniranog modela izvršiti segmentaciju slika. Potrebno je izraditi virtualnog asistenta s kojim korisnik može komunicirati nakon provedene segmentacije. Virtualni asistent mora na temelju korisnikovih upita, a iz pouzdanih izvora i znanstvenih članaka, pružiti dodatne informacije o segmentiranom području, dijagnozi i mogućnostima liječenja.

Rok za predaju rada: 28. lipnja 2024.

Zahvaljujem se svojoj mentorici Jeleni Božek na pomoći, strpljenju i pozitivnim mislima prilikom pisanja diplomskog rada. Posebno se zahvaljujem svojoj obitelji na iskazanoj podršci i ohrabrenju tijekom studiranja.

Sadržaj

1. Uvod	3
2. Virtualni medicinski asistent	5
3. MRI: Osnovni koncepti i metodologija	8
3.1. Skup MRI podataka	11
4. U-net: mreža za semantičku segmentaciju MRI slika	14
4.1. NnU-Net: samo-konfigurirajuća U-net mreža	16
4.1.1. Karakteristike skupa podataka	17
4.1.2. Fiksni parametri	17
4.1.3. Treniranje	18
4.1.4. Parametri temeljeni na pravilima	19
4.1.5. Empirijski parametri	20
5. Veliki jezični modeli	21
5.1. Halucinacije velikih jezičnih modela: izazovi i rješenja	23
5.1.1. Inženjering upita	24
5.2. RAG	26
5.3. Napredni RAG	27
6. Implementacija	31
7. Zaključak	37
Literatura	39
Sažetak	43

Abstract **44**

1. Uvod

Napredak u tehnologijama umjetne inteligencije omogućio je razvoj alata koji značajno unapređuju medicinsku dijagnostiku, osobito u području obrade medicinskih slika. Segmentacija slika magnetske rezonancije je proces izdvajanja i označavanja specifičnih anatomske struktura ili patoloških promjena poput tumora, krvnih žila, organa i raznih oštećenja. Ručna segmentacija medicinskih slika je spor i zahtjevan proces koji traži visoku razinu stručnosti i znanja, stoga potreba za automatiziranim, bržim i pouzdanijim metodama segmentacije postaje sve nužnija. Razvijaju se različiti modeli koji značajno smanjuju potrebu za ljudskom intervencijom, no neki izazovi i dalje ostaju [1]. Svaki skup podataka često zahtijeva prilagodbu ili dodatno treniranje modela kako bi se postigla optimalna točnost, što komplicira široku primjenu u kliničkoj praksi. U ovom radu istražuje se primjena U-net mreže koja se automatski konfigurira i prilagođava različitim skupovima podataka bez potrebe za dodatnim ručnim podešavanjem. Mreža je unaprijed trenirana na raznolikim podacima koji variraju u svojim značajkama poput broja vokselâ, razmaka među vokselima, broja primjera u skupu za treniranje i broja klasa, čime se pokazuje njezina prediktivna moć.

Segmentacijski postupak maksimalno je pojednostavljen te se svi koraci odvijaju kroz interakciju putem *chat* sučelja koje korisniku pruža personalizirano iskustvo, smanjuje potrebu za tehničkim znanjem i daje osjećaj kao da razgovara s nekime tko ga vodi kroz cijeli proces. Možemo reći da korisnik komunicira sa svojim virtualnim medicinskim asistentom koji koristi napredne algoritme dubokog učenja kako bi riješio dane zadatke. Zahvaljujući razvitku velikih jezičnih modela, ovaj asistent može razumjeti složene upite, interpretirati ih i pružiti relevantne informacije [2]. Također, uz dane alate, asistent može pristupiti bazi podataka, spremati podatke, pokrenuti segmentaciju slika, prikazati rezultate i slično. Osim segmentacije slika magnetske rezonancije, asistent koristi napredne tehnike pretraživanja vanjske baze znanstvenih članaka za odgovaranje na korisnikova pitanja, što dodatno unapređuje točnost i korisnost njegovih odgovora.

Fokus rada je na integraciji postojećih tehnologija umjetne inteligencije, uključujući algoritme dubokog učenja i obrade prirodnog jezika, kako bi se stvorio učinkovit sustav za dijagnostiku i automatizaciju segmentacije. Umjesto razvijanja novih tehnologija, cilj je povezivanje i primjena već usavršenih rješenja unutar intuitivnog sučelja.

2. Virtualni medicinski asistent

Virtualni asistent predstavlja sustav koji koristi napredne AI (engl. *Artificial Intelligence*) tehnologije uključujući obradu prirodnog jezika (engl. *Natural Language Processing - NLP*), algoritme strojnog i dubokog učenja i velike jezične modele (engl. *Large Language Models - LLMs*). Virtualni asistent “razumije” i analizira upite korisnika te sukladno upitima generira relevantne odgovore. Osim konverzijskih sposobnosti, virtualni asistent nudi automatizaciju rutinskih zadataka, ali i mogućnost rješavanja naprednijih logičkih zadataka ukoliko mu se za to pruže odgovarajući alati [3]. Primjeri takvih zadataka mogu biti:

- Rutinski zadaci, npr. slanje e-mailova, naručivanje namirnica ili plaćanje računa;
- Organizacijski zadaci, npr. upravljanje kalendarima, zakazivanje sastanaka ili postavljanje podsjetnika;
- Analiza korisničkih podataka, prepoznavanje obrazaca i generiranje personaliziranih preporuka;
- Korištenje naprednih algoritama za obradu i analizu velikih količina podataka s ciljem predviđanja kretanja tržišta i donošenja poslovnih odluka.

Kao što možemo vidjeti, virtualni asistent koristi se na različitim domenama primjene uključujući medicinu, gdje je on poznatiji pod nazivom *Medibot* [4]. Njegov zadatak može biti primjerice:

- Podrška u dijagnostici: Analizirajući medicinske slike i podatke, asistent može pomoći liječnicima u identifikaciji bolesti.
- Automatizacija administrativnih zadataka: Asistent ubrzava procese poput unosa podataka ili zakazivanja pregleda.
- Edukacija o zdravlju: Asistent pruža informacije o zdravlju, uključujući savjete za prevenciju bolesti i upute za zdrave životne navike.
- Procjena simptoma: Asistent analizira simptome koje korisnici opisuju i pruža informacije o mogućim uzrocima i mogućnostima liječenja.

- Informacije o lijekovima: Asistent pruža detaljne informacije o lijekovima, uključujući preporučene doze, moguće nuspojave i interakcije s drugim lijekovima.

U sklopu ovog rada razvijen je virtualni medicinski asistent s kojim korisnik može komunicirati putem jednostavnog i intuitivnog *chat* sučelja. Iako korisnik koristi *chat* za interakciju s aplikacijom, važno je napomenuti da se ne radi o klasičnom *chatbotu*. *Chatbotovi* obično temelje svoje odgovore na unaprijed definiranim obrascima i specijalizirani su za odgovaranje na konkretna, često postavljana pitanja (FAQ), dok agenti nude naprednije mehanizme kako bi analizirali i interpretirali složene upite i zadatke [5]. Također, agenti imaju “svijest” o kontekstu i povezuju informacije iz prethodnih odgovora kroz cijeli razgovor dok se *chatbotovi* često nisu u mogućnosti referirati na prethodne interakcije. Medicinski asistent, razvijen u sklopu ovog rada, ima dvije ključne funkcionalnosti:

1. Segmentacija slika magnetske rezonancije (engl. *Magnetic Resonance Imaging - MRI*): Korisnik stvara novi projekt kroz *chat* gdje u razgovoru s asistentom unosi potrebne informacije poput imena projekta, dobi, spola, bilješki te MRI datoteke. Asistent, pomoću alata kojim može pristupiti bazi podataka, sprema korisnikov projekt te traži dodatne informacije o odabiru skupa podataka na kojem je mreža za segmentaciju trenirana i odabiru arhitekture mreže. Nakon prikupljenih podataka, asistent koristi alat za pristup modelu za segmentaciju, predaje prikupljene podatke tom modelu i pokreće segmentaciju korisnikove MRI slike. Korisnik može pratiti napredak procesa segmentacije i vidjeti u kojem je koraku (pretprocesiranje, predikcija, postprocesiranje). Nakon što je segmentacija završena, rezultati segmentacije (segmentacijska maska) vidljivi su u *chatu*.
2. Odgovaranje na medicinska pitanja: Odgovaranje na pitanja bazira se na naprednom RAG (engl. *Retrieval-Augmented Generation*) sustavu. Korisnik unosi ključne riječi za pretraživanje vanjske baze znanstvenih radova. Asistent pristupa ovoj bazi podataka, pretražuje članke koji odgovaraju unosu i prikazuje korisniku rezultate pretrage. Korisnik zatim može odabrati koje članke želi uključiti u daljnje istraživanje. Za odabrane članke, asistent preuzima cijeli tekst i pohranjuje ga u internu vektorsku bazu podataka. Kada korisnik postavi pitanje, asistent pretražuje odabrane članke. Koriste se različite napredne tehnike pretraživanja kako bi se pronašli relevantni dijelovi teksta i na temelju njih i početnog upita generirao ispravan

odgovor. Ovaj pristup omogućava asistentu da generira odgovore na temelju najrelevantnijih informacija iz odabranih izvora, čime se poboljšava točnost i korisnost odgovora.

U nastavku ovog rada bit će detaljno objašnjena terminologija i koncepti potrebni za razumijevanje funkcionalnosti medicinskog asistenta. Također, bit će razjašnjeni tehnički aspekti aplikacije, uključujući arhitekturu sustava, odabrane tehnologije i pristupe koji su primjenjeni u razvoju.

3. MRI: Osnovni koncepti i metodologija

Iako bi se magnetskoj rezonanciji mogao posvetiti čitav jedan novi rad, u ovom će se poglavlju spomenuti osnovni koncepti potrebni za razumijevanje “putovanja” snimke iz skenera do skupa podataka spremnog za segmentaciju. Nedugo nakon što je ušla u kliničku praksu ranih 1980-ih, MRI je postala jedna od najvažnijih tehnika medicinskog oslikavanja, a tu ulogu zadržava do danas [6]. MRI je neinvazivna tehnika koja pruža širok spektar informacija uključujući ne samo strukturne i anatomske karakteristike, već i one funkcionalne i metaboličke. S mogućnosti snimanja cijelog tijela primjena je neograničena - od praćenja metaboličkih promjena oko tumora, predoperativnog planiranja, praćenja kroničnih bolesti pa sve do kardioloških, mišićno-koštanih ili neuroloških primjena.

Ono što nam omogućuje dobivanje slike, gledajući s hardverske strane, su magneti unutar skenera u čije magnetno polje postavljamo željene dijelove tijela. Cilindrični glavni magnet, jačine iznad 1 T (Tesla), centriran je na središnjoj osi skenera i proizvodi snažno homogeno magnetsko polje koje usmjerava magnetske momente protona u smjeru jednakom ili suprotnom orijentaciji glavnog polja. Kako bismo mogli razlikovati protone na različitim lokacijama u tijelu, dodaju se tri gradijente zavojnice koje proizvode prostorno promjenjiva magnetna polja za svaku os. Njihov glavni učinak je mijenjanje jakosti, ali ne i smjera glavnog polja. Posljedično, zbog promjene jakosti polja, protoni na različitim lokacijama osciliraju različitim frekvencijama i te frekvencijske informacije možemo koristiti za određivanje položaja protona i time lokalizaciju signala.

