

Razvoj algoritma za detekciju Gaucherove bolesti

Grebenar, Lea

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://um.nsk.hr/um:nbn:hr:168:341212>

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. 592

**RAZVOJ ALGORITMA ZA DETEKCIJU GAUCHEROVE
BOLESTI**

Lea Grebenar

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 592

**RAZVOJ ALGORITMA ZA DETEKCIJU GAUCHEROVE
BOLESTI**

Lea Grebenar

Zagreb, lipanj 2024.

DIPLOMSKI ZADATAK br. 592

Pristupnica: **Lea Grebenar (0036522488)**

Studij: Računarstvo

Profil: Znanost o podacima

Mentorica: izv. prof. dr. sc. Mihaela Vranić

Zadatak: **Razvoj algoritma za detekciju Gaucherove bolesti**

Opis zadatka:

Rijetke bolesti definirane su kao bolesti od kojih boluje do 5 osoba u 10 000 ljudi. One mogu biti opasne po život ili smanjiti kvalitetu života. Veliki problem kod ovakvih bolesti je teško i dugotrajno postavljanje dijagnoze što smanjuje vjerojatnost uspješnog liječenja. Ovaj problem prisutan je i kod dijagnostike Gaucherove bolesti koja je autosomno recesivna bolest koju karakteriziraju snižene vrijednosti enzima glukocerebrosidaze u lizosomima. Kod ove bolesti smanjeni su očekivani životni vijek i kvaliteta života, a kasna dijagnostika može dovesti do nepovratnih zdravstvenih posljedica. Pretpostavlja se da je u Hrvatskoj Gaucherova bolest dijagnosticirana u tek 34 % slučajeva. Razlozi tome su manjak znanja o ovoj bolesti i malen broj doktora koji su stručni u ovom području. Vaš je zadatak razviti model koji će liječnicima olakšati ispravnu dijagnozu Gaucherove bolesti tako da će između njihovih pacijenata izdvojiti one za koje bi bilo vjerojatnije da boluju od ove bolesti. Time će testiranja na ovu bolest biti bolje usmjerena te se očekuje povećanje vjerojatnosti detektiranja ove bolesti. Pri razvoju modela trebate se osloniti na postojeće studije o navedenoj bolesti te razraditi neke od metoda pripreme medicinskih podataka kako bi se u konačnici razvio što bolji model. U radu trebate razraditi i metodologiju validacije razvijenog modela uzimajući u obzir da će za konačnu validaciju biti potrebno dulje vrijeme te da se primjenom dobrog modela može očekivati relativno mali broj pozitivno dijagnosticiranih bolesnika. Istraživanja i razvoj modela provodit ćete u suradnji s globalnom biofarmaceutskom kompanijom Takeda.

Rok za predaju rada: 28. lipnja 2024.

Sadržaj

Uvod	1
1. Motivacija za izradu diplomskog rada.....	3
2. Metodologija.....	5
2.1. Sudionici projekta.....	5
2.2. Zaštita podataka	6
2.3. Metodologija rada u užem smislu.....	6
2.4. Validacija algoritma i podešavanje algoritma	10
3. Gaucherova bolest	11
3.1. Rijetke bolesti	11
3.2. Općenito o Gaucherovoj bolesti	11
3.3. Uzroci Gaucherove bolesti	12
3.4. Dijagnosticiranje Gaucherove bolesti.....	13
3.5. Liječenje Gaucherove bolesti	14
4. Izvori podataka	15
4.1. Prikupljanje podataka	16
4.2. Strukturirani podatci	17
4.3. Nestrukturirani podatci	20
5. Baze podataka.....	21
5.1. Baze podataka u medicini u Hrvatskoj	21
5.2. ER model djela baze podataka koji je relevantan za istraživanje.....	23
5.3. Relacijska shema djela baze relevantnog za istraživanje	29
5.4. Razrada – ER dijagram.....	30
6. Analiza i ekstrakcija podataka iz specijalističkih nalaza.....	31
6.1. Obrada prirodnog jezika	31
6.2. Specijalistički nalazi	34

6.3.	Pronalazak dijagnoza	34
7.	Model za detekciju.....	38
7.1.	Pregled rada po kojemu je napravljen algoritam	38
7.2.	Kriteriji	39
7.3.	Korišteni alati	42
7.4.	Implementacija	42
8.	Rasprava	46
8.1.	Izazovi s dostupnošću i obradom podataka	46
8.2.	Validacija i unaprjeđenje algoritma.....	46
8.3.	Izgradnja novih algoritama.....	47
	Zaključak	48
	Literatura	49
	Sažetak.....	52
	Summary.....	53
	Skraćenice.....	54

Uvod

Rijetke bolesti definirane su kao bolesti od kojih boluje do 5 osoba u 10 000 ljudi. One često predstavljaju ozbiljan problem jer mogu biti opasne po život ili uzrokovati značajno smanjenje kvalitete života. Veliki problem kod ovakve vrste bolesti je kompleksan i dugotrajan proces postavljanja ispravne dijagnoze. U prosjeku, potrebno je 5 godina kako bi se postavila ispravna dijagnoza, što smanjuje vjerojatnost uspješnog liječenja i donosi značajan rizik od nastajanja nepovratnih zdravstvenih posljedica [1].

Jedna od najčešćih rijetkih bolesti je Gaucherova bolest (*GB, eng. Gaucher disease - GD*). To je autosomno recesivna bolest koju karakteriziraju snižene vrijednosti enzima glukocerebrozidaze u lizosomima, zbog čega se glukocerebrozid nakuplja u različitim organima [2]. Posljedice ove bolesti su smanjeni očekivani životni vijek i smanjena kvaliteta života, a zakašnjelo uspostavljanje dijagnoze može dovesti do nastanka ozbiljnih i nepovratnih zdravstvenih posljedica. Gaucherova bolest prisutna je kod oko 1 od 50 000 osoba u svijetu. Pretpostavlja se da je u Hrvatskoj Gaucherova bolest dijagnosticirana u tek 34 % slučajeva [1]. Razlozi tako malom postotku postavljenih dijagnoza su manjak znanja o bolesti i nedovoljan broj specijalista koji su obučeni za detekciju Gaucherove bolesti.

Cilj ovog diplomskog rada, kako je navedeno u naslovu, je razviti algoritam za detekciju Gaucherove bolesti. Algoritam se razvija u sklopu „Gaucher screening“ projekta u suradnji s globalnom biofarmaceutskom kompanijom Takeda, MCS Grupom te Domom zdravlja Zagreb – Centar. Razvijeni algoritam temeljiti će se na kombinaciji domenskog znanja, znanstvenih publikacija i dostupnih podataka, s naglaskom na prepoznavanje rizičnih pacijenata za daljnje pretrage. Glavni cilj projekta je povećati dijagnostičku stopu i time povećati kvalitetu života oboljelih od Gaucherove bolesti u Republici Hrvatskoj. Uz to projekt uvelike doprinosi digitalizaciji hrvatskog zdravstvenog sustava i optimizaciji korištenja zdravstvenih resursa.

Implementirani algoritam koristit će dvije vrste podataka: strukturirane i nestrukturirane. Strukturirani podatci, poput laboratorijskih nalaza u tabličnom obliku, lako se obrađuju i analiziraju, dok nestrukturirani, poput specijalističkih nalaza koji sadrže slobodan tekst, predstavljaju izazov za analizu i obradu. Algoritam će korištenjem SQL upita moći

identificirati pacijente koji imaju određene simptome Gaucherove bolesti na bazi rezultata laboratorijskih nalaza, postavljenih dijagnoza i ključnih riječi u slobodnom tekstu.

U sklopu rada opisano je kako uz pomoć obrade prirodnog jezika pravilno pristupiti analizi i obradi specijalističkih nalaza.

Još jedna od ključnih stavki spomenutog projekta je zaštita privatnosti pacijenata. Kako se ne bi narušila usklađenost s Općom uredbom o zaštiti podataka (*eng. General Data Protection Regulation - GDPR*) pristup osobnim podacima omogućen je samo liječniku opće prakse, što je i do sada bio slučaj. Ostali sudionici u projektu će imati pristup samo agregiranim rezultatima potrebnim za validaciju algoritma. Za pisanje rada na temu razvoja opisanog algoritma dobiveno je odobrenje Etičkog povjerenstva Fakulteta elektrotehnike i računarstva Sveučilišta u Zagrebu.

Ovaj diplomski rad ima za cilj ne samo razvoj učinkovitog alata za detekciju Gaucherove bolesti, već i postavljanje temelja za buduće slične projekte koji bi mogli koristiti slične metode za prepoznavanje drugih, kako rijetkih, tako i ostalih bolesti. Time se očekuje dodatno unaprjeđenje zdravstvene skrbi i zdravstvenog sustava Republike Hrvatske.

Rad je strukturiran na sljedeći način: nakon uvoda prvo je predstavljena motivacija za izradu diplomskog rada, koji su sve benefiti za razvoj algoritma za detekciju Gaucherove bolesti. Zatim je opisana metodologija, tko sve sudjeluje u projektu te kako algoritam treba uklopiti u zdravstveni sustav. Nakon toga dan je opći pregled Gaucherove bolesti, njezini simptomi, uzroci i mogućnosti liječenja. Posebna poglavlja posvećena su izvorima i bazama podataka. Analiza i ekstrakcija podataka iz specijalističkih nalaza predstavljaju poseban izazov te su razrađeni mogući pristupi ovome problem. U konačnici predstavljen je razvijeni model za detekciju osoba u riziku od Gaucherove bolesti. Dana je rasprava u kojoj su opisani mogući daljnji koraci te unaprjeđenja. Na kraju je dan zaključak.

1. Motivacija za izradu diplomskog rada

Kao što je već navedeno u Uvodu, Gaucherova bolest je rijetka genetska bolest koja može imati ozbiljne posljedice na zdravlje oboljelih ako se ne krene liječiti na vrijeme. Bolest se pojavljuje kod 1 od 50 000 osoba, ali zbog svojih nespecifičnih simptoma često ostaje nedijagnosticirana. Pravovremeno otkrivanje i pravovremeni početak liječenja ključni su za sprječavanje napredovanja bolesti i poboljšanje života oboljelih.

Postoje mnogi razlozi i brojne motivacije za razvoj algoritma za detekciju Gaucherove bolesti, neki od razloga bit će navedeni u nastavku.

Prvi i najvažniji razlog već je nekoliko puta spomenut, a to je **rano otkrivanje i pravovremeno liječenje**. Gaucherova bolest manifestira se kroz kombinaciju brojnih simptoma kao što su splenomegalija (povećana slezena), hepatomegalija (povećana jetra), trombocitopenija (sniženi broj trombocita u krvi), anemija i mnogi drugi. Ovi simptomi su jako nespecifični i mogu biti povezani s brojnim drugim bolestima, što otežava uspostavljanje dijagnoze. Algoritam koji može brzo i točno povezati skup simptoma koji karakteriziraju Gaucherovu bolest omogućit će pravovremeno uspostavljanje dijagnoze i rani početak liječenja, što može značajno poboljšati rezultate liječenja.