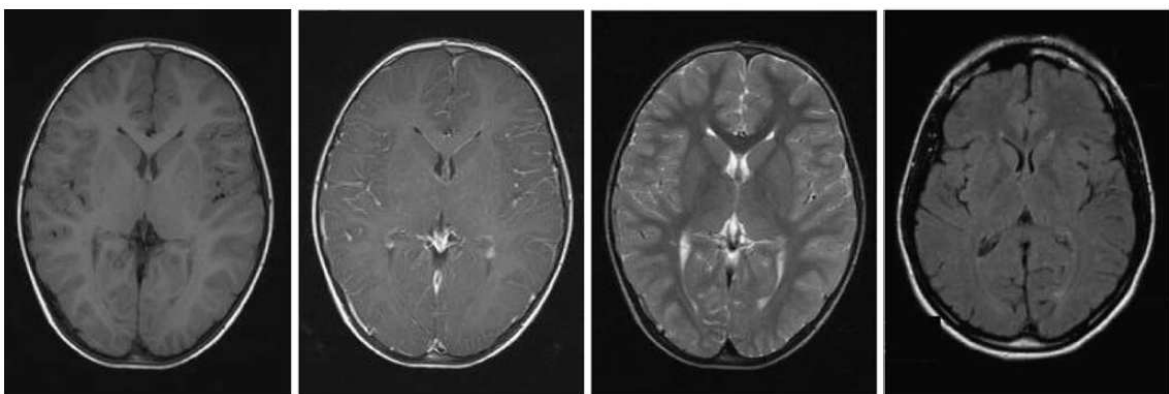
Nakon što se tijelo, koje se oslikava, nalazi u vanjskom magnetnom polju, primjenjuje se RF puls (engl. *Radiofrequency Puls*) koji zakreće os magnetizacije i pobuđuje spinove protona te oni prelaze u viši energetski nivo. Nakon djelovanja RF pulsa protoni se vraćaju u ravnotežno stanje i taj proces nazivamo relaksacijom. Ovdje su ključna dva parametra:

- T1 relaksacija (longitudinalna relaksacija): Opisuje proces povratka longitudinalne komponente magnetizacije (paralelne s glavnim magnetskim poljem) u svoju ravnotežnu vrijednost.
- T2 relaksacija (transverzalna relaksacija): Opisuje proces povratka transverzalne komponente magnetizacije (okomite na glavno magnetsko polje) u svoju ravnotežnu vrijednost.

Vrijeme između dva RF pulsa naziva se vrijeme ponavljanja (engl. *Time to Repeat - TR*), a vrijeme između primjene RF pulsa i mjerenja signala naziva se vrijeme odjeka (engl. *Time to Echo - TE*). S obzirom na spomenutu terminologiju, dobivamo različite vrste MRI sekvenci:

- T1 slike (engl. *T1 weighted - T1w*): Nastaju kada su TR i TE vremena kratka. Korisne su za prikazivanje anatomskih detalja i strukture tkiva. Tkiva s dugim vremenom T1 relaksacije biti će tamnija.
- T2 slike (engl. *T2 weighted - T2w*): Nastaju kada su TR i TE vremena duga. Korisne su za prikazivanje tekućine i edematoznih područja. Tkiva s dugim vremenom T2 relaksacije bit će svjetlija.
- T2-FLAIR (engl. *Fluid Attenuated Inversion Recovery*): Nastaju kada su TR i TE vremena duga, uz dodatan RF puls. Korisne su za prikazivanje malih lezija i abnormalnosti.

Dodatno, u literaturi još susrećemo i T1w-Gd (engl. *T1 weighted with gadolinium*) slike koje se odnose na T1 slike koje koriste gadolinij kao kontrastno sredstvo [7] [8].



Slika 3.1. MRI sekvence: T1w, T1w s kontrastom, T2w i FLAIR [6]

Prilikom povratka u niže energetska stanje protoni emitiraju signale koje zatim primljene zavojnice "hvataju" pojačavajući i bilježeći signal. Jakost primljenog signala se

zatim prevodi u nijanse sive boje čime se postupno slažu dvodimenzionalni presjeci tijela (engl. *slice*), a kombinacijom tih presjeka dobiva se trodimenzionalna slika, odnosno 3D volumen podataka. Podaci su podijeljeni u voksele koji su ekvivalent pikselima u 2D prostoru. Svaki voksel ima određenu dimenziju, primjerice $1 \times 1 \times 1 \text{ mm}^3$, te sadrži jednu vrijednost koja predstavlja prosječan signal izmjeren na toj lokaciji. MRI skeneri generiraju podatke u sirovom (engl. *raw*) formatu. DICOM (engl. *Digital Imaging and Communications in Medicine*) je uobičajen, standardiziran format, međutim s njim većina programa ne može raditi pa se podaci najčešće pretvaraju u NIFTI (engl. *Neuroimaging Informatics Technology Initiative*).

Nakon akvizicije podataka potrebno je provesti nekoliko koraka pretprocesiranja kako bi se MRI slike pripremile za segmentaciju. Najvažniji koraci uključuju korekciju polja, normalizaciju i registraciju. Opcionalno, mogu se ukloniti dijelovi slike koji se neće koristiti, poput lubanje u slučaju kada želimo promatrati isključivo mozak.

Zbog prostorne nehomogenosti magnetskog polja, varijacija u osjetljivosti prijemnih zavojnica i interakcije između čovjeka i magnetskog polja, mogu nastati nepravilnosti u intenzitetu signala tkiva s istim fizičkim svojstvima. Kako bi se sa slike uklonili artefakti primjenjuje se zaglađivanje filterom koji uklanja visoke frekvencije. Zaglađivanje će povećati omjer signala i šuma pa će promjene na slici biti naglašenije, no istovremeno će se rezolucija malo smanjiti. Za korekciju polja u literaturi se mogu pronaći i druge metode poput minimizacije entropije slike, usporedbe histograma lokalnog okruženja s globalnim histogramom ili korištenja registriranih predložaka [9].

Predložak se također koristi kod normalizacije i tim pojmom opisujemo standardni referentni prostor na koji preslikavamo podatke. Da bi se slike različitih subjekata međusobno mogle usporediti, moraju se uskladiti i uprosječiti na zajednički oblik i veličinu što upravo zovemo normalizacijom. U kontekstu mozga, referentni predložak koji sadrži informacije o anatomskej strukturi naziva se atlas.

Registracija slika je proces prostornog usklađivanja dviju ili više slika istog subjekta ili različitih subjekata. Registracija podrazumijeva pronalaženje transformacija između slika kako bi se odgovarajuće značajke “poklopile”. Ukoliko se radi samo o jednom subjektu najčešće je dovoljna linearna odnosno rigidna transformacija koja se sastoji od translacije i rotacije. Kada je u pitanju više subjekata, često se moraju primijeniti afina transformacija koja dodatno uključuje skaliranje i nakošenje ili nelinearna transforma-

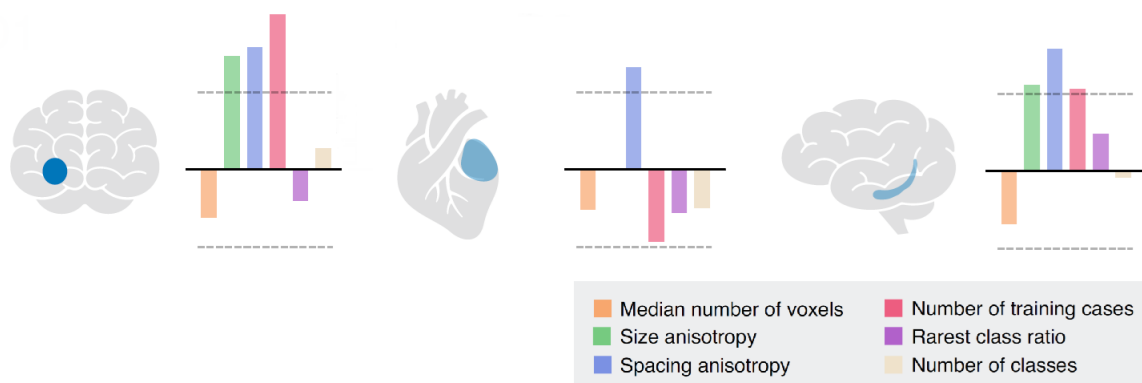
cija.

Ukoliko želimo imati označeni skup podataka slike je dodatno potrebno ručno segmentirati “sloj po sloj”. Ručna segmentacija je i dalje najpreciznija metoda, ali i vremenski najiscrpnija, no potrebna je kako bi se automatizirana segmentacija ispravno mogla evaluirati.

3.1. Skup MRI podataka

Da bi uopće mogli razmišljati o pokušajima automatizirane semantičke segmentacije medicinskih slika, potreban je ručno segmentiran, javno dostupan i visokokvalitetan skup podataka. Institucije nerado dijele medicinske podatke zbog zakonskih regulativa o privatnosti podataka, a kada su oni i dostupni, moguće je da će nedostajati odgovarajuće oznake ili će podataka biti nedovoljno mnogo što dobrom algoritmu za segmentaciju smanjuje prediktivnu moć i sposobnost generalizacije [10]. Nije nova vijest da algoritama za segmentaciju u medicinskom području ima sve više, ali većina arhitektura tih algoritama je dobro prilagođena za specifični problem [11]. Izazov je dizajnirati takvu osnovnu arhitekturu koja bi se mogla prilagoditi različitim zadacima, odnosno u ovom kontekstu, segmentirati različite anatomske strukture dobivene različitim modalitetima snimanja. Za potrebe adresiranja navedenih problema, razvijen je koncept pod nazivom *Medical Segmentation Decathlon* (MSD) [12]. MSD je internacionalni izazov u kojem se pokušava pronaći najbolji algoritam opće namjene za segmentaciju medicinskih slika. U svrhu rješavanja tog izazova prikupljen je veliki, javno dostupan, ručno anotiran skup podataka koji se sastoji od deset zadataka za segmentaciju različitih anatomskih struktura. Izazov obuhvaća segmentaciju regija u području mozga, srca, hipokampusa, jetre, pluća, gušterače, prostate, debelog crijeva, krvnih žila jetre i slezene u dva različita modaliteta - MRI i kompjuterizirana tomografija (CT). Ipak, većina zadataka snimana je CT modalitetom snimanja, čime se ovaj rad ne bavi, pa su za potrebe aplikacije preuzeta tri podskupa podataka pod nazivima: *Task01_BrainTumor* (tumor na mozgu), *Task03_Heart* (srce) i *Task04_Hippocampus* (hipokampus). U svakom od podskupova susrećemo se s različitim izazovom - segmentacija kompleksnih i neuniformno razmještenih lokacija tumora, segmentacija dva jako mala susjedna područja s visokom preciznošću u hipokampusu i mali skup podataka za segmentaciju lijevog atrija s velikom varijabilnosti.

Zadatak *Task01_BrainTumor* obuhvaća združene podatke iz izazova pod nazivom



Slika 3.2. Raznolikost svojstva odabranih skupova podataka [13]

The International Brain Tumor Segmentation (BraTS) iz 2016. i 2017. godine. Sastoji se od 750 snimaka magnetske rezonancije (484 u skupu za treniranje i 266 u skupu za testiranje) u četiri sekvence (T1w, T1w-Gd, T2w i T2-FLAIR). Cilj je segmentacija tri područja tumora - peritumoralni edem, pojačavajući tumor i nekrotična jezgra tumora bez povećanja. Korišteni skeneri jačinom variraju od 1 T do 3 T. Sve snimke su registrirane na isti anatomske predložak (SRI24 atlas), napravljeno je ponovno uzorkovanje (engl. *resample*) na rezoluciju voksela od 1 mm^3 te se mozak ručno ekstrahirao.

Zadatak *Task03_Heart* prvobitno je objavljen kroz izazov *Left Atrial Segmentation Challenge (LASC)* i uključuje svega 30 snimaka (20 u skupu za treniranje i 10 u skupu za testiranje) na kojima se vidi cijelo srce. Korišten je skener jačine 1.5 T, a rezolucija voksela je $1.25 \times 1.25 \times 2.7 \text{ mm}^3$. Lijevi atrij segmentirao je ekspert pomoću automatiziranog alata uz naknadnu manualnu korekciju.

Zadatak *Task04_Hippocampus* sastoji se od 195 snimaka, a uključuje 90 odraslih zdravih osoba i 105 osoba s nekom vrstom shizofrenije. Podaci su preuzeti iz repozitorija podataka *Psychiatric Genotype/Phenotype Project* sa sveučilišta Vanderbilt University Medical Center (SAD). Korištena je T1 sekvenca snimanja na skeneru jačine 1.5 T, a veličina voksela je 1 mm^3 .

Svi podaci su u NIfTI formatu i mogu se preuzeti u obliku TAR datoteke u kojoj se nalaze slike i oznake iz skupa podataka za treniranje, slike iz skupa podataka za testiranje i JSON datoteka s informacijama o skupu podataka ¹. Svi podaci sistematično su organizirani i imenovani. Primjerice, *hippocampus_367.nii* predstavlja osobu 367 iz skupa podataka za segmentaciju hipokampusa, a nalazi se u *imagesTr* direktoriju, dok

¹<http://medicaldecathlon.com/>

se odgovarajuća oznaka nalazi u *labelsTr* direktoriju pod istim nazivom. Pritom se pod terminom "oznaka" (engl. *label*), u kontekstu segmentacije, misli na ručno segmentiranu masku ciljane anatomske strukture.

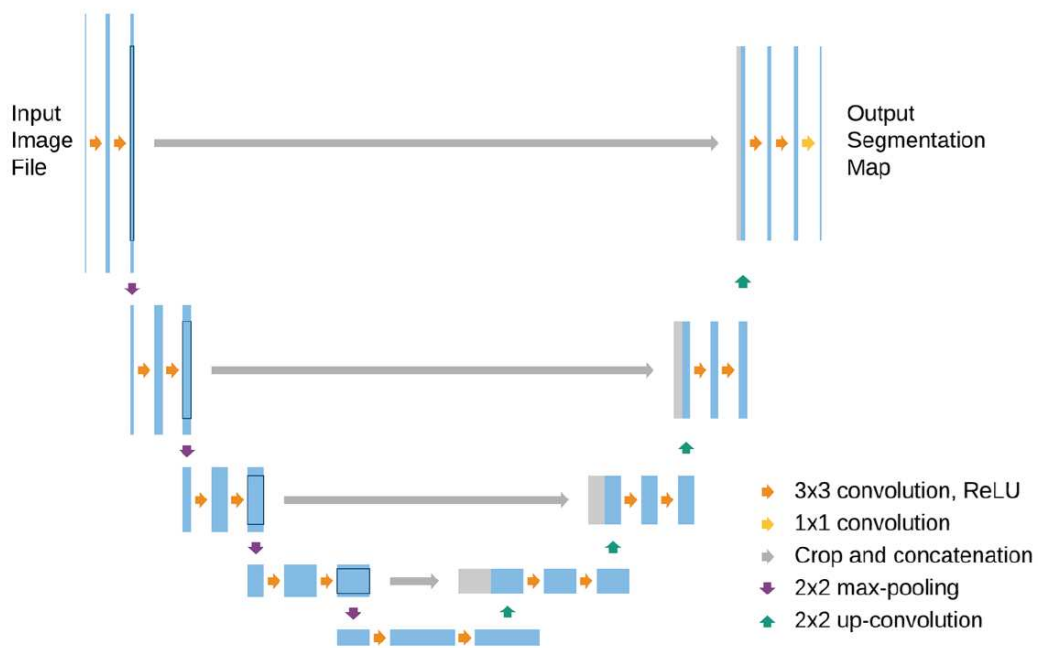
4. U-net: mreža za semantičku segmentaciju MRI slika

Semantička segmentacija je postupak pridjeljivanja semantičke oznake svakom pikselu slike pri čemu oznaka predstavlja kategoriju ili klasu kojoj taj piksel pripada. Rezultat semantičke segmentacije je matrica, koju još nazivamo i maska, u kojoj svaki element odgovara pikselu u izvornoj slici i označava njegovu klasu. U kontekstu medicinskih slika segmentacija se koristi za identifikaciju anatomskih struktura, lezija ili tkiva. U ovom poglavlju biti će objašnjena korištena arhitektura mreže, uz pretpostavku da čitatelj već poznaje osnovne termine poput konvolucijskog sloja, aktivacijske funkcije, regularizacije i sličnih pojmova.

U klasifikacijskim zadacima izlaz iz mreže je vektor veličine k gdje je k broj klasa. U zadacima detekcije izlaz iz mreže je vektor koji uključuje parametre potrebne za definiciju okvira oko objekta: x koordinata centra okvira, y koordinata centra okvira, visina, širina i klasa. Za segmentacijski zadatak izlaz iz mreže je čitava slika istih dimenzija kao i ulazna, što je puno složenije od prethodno spomenutih vektora te nam stoga obična konvolucijska mreža (engl. *Convolutional Neural Network - CNN*) neće biti od pomoći [14]. CNN mreže koriste slojeve sažimanja (engl. *pooling layer*) kako bi postupno smanjile prostornu dimenziju značajki (engl. *feature*). To će nam pomoći za “hvatanje” značajki visoke razine, ali dovodi do gubitka prostornih informacija što otežava precizno otkrivanje i razdvajanje objekata na razini piksela. Također, CNN mreže često na kraju koriste potpuno povezane slojeve da bi izlaz bio vektor fiksne veličine za klasifikaciju, a ne zadržavaju prostorne značajke potrebne za lokalizaciju objekata.

U-net arhitektura neuronske mreže dizajnirana je specifično za zadatak segmentacije. Osnovna arhitektura U-net mreže sastoji se od dva dijela - koder (engl. *encoder*) i dekoder, koji omogućuju prikupljanje lokalnog konteksta, ali i globalnih informacija. Koder je sličan uobičajenoj konvolucijskoj mreži, on “skenira” sliku, izvlači relevantne značajke i sažima ih u reprezentacije s većom apstrakcijom. Dekoder uzima značajke

i uči informacije potrebne za lokalizaciju objekta te rekonstruira segmentacijsku mapu. Preskočne veze između kodera i dekodera (engl. *skip connections*) čuvaju ključne značajke iz prethodnih slojeva i kombiniraju ih sa značajkama u kasnijim slojevima što omogućuje precizniju segmentaciju. Rezultirajuća mreža je gotovo simetrična što joj daje oblik slova “U”. U-net mreža je posebice korisna kod medicinskih slika, čiji skupovi podataka mogu biti mali, jer daje dobre rezultate čak i na vrlo ograničenom broju podataka [15].



Slika 4.1. Arhitektura U-net mreže [11]

Proces kodiranja započinje nizom konvolucijskih slojeva. Konvolucijski slojevi ekstrahiraju značajke primjenom filtra, primjerice veličine 3 x 3, i djeluju na lokalno receptivno polje što omogućuje mreži da prepozna sitne značajke. Kako raste dubina konvolucijskih slojeva, mape značajki se povećavaju, a mreža uči sve veće, složenije objekte. Nakon konvolucijskih slojeva primjenjuje se nelinearna aktivacija ReLU (engl. *Rectified Linear Unit*) i sloj sažimanja. Slojevi sažimanja smanjuju prostornu dimenzionalnost mapa značajki kako bi se mreža mogla fokusirati na sve veće objekte. Kako se prostorne dimenzije smanjuju, proces kodiranja usredotočuje se na ekstrakciju globalnog konteksta.

S obzirom da prolaskom kroz niz konvolucijskih slojeva i slojeva sažimanja gubimo dobru rezoluciju slike s početka mreže, u procesu dekodiranja koriste se transponirane konvolucije koje omogućuju povećanje uzorkovanja (engl. *upsampling*) i malo po malo

vraćaju prostornu rezoluciju slike. Tijekom ovog procesa mreža uči kako popuniti praznine između trenutnih prostornih lokacija čime se povećava razlučivost značajki. Značajke iz slojeva kodiranja konkatenuiraju se s uvećanim značajkama iz slojeva dekodiranja pomoću preskočnih veza. Takva kombinacija značajki s više razina omogućuje precizniju segmentaciju. Funkcija dekodera je da obnovi prostorne informacije izgubljene tijekom kodiranja.

Da zaključimo, koder je odgovoran za dobivanje značajki visoke apstrakcije i učenje globalnog konteksta. Preskočne veze čuvaju prostorne informacije koderu i prenose ih do kasnijih slojeva dekodera te tako spajaju informacije s više skala iz različitih slojeva mreže. Dekoder integrira uvećane značajke sa značajkama dobivenim preko preskočnih veza i uči semantičke reprezentacije. Ova struktura pruža optimalnu kombinaciju informacija: u kasnijim slojevima koderu koriste se globalni kontekst i bogate semantičke informacije, dok se u ranijim slojevima čuvaju precizni detalji i lokalne značajke. Na taj način mreža koristi globalne i lokalne informacije za postizanje maksimalne točnosti u segmentaciji spajajući široki kontekst s finim detaljima slike.

4.1. NnU-Net: samo-konfigurirajuća U-net mreža

Unatoč uspješnosti U-net mreža u segmentacijskim zadacima, njihova primjena na specifične zadatke i dalje zahtijeva visok stupanj stručnosti u konfiguraciji mreže pri čemu male pogreške mogu dovesti do velikog pada u performansi [13]. Ovo je posebice izraženo u biomedicinskim podacima gdje svojstva skupova podataka, poput vrste modaliteta, veličine slike, razmaka između voksela i omjera klasa, značajno variraju. Uspješna konfiguracija mreže na jednom skupu podataka često se ne može prenijeti na drugi već je potrebno mrežu ponovno prilagoditi. Dodatan sloj složenosti u cijeloj priči predstavlja i limitiranost hardvera dostupnog za treniranje i predikciju. U ovom je radu za segmentaciju medicinskih slika preuzeta pretrenirana nnU-Net mreža koja se automatski konfigurira uključujući pretprocesiranje podataka, arhitekturu mreže, parametre treniranja i postprocesiranje za bilo koji zadatak u biomedicinskom području. Sveukupno, mreža je trenirana na 53 segmentacijska zadatka pokrivajući veliku raznolikost različitih tipova i svojstava slika. Model je javno dostupan (engl. *open source*) te se može koristiti za jednostavno *out-of-the box* treniranje bez dodatnih prilagodbi ili uzeti s već pretreniranim

težinama za predikciju na neviđenim primjerima ¹ ². Autori ove mreže definiraju “recept” za pronalaženje najbolje konfiguracije, a parametri su podijeljeni u tri grupe [13]:

- Fiksni parametri: Ne zahtijevaju prilagodbu između različitih skupova podataka.
- Parametri temeljeni na pravilima: Ovisе o specifičnim svojstvima skupa podataka (engl. *dataset fingerprint*) i mogu se definirati pomoću heurističkih pravila.
- Empirijski parametri: Preostali parametri koji se ne mogu smjestiti u jednu od prve dvije kategorije i moraju se empirijski naučiti.

4.1.1. Karakteristike skupa podataka

Kao prvi korak pretprocesiranja sa slike se miču rubne regije ispunjene nulama. Na većinu podataka ovo nema utjecaja, ali primjerice kod skupa podataka *Task01_BrainTumor* značajno su se smanjile veličine slika. Za izrezane slike bilježe se važne karakteristike:

- Veličina slike: Broj voksela u svakoj prostornoj dimenziji slike prije i nakon izrezivanja.
- Fizička veličina voksela (engl. *image spacing*): Stvarna udaljenost između susjednih voksela u svakoj prostornoj dimenziji. Ova vrijednost ostaje ista prije i nakon izrezivanja.
- Modalitet: Čita se iz metapodataka i u ovom radu je to MRI za svaki korišteni skup podataka.
- Broj klasa
- Ukupan broj primjera za treniranje
- Srednja vrijednost intenziteta svih voksela
- Standardna devijacija intenziteta svih voksela
- 0.5 i 0.95 percentili intenziteta: Ove vrijednosti pokazuju donje i gornje ekstreme u raspodjeli intenziteta.

4.1.2. Fiksni parametri

Sve nnU-Net arhitekture temelje se na istoj jednostavnoj osnovi. Autori nisu koristili ni jednu nedavno predloženu kompliciraniju varijantu mreže poput rezidualnih veza,

¹<https://github.com/MIC-DKFZ/nnUNet>

²<https://zenodo.org/records/3734294>

mehanizma pažnje ili dilatiranih konvolucija [13]. Koristi se uobičajena konfiguracija s dva bloka po koraku rezolucije u koderu i dekoderu, pri čemu svaki blok sadrži konvoluciju, normalizaciju i nelinearnu funkciju. Smanjivanje rezolucije (engl. *downsampling*) provodi se pomoću konvolucije s korakom (engl. *strided convolution*), a povećanje rezolucije (engl. *upsampling*) pomoću transponirane konvolucije. Početni broj značajki je 32 i udvostručen je (ili prepolovljen) pri svakom smanjenju (ili povećanju) rezolucije. Broj značajki dodatno je ograničen na 320 za 3D U-Net i 512 za 2D U-Net mrežu. Napravljene su samo manje promjene u odnosu na osnovnu U-Net arhitekturu:

- Kako bi segmenti slike (engl. *patch*) koji se obrađuju u jednom prolazu kroz mrežu bili veći, koriste se male grupe podataka (engl. *batch size*). Većina mreža trenirana je u grupama po 2 primjera.
- ReLU aktivacijska funkcija zamijenjena je s Leaky ReLU funkcijom koja ipak propušta neke negativne gradijente umjesto da ih sve postavlja na nulu.
- Normalizacija po grupama (engl. *batch normalization*) zamijenjena je s normalizacijom po primjeru (engl. *instance normalization*).