Zamišljeni sustav čiji će jedan od modula biti algoritam razvijen u ovome radu, bit će odlična **podrška liječnicima** u donošenju odluka. Liječnici često moraju donositi odluke na temelju velikih količina podataka koji dolaze iz različitih izvora kao što su medicinski kartoni, laboratorijski i specijalistički nalazi te klinički simptomi. Računalo može automatizirati proces traženja relevantnih simptoma i prezentirati ih liječniku, što će mu pomoći u donošenju daljnjih odluka. Ovime se također smanjuje i rizik od pojave pogreške i/ili propusta.

Još jedan benefit razvoja ovog algoritma je **povećanje efikasnosti i ušteda resursa**. Pregledavanje medicinskih podataka može biti izuzetno skup i vremenski zahtjevan proces ako ga obavlja čovjek, a može biti efikasniji i brži korištenjem računala. Računalo može u kratkom vremenskom roku obraditi i analizirati velike količine podataka što smanjuje potrebu za ručnim radom visoko obrazovanih radnika i optimizira upotrebu resursa u zdravstvenom sustavu. Liječnici, umjesto da pregledavaju brojne stare nalaze, mogu raditi druge korisne poslove.

Razvoj algoritma donosi **višestruke benefite** te uz poboljšanje kvalitete zdravstvene skrbi omogućava **bolju edukaciju i podizanje svijesti**. Kako je već više puta naglašeno rano otkrivanje bolesti omogućuje pravovremeno započinjanje specifične terapije, što sprječava komplikacije i poboljšava ukupnu kvalitetu života pacijenata. Ovo ne samo da ima brojne benefite za pacijente, nego **smanjuje i troškove dugotrajne zdravstvene skrbi**. Implementacija ovog algoritma i njegovi rezultati mogu se koristiti za daljnja istraživanja i edukaciju zdravstvenih djelatnika ne samo o Gaucherovoj bolesti, već i o općenitim mogućnostima upotrebe sličnih tehnologija u unaprjeđenju zdravstvenog sektora.

Iako je glavni fokus projekta na Domu zdravlja Zagreb – Centar, razvijeni algoritam ima veliki potencijal za širu primjenu u različitim zdravstvenim ustanovama kako diljem Hrvatske tako i svijeta. Partnerstvo s globalnim liderom u biofarmaceutici, kompanijom Takeda, omogućuje plasiranje ideje, koncepta i algoritma na strana tržišta i prilagodbu algoritma stranim zdravstvenim sustavima.

Razvoj algoritma za detekciju Gaucherove bolesti važan je korak prema unaprjeđenju hrvatskog zdravstvenog sustava, s mogućnošću šire primjene i globalnog utjecaja.

2. Metodologija

U ovom poglavlju bit će opisana metodologija korištena za razvoj algoritma za detekciju Gaucherove bolesti. Kao što je već navedeno, algoritam se razvija u sklopu „Gaucher screening“ projekta u suradnji s kompanijom Takeda, MCS Grupom i Domom zdravlja Zagreb – Centar.

2.1. Sudionici projekta

U ovom potpoglavlju bit će detaljnije opisane uloge svakog od sudionika projekta.

Takeda

Takeda je globalna biofarmaceutska kompanija čiji je fokus stavljen na pacijente, temelji se na vrijednostima, istraživanju i razvoju te je posvećena stvaranju boljeg zdravlja i svjetlije budućnosti za ljude diljem svijeta [3]. Ova kompanija financira projekt i pruža stručno znanje iz područja medicine i farmacije. Oni su naručitelji algoritma i idejni začetnici cijeloga projekta.

MCS Grupa

MCS Grupa je vodeći hrvatski proizvođač informacijskih sustava u zdravstvenom sektoru od 1999. godine. Oni se specijaliziraju isključivo za složena cloud rješenja u zdravstvu, uključujući sustave za elektroničko naručivanje, napredne analize medicinskih podataka, sustave potpore u odlučivanju i slične djelatnosti [4]. Njihova uloga u ovome projektu je napraviti konačnu realizaciju tehničkog rješenja i implementirati ga u postojeći sustav koji se već koristi u Domu zdravlja Zagreb – Centar.

Dom zdravlja Zagreb - Centar

Dom zdravlja Zagreb – Centar je ustanova čija je primarna zadaća zdravstvena zaštita koja stanovnicima Grada Zagreba osigurava lako dostupnu, kontinuiranu, cjelovitu i integriranu zdravstvenu zaštitu. Ovo je najveći dom zdravlja u Republici Hrvatskoj kojemu gravitira 350 000 stanovnika Grada Zagreba [5]. Oni daju

dozvolu za obradu podataka svojih pacijenata, provode daljnja testiranja i daju povratnu informaciju o točnosti algoritma.

Fakultet elektrotehnike i računarstva (FER)

Fakultet elektrotehnike i računarstva Sveučilišta u Zagrebu (FER) je najutjecajnija i najveća znanstvena i obrazovna ustanova u Republici Hrvatskoj u području elektrotehnike, računarstva te informacijskih i komunikacijskih tehnologija [6]. Njegova uloga na ovom projektu je razvijanje metodologije i algoritma koji će pomoći prepoznati rizične pacijente. Ovaj diplomski rad dio je spomenutog projekta i izrađuje se na ovoj ustanovi.

2.2. Zaštita podataka

Kako je već istaknuto, zaštita podataka jedna je od ključnih komponenti ovoga projekta. Sve aktivnosti vezane za projekt i pisanje diplomskog rada odrađene su u skladu s Općom uredbom o zaštiti podataka (GDPR-om). Osobni podatci bit će zaštićeni od bilo kakve mogućnosti zlouporabe. Bitno je još jednom napomenuti kako je to osigurano tako što je pristup osobnim podacima stvarnih pacijenta kao i rezultatima rada algoritma omogućen isključivo pacijentovom liječniku opće prakse. Svi ostali sudionici imaju pristup isključivo agregiranim rezultatima algoritma.

Za pisanje ovog diplomskog rada traženo je mišljenje i dobiveno je odobrenje Etičkog povjerenstva Fakulteta elektrotehnike i računarstva Sveučilišta u Zagrebu. Odluka etičkog povjerenstva može se pronaći kao prilog ovog diplomskog rada.

2.3. Metodologija rada u užem smislu

Algoritam će biti integriran u postojeći MCS-ov liječnički sustav koji se trenutno koristi u Domu zdravlja Zagreb – Centar. Funkcionalnost algoritma ukratko će biti opisana u nastavku. Algoritam se oslanja na određene ulazne podatke koji se obrađuju i na osnovu kojih se dobivaju određeni izlazni podaci. Slijedi njihov pregledan prikaz:

1. Ulazni podatci:

- MBO pacijenta
- laboratorijski nalazi
- tekstualni nalazi liječnika
- uspostavljene MKB-10¹ dijagnoze

2. Obrada podataka

- SQL upiti koji pronalaze sve tražene simptome kod pacijenata
- BODOVANJE svakog pronađenog simptoma

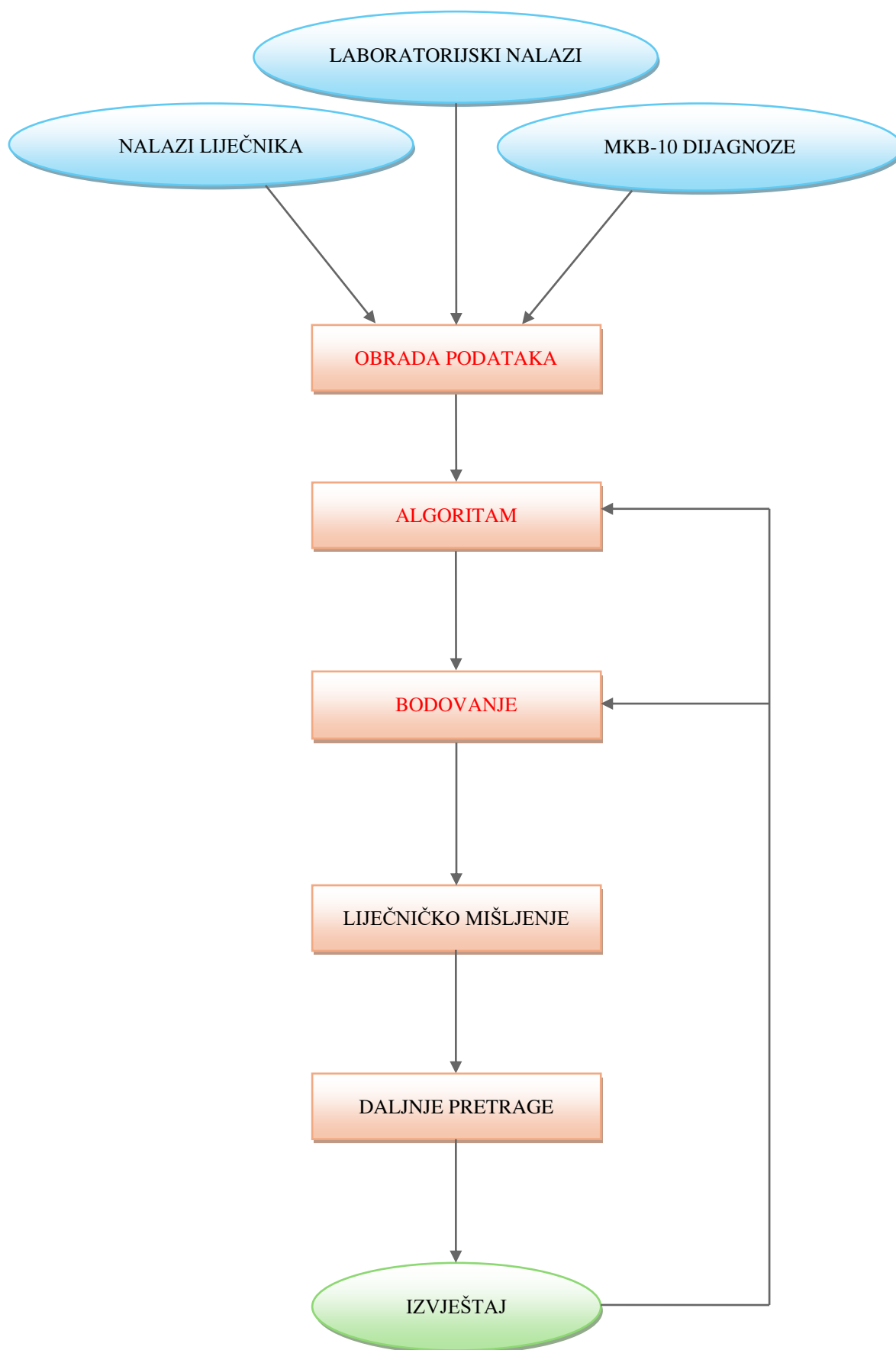
3. Izlazni podatci:

- određen broj bodova koji korelira s rizikom od obolijevanja od Gaucherove bolesti
- usporedba s postavljenom granicom
- agregirani podatci o broju ispravno detektiranih pacijenata s vremenskim odmakom.

Za kreiranje ulaznih podataka nužan je niz SQL upita koji analiziraju medicinske podatke pacijenata. Taj niz SQL upita popunjava dio konačne tablice koja će poslužiti kao podloga konačno izvedenom algoritmu. Po dogovoru s hematolozima i smjericama danim u znanstvenim radovima [7], na temelju prisutnih simptoma algoritam dodjeljuje bodove svakom od pacijenata i pohranjuje ih u bazu. Pacijenti s određenim brojem bodova bit će označeni kao kandidati za daljnje testiranje. Bodovanje svih pacijenata bit će dostavljeno liječniku koji će na bazi pronađenih simptoma odlučiti hoće li pacijenta zaista poslati na daljnje pretrage ili ne. Tek nakon testiranja odabranih pacijenata na GB bit će dobiven izlazni podatak o broju točno klasificiranih pacijenata. Očekuje se i usklađenje algoritma i bodovanja pojedinih simptoma nakon dobivanja povratnih informacija o točnosti dijagnosticiranja.