4.1.3. Treniranje

Sve mreže treniraju se tijekom 1000 epoha, a jedna epoha sadrži 250 grupa primjera. Koristi se stohastički gradijentni spust s Nasterovim momentom ($\mu = 0.99$), a početna stopa učenja (engl. *learning rate*) iznosi 0.01. Početna stopa učenja smanjuje se prema izrazu $\left(1 - \frac{\text{epoch}}{\text{epoch}_{\max}}\right)^{0.9}$.

Funkcija gubitka je zbroj unakrsne entropije (engl. *cross entropy loss*) i Dice gubitka (engl. *Dice loss*). Za računanje gubitka koristi se odgovarajuća ručno segmentirana maska. Gubitci se zbrajaju na svim razinama rezolucije i normaliziraju se tako da njihov zbroj bude 1.

Uzorci za mini grupe biraju se nasumično. Tijekom treniranja primjenjuju se različite tehnike augmentacije podataka: rotiranje, skaliranje, GausSovo zaglađivanje, podešavanje svjetline i kontrasta, gamma korekcija i zrcaljenje. Predikcija se vrši pomoću pristupa kliznog prozora (engl. *sliding window*), a slika se dijeli na manje segmente koji se obrađuju zasebno. Prozori se preklapaju na polovici svoje veličine, a s obzirom da je točnost smanjena na rubovima prozora, središnji dijelovi prozora dobivaju veću težinu u konačnom izračunu, dok se utjecaj rubova smanjuje.

4.1.4. Parametri temeljeni na pravilima

U procesu treniranja ili predikcije potrebno je napraviti nekoliko koraka čije se konkretno izvođenje bazira na parametrima izračunatim zasebno za svaku sliku. Primjenjuje se z-normalizacija tako da se od slike oduzima njezina srednja vrijednost, a zatim se rezultantna slika dijeli standardnom devijacijom. Dodatno, ukoliko su rezultati rezanja slike doveli do smanjenja slike za više od 25%, kreira se maska za središnje ne-nulte voksele i normalizacija se primjenjuje samo unutar te maske. Za CT modalitet normalizacija se izvodi drugačije od prethodno opisanog načina, no taj dio će se ovdje preskočiti.

NnU-Net koristi ponovno uzorkovanje kako bi se mreža bolje nosila s neujednačenim razmakom između voksele. Ovo je posebno važno za CNN mreže koje se temelje na rešetkama i zanemaruju informacije o fizičkom prostoru. Sve slike, ali i segmentacijske maske, se ponovno uzorkuju na isti razmak između voksele kako bi se postigla konzistentnost. Potrebno je pažljivo odabrati ciljni razmak između voksele. Veći razmaci rezultiraju s manjim slikama te gubitkom detalja, a manji razmaci rezultiraju s većim slikama, ali mreža teže akumulira kontekstualne informacije. Za 3D mrežu računa se medijan vrijednosti razmaka između voksele pronađenih u primjerima za svaku os. 2D mreža radi s dvije osi slike s najvećom rezolucijom, a također se za ciljni razmak odabire medijan za svaku os.

NnU-net mreža daje prednost tome da dijelovi slike koji se obrađuju budu veći kako bi se zadržale kontekstualne informacije. Cijena toga je smanjenje veličine grupe koje dovodi do više šuma u gradijentima tijekom unazadne propagacije. Početna veličina dijela slike koji se obrađuje u konvolucijskoj mreži postavlja se na medijan veličine slike dobivene nakon ponovnog uzorkovanja. U većini slučajeva ova veličina je prevelika da bi stala u GPU memoriju te ju je potrebno iterativno dodatno ažurirati sve dok ne odgovara zadanom memorijskom proračunu. U tom slučaju veličina grupe postavlja se na 2. Ukoliko nije bilo potrebe za dodatnim ažuriranjem, veličina grupe se može povećati sve dok se ne iskoristi sva dostupna GPU memorija.

Arhitektura mreže konfigurira se u ovisnosti o potrebnom broju operacija smanjenja rezolucije. Smanjenje rezolucije vrši se za svaku os zasebno na temelju veličine dijela slike i udaljenosti između voksele. Proces smanjenja rezolucije zaustavlja se kada bi daljnje smanjenje mapu značajki svelo na manje od 4 voksele ili kada bi mape značajki postale anizotropne (izraženo nejednake). Veličina konvolucijske jezge je $3 \times 3 \times 3$ za 3D

mrežu te 3 x 3 za 2D mrežu.

4.1.5. Empirijski parametri

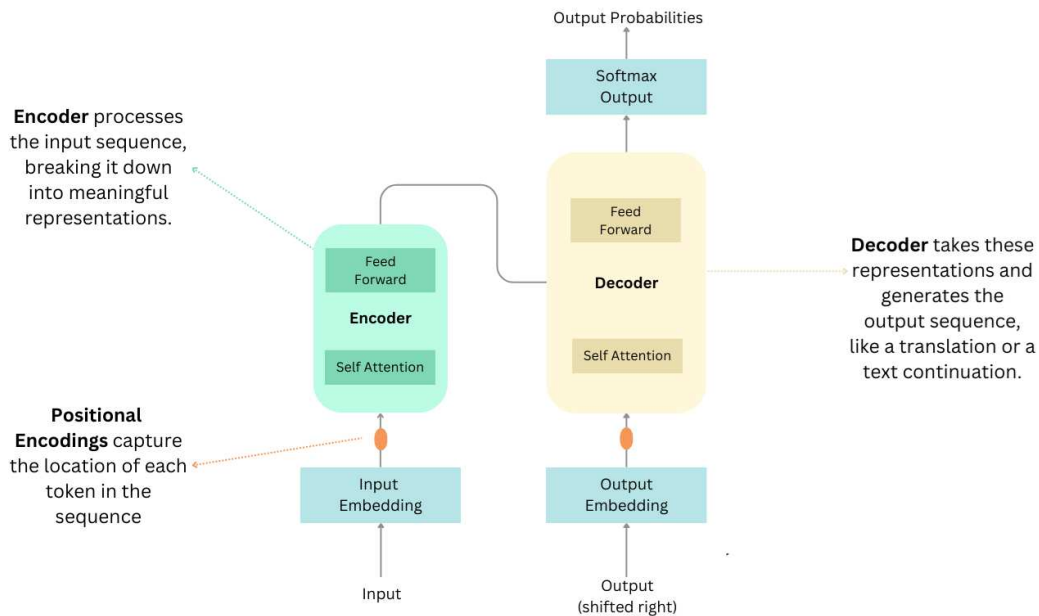
Na temelju prosječnog Dice koeficijenta učenog tijekom unakrsne validacije na treniranim podacima, autori nude mogućnost usporedbe različitih modela i odabira najboljeg između 2D modela, 3D modela s punom rezolucijom, 3D modela s nižom rezolucijom ili kaskada U-net mreža. S obzirom da proces usporedbe traje, za potrebe aplikacije korisnik je ograničen samo na izbor između 2D i 3D modela s punom rezolucijom.

Kao dio postprocesiranja koristi se tehnika povezanih komponenti (engl. *connected component-based post-processing*) kako bi se eliminirali lažno pozitivni rezultati. Identificiraju se sve grupacije vokseli koje su međusobno povezane te se uklanjaju one manje od zadane veličine ili se zadržava samo najveća pronađena grupacija za koju pretpostavljamo da predstavlja cilj segmentacije.

5. Veliki jezični modeli

Zahvaljujući velikim jezičnim modelima pojam umjetne inteligencije u posljednje je vrijeme privukao ogromnu pažnju [16] [17]. Gotovo svi su se u nekom trenutku koristili nezaobilaznim alatom ChatGPT-jem, najpoznatijim LLM-om, bilo da su ga integrirali u vlastite aplikacije ili koristili kao pomoć pri pisanju eseja. Prvo, pokušajmo razumjeti gdje se LLM-ovi nalaze u svijetu umjetne inteligencije. Umjetna inteligencija (engl. *Artificial Intelligence - AI*) je široko polje računalne znanosti koje se bavi razvojem sustava koji obavljaju zadatke koji bi inače zahtijevali ljudsku inteligenciju poput donošenja odluka ili prepoznavanja slika ili govora. Strojno učenje (engl. *Machine Learning - ML*) je grana umjetne inteligencije koja se bavi oblikovanjem algoritama koji svoju učinkovitost poboljšavaju na temelju empirijskih podataka. Unutar strojnog učenja nalazi se duboko učenje (engl. *Deep Learning - DL*) koji koristi višeslojne neuronske mreže za učenje iz velikih količina podataka. LLM-ovi su specifična vrsta dubokih modela treniranih na ogromnim količinama tekstualnih podataka, a mogu obavljati različite zadatke obrade prirodnog jezika poput generiranja teksta, sažimanja ili odgovaranja na pitanja. U akronimu GPT (engl. *Generative Pre-trained Transformer*) već možemo vidjeti o kakvoj se arhitekturi dubokog modela zapravo radi. LLM-ovi su generativni modeli jer mogu stvarati novi tekstualni sadržaj, ujedno su i pretrenirani te se baziraju na transformator arhitekturi. Transformator je temeljna gradivna jedinica svih jezičnih modela, a njegova osnovna sposobnost je predviđanje sljedeće riječi, odnosno *tokena*, na temelju dosad viđenog teksta. U transformator arhitekturi postoji nekoliko komponenti:

- **Ulaz: *Tokeni***, koje korisnik unosi, ulaz su u model. Ulazi se pretvaraju u vektore, odnosno ugrađuju (engl. *embedding*) u matematički prostor. U ovom prostoru semantički slične riječi predstavljene su vektorima koji su međusobno blizu. Ugradnje su poput rječnika koji model uči tokom treniranja, a pomažu mu da razumije značenje riječi.
- **Pozicijsko kodiranje**: U prirodnom jeziku redoslijed riječi mijenja značenje reče-



Slika 5.1. Transformator arhitektura [18]

nice. Neuronske mreže nemaju ugrađeno razumijevanje redoslijeda riječi pa se zbog toga treba dodatno ostvariti pozicijsko kodiranje. Modelu se dodaje informacija o poziciji svake riječi u ulaznoj sekvenci.

- **Koder:** Ključni dio transformator arhitekture koristi mehanizam pažnje (engl. *attention mechanism*) kako bi model procijenio važnost svake riječi u odnosu na druge riječi u tekstu i razumio kontekstualne veze između njih. Koder obrađuje tekst i generira niz skrivenih stanja (engl. *hidden states*) koje možemo zamisliti kao reprezentacije teksta na različitim nivoima apstrakcije. U transformator arhitekturi koristi se više slojeva, a svaki sloj generira svoja skrivena stanja koja se šalju u sljedeći sloj. Višeslojna arhitektura omogućava modelu da nauči složene obrasce i odnose u tekstu.
- **Izlaz:** Tijekom treniranja izlazna sekvenca pomiče se udesno kako bi se mogla predvidjeti sljedeća riječ na temelju prethodnih. Kao i ulazi, izlazi se također pretvaraju u vektore.
- **Dekoder:** Koristeći ulazni niz i kontekstualne informacije koje je prikupio koder, dekodeur uči predvidjeti sljedeću riječ i generira izlazni niz.
- **Linearni sloj i softmax:** U završnom koraku izračunati vektori prolaze kroz niz linearnih operacija kako bi se transformirali u *tokene*, pri čemu svaki *token* dobiva pridruženu vjerojatnost koja predstavlja šansu tog *tokena* da bude odabran kao sljedeći u nizu s obzirom na prethodni kontekst.

5.1. Halucinacije velikih jezičnih modela: izazovi i rješenja

Unatoč impresivnim sposobnostima, LLM-ovi ponekad pokazuju neobično ponašanje poput iznošenja netočnih ili besmislenih informacija [19] [20]. Takvo ponašanje naziva se halucinacijom, a postoji nekoliko vrsta:

- Kontradikcije u rečenici: Model generira rečenicu koja je proturječna nekoj prethodno generiranoj rečenici.
- Kontradikcija u upitu: Model generira tekst koji je proturječan postavljenom pitanju ili upitu.
- Činjenična kontradikcija: Model iznosi lažnu tvrdnju kao da je istinita.
- Besmisleni izlaz: Model generira tekst koji nema smisla ili logičke povezanosti.

Budući da je skup podataka na kojem je model treniran iznimno velik, vrlo je nepraktično provjeriti točnost svake pojedine informacije. Posljedično, u skupu podataka za treniranje osim pouzdanih izvora, tamo mogu završiti i podaci iz neprovjerenih izvora. Zbog ove raznolikosti u podacima postoji mogućnost da model stvori neočekivane poveznice između informacija što je glavni izvor nastanka halucinacija. Također, LLM-ovi nemaju sposobnost logičkog razmišljanja već funkcioniraju tako da oponašaju ljudski jezik na temelju vjerojatnosti generiranja riječi u određenom kontekstu. Model ne raspolaže mehanizmom provjere točnosti informacija i ne zna izraziti nesigurnosti ili “priznati” kad mu nedostaje znanja. Do halucinacija može doći i kada model naiđe na nejasan upit, neodređeno pitanje ili zadatak koji nije dovoljno precizan. Primjerice, ako upit sadrži skraćenicu koja se ovisno o kontekstu može tumačiti na različite načine, model može generirati irelevantan, netočan ili previše općenit odgovor. Ukoliko je upit vezan za vrlo specifičnu temu iz određenog područja, također je vrlo vjerojatno da će doći do halucinacije, s obzirom da model nije specijaliziran za određenu domenu već pokriva širok spektar znanja. Krajnji korisnici i programeri mogu koristiti nekoliko različitih strategija kako bi se smanjila sklonost modela halucinacijama i postigla veća sposobnost generiranja specifičnih odgovora.

Fino podešavanje (engl. *fine-tuning*) je tehnika prilagodbe već pretreniranog LLM-a kako bi se optimizirao za specijalizirane zadatke. Za dodatno treniranje potrebno je odabrati podatke koji što bolje odražavaju zadatke s kojima će se model susresti te po-

krivaju što više scenarija unutar ciljane domene. Priprema podataka uključuje čišćenje, označavanje i moguće proširivanje podataka kako bi se osiguralo da model ima robustan skup podataka za učenje. U ovom procesu često je potrebno dodatno prilagoditi hiperparametre modela ili adaptirati zadnjih nekoliko slojeva modela. Međutim, takve prilagodbe treba provoditi vrlo oprezno kako bi se izbjeglo prekomjerno prilagođavanje (engl. *overfitting*) i očuvala sposobnost modela za generalizaciju. Iako fino podešavanje može značajno poboljšati točnost modela, ono također zahtijeva napredno tehničko znanje, vrijeme i značajne računalne resurse za treniranje.

5.1.1. Inženjering upita

Inženjering upita (engl. *prompt engineering*) je metoda koja ne zahtijeva ponovno treniranje modela, čime se štede računalni resursi, a također je fleksibilna i omogućuje brze eksperimente s različitim zadacima i stilovima izlaznih podataka. Inženjering upita uključuje skup tehnika i metoda za dizajniranje, pisanje i optimizaciju upita, koje još nazivamo *promptovi*, kako bi odgovori velikih jezičnih modela bili precizni, konkretni, točni i činjenično ispravni. U ovom radu je LLM, koji komunicira s korisnikom, vođen nizom različitih *promptova* koji se mijenjaju o ovisnosti o zahtjevima pojedinog koraka. Više detalja biti će izneseno u poglavlju o tehničkim specifikacijama.

Promptovi se mogu smatrati oblikom programiranja jer omogućuju prilagodbu izlaza i interakciju s LLM-om, a obuhvaćaju prilagodbu uputa na prirodnom jeziku kako bi se dobili željeni odgovori [21]. Prema strukturi upiti se mogu formulirati koristeći otvorena ili zatvorena pitanja. Otvorena pitanja omogućuju opsežan i detaljan odgovor, a nasuprot tome zatvorena pitanja imaju specifične, ograničene odgovore. Prema dostupnim informacijama upiti se mogu kategorizirati u razine od 1 do 4. Prva razina obuhvaća jednostavna i direktna pitanja koja ne pružaju dodatan kontekst ili upute jezičnom modelu. Druga razina pruža dodatan kontekst koji modelu pomaže u generiranju točnog odgovora. Kontekst može uključivati informacije o piscu (npr. tko postavlja pitanje i zašto) te informacije o samom modelu (npr. kako model treba pristupiti odgovoru). Treća razina uključuje primjere koji modelu pomažu da razumije očekivani oblik i stil odgovora. I konačno, četvrta razina omogućuje modelu da razloži složenije upite na pojedinačne korake ili komponente. Slijedi nekoliko primjera tehnika inženjeringa upita:

- *Zero-shot prompting*: Model dobiva uputu koja ga vodi prema rješavanju novog za-

datka. Model koristi svoje postojeće znanje i zadanu uputu kako bi generirao odgovor. Primjer *prompta*: "Classify the sentiment of this review as positive or negative: *The movie was absolutely fantastic, with stunning visuals and a gripping story.*"

- *Few-shot prompting*: Tehnika u kojoj se modelu daje nekoliko primjera ulaz-izlaz kako bi imao jasniju predodžbu o tome kakav se odgovor očekuje. Negativne strane ovog pristupa su povećanje broja ulaznih *tokena* u model te pristranost modela, poput favoriziranja češćih riječi ili obrazaca u primjerima. Primjer *prompta*: "Classify the following reviews as positive or negative. Example 1: *The movie was thrilling and captivating from start to finish. (Positive)* Example 2: *The service was slow and the food was cold. (Negative)* Now, classify this review: *The book was well-written and engaging throughout.*"
- *Chain-of-Thought (CoT) prompting*: Ova tehnika omogućuje modelu da daje odgovore "korak po korak". Model razlaže složen problem na slijed logičkih faza koje pokazuju cijeli proces "razmišljanja" koji je potreban da bi se došlo do konačnog rješenja. Primjer *prompta*: "Solve the following math problem using step-by-step reasoning: $2x + 3 = 7$ "
- *Prompt chaining*: Tehnika u kojoj se koristi niz upita, a svaki se upit nadograđuje na prethodni. Primjer *promptova*: "Calculate the total cost of a shopping list." i "If I have a 10% discount on the total cost calculated in the previous step, how much will the discounted total be?"

Cilj koji se želi postići određuje strukturu i dizajn samog upita, no postoji nekoliko generalnih sugestija za stvaranje učinkovitih upita koji omogućuju generiranje željenih odgovora [22]:

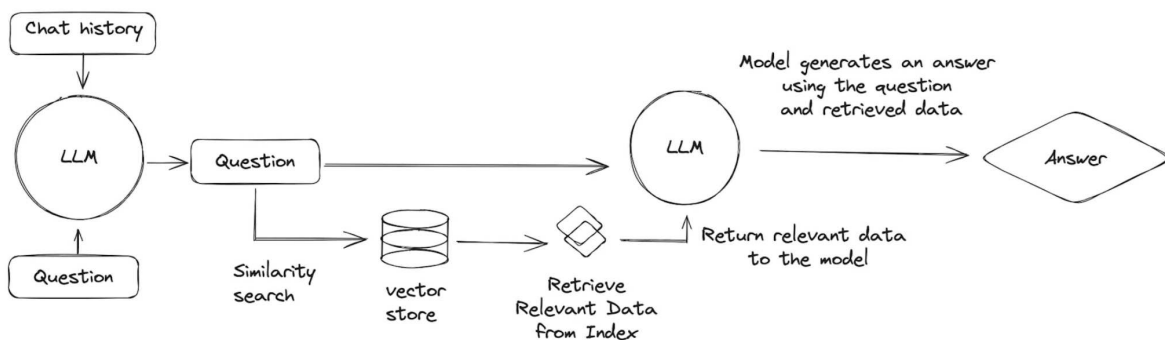
- Definiranje uloge, npr. "You are..." ili "You act as...";
- Definiranje zahtjeva, npr. "Your task is to...";
- Opis formata izlaza, npr. JSON, paragraf, popis s oznakama..;
- Opis ograničenja za očekivane rezultate;
- Proširenje *prompta* s izrazima kao što su "Within the scope..." ili "Let's think step by step ... to reach conclusions";
- Uključivanje strategije za složene odgovore koja omogućuje razbijanje rezultata na manje dijelove;
- Dodavanje primjera za izgled željenog odgovora;

- Razmatranje upita u kontekstu računalnog programiranja te izbjegavanje fraza poput “Please, give me...”.

Ostvarenje zadanog cilja uvijek započinje osnovnim *promptom* koji definira opću ideju zadatka, nakon čega se provjeravaju dobiveni rezultati. Ako rezultati nisu zadovoljavajući, *prompt* se finije podešava. Ovaj proces se ponavlja iterativno sve dok se ne postigne željeni odgovor.

5.2. RAG

RAG (engl. *Retrieval-Augmented Generation*) je tehnika koja se koristi za pronalaženje najrelevantnijih informacija iz vanjskih podataka, koje je primjerice unio korisnik te umetanja istih u upit prije nego što se on pošalje velikom jezičnom modelu. Model prilikom generiranja odgovora koristi ne samo svoje vlastito znanje, već i dodatne kontekstualne podatke. RAG spaja najbolje od oba svijeta: sposobnost LLM-a da razumije i generira tekst s mogućnošću pristupa najnovijim i najrelevantnijim vanjskim informacijama, čime se osiguravaju točni, kontekstualni i aktualni odgovori [23]. RAG pristup sastoji se od tri ključna koraka: indeksiranje (engl. *indexing*), dohvaćanje (engl. *retrieval*) i generiranje (engl. *generation*).



Slika 5.2. RAG sustav [24]

Prvi korak indeksiranja započinje čišćenjem podataka. Korisnik može učitati datoteke različitih formata poput PDF-a, Worda ili Markdowna koje je potrebno pretvoriti u običan “sirov” tekst i sačuvati samo esencijalni sadržaj. U ovom koraku korisno je ukloniti sve nepotrebne elemente koji mogu otežati kasnije pretraživanje poput višestrukih razmaka, posebnih znakova, simbola i slično. Nakon čišćenja slijedi segmentacija teksta na manje dijelove (engl. *chunks*) poput rečenica ili odlomaka, ovisno o količini teksta

koju model može obraditi. Svaki segment pretvara se u vektorsku reprezentaciju te se dobiveni vektori spremaju u vektorsku bazu i čuvaju za kasnije pretraživanje.

Dohvaćanje započinje korisnikovim upitom u obliku prirodnog jezika u kojem on traži informaciju, odnosno odgovor na svoje pitanje. Upit se pretvara u vektorski oblik pomoću istog modela korištenog tijekom faze indeksiranja. Korištenjem kosinusne sličnosti ili neke druge metrike izračunava se koliko je vektor upita sličan vektoru svakog segmenta unutar vektorske baze podataka. Ovaj korak identificira segmente koji su najrelevantniji za korisnikov upit, odnosno koji s najvećom vjerojatnosti sadrže odgovor na postavljeno pitanje. Segmenti se rangiraju po izračunatoj sličnosti i dohvaća se prvih k segmenata, gdje je k unaprijed zadan parametar.

Korak generiranja prvo uključuje pripremu *prompta*, koji povezuje korisnikov upit i pronađeni kontekst. Primjerice, *prompt* može izgledati ovako:

“Context information is provided here: {context}. Using only the provided context information and without relying on any prior knowledge or inventing additional information, answer the query accurately and concisely. If the information needed to answer the query is not present in the context, state that you have no answer. Query: {query}”.

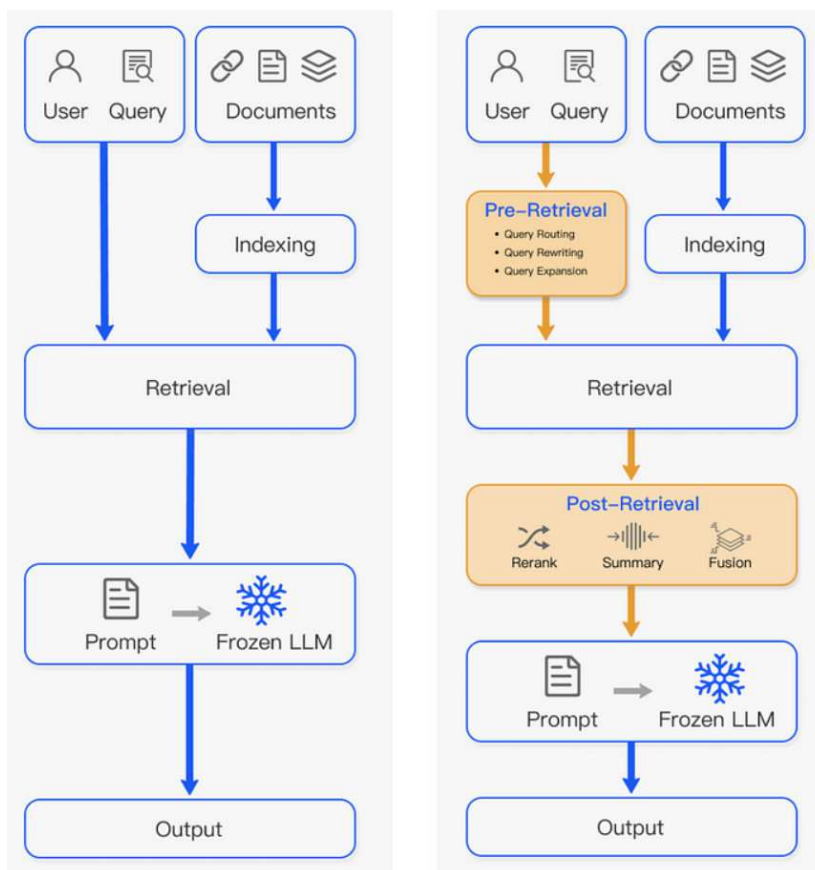
Na temelju upita i prosljeđenog konteksta, LLM generira svoj odgovor.