¹ 10. revizija Međunarodne klasifikacije bolesti i srodnih zdravstvenih problema MKB-10 (*engl. International Statistical Classification of Diseases and Related Health Problems ICD-10*)

Na slici (Slika 2.1) dan je prikaz svih potrebnih koraka za dijagnosticiranje Gaucherove bolesti. U plavo obojenim poljima prikazani su potrebni ulazni podaci. Oni se najprije obrađuju te se zatim predaju algoritmu. Algoritam na bazi zadanih parametara i dobivenih ulaznih podataka pacijentu dodjeljuje određeni broj bodova. Broj bodova pojedinog pacijenta prosljeđuje se njegovom liječniku opće prakse koji na bazi pronađenih simptoma donosi odluku je li određenog pacijenta potrebno slati na daljnje pretrage. Nakon odrađenih pretraga stvara se izvještaj o radu algoritma, zeleno obojeno polje na slici. Izvješta treba sadržavati podatke kao što su; koliko je algoritam pacijenta predložio za testiranje, koliko je pacijenta uistinu testirano, koliko ih zaista ima Gaucherovu bolest a koji su pogrešno prepoznati kao pacijenti u riziku. Pomoću informacija dobivenih iz izvještaja moguće je prilagođavati parametre algoritma te uvjete bodovanja što će dodatno poboljšati rezultate rada algoritma u budućnosti. Bitno je napomenuti kako će se izvještaj dobiti s velikim vremenskim odmakom jer je potrebno algoritam ispitati na većem broju pacijenata nakon što su odrađene sve daljnje pretrage. Dijelovi dijagnosticiranja Gaucherove bolesti u kojima direktno sudjeluje FER na slici su napisani crvenom bojom. To su obrada podataka, algoritam i bodovanje.



Slika 2.1 Model za pronalazak pacijenata s Gaucherovom bolesti

2.4. Validacija algoritma i podešavanje algoritma

Bitno je još jednom napomenuti kako u ovoj fazi projekta neće biti moguće validirati algoritam jer ne postoji baza s već dijagnosticiranim pacijentima na kojoj bi uspješno mogli odrediti točnost algoritma. Validaciju algoritma bit će moguće napraviti tek nakon duljeg korištenja razvijenog produkta u domu zdravlja i kada budu prikupljeni agregirani rezultati algoritma te napravljen izvještaj. Nakon što se prikupi dovoljno podataka o tome koliko je pacijenata s određenim minimumom broja bodova algoritam pronašao te koliko njih uistinu ima Gaucherovu bolest, bit ćemo u mogućnosti odrediti točnost algoritma i dodatno podešavati potrebne parametre.

3. Gaucherova bolest

Gaucherova bolest rijedak je i nasljedan poremećaj skladištenja lipida. To je autosomno recesivni poremećaj kojeg karakterizira nedostatak lizosomskog enzima β -glukocerebrozidaze (GBA), a izazvan je patogenom varijacijom u genu GBA1 [8]. Ova bolest može izazvati široki spektar simptoma koji variraju u težini od blagih do izuzetno ozbiljnih. Dubinsko razumijevanje Gaucherove bolesti ključno je za postavljanje ispravne dijagnoze te pružanje adekvatnog liječenja pacijentima.

3.1. Rijetke bolesti

Rijetke bolesti su one bolesti koje pogađaju mali broj ljudi. Preciznije rijetkim bolestima se smatraju one od kojih je u svijetu oboljelo manje od 5 u 10 000 osoba [9]. Danas nam je poznato oko 6 000–8 000 rijetkih bolesti, te ih se godišnje pronađe 250-280 novih [10]. Iako od svake rijetke bolesti zasebno boluje vrlo mali broj ljudi od svih rijetkih bolesti zajedno boluje oko 6-7 % svjetske populacije [11]. Postoje i vrlo rijetke bolesti koje su definirane kao one koje se javljaju kod jedne osobe od njih 100 000 ili manje [12].

Jedna od najčešćih karakteristika rijetkih bolesti je teško i vrlo kasno uspostavljanje dijagnoze. Budući da su one često kronične, degenerativne i smrtonosne ta činjenica predstavlja veliki problem, jer kasno dijagnosticiranje i kasno uvođenje terapije dovodi do smanjena kvalitete života oboljelih.

3.2. Općenito o Gaucherovoj bolesti

Gaucherova bolest jedna je od najčešćih rijetkih bolesti u svijetu. Ona se javlja kod 1-2 osobe na njih 100 000 i jednako je učestala kod žena i muškaraca.

Gaucherova bolest dijeli se na tri tipa [13].

- **Tip 1 - Neneuropatska Gaucherova bolest**

Ovo je najčešći tip Gaucherove bolesti koji ne zahvaća centralni živčani sustav. Karakteristike su bezbolna splenomegalija (povećana slezena), anemija ili trombocitopenija. Česti simptomi su kronični umor,

hepatomegalija (povećana jetra), bol u kostima ili česti prijelomi kostiju kao i učestala pojava modrica. Zanimljiva je činjenica kako je ovaj tip bolesti češće prisutan kod osoba židovskog podrijetla.

- **Tip 2 – Akutna neuropatska Gaucherova bolest**

Ovaj tip moguće je dijagnosticirati vrlo rano u životu, prenatalno, pri samom rođenju ili u dojenačkoj dobi. Znakovi prisutnosti ovoga tipa su povišeni tonus, razni napadaji, strabizam² i organomegalija (povećani organi). Tipični simptomi su i zastoj u razvoju, abnormalnosti gutanja, hepatosplenomegalija i okulomotorna apraksija. Akutna neuropatska Gaucherova bolest je rijetka i ozbiljna, a zahvaća mozak što često dovodi do smrti u ranom djetinjstvu.

- **Tip 3 – Kronična neuropatska Gaucherova bolest**

Simptomi ovog tipa slični su onima koji se javljaju kod tipa 1 uz dodatak neuroloških problema koji se najčešće razvijaju u djetinjstvu ili adolescenciji. Karakteristike ovog tipa su organomegalije, zahvaćenost kostiju, usporavanje horizontalnih pokreta očiju.

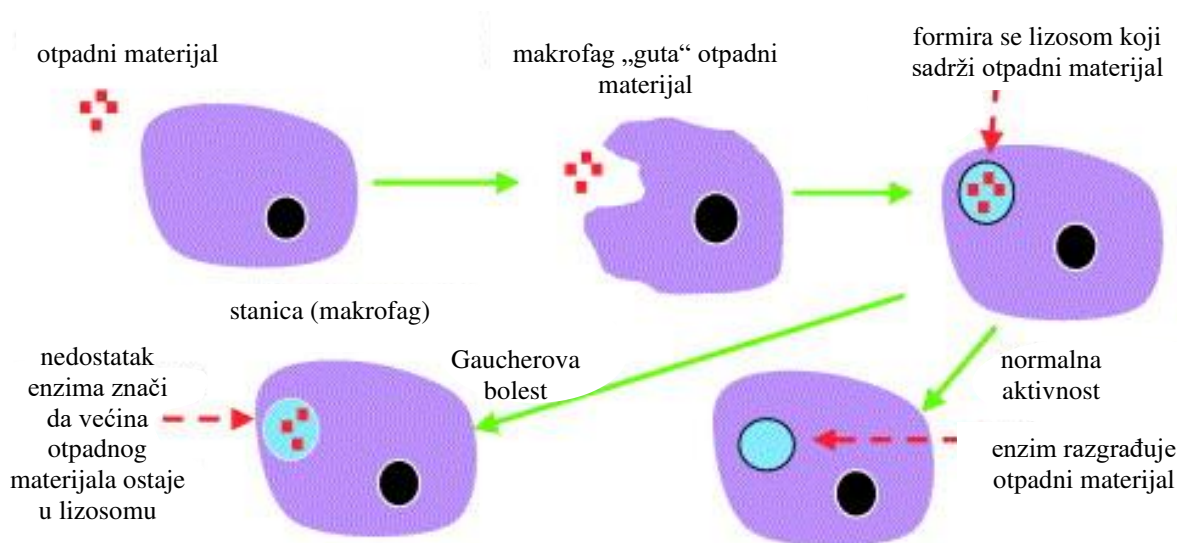
Dijagnozu je često jako teško postaviti zbog nespecifičnih simptoma koji mogu upućivati na mnoge druge bolesti. Činjenica da različiti tipovi ove bolesti dijele neke od simptoma ne pomaže procesu postavljanja ispravne dijagnoze.

3.3. Uzroci Gaucherove bolesti

Sva tri tipa Gaucherove bolesti uzrokovane su sniženom aktivnošću enzima glukocerebrozidaze do koje dolazi zbog mutacije u genu GBA1 čija je uloga kodirati taj enzim. Zbog nedostatka navedenog enzima dolazi do nakupljanja glukozilceramida u makrofagima što rezultira nakupljanjem tih lipida u tijelu [13]. Akumulirani glukocerebrozid unutar lizosoma makrofaga nazivaju se Gaucherovim stanicama. Njihovo nakupljanje u različitim tkivima, poput jetre, slezene i koštane srži, dovodi do kliničkih simptoma GB. Shematski prikaz ovog procesa prikazan je na slici (Slika 3.1), koja je preuzeta iz izvora [14] te prevedena na hrvatski jezik.

² Strabizam je poremećaj položaja ili pokretljivosti očiju kod kojeg oba oka nemaju zajednički smjer gledanja [15]

Kod apatičnih slučajeva uzrok bolesti je nešto drugačiji. Uzrok ovakvog oblika bolesti je nedostatak Saposina C, uniparentalne izodizomije ili nove mutacije [13].



Slika 3.1 Proces nastajanja Gaucherove bolesti [14]

3.4. Dijagnosticiranje Gaucherove bolesti

Pravovremena dijagnoza izuzetno je bitna jer može spriječiti pojavu određenih simptoma i daljnje napredovanje bolesti. Nažalost, kako je već navedeno, Gaucherovu bolest izuzetno je teško dijagnosticirati.

Dijagnoza se potvrđuje mjerenjem aktivnosti glukocerebrozidaze u bijelim krvnim stanicama periferne krvi [13]. Još jedan česti faktor u dijagnostici su manja povišenja razine jetrenih i angiotenzin-konvertirajućih enzima. Često se koristi i genetsko testiranje, analiza DNA, za uspostavljanje dijagnoze. Proučava se postoji li prisutnost 2 mutirana alela u GBA genu [16].

Postoje i slikovne tehnike koje pomažu u dijagnosticiranju GB poput MRI-a i CT skenova i ultrazvuka koji se koriste za procjenu oštećenja nastalih na organima i kostima. Ovim testovima otkriva se splenomegalija, hepatomegalija, te promjene u kostima, svi simptomi koji upućuju na GB.

3.5. Liječenje Gaucherove bolesti

Liječenje Gaucherove bolesti ovisi o tipu bolesti i prisutnim simptomima. Cilj liječenja je smanjenje postojećih simptoma, poboljšanje kvalitete života i sprječavanje pojave novih simptoma.

Trenutno postoje dvije glavne terapije [17][18]:

- **Enzimski nadomjesna terapija**

Ova tip terapije koristi se kod pacijenata s tipovima GB 1 i 3. Princip na kojemu se zasniva ova terapija je intravenozno davanje glukocerebrozidaze kako bi se nadomjestio nedostatak tog enzima u stanicama. Rezultati terapije su smanjenje simptoma povezanih s splenomegalijom, hepatomegalijom, anemijom i trombocitopenijom.

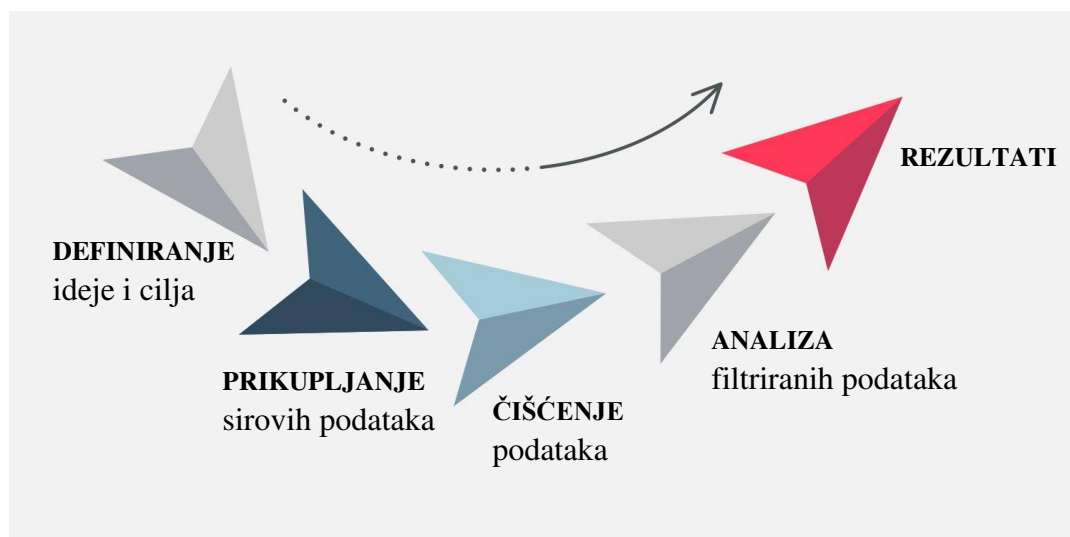
- **Supstrat redukcijaska terapija**

Ova terapija smanjuje proizvodnju staničnog glukozilceramida čime se smanjuje njegov višak nakupljen u tijelu. Ovom terapijom smanjuje se veličina jetre i slezene te smanjuje simptome povezane s anemijom i trombocitopenijom.

4. Izvori podataka

U današnje digitalno doba, podaci su postali jedan od najvrjednijih resursa. Podatci su prisutni u svakom aspektu modernog života, od poslovanja i zdravstva do obrazovanja i zabave. Njihovo prikupljanje te naknadna analiza ključni su za rast i razvoj tehnologija kao što su umjetna inteligencija (*engl. Artificial Intelligence – AI*), strojno učenje (*engl. Machine Learning– ML*), obrada prirodnog jezika (*engl. Natural Language Processing - NLP*) i napredne analize podataka. Osim za napredak tehnologija, analiza podataka pokazuje se vrlo vrijednom za brojne organizacije, institucije i poslovne objekte koji žele unaprijediti svoje usluge. Postupku analize podataka prethodi postupak prikupljanja te organizacije potrebnih podataka u skupove podataka.

Na slici (Slika 4.1), preuzetoj iz izvora [19] te prevedenoj na hrvatski jezik, prikazan je životni vijek podataka od začetka ideje do dobivanja rezultata iz podataka. Nakon definiranja ideje i cilja obrade podataka, provodi se prikupljanje podataka. Prikupljeni podatci se zatim čiste i organiziraju u prikladne skupove podataka, uklanjaju se neupotrebljivi podatci, nad kojima se zatim provode potrebni procesi transformacije i analize. Iz tako obrađenih podataka dobivaju se traženi rezultati.



Slika 4.1 Životni vijek podataka [19]

Skup podataka je zbirka podataka koji su najčešće organizirani u nekom standardiziranom formatu radi lakše analize i obrade. Kao što je već navedeno, skupovi podataka imaju vrlo široku primjenu u analizi podataka, poslovnoj inteligenciji, raznim istraživanjima,

treniranju modela umjetne inteligencije, te brojnim drugim područjima [20]. Raznovrsnost skupova podataka očituje se u njihovoj veličini, strukturi i vrsti informacija koje sadrže. S obzirom na strukturu, razlikujemo dvije vrste skupova podataka: strukturirane i nestrukturirane. Ove dvije vrste i razlika između njih bit će detaljnije opisane u nastavku.

Jedan od primjera primjene skupova podataka je u meteorologiji. Podatci o temperaturama oceana, atmosferskim tlakovima i klimatskim promjenama organiziraju se u skupove podataka koji se zatim koriste za praćenje klimatskih obrazaca i predviđanje vremenskih prognoza. Drugi primjer, puno značajniji za ovaj rad, su skupovi podataka iz zdravstvene domene. Ovi podatci mogu sadržavati informacije o bolestima, dijagnozama, terapijama i uspješnosti liječenja pacijenata. Analizom ovih podataka moguće je lakše postavljanje ispravne dijagnoze te odabir prikladne terapije. Podatci korišteni u ovom radu dolaze upravo iz domene zdravstva te su detaljnije opisani u idućem potpoglavlju.

4.1. Prikupljanje podataka

Podatci na kojima će se izvoditi algoritam izrađen u ovom radu prikupljeni su iz više različitih izvora podataka kao što su stari kartoni pacijenata, elektronski kartoni pacijenata te rezultati raznih laboratorijskih i specijalističkih pretraga. Podatci dolaze iz Doma zdravlja Zagreb – Centar, koji koristi CEZIH (Centralni zdravstveni informacijski sustav Hrvatske) za upravljanje elektronskim zdravstvenim kartonima pacijenata. Pristup podacima FER-u omogućili su MCS Grupa i Dom zdravlja Zagreb – Centar, koji su osigurali podatke u skladu s pravilima o zaštiti privatnosti. Za razvoj algoritma bio je omogućen pristup samo kopiji baze koji ne sadrži podatke stvarnih pacijenata, već sadrži umjetno generiran skup podataka.

Isprva se čini kako uz svakodnevno obavljanje zdravstvenih pregleda ovih podataka ima izrazito puno te da pronalazak kvalitetnog skupa podataka ne bi trebao biti problem, no to ipak nije tako. Dok je istina da postoji velika količina podataka potrebno je uzeti u obzir i njihovu kvalitetu, balansiranost podataka te određena etička pitanja poput privatnosti pacijenta. Ovi nedostaci predstavljaju veliku prepreku u radu s podacima u zdravstvu.

Već je spomenuto kako su dva glavna tipa podataka strukturirani i nestrukturirani podatci. Za razvoj ovog algoritma korištena su oba tipa podataka. Njihova razlika i sadržaj pojedinog tipa podataka bit će detaljnije opisani u nastavku ovog poglavlja.

4.2. Strukturirani podatci

Strukturirani podatci su organizirani na točno definirani način. Iako ne postoji formalna definicija strukturiranih podataka jedna od mogućih definicija dana je u nastavku [21].

„Strukturirani podaci su podaci koji odgovaraju formalno definiranom podatkovnom modelu koji određuje kako podaci trebaju izgledati, koje operacije se nad njima mogu izvršavati te koja pravila podaci moraju zadovoljavati.“

Strukturiranim podacima obično zovemo one podatke koji prate određenu formu. Oni se pohranjuju u obliku stupaca i ovako organiziranim podacima lako je upravljati s pomoću različitih programskih jezika pa tako i SQL-a [22]. Najčešći primjer strukturiranih podataka su relacijske tablice, Excel dokumenti i CSV datoteke. One imaju točno definirane attribute te tip podatka za svaki od atributa. Ovo omogućava lakšu analizu i obradu podataka zbog čega je uvijek poželjno raditi sa strukturiranim podacima.

U zdravstvenoj domeni postoje brojni strukturirani podatci. Najvažniji su laboratorijski nalazi. To su strukturirani podatci koji sadrže rezultate raznih medicinskih testova i najčešće se izražavaju brojčano. Oni igraju važnu ulogu u razvoju algoritma jer omogućavaju preciznu identifikaciju simptoma karakterističnih za GB. Laboratorijski nalazi dobiveni od doma zdravlja najčešće se pohranjuju u obliku PDF formata.

Na slici (Slika 4.2) prikazan je primjer izgleda laboratorijskog nalaza u PDF formatu, koji služi kao izvor strukturiranih podataka. Svi osobni podatci pacijenta prikazani na nalazu su umjetno generirani. Bitno je naglasiti kako se svaki mjerni pokazatelj, na ovakvom tipu laboratorijskog nalaza, izražava u određenoj mjernoj jedinici. Unaprijed je određen raspon očekivanih, „normalnih“, vrijednosti. Time nam je poznat minimum i maksimum očekivanog raspona.

Ime i prezime: Ivana Horvat
 Spol: Ženski
 Rođen: 11.08.2003.
 Naručitelj: Ivan Horvat

Broj protokola

2082

Naručitelj: (123456)
 KEM/Odjel za prijem pacijenata i upravljanje kvalitetom
 Vrijeme uzorkovanja: 06.06.2024
 Vrijeme primitka: 06.06.2024 07:51
 Vrijeme izdavanja: 06.06.2024 08:44

Odjel za medicinsku biokemiju i analitičku toksikologiju

Uzorak	Biokemijske pretrage	Rezultat	Jedinica	Referentni interval
vS	Bilirubin (ukupni)	17	μmol/L	3 do 20
vS	Ureja	4,1	mmol/L	2,8 do 8,3
vS	Kreatinin	68	μmol/L	49 do 90
vS	Procjena glomerularne filtracije (CKD-EPI)	109	mL/min/1,73m ²	Glomerularna filtracija: G1 (normalna): >=90; G2 (blago snižena): 60-89; G3a (blago do umjereno snižena): 45-59; G3b (umjereno do jako snižena): 30-44; G4 (jako snižena): 15-29; G5 (zatajenje bubrega): <15
vPo	Glukoza (P)	5,1	mmol/L	4,4 do 6,4
vS	Aspartat-aminotransferaza (AST)	18	U/L 37°C	8 do 30
vS	Alanin-aminotransferaza (ALT)	13	U/L 37°C	10 do 36
vS	Alkalna fosfataza (ALP)	54	U/L 37°C	54 do 119
vS	Gama-glutamilttransferaza (GGT)	12	U/L 37°C	9 do 35