5.3. Napredni RAG

Napredni RAG je poboljšanje osnovnog RAG pristupa uključivanjem dodatnih tehnika optimizacije u proces indeksiranja, pretraživanja i generiranja odgovora. Postoji nebrojeno različitih metoda optimizacije i svakim danom “izmišljaju se” nove, najčešće kombinacijom postojećih metoda i pukim eksperimentiranjem [25]. Ovo poglavlje donosi pregled najznačajnijih tehnika od kojih su neke i upotrijebljene u ovom radu.

Prije samog dohvaćanja teksta iz vektorske baze, cilj je optimizirati početni upit korisnika te sam proces indeksiranja. RAG tehnika oslanja se na zadani upit koji, ako je loše napisan, može uvelike utjecati na točnost odgovora. Ukoliko je upit prejednostavan može ga se proširiti dodatnim kontekstom i sinonimima ili obogatiti primjerima. S druge strane, ukoliko je upit presložen potrebno ga je razložiti na niz jednostavnijih dijelova i obraditi zasebno, a zatim sintetizirati rezultate u jedan koherentan odgovor. Također, ako je upit nejasan, dvosmislen ili naprosto pogrešan potrebno ga je preformulirati.

Segmentiranje teksta na manje dijelove ključno je za poboljšanje preciznosti pretra-



Slika 5.3. Usporedba RAG-a i naprednog RAG-a [26]

živanja i izbjegavanje problema s duljinom teksta koju LLM-ovi mogu dobiti na ulazu. Proces segmentacije može se primijeniti na različitim razinama granularnosti poput *tokena*, rečenice ili semantičkih cjelina. U najjednostavnijoj varijanti, tekst se može podijeliti na segmente fiksne veličine s malim preklapanjem između segmenata. Iako se kontekst donekle prenosi zbog preklapanja, ovakav pristup može razbiti semantičke cjeline i pogoršati rezultat pretraživanja. Suprotno tome, tekst se može podijeliti na temelju značenja osiguravajući da svaki segment predstavlja smislen dio informacije. Ovaj pristup pruža očuvanje konteksta i koherentnosti, ali je potrebno provući tekst kroz LLM, koji će odrediti prijelomne točke, što konzumira značajno više vremena. Također, tekst se može podijeliti na semantičke cjeline poput rečenica ili odlomaka. Ovaj pristup ne zahtijeva korištenje LLM-a jer se granice mogu jednostavno odrediti te predstavlja kompromis između lake obrade i očuvanja značenja. Veličina segmenta značajno utječe na performanse. Veći segmenti pružaju više informacija što poboljšava razumijevanje, ali je to skuplje i povećava vrijeme obrade. Manji segmenti poboljšavaju preciznost pretraživanja, ali im može nedostajati dovoljno konteksta. Pokazalo se da segmenti veličine

između 256 i 512 *tokena* daju najbolje rezultate [25].

Prilikom indeksiranja podataka, uz tekst koji je pretvoren u vektore, mogu se dodati korisni metapodaci koji pružaju dodatan kontekst. Primjerice, ukoliko se radi o medicinskim podacima, to mogu biti dob pacijenta, datum posjeta ili vrsta bolesti. Kada se podaci pretražuju, metapodaci pomažu filtrirati rezultate.

Pretraživanje podataka oslanja se na model kojim se tekst ugrađuje u vektorski prostor. Pretrenirani modeli su solidna osnova za pretraživanje, no moguće ih je dodatno trenirati na vlastitim podacima kako bi model bolje razumio specifične odnose i semantiku. Vektore treba pohraniti u prikladnu bazu podataka, a kod odabira baze mogu se gledati kriteriji poput vrste indeksiranja, podrške za velike vektore, vrste pretraživanja i mogućnosti pohrane u oblaku. Prilikom pretraživanja mogu se koristiti tri metode:

- Vektorsko pretraživanje (engl. *vector search*): Rezultati pretrage biti će isključivo semantički povezani sa značenjem upita.
- Pretraživanje po ključnim riječima (engl. *keyword search*): Osigurava se točna podudarnost za specifične pojmove što može biti značajno kod medicinskih termina.
- Hibridno pretraživanje (engl. *hybrid search*). Kombinira obje vrste pretraživanja, a rezultati se težinski sumiraju.

Nakon pretraživanja može se upotrijebiti tehnika ponovnog rangiranja (engl. *re-ranking*) koja koristi dodatan, često sporiji model, za dodatno procjenjivanje dohvaćenih segmenata. Dakle, relevantni segmenti dohvaćaju se brzim modelom, ali se dio informacija gubi zbog procesa kompresije dokumenata u vektore. Zatim se koristi napredniji model, poznatiji kao unakrsni koder (engl. *cross-encoder*), za procjenjivanje sličnosti između upita i svakog dohvaćenog segmenta. Segmenti se ponovno sortiraju kako bi oni najrelevantniji bili na vrhu.

Rezultati pretraživanja mogu sadržavati suvišne informacije pa ih je po potrebi moguće sažeti. Primjerice, dohvaćeni tekst može se segmentirati u rečenice i svakoj rečenici dodijeliti ocjena, a zatim se odabiru najbolje ocijenjene rečenice i stavljaju u sažetak koji zadržava ključne informacije iz izvornog teksta.

Obično nakon koraka pretraživanja informacija slijedi generiranje odgovora, no moguće je koristiti i složenije pristupe koji pretražuju u nekoliko koraka:

- Iterativno preuzimanje (engl. *iterative retrieval*): Proces u kojem se vektorska baza više puta pretražuje na temelju početnog upita i dosad generiranog teksta.

- Rekurzivno preuzimanje (engl. *recursive retrieval*): Prvo se preuzmu početni segmenti na temelju zadanog upita. Na temelju rezultata prvog preuzimanja, generiraju se novi upiti za daljnje pretraživanje, a zatim se ponovno preuzimaju dodatni segmenti teksta.
- Adaptivno preuzimanje (engl. *adaptive retrieval*): Koriste se povratne informacije iz ranijih faza pretraživanja kako bi se prilagodile metode i strategije pretraživanja. Ova prilagodba može uključivati promjenu ključnih riječi, filtriranje ili čak promjenu izvora podataka.

6. Implementacija

Medicinski asistent razvijen je u Pythonu koristeći Streamlit biblioteku i FastAPI radni okvir (engl. *framework*). Streamlit je *open-source*, besplatna Python biblioteka koja omogućuje brzu izradu interaktivnih komponenti, posebice onih za vizualizaciju podataka, što ga čini pogodnim za projekte iz područja strojnog učenja i znanosti o podacima. Zbog jednostavnosti korištenja, smanjena je fleksibilnost u dizajnu korisničkog sučelja u usporedbi s tradicionalnim radnim okvirima poput Reacta. Komponente je moguće dodatno stilizirati po želji, ali je prilagodba ograničena. S obzirom da se ovaj rad ne fokusira na izradu web sučelja, Streamlit je prikladan izbor jer omogućuje puno bržu implementaciju. FastAPI je relativno nov, brz radni okvir za izgradnju *backend* dijela aplikacije. FastAPI nudi sjajne performanse, intuitivan je i potkrijepljen dobrom dokumentacijom. Integracija ovih dviju tehnologija omogućuje brz razvoj robusnih aplikacija [27].

Korisnikovi projekti pohranjeni su u PostgreSQL relacijsku bazu podataka, a za pristup podacima koristi se Tortoise ORM (engl. *Object-Relational Mapping*). MRI slike pohranjene su u AWS S3 (engl. *Amazon Simple Storage Service*) spremnicima, a putanja do svake slike spremljena je uz odgovarajući projekt u bazu podataka. Osim “obične” baze podataka, medicinski asistent koristi i vektorsku bazu u koju se spremaju tekstovi svih znanstvenih članaka koje je korisnik odabrao za daljnje pretraživanje. Kao vektorska baza odabrana je *open-source* Weaviate baza koja ima ugrađene sve vrste pretraživanja - vektorsko, hibridno i pretraživanje po ključnim riječima. Weaviate se fokusira na sigurnost i zaštitu podataka te laku integraciju s različitim *embedding* modelima pomoću kojih se podaci pretvaraju u vektore [28].

Korisnik kroz *chat* komunicira s velikim jezičnim modelom ChatGPT-4o koji može primiti 128,000 *tokena* (oko 96,000 riječi) na ulazu. Ovaj model ima puno više parametara i sposobnosti da rješava složenije zadatke od njegovog prethodnika ChatGPT-3.5, a nije pretjerano skuplji [29]. Model je kroz različite faze projekta vođen odgovarajućim *promptovima*. Primjerice, u prvoj fazi, kada korisnik tek kreće u kreiranje projekta,

model ima zadatak da prikupi sve potrebne informacije za kreiranje projekta pa njegov *prompt* može izgledati ovako:

"You are a medical expert, tasked with collecting data for MRI segmentation for a patient. You will be speaking with a user to gather the necessary information about the patient. You need to collect the following details: project name, age, gender (male or female) and clinical notes (optional). Ask the user specific questions to gather each piece of information. Ensure all the required information is collected. When you have all information needed, you will return a JSON object with the following design:

```
"project_name": <type:string, collected project name>,
"age": <type:number, collected age>,
"gender": <type:string, collected gender>,
"clinical_notes": <type:string, collected clinical notes>"
```

Što se tiče segmentacijskog dijela, korištena je spomenuta samo-konfigurirajuća nnU-net mreža s mogućnosti odabira između 2D ili 3D arhitekture. Mreža je pretrenirana te se korisniku nudi izbor između tri različita skupa podataka: tumori na mozgu, lijevi atrij na srcu i hipokampus. Ovi skupovi podataka odabrani su za početnu MVP (engl. *Minimum Viable Product*) verziju aplikacije zbog različitih značajki. Ovim se demonstrira sposobnost mreže da se nosi s različitim vrstama podataka, dok je u budućnosti planirano dodavanje novih. Također, predviđeno je proširenje funkcionalnosti aplikacije koje uključuje mogućnost da korisnici učitaju vlastite skupove podataka i treniraju mrežu prema specifičnim potrebama.

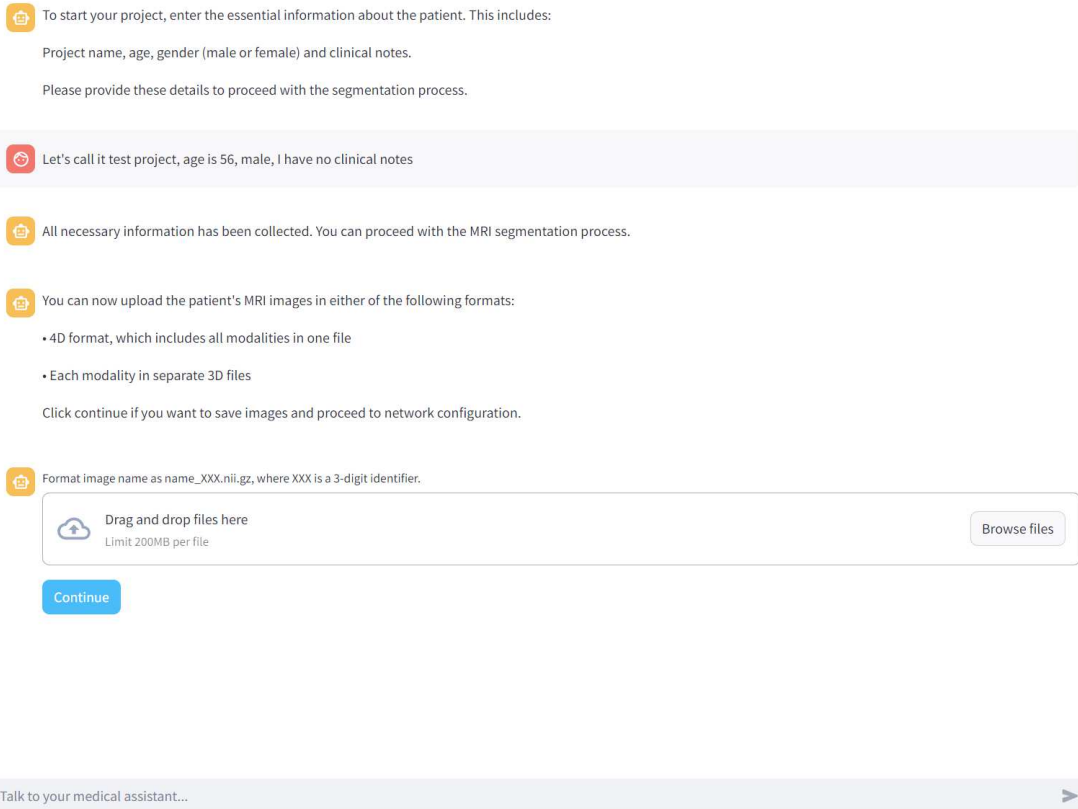
Članci se dohvaćaju iz *open-source* baze podataka OpenAlex koja uključuje velik broj istraživačkih radova i znanstvenih publikacija ¹. Neki članci nalaze se iza zatvorenog pristupa pa njihov tekst nije moguće dohvatiti. Ovdje je također planirano proširenje aplikacije tako da korisnik može sam učitati PDF dokument članka. Željeni članci ugrađuju se u vektore pomoću modela *text-embedding-ada-002*. Za pretraživanje članaka koristi se napredni RAG sustav s hibridnom metodom pretraživanja koja uključuje 50% vektorsko pretraživanje i 50% pretraživanje po ključnim riječima. Rezultati pretrage su ponovno rangirani pomoću modela *rerank-multilingual-v3.0*.

Korisničko sučelje sastoji se od bočnog izbornika i glavnog dijela aplikacije. Moguće je odabrati pregled dosadašnjih projekata ili pokrenuti novi. Prilikom pokretanja novog

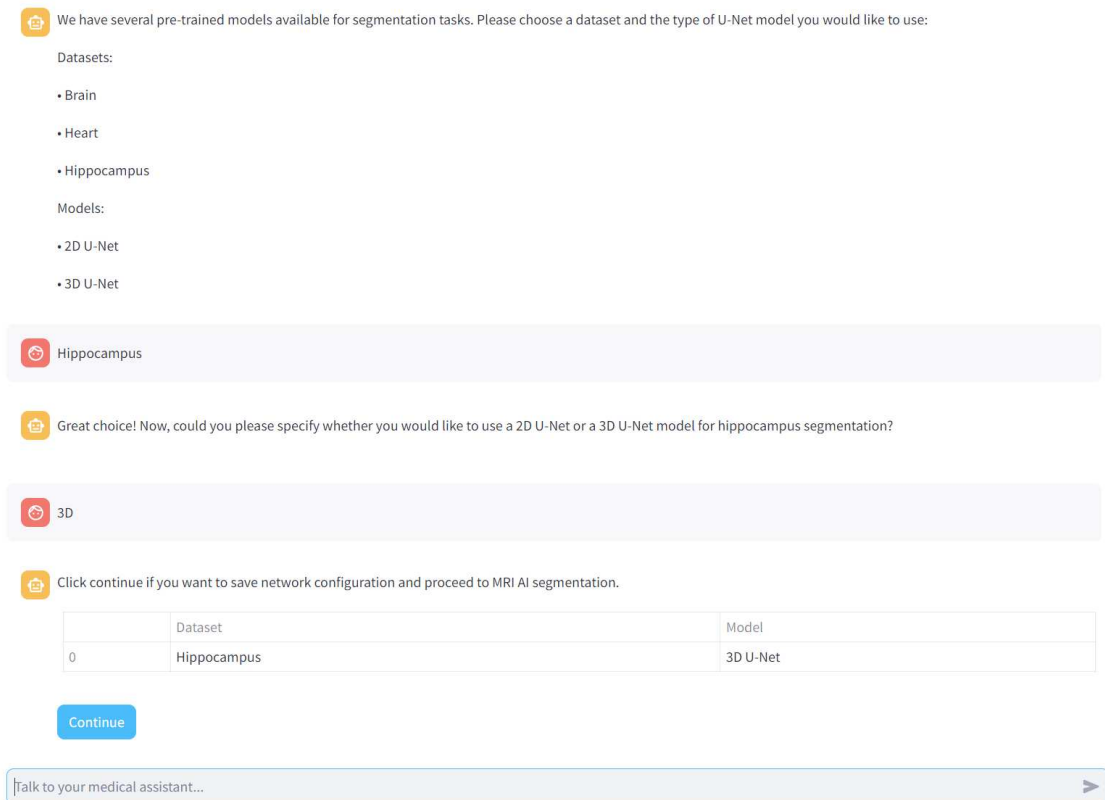
¹<https://openalex.org/>

projekta, sva komunikacija korisnika s aplikacijom odvija se kroz *chat*, a korisnik može pratiti u kojem je statusu projekt. Projekt prolazi kroz sljedeće faze:

- Inicijalizacija projekta: Agent od korisnika traži da unese potrebne podatke za stvaranje novog projekta - ime projekta, dob i spol osobe za koju se vrši segmentacija, opcionalne dodatne bilješke i učitavanje MRI datoteke u NIfTI formatu. U aplikaciji je očuvana privatnost podataka pa se ne pohranjuju osobne informacije koje bi mogle omogućiti identifikaciju osobe. Kada su svi podaci uneseni, agent od korisnika traži da potvrdi ispravnost podataka kako bi ih mogao unijeti u bazu podataka (slika 6.1.).
- Odabir skupa podataka i arhitekture mreže: Agent od korisnika traži da odabere skup podataka na kojem je mreža trenirana i arhitekturu mreže koju želi koristiti za segmentaciju (slika 6.2.).
- Segmentacija MRI slike i rezultati: Agent pokreće segmentaciju učitane MRI slike. Ovo je proces koji traje 40-60 sekundi te korisnik može pratiti u kojoj je fazi. Nakon završetka segmentacije, korisniku se prikazuje segmentirana maska koja sadrži izdvojene regije od interesa. Također, prikazuje se točnost segmentacije izražena pomoću Dice koeficijenta (slika 6.3.).
- Zadavanje ključnih riječi i pretraga članaka: Agent od korisnika traži da unese ključne riječi koje će se koristiti za pretraživanje baze znanstvenih članaka OpenAlex. Agent prikazuje listu pronađenih članaka uključujući naslove i sažetke. Korisnik može pregledati listu i odabrati članke koje želi zadržati. Kada korisnik napravi svoj izbor, agent započinje proces dohvaćanja cijelog teksta odabranih članaka. Ovaj proces može trajati neko vrijeme, ovisno o broju odabranih članaka i dostupnosti teksta. Nakon što su tekstovi dohvaćeni, segmentiraju se na manje dijelove i pohranjuju u vektorsku bazu podataka za daljnju analizu i pretraživanje (slika 6.4.).
- Postavljanje upita i generiranje odgovora: Kada korisnik postavi pitanje agentu, agent koristi prethodno dohvaćene i pohranjene članke kako bi generirao odgovor. Na temelju korisnikovog upita, agent pretražuje vektorsku bazu podataka i pronalazi najprikladnije podatke. Generirani odgovor temelji se na analizi informacija iz relevantnih članaka, pružajući korisniku precizne i točne odgovore na postavljeno pitanje (slika 6.5.).



Slika 6.1. Inicijalizacija projekta



Slika 6.2. Odabir skupa podataka i arhitekture mreže

Please wait while I processes the data and generate the segmented images.

You can keep track of the progress on the progress bar.

You will be notified once the process is complete.

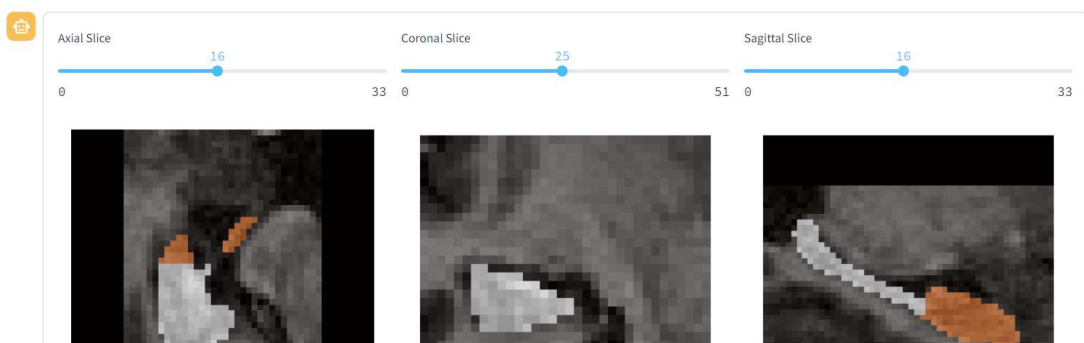
The AI segmentation process is complete.

Here are the results of the segmentation:

- Segmented images
- DC coefficients

Please review the results and click continue.

	Class ID	Class Name	Dice coefficient (Post processed)	Dice coefficient (Raw)
0	1	Anterior	0.8789	0.8787
1	2	Posterior	0.8592	0.8595



Slika 6.3. Prikaz rezultata segmentacije hipokampusa

To help you with your project, please enter keywords related to your topic.

I will use these keywords to search for relevant research papers. Based on the findings from these papers,

I will provide you with detailed and well-informed answers to your questions.

Hm... something about hippocampus

Great! The hippocampus is a broad topic. Could you specify what aspect of the hippocampus you are interested in? For example, are you looking into its role in memory, its structure, its involvement in diseases, or something else?

Let's search for hippocampus and schizophrenia

I have collected the keywords 'hippocampus' and 'schizophrenia'. If you have any additional keywords or specific aspects you want to explore, please let me know!

I found several articles based on your keywords. Please review the list and select which articles you would like to include in the search results and click continue.


title	doi	include
Neuroimaging studies of the hippocampus in schizophrenia	https://doi.org/10.1002/hipo.1068	<input type="checkbox"/>
The hippocampus in schizophrenia: a review of the neuropathological evidence and its pathophysiological implications	https://doi.org/10.1007/s00213-003-1761-y	<input type="checkbox"/>
A selective decrease in the relative density of parvalbumin-immunoreactive neurons in the hippocampus in schizophrenia	https://doi.org/10.1016/s0920-9964(01)00188-8	<input type="checkbox"/>
Hippocampus-amygdala volumes and psychopathology in chronic schizophrenia	https://doi.org/10.1016/0006-3223(93)90289-p	<input checked="" type="checkbox"/>
Reduction in Reelin immunoreactivity in hippocampus of subjects with schizophrenia, bipolar disorder and major depression	https://doi.org/10.1038/sj.mp.4000783	<input type="checkbox"/>
Altered resting-state functional connectivity and anatomical connectivity of hippocampus in schizophrenia	https://doi.org/10.1016/j.schres.2007.11.039	<input type="checkbox"/>
Boundary and medial shape analysis of the hippocampus in schizophrenia	https://doi.org/10.1016/j.media.2004.06.004	<input checked="" type="checkbox"/>
Probing the human hippocampus using rCBF: Contrasts in schizophrenia	https://doi.org/10.1002/hipo.1070	<input type="checkbox"/>
Schizophrenia and the Hippocampus: The Embryological Hypothesis Extended	https://doi.org/10.1093/schbul/13.4.577	<input type="checkbox"/>

Slika 6.4. Zadavanje ključnih riječi i pretraga članaka

 Your articles have been successfully collected and indexed.

You can now proceed to ask a question related to your project.

Please enter your query, and I will provide you with relevant information based on the collected articles.


 Compare volume of hippocampus in normal people and patients with schizophrenia

 Based on article **Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium**, my answer is:

Individuals with schizophrenia have significantly smaller hippocampus volumes compared to healthy controls. The mean difference from the control mean is -4.10%, with an effect size of Cohen's $d = -0.46$.