Uzorak	Lipidi i lipoproteini	Rezultat	Jedinica	Preporučena vrijednost
vS	Trigliceridi	0,7	mmol/L	Natašte: < 1,7 mmol/L Postprandijalno: < 2,0 mmol/L
vS	Kolesterol	5,1 H	mmol/L	< 5,0
vS	HDL kolesterol	1,2	mmol/L	> 1,2
vS	Ne-HDL kolesterol	3,9 H	mmol/L	Natašte: < 3,8 mmol/L Postprandijalno: < 3,9 mmol/L
vS	Ostatni kolesterol	< 0,1	mmol/L	Natašte: < 0,8 mmol/L Postprandijalno: < 0,9 mmol/L
vS	LDL kolesterol	3,7 H	mmol/L	Za osobe s: vrlo velikim rizikom: <1,4 mmol/L velikim rizikom: <1,8 mmol/L umjerenim rizikom: <2,6 mmol/L malim rizikom: <3,0 mmol/L

Slika 4.2 Primjer PDF-a laboratorijskog nalaza

Kako bi podatci iz PDF dokumenta bili u obliku prikladnom za daljnju obradu, u bazi podataka pohranjene su samo iščitane vrijednosti, a ne cijele PDF datoteke. U sklopu ovog diplomskog rada napravljena je skripta koja prikazuje kako se obrađuju PDF datoteke laboratorijskih nalaza. Skripta nije korištena u svrhe razvoja algoritma, jer je taj proces već napravljen od strane MCS Grupe. Skripta je napravljena u Pythonu i prikazana je u nastavku na slici (Slika 4.3).

```

import PyPDF2
import re

def extract_data_from_lab_report(pdf):
    with open(pdf, 'rb') as f:
        reader = PyPDF2.PdfReader(f)
        text = ''
        for page in reader.pages:
            text += page.extract_text()
    print(text)
    # Regular expressions for extracting specific information
    name_regex = r'Ime \w+: (\w+\s\w+)'
    gen_regex = r'Spol :[\s\w\S]*(\s[MMŽŽ]\s)'
    date_regex = r'Datum rođenja:[\s\w\S]*(\d{1,2}\.\d{1,2}\.\d{4})'
    mbo_regex = r'MBO: (\d+)'
    values_regex = r'(Nalaz\s+\d+[\s\S]*?)(?=(?:Nalaz\s+\d+)|\Z)'

    # Extracting information using regex
    name = re.search(name_regex, text)
    gen = re.search(gen_regex, text)
    date = re.search(date_regex, text)
    mbo = re.search(mbo_regex, text)
    values = re.findall(values_regex, text)
    # Process and print extracted data
    if name:
        print("Ime i prezime:", name.group(1))
    if gen:
        print("Spol:", gen.group(1))
    if date:
        print("Datum:", date.group(1))
    if mbo:
        print("MBO:", mbo.group(1))
    if values:
        print("Detektirane vrijednosti i intervali:")
        for val in values:
            print(val)

# Poziv funkcije za ekstrakciju podataka iz PDF-a
pdf_file_path = '1.PDF'
extract_data_from_lab_report(pdf_file_path)

```

Slika 4.3 Skripta za izvlačenje podataka iz laboratorijskih nalaza

Zbog njihove pouzdanosti i jednostavnosti obrade strukturirani podatci iz laboratorijskih nalaza korišteni su kao ključni dio u razvoju algoritma za dijagnosticiranje GB. Za razliku od laboratorijskih nalaza korišteni su i nestrukturirani nalazi o kojima je više napisano u potpoglavlju 4.3.

4.3. Nestrukturirani podatci

Nestrukturirani podatci, za razliku od strukturiranih, nemaju točno definiranu formu. Neki od primjera koji spadaju u ovaj tip podataka su slikovni, video i audio zapisi te slobodan tekst. Čak je 90 % podataka, čije se količine neprestano povećavaju jako velikom brzinom, nestrukturirana [22]. Ovi podaci nisu prikladni za relacijske baze podataka zbog čega ih je znatno teže analizirati i obraditi. Rukovanje nestrukturiranim podacima obično zahtijeva mnogo resursa, uključuje mnoge korake i nosi visoku razinu nesigurnosti, ali nestrukturirani podaci često sadrže vrlo korisne i ključne informacije koje se ne mogu pronaći u strukturiranim podacima [23].

U našoj domeni glavni izvor nestrukturiranih podataka su medicinske snimke, liječnikove bilješke i specijalistički nalazi koji sadrže slobodni tekst. Budući da naša baza podataka ne sadrži medicinske snimke (nisu ključne za dijagnosticiranje GB) mi ćemo se baviti specijalističkim nalazima i bilješkama liječnika kao izvorom nestrukturiranih podataka. U njima se obično navodi ime i prezime pacijenta, svi simptomi, sve provedene pretrage i sve postavljene dijagnoze. Ovakav tip podataka znatno je teže obraditi nego laboratorijske nalaze što je predstavljalo problem u razvoju algoritma za detekciju GB. U trenutnoj fazi projekta bit će korištene informacije izvađene samo iz bilježaka liječnika jer one sadrže znatno manju količinu teksta. Način na koji se izvlače informacije iz ovakvog tipa podataka bit će opisan u poglavlju 7.

Ekstrakcija relevantnih informacija iz složenijih i duljih tekstova znatno je zahtjevnija od parsiranja kraćih tekstualnih zapisa. S razvojem umjetne inteligencije i strojnog učenja, ogromne količine nestrukturiranih podataka mogu se analizirati i koristiti za donošenje odluka. Jedna od vodećih tehnologija u ovom području je obrada prirodnog jezika. NLP koristi algoritme strojnog učenja kako bi ljudski govor (slobodan tekst) pretvorio u podatke razumljive strojevima. Ovaj proces kombinira snagu računalne znanosti i lingvistike kako bi se istražila struktura i pravila jezika [24]. Više o NPL-u bit će rečeno u poglavlju 6.

5. Baze podataka

Baze podataka su organizirane zbirke logički povezanih, pretražljivih i međusobno ovisnih podataka (informacija), pohranjene u nekom od računalno čitljivih medija [25]. Jednostavnije rečeno baza podataka je alat za prikupljanje i organizaciju podataka u koji se mogu pohraniti podatci o osobama, proizvodima, narudžbama, poslovanju i bilo čemu drugome [26]. Milijuni baza podataka koriste se u raznim područjima, naprimjer u poslovnom upravljanju, državnim upravama, upravljanju znanstvenim i inženjerskim podacima [27].

Osim već navedenih područja, baze podataka igraju ključnu ulogu i u medicini, one omogućuju pohranu i obradu velikih količina podataka, čime se smanjuje rutinski posao liječnika. U medicinske baze podataka pohranjuju se razne informacije poput laboratorijskih nalaza, dijagnoza, propisanih terapija medicinska povijest pacijenata te mnoge druge. Pravilno upravljanje navedenim informacijama može pomoći unaprijediti kvalitetu zdravstvenih sustava i provođenje raznih inovativnih istraživanja.

5.1. Baze podataka u medicini u Hrvatskoj

U Hrvatskoj, kao i u mnogim drugim zemljama, zdravstveni sustavi koriste razne vrste baza podataka za upravljanje medicinskim informacijama. Ove baze podataka mogu biti specifične za bolnice, klinike, laboratorije ili nacionalne zdravstvene sustave.

CEZIH, Centralni zdravstveni informacijski sustav Republike Hrvatske, je glavni zdravstveni informacijski sustav u Hrvatskoj koji centralizira podatke iz različitih dijelova zdravstvenog sustava. CEZIH omogućava razmjenu informacija između liječnika, bolnica, ljekarni i drugih zdravstvenih ustanova, čime se poboljšava koordinacija zdravstvene skrbi i smanjuju mogućnosti ljudske pogreške.

Neke od vrsta informacijskih sustava korištenih u zdravstvenoj domeni

- **Elektronički zdravstveni kartoni (EHR):**

EHR su digitalni oblici konvencionalnih papirnatih zapisa koje zdravstvene ustanove vode kako bi olakšale pružanje usluga [28]. Digitalni zapisi o pacijentima uključuju medicinsku povijest, dijagnoze, terapije i laboratorijske rezultate.

- **Laboratorijski informacijski sustavi (LIS):**

LIS je program koji omogućava unos, obradu i pohranu podataka nastalih kao rezultat laboratorijskih pretraga [29]. Omogućava elektroničku komunikaciju između analitičkog i administrativnog rada laboratorija [30]. Ovaj sustav omogućuje brzu i točnu analizu laboratorijskih podataka, čime se poboljšava dijagnostika i liječenje.

- **Sustavi za upravljanje radiološkim informacijama (RIS):**

RIS čine baze podataka koje pohranjuju slike i podatke iz radioloških pretraga, omogućujući liječnicima da pristupe i analiziraju radiološke pretrage [31].

MCS Grupa je tvrtka koja razvija informacijske sustave za zdravstvene ustanove u Hrvatskoj. Oni su jedan od certificiranih proizvođača programske podrške za CEZIH. Njihovi sustavi koriste različite baze podataka za pohranu i upravljanje medicinskim informacijama. Općenito, u medicini se koriste različite vrste baza podataka, ovisno o potrebama i specifičnostima zdravstvene ustanove. Neke od najčešće korištenih vrsta baza podataka uključuju:

- **Relacijske baze podataka:**

Ove baze podataka organizirane su u relacije i koriste SQL (Structured Query Language) za upravljanje podacima. Primjeri uključuju MySQL, PostgreSQL, Oracle Database i Microsoft SQL Server. Relacijske baze podataka često se koriste za pohranu strukturiranih podataka kao što su laboratorijski rezultati.

- **NoSQL baze podataka:**

Ove baze podataka dizajnirane su za rad s velikim količinama nestrukturiranih podataka. Primjeri uključuju MongoDB, Cassandra i Elasticsearch. NoSQL baze podataka često se koriste za pohranu teksta, slika i drugih nestrukturiranih podataka.

- **Objektno-relacijske baze podataka:**

Kombiniraju značajke relacijskih i objektno-orijentiranih baza podataka. Omogućuju pohranu kompleksnih podataka i struktura podataka. Primjer je Oracle Database.

Baze podataka u medicini ključne su za učinkovito upravljanje informacijama o pacijentima i podršku u donošenju kliničkih odluka. U Hrvatskoj, sustavi poput CEZIH-a i rješenja razvijena od strane MCS Grupe igraju ključnu ulogu u centralizaciji i upravljanju medicinskim podacima. Korištenjem modernih tehnologija i pridržavanjem propisa o zaštiti podataka, zdravstveni sustavi mogu poboljšati kvalitetu skrbi i sigurnost pacijenata.

U projektu za koji se razvijao algoritam opisan u ovome radu koristila se relacijska baza podataka koja će biti opisna u nastavku ovog poglavlja.

5.2. ER model djela baze podataka koji je relevantan za istraživanje

Zbog tipa projekta bit će prikazan samo dio stvarne baze koji je dovoljan za demonstraciju kako sustav i algoritam funkcioniraju. Prikazani su samo najbitniji entiteti i njihovi najvažniji atributi koji su potrebni za rad algoritma. ER model dijela baze koji je relevantan za istraživanje, napravljen u alatu ERDPlus, prikazan je na slikama u nastavku (Slika 5.1 i Slika 5.2).

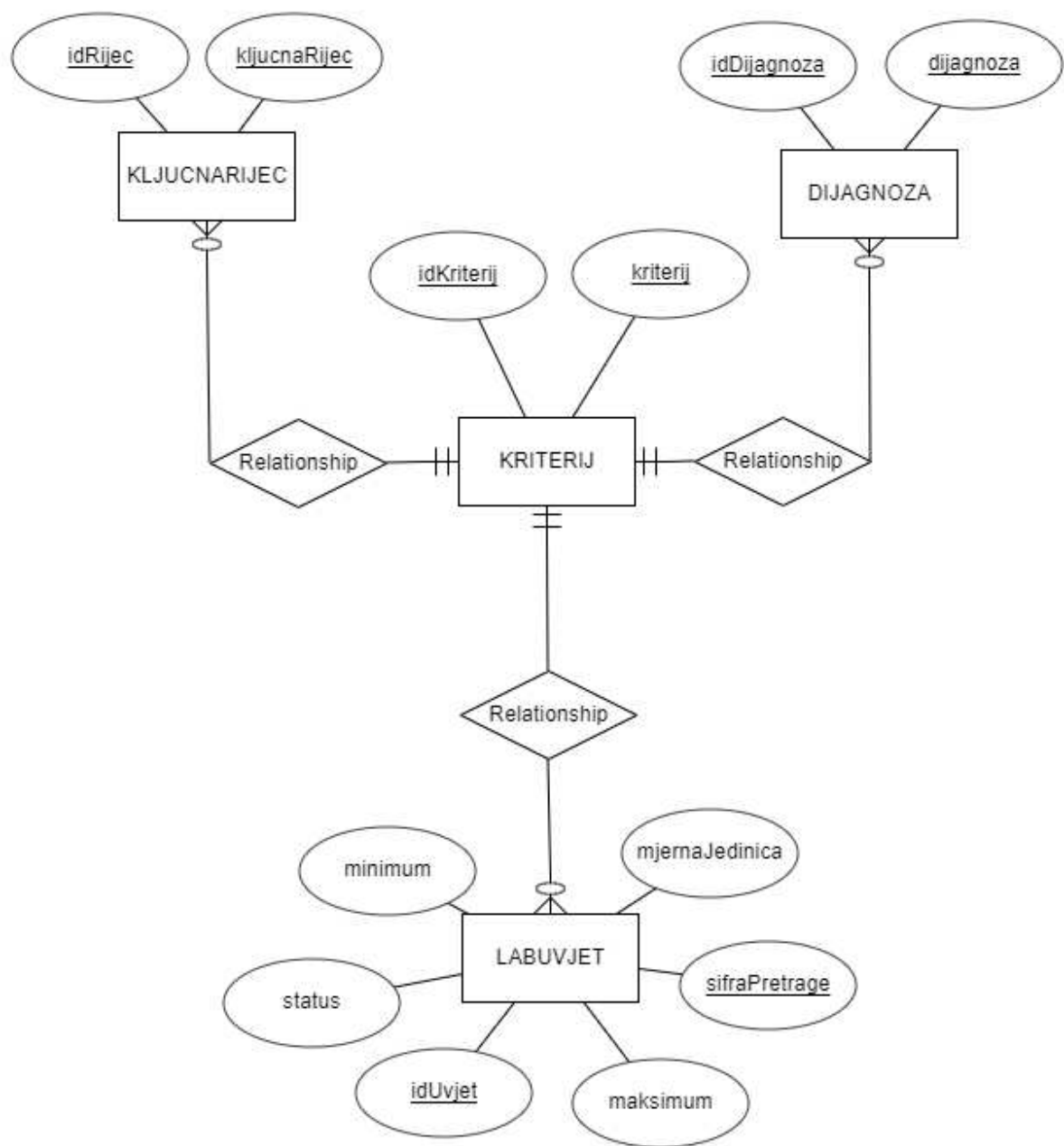
U dijelu baze prikazanom na slici (Slika 5.1) pohranjeni su parametri za svaki od kriterija koje je potrebno ispitati kako bi utvrdili je li osoba kandidat za daljnje testiranje za Gaucherovu bolest ili nije. Ovaj dio baze sastoji se od 4 različita entiteta koji su opisani u nastavku.

Entitet **KRITERIJ** ima dva jedinstvena atributa *idKriterij* i *kriterij*. U atributu *idKriterij* pohranjen je jedinstveni identifikator pojedinog kriterija, dok je u atributu *kriterij* pohranjen jedinstveni naziv svakog kriterija.

Entitet **KLJUCNARIJEC** također ima dva jedinstvena atributa *idRijec* i *kljucnaRijec*. Atribut *idRijec* predstavlja jedinstveni identifikator svake ključne riječi, dok *kljucnaRijec* sadrži tekstualni izraz koji se obično sastoji od jedne do tri riječi. Svakoj od ključnih riječi pripada točno jedan kriterij, više ključnih riječi može pripadati istom kriteriju.

Entitet **DIJAGNOZA**, kao i prijašnja dva, ima dva jedinstvena atributa *idDijagnoza* i *dijagnoza*. U atribut *idDijagnoza* pohranjuje se jedinstveni identifikator svake MKB-10 dijagnoze, a u atribut *dijagnoza* MKB-10 šifra dijagnoze. Svaka od dijagnoza povezana je s kriterijima na isti način kao i ključne riječi. Svakoj od dijagnoza pripada točno jedan kriterij, više dijagnoza može pripadati istom kriteriju.

Entitet **LABUVJET** sadrži šest različitih atributa u kojima su pohranjene informacije o različitim laboratorijskim pretragama. Jedinstveni atribut *idUvjeta* predstavlja jedinstveni identifikator svakog laboratorijskog uvjeta. Postoji još jedan jedinstveni atribut, a to je *sifraPretrage*. U *sifruPretrage* sprema se jedinstvena šifra koja odgovara nekoj pretrazi, na primjer šifra 1-1010100 K pripada pretrazi broja leukocita u krvi. Atributi *minimum* i *maksimum* prikazuju ispod odnosno iznad koje granice određena vrijednost mora biti da bi taj parametar bio izvan referentnog intervala. Atribut *mjernaJedinica* pokazuje u kojoj mjernoj jedinici pretraga mora biti izražena kako bi pripadajuće minimalne i maksimalne vrijednosti bile ispravne. Posljednji atribut je *status*. U *status* se obično upisuju dva moguća stanja L (low) ili H (high). Status nam govori da li snižena ili povišena vrijednost određenog parametra ukazuju na prisutnost nekog od kriterija za dijagnosticiranje Gaucherove bolesti. Svaki laboratorijski uvjet povezan je s kriterijem na potpuno isti način kao i ključne riječi i dijagnoze. Svakom od laboratorijskih uvjeta pripada točno jedan kriterij, više laboratorijskih uvjeta može pripadati istom kriteriju.



Slika 5.1 ER model dijela baze relevantnog za razvoj algoritma

U dijelu baze prikazanom na slici (Slika 5.2) pohranjeni su svi podatci o pacijenti koji su potrebni za odrediti je li osoba kandidat za daljnje testiranje za Gaucherovu bolest ili nije. Ovaj dio baze sastoji se od šest različitih entiteta koji su opisani u nastavku.

Entitet **PACIJENT** ima tri jedinstvena atributa *idPacijenta*, *JMBG* i *MBO*. U atributu *idPacijenta* pohranjen je jedinstveni identifikator pojedinog pacijenta, u atributu *JMBG* pohranjen jedinstveni matični broj građana te atribut *MBO* predstavlja jedinstveni matični broj osigurane osobe. Svaki građanin Republike Hrvatske ima svoj JMBG i MBO koji se koriste za identifikaciju u raznim državnim ustanovama pa tako i domovima zdravlja.

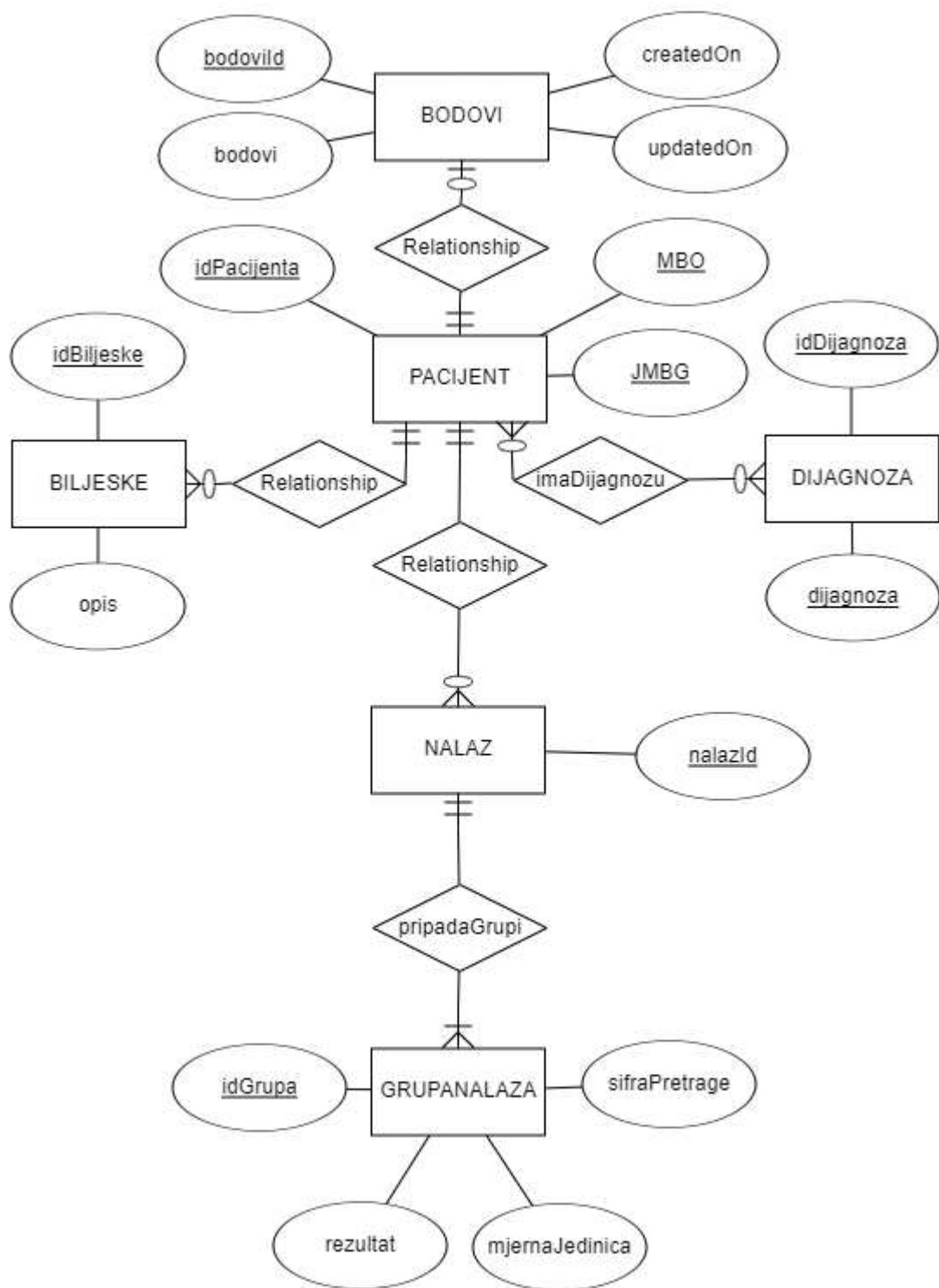
Entitet **NALAZ** ima samo jedan atribut i to je *idNalaza*. Atribut *idNalaza* jedinstveni je atribut u koji se pohranjuje jedinstveni identifikacijski broj pojedinog nalaza. Svaki od nalaza povezan je s točno jednim pacijentom, dok svaki pacijent može biti povezan s bilo kojim brojem nalaza.

Entitet **GRUPANALAZA** ima četiri različita atributa. U atribut *idGrupa* pohranjuje se jedinstveni identifikator grupe nalaza. Atribut *sifraPretrage* pohranjuje šifru koja pripada određenoj pretrazi. U atribut *rezultat* pohranjuje se rezultat napravljene pretrage, a u atribut *mjernaJedinica* mjerna jedinica u kojoj je mjerena provedena pretraga. Svakoj grupi nalaza pripada točno jedan nalaz, a svakom od nalaza pripada barem jedna grupa nalaza.

Entitet **BILJESKE** ima dva atributa. Jedan je *idBiljeske* koji prikazuje jedinstveni identifikator svake liječničke bilješke te atribut *opis* koji sadrži bilješku u obliku slobodnog teksta koju je napravio liječnik tijekom pregleda. Svaka bilješka pripada točno jednom pacijentu, a svaki pacijent može imati nekoliko bilježaka, ali ne mora imati ni jednu.

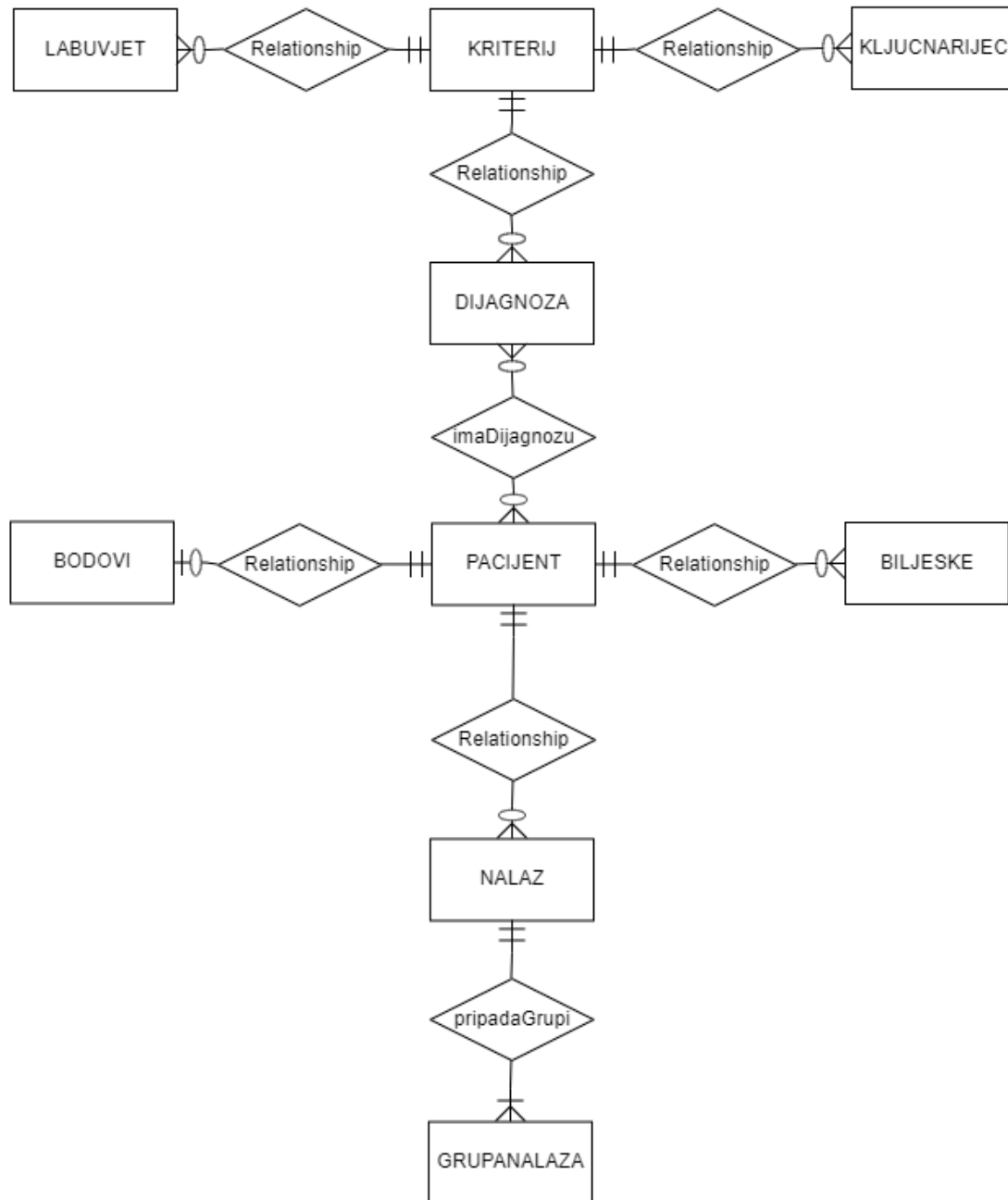
Entitet **DIJAGNOZA** prikazan je već na slici (Slika 5.1) i opisan u objašnjenju te slike. Na slici (Slika 5.2) grafički je prikazana veza između pacijenata i dijagnoza. Svaki pacijent ne mora imati dijagnozu, ali ih može imati i više. Isto tako dijagnoza ne mora biti prisutna ni kod jednog pacijenta, ali je moguće da bude prisutna kod više njih.

Entitet **BODOVI** ima četiri atributa. Ovaj entitet najvažniji je za rad algoritma i u njega se pohranjuje konačno bodovanje za svakog od pacijenata. Atribut *bodoviId* jedinstveni je atribut u koji se pohranjuje jedinstveni identifikacijski broj bodova dodjeljenih određenom pacijentu. Ima atribut *bodovi* u koji se sprema broj kriterija koje određeni pacijent zadovoljava. Određeni broj zadovoljenih kriterija uputit će na potrebu za daljnjim testiranjem. Atributi *createdOn* i *updatedOn* sadrže datume kada su bodovi za određenog pacijenta prvi put upisani u bazu te kada se njihov rezultat zadnji puta ažurirao.



Slika 5.2 ER model dijela baze relevantnog za razvoj algoritma

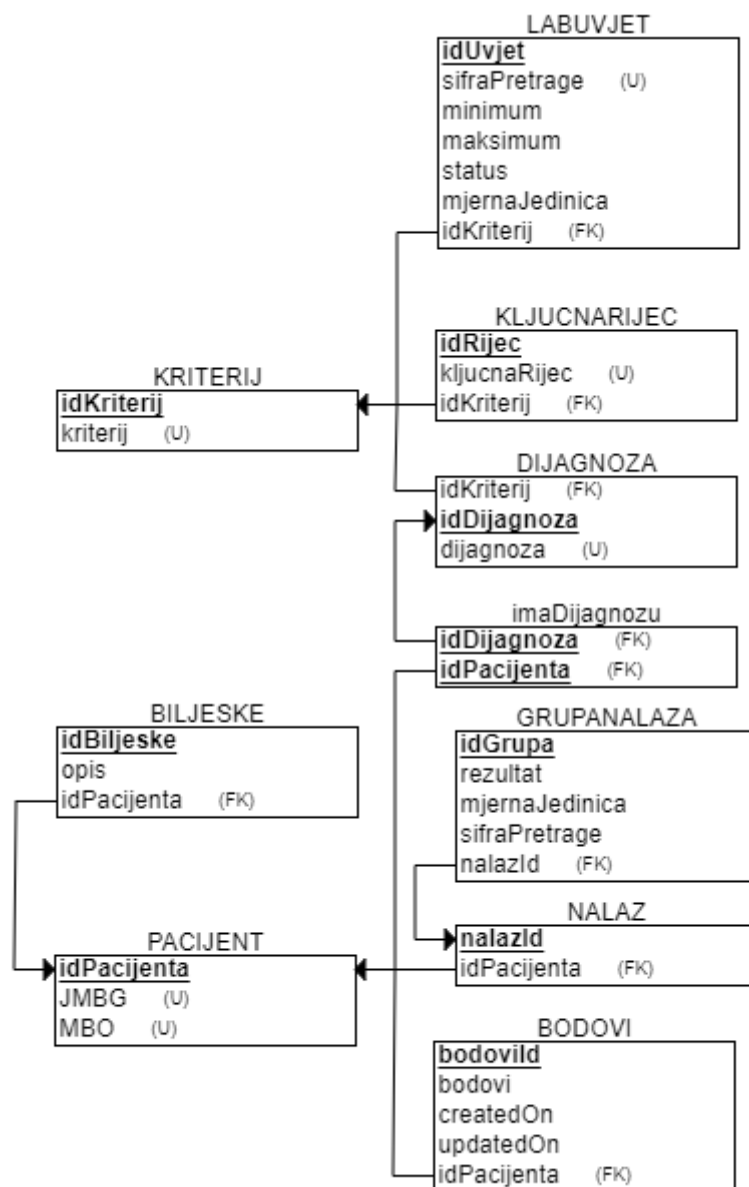
Na slici (Slika 5.3) prikazan je cijeli ER model dijela baze relevantnog za razvoj algoritma. U ovom modelu prikazani su svi entiteti bez pripadajućih atributa sa svim postojećim vezama.



Slika 5.3 ER model dijela baze relevantnog za razvoj algoritma na kojemu su prikazani svi entiteti

5.3. Relacijska shema djela baze relevantnog za istraživanje

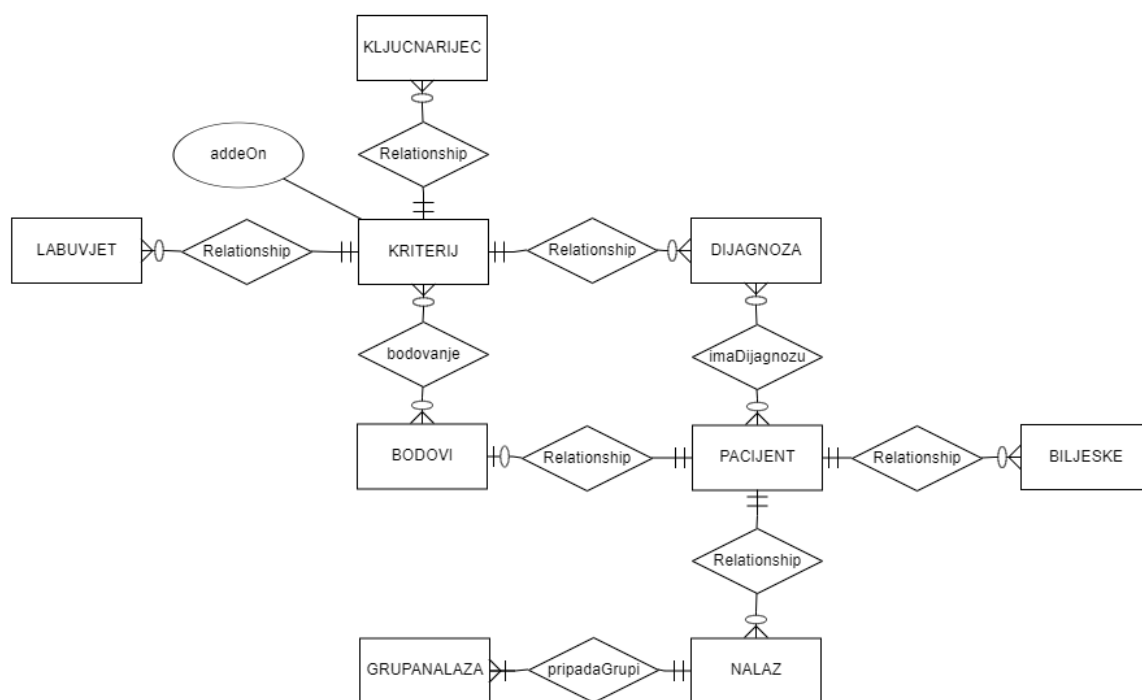
Na slici (Slika 5.4) prikazana je relacijska shema djela baze podataka relevantnog za istraživanje, napravljena u alatu ERDPlus.



Slika 5.4 Relacijska shema djela baze relevantnog za razvoj algoritma

5.4. Razrada – ER dijagram

U ovom potpoglavlju opisana je baza s mogućim unaprjeđenjima. Na slici (Slika 5.5) prikazano je kako se u bazi može pohraniti dodatna informacija o tome koji su to kriteriji donijeli bodove za određenog pacijenta. To može biti realizirano dodavanjem veze između entiteta **KRITERIJ** i entiteta **BODOVI**. Svim bodovima može pripadati više različitih kriterija te svakom kriteriju može pripadati više različitih bodova. Ovom vezom moguće je liječniku prikazati na temelju kojih pronađenih kriterija je algoritam došao do zaključka da je pacijent dobar kandidat za daljnje testiranje. U novo prikazanom ER modelu dodana je još jedna funkcionalnost. Entitetu **KRITERIJ** dodan je atribut *addedOn* u koji se upisuje datum od kada koji kriterij vrijedi.



Slika 5.5 ER model nadograđene baze podataka

Za ovakvo rješenje, pohrane informacije o tome koji je kriterij bodovan, ima neke redundantne podatke. Atribut *bodovi* u entitetu **BODOVI** redundantan je podatak jer je do te informacije moguće doći jednostavnim prebrojavanjem kriterija koji pripadaju određenom identifikacijskom broju bodova.

6. Analiza i ekstrakcija podataka iz specijalističkih nalaza

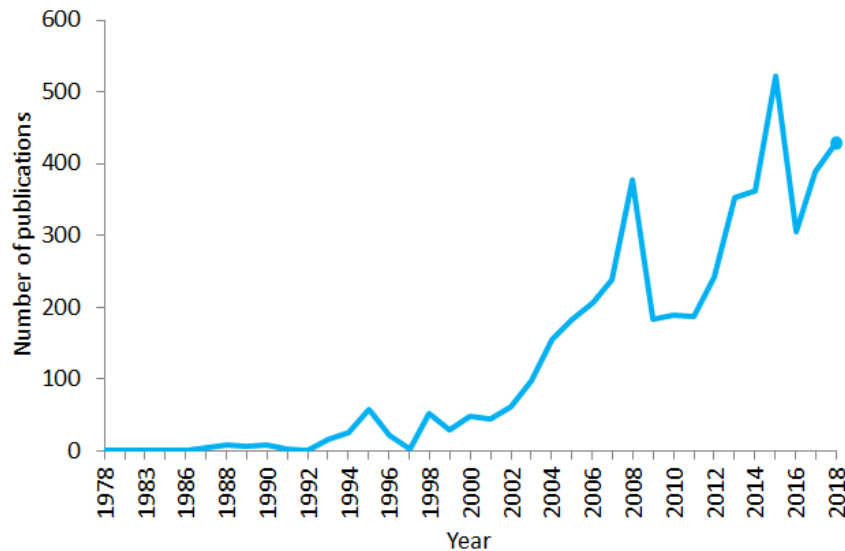
Veliki problem u provođenju ovog projekta predstavljala je analiza i ekstrakcija podataka iz specijalističkih nalaza, odnosno slobodnog teksta. Jako velika količina korisnih informacija nalazi se u slobodnom tekstu specijalističkih nalaza. To je korpus koji može biti vrlo izazovan za obradu. U nastavku ovog poglavlja opisane su neke od metoda kojima tekst navedenog oblika može bit pretvoren u obradive podatke te kako iz teksta izvaditi korisne informacije.

6.1. Obrada prirodnog jezika

Obrada prirodnog jezika (*engl. Natural Language Processing - NLP*) može biti opisana kao interakcija između računala i ljudi koji koriste prirodan jezik, naglasak je na sposobnosti računala da razumije nestrukturirani prirodni jezik [32]. Još jedan dobar opis NLP-a je da je to polje umjetne inteligencije koje strojevima daje mogućnost čitanja, razumijevanja i izvlačenja značenja iz ljudskog jezika [33]. Ovo područje ima brojne primjene poput automatskog prevođenja, sažimanja teksta, ekstrakcije informacija, filtriranja neželjene pošte, pisanja i uređivanja teksta te brojne druge. Sve od navedenog može biti korisno u raznim područjima poput financija, marketinga, odnosa s javnošću, nabavi i prodaji, kibernetičkoj sigurnosti, prometu i brojnim drugim [34].

Primjena NLP-a doživjela je poseban procvat i u zdravstvenoj industriji. Tome svjedoči i graf prikazan na slici (Slika 6.1) koji prikazuje porast broja PubMed³ publikacija koje u sebi sadrže sklop pojmova „obrada prirodnog jezika“ [33].

³ PubMed je besplatna baza podataka koja sadrži reference i sažetke o temama iz znanosti o životu i biomedicine



Slika 6.1 Broj PubMed publikacija koje u sebi sadrže sklop pojmova „obrada prirodnog jezika“

[33]

Iako NLP pronalazi brojne primjene u zdravstvenoj domeni, za problem opisan u ovom radu iznimno je važna mogućnost NLP-a za ekstrakciju informacija iz nestrukturiranih podataka kao što su specijalistički nalazi.

U nastavku rada razmotrit ćemo dva različita pristupa obrade prirodnog jezika.

- NLP modeli koji se temelje na treniranju anotiranih podataka
- NLP modeli za koji nije potrebna velika količina anotiranih podataka

NLP modeli temeljeni na treniranju podataka

Klasični NLP modeli zahtijevaju treniranje na velikim skupovima podataka. Proces uključuje nekoliko koraka:

- Tokenizacija: Razbijanje teksta na manje jedinice (riječi, rečenice).
- Lematizacija i stemming: Svođenje riječi na njihov osnovni oblik.
- Prepoznavanje entiteta (NER): Identifikacija i klasifikacija ključnih entiteta (npr. dijagnoze, simptomi).
- Parsiranje zavisnosti: Analiza gramatičke strukture rečenica.

Neki od klasičnih algoritama NLP-a:

- Bag of Words (BoW): Pretvaranje teksta u vektorski prostor bez obzira na redoslijed riječi.
- Term Frequency-Inverse Document Frequency (TF-IDF): Mjerenje važnosti riječi u dokumentu.

Glavni problem kod klasičnih modela poput Bag of Words, Term Frequency-Inverse Document Frequency i Naive Bayes je to što je za njihovo treniranje potrebna velika količina označenih podataka. U ovoj fazi projekta ne postoji baza podataka koja ima označeni slobodan tekst te bi njegovo stvaranje iziskivalo veliki broj resursa. Bilo bi potrebno zaposliti veliki broj stručnjaka čime bi se potrošilo vrijeme koje može biti potrošeno na važnije i kompleksnije zadatke te su potrebni veliki financijski resursi.

Zbog toga bi za ovaj tip problema mogli koristiti predtreniran QA (*engl. Question answering - QA*) model kako bi saznali je li iz određenih slobodnih tekstova moguće uspostaviti određenu dijagnozu. Odgovaranje na pitanja popularan je i izazovan zadatak u obradi prirodnog jezika, gdje je cilj pronaći najbolji odgovor na pitanje postavljeno na prirodnom jeziku. NLP modeli za odgovaranje na pitanja općenito mogu biti podijeljeni u dvije kategorije: **ekstraktivni** i **generativni** [35]. Njihova razlika ukratko je opisana u nastavku [35].

Ekstraktivni modeli

Ovi modeli nastoje pronaći odgovor kao dio teksta iz zadanog odlomka ili dokumenta koji odgovara postavljenom pitanju.

Generativni modeli

Ovi modeli stvaraju odgovor kao novi tekst koji sintetizira informacije iz više izvora.

Obije vrste modela koriste duboke neuronske mreže (*engl. deep neural network*), poput rekurentnih neuronskih mreža (*engl. recurrent neural network - RNN*), konvolucijskih neuronskih mreža ili transformera, kako bi interpretirale pitanje, analizirale izvor informacija i generirale odgovor na postavljeno pitanje [35].

Postoje brojni generativni QA modeli, ali u nastavku je prikazano kako koristeći Chat GPT-4 i Llama3, izvući informaciju o pronađenim kriterijima iz specijalističkih nalaza.

6.2. Specijalistički nalazi

Na slici (Slika 6.2) prikazan je umjetno generiran specijalistički nalaz. Nalaz je napravljen u Wordu, a sadržaj nalaza prikazuje na koji način liječnici obično opisuju dijagnoze i simptome kod određenog pacijenta. Na ovom primjeru bit će demonstriran rad generativnih QA modela.

Ime i prezime: Ivana Horvat
Spol: Ženski
Rođen: 11.08.2003.
Naručitelj: Ivan Horvat

SPECIJALISTIČKI NALAZ

ANAMNEZA

Pacijentica se žali na bolove u kostima. Ne osijeća pretjerani umor i nije prisutan zastoj u rastu. Napravljene su razne pretrage. Ustanovljeno je da pacijentica ima uvećau slezenu i povišeni IGG. Jetra je uredne veličine. Pacijentica navodi kako nije primijetila nastajanje modrica. Pije 10 godina dijagnosticirana joj je anemija za koju trenutno uzima propisanu terapiju.

Majka pacijentice boluje od Parkinsonove bolesti.

Slika 6.2 Umjetno generiran specijalistički nalaz

6.3. Pronalazak dijagnoza

Na slici (Slika 6.3) prikazano je kako Chat GPT-4, generativni model, iz umjetno generiranog specijalističkog nalaza prikazanog na slici (Slika 6.2) može zaključiti koji su kriteriji prisutni kod pacijenta. Bitno je napomenuti kako je ovakav tip modela otporan na pravopisne pogreške. Bez obzira na to što je napisano „uvećau“ umjesto „uvećanu“, model je prepoznao o kojem se simptomu radi, te uspješno detektira negacije u složenijim rečenicama. Model je ispravno detektirao sve prisutne kriterije. Bitno je napomenuti kako ovaj model sam prepoznaje koji su simptomi određenih kriterija, na primjer nije mu u upitu dana informacija kako je gamapatija opisana povišenim IGG-om niti da na asteniju upućuje umor. Model je već od prije znao tu informaciju.

Na osnovi danog specijalističkog nalaza napravi tablicu u kojoj jedan stupac sadrži kriterij, a drugi u obliku 0 i 1 informaciju je li taj kriterij pronađen u danom nalazu. Kriteriji su splenomegalija, trombocitopenija, bolovi u kostima, anemija, hepatomegalija, gamapatija, žuči kamenci, modrice, zastoj u rastu, astenija i Parkinsonova bolest u obiteljskoj anamnezi. Simptomi određenih kriterija opisani su u nastavku.

ANAMNEZA

Pacijentica se žali na bolove