 Based on article **Abnormal asymmetries in subcortical brain volume in schizophrenia**, my answer is:

Patients with schizophrenia demonstrated smaller bilateral hippocampus volumes compared with healthy controls.


 On what sample size is this conclusion based?

 Based on article **Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium**, my answer is:

The conclusion is based on a sample size of 2028 individuals with schizophrenia and 2540 healthy controls.

 Based on article **Abnormal asymmetries in subcortical brain volume in schizophrenia**, my answer is:

The conclusion is based on a sample size of 1680 healthy individuals and 884 patients with schizophrenia from 15 protocols.

Talk to your medical assistant... 

Slika 6.5. Postavljanje upita i generiranje odgovora na temelju odabranih članaka

7. Zaključak

Ovaj rad demonstrira kako se postojeća AI rješenja mogu uspješno kombinirati i integrirati u jedinstven sustav kako bi se stvorio učinkovit alat za medicinsku dijagnostiku. Tijekom razvoja virtualnog medicinskog asistenta prevladani su brojni izazovi:

- Pronalaženje skupova podataka s velikom varijabilnosti: Identifikacija skupova podataka s dovoljnom varijabilnosti kako bi se pokazala prilagodljivost segmentacijskog modela.
- Odabir arhitekture segmentacijske mreže: Pronalazak arhitekture koja se prilagođava različitim vrstama podataka.
- Prevladavanje halucinacija velikih jezičnih modela: Upotreba različitih tehnika inženjeringa upita koje ne zahtijevaju ponovno treniranje modela i trošenje resursa.
- Integracija s vanjskom bazom podataka: Pronalazak adekvatne baze znanstvenih članaka s mogućnosti besplatne integracije.
- Unapređenje RAG sustava: Upotreba naprednih tehnika i metoda poput hibridnog pretraživanja i ponovnog rangiranja kako bi se unaprijedio sustav pretraživanja članaka i odgovaranja na korisnikove upite.
- Implementacija intuitivnog korisničkog sučelja: Razvijanje intuitivnog *chat* sučelja koje omogućuje jednostavnu interakciju s asistentom.
- Upotreba različitih tehnologija: Korištenje pouzdanih i dobro dokumentiranih alata za učinkovito rješavanje složenih zadataka bez potrebe za poznavanjem sitnih implementacijskih detalja.

Za budući rad preporučuje se dodatna evaluacija ključnih komponenti sustava, posebno točnosti segmentacije i učinkovitosti korištenih RAG tehnika. Postoji mogućnost proširenja skupova podataka na druge modalitete, što bi omogućilo veću fleksibilnost i primjenjivost sustava u različitim dijagnostičkim scenarijima. U budućnosti, platforma bi mogla biti proširena tako da omogućí korisnicima treniranje i segmentaciju pomoću vlastitih skupova podataka. Također, korisnik bi mogao graditi vlastitu bazu znanstve-

nih članaka na kojima bi LLM-ovi mogli učiti te se prilagođavati specifičnim zahtjevima korisnika.

Literatura

- [1] S. Gaj, D. Ontaneda, i K. Nakamura, “Automatic segmentation of gadolinium-enhancing lesions in multiple sclerosis using deep learning from clinical MRI”, *PloS one*, sv. 16, br. 9, str. e0255939, 2021.
- [2] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, “A survey of large language models”, *arXiv preprint arXiv:2303.18223*, 2023.
- [3] A. S. Tulshan i S. N. Dhage, “Survey on virtual assistant: Google Assistant, Siri, Cortana, Alexa”, u *Advances in Signal Processing and Intelligent Recognition Systems: 4th International Symposium SIRS 2018, Bangalore, India, September 19–22, 2018, Revised Selected Papers*, sv. 4. Springer Singapore, 2019. <https://doi.org/10.1007/978-981-13-5758-9>
- [4] K. A. Kumar, J. F. Rajan, C. Appala, S. Balurgi, i P. R. Balaiahgari, “Medibot: Personal medical assistant”, u *2024 2nd International Conference on Networking and Communications (ICNWC)*. IEEE, 2024., str. 1–6.
- [5] Upwork, “AI chatbot vs virtual assistant: What’s the difference?” <https://www.upwork.com/resources/ai-chatbot-vs-virtual-assistant>, 2024., pristupljeno: 28. kolovoza 2024.
- [6] E. Moser, A. Stadlbauer, C. Windischberger, H. H. Quick, i M. E. Ladd, “Magnetic resonance imaging methodology”, *European journal of nuclear medicine and molecular imaging*, sv. 36, str. 30–41, 2009.
- [7] S. D. Serai, “Basics of magnetic resonance imaging and quantitative parameters T1, T2, T2*, T1rho and diffusion-weighted imaging”, *Pediatric radiology*, sv. 52, br. 2, str. 217–227, 2022.
- [8] E.-S. H. Ibrahim i R. E. Gabr, “MRI basics”, *Heart Mechanics: Magnetic Resonance Imaging-Mathematical Modeling, Pulse Sequences, and Image Analysis*, str. 81–120, 2017.

- [9] I. Despotović, B. Goossens, i W. Philips, “Review article: MRI segmentation of the human brain: challenges, methods, and applications”, *Computational and Mathematical Methods in Medicine, Hindawi*, str. 25, 2015.
- [10] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, “A large annotated medical image dataset for the development and evaluation of segmentation algorithms”, *arXiv preprint arXiv:1902.09063*, 2019.
- [11] N. Siddique, S. Paheding, C. P. Elkin, i V. Devabhaktuni, “U-net and its variants for medical image segmentation: A review of theory and applications”, *IEEE access*, sv. 9, str. 82 031–82 057, 2021.
- [12] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, M. Bilello, P. Bilic, P. F. Christ, R. K. G. Do, M. J. Gollub, S. H. Heckers, H. Huisman, W. R. Jarnagin, M. K. McHugo, S. Napel, J. S. G. Pernicka, K. Rhode, C. Tobon-Gomez, E. Vorontsov, J. A. Meakin, S. Ourselin, M. Wiesenfarth, P. Arbeláez, B. Bae, S. Chen, L. Daza, J. Feng, B. He, F. Isensee, Y. Ji, F. Jia, I. Kim, K. Maier-Hein, D. Merhof, A. Pai, B. Park, M. Perslev, R. Rezaiifar, O. Rippel, I. Sarasua, W. Shen, J. Son, C. Wachinger, L. Wang, Y. Wang, Y. Xia, D. Xu, Z. Xu, Y. Zheng, A. L. Simpson, L. Maier-Hein, i M. J. Cardoso, “The Medical Segmentation Decathlon”, *Nat Commun*, sv. 13, str. 4128, srpanj 2022. <https://doi.org/10.1038/s41467-022-30695-9>
- [13] F. Isensee, P. F. Jaeger, P. J. Jäger, S. A. A. Kohl, J. Petersen, i K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”, *Nature Methods*, sv. 18, str. 203–211, veljača 2021. <https://doi.org/10.1038/s41592-020-01008-z>
- [14] W. Baccouch, S. Oueslati, B. Solaiman, i S. Labidi, “A comparative study of CNN and U-Net performance for automatic segmentation of medical images: Application to cardiac MRI”, *Procedia Computer Science*, sv. 219, str. 1089–1096, 2023.
- [15] P. J. R. Prasad, O. J. Elle, F. Lindseth, F. Albregtsen, i R. P. Kumar, “Modifying U-Net for small dataset: a simplified U-Net version for liver parenchyma segmentation”, u *Medical Imaging 2021: Computer-Aided Diagnosis*, sv. 11597. SPIE, 2021., str. 396–405.

- [16] M. D. S. Team, “How large language models work”, <https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f>, 2024., pristupljeno: 19. kolovoza 2024.
- [17] R. P. Menon, “Introduction to large language models and the transformer architecture”, <https://rpradeepmenon.medium.com/introduction-to-large-language-models-and-the-transformer-architecture-534408ed7e61>, 2024., pristupljeno: 19. kolovoza 2024.
- [18] T. Gumption, “Transformer architecture simplified”, <https://medium.com/@tech-gumptions/transformer-architecture-simplified-3fb501d461c8>, 2024., pristupljeno: 1. rujna 2024.
- [19] M. Siino i I. Tinnirello, “GPT hallucination detection through prompt engineering”, *Working Notes of CLEF*, 2024.
- [20] G. Perković, A. Drobnjak, i I. Botički, “Hallucinations in LLMs: Understanding and addressing challenges”, u *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*. IEEE, 2024., str. 2084–2088.
- [21] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, i A. Chadha, “A systematic survey of prompt engineering in large language models: Techniques and applications”, *arXiv preprint arXiv:2402.07927*, 2024.
- [22] J. D. Velásquez-Henao, C. J. Franco-Cardona, i L. Cadavid-Higuaita, “Prompt engineering: A methodology for optimizing interactions with AI-Language Models in the field of engineering”, *Dyna*, sv. 90, br. 230, str. 9–17, 2023.
- [23] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, i B. Cui, “Retrieval-augmented generation for AI-generated content: A survey”, *arXiv preprint arXiv:2402.19473*, 2024.
- [24] Koyeb, “What is RAG? Retrieval-augmented generation for AI”, <https://www.koyeb.com/blog/what-is-rag-retrieval-augmented-generation-for-ai>, 2024., pristupljeno: 1. rujna 2024.
- [25] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian *et al.*, “Searching for best practices in retrieval-augmented generation”, *arXiv preprint arXiv:2407.01219*, 2024.
- [26] J.-F. Rezkallah, “A deep dive into retrieval-augmented generation”, https://medium.com/@jf_rezkallah/a-deep-dive-into-retrieval-augmented-generation-

8ecff8f9093f, 2024., pristupljeno: 5. rujna 2024.

- [27] S. Jain, “Streamlit + fastapi: The ingredients you need for your next data science recipe”, 2022., pristupljeno: : 25. kolovoza 2024. [Mrežno]. Adresa: <https://medium.com/codex/streamlit-fastapi-%EF%B8%8F-the-ingredients-you-need-for-your-next-data-science-recipe-ffbeb5f76a92>
- [28] Weaviate, “Weaviate: Open-source vector database”, 2024., pristupljeno: : 25. kolovoza 2024. [Mrežno]. Adresa: <https://weaviate.io/>
- [29] Talkative, “GPT-4 vs GPT-3.5: What’s the difference?” 2024., pristupljeno: : 25. kolovoza 2024. [Mrežno]. Adresa: <https://gettalkative.com/info/gpt-4-vs-gpt-3-5>

Sažetak

Segmentacija slika magnetske rezonancije i dijagnostika pomoću virtualnog medicinskog asistenta

Mia Jurdana

U ovom radu razvijen je virtualni medicinski asistent za automatiziranu segmentaciju organa i promjena u tkivu iz slika magnetske rezonancije. Korisnici mogu jednostavno komunicirati s asistentom putem intuitivnog *chat* sučelja. Istražuje se primjena pretrenirane samo-konfigurirajuće U-net mreže za segmentaciju na raznolikim podacima koji variraju u svojim značajkama poput broja vokseli, razmaka među vokselima, broja primjera u skupu za treniranje i broja klasa. Dodatno, asistent koristi napredne tehnike pretraživanja vanjske baze znanstvenih članaka za odgovaranje na postavljena pitanja, čime se dodatno unapređuje preciznost i korisnost njegovih odgovora i smanjuje mogućnost pojave halucinacija. Fokus rada je na integraciji postojećih tehnologija umjetne inteligencije kako bi se stvorio učinkovit sustav za medicinsku dijagnostiku.

Ključne riječi: U-net mreža za segmentaciju; MR segmentacija; virtualni medicinski asistent; analiza medicinskih slika; RAG;

Abstract

Magnetic resonance images segmentation and diagnostics using a virtual medical assistant

Mia Jurdana

This work presents the development of a virtual medical assistant designed for the automated segmentation of organs and tissue changes from magnetic resonance scans. The assistant allows users to interact through an intuitive chat interface, enabling seamless communication and access to advanced diagnostic tools. It utilizes a pretrained self-configuring U-net model for segmentation, capable of adapting to diverse datasets with varying characteristics such as voxel count, voxel spacing, training set size, and the number of classes. Additionally, the assistant uses advanced retrieval-augmented generation techniques to search external scientific literature databases, providing precise and reliable responses to medical and diagnostic inquiries, while minimizing the risk of misinformation or hallucinations. The focus of the work is on integrating existing artificial intelligence technologies to create an effective system for medical diagnostics.

Keywords: U-net segmentation network; MRI segmentation; virtual medical assistant; medical image analysis; RAG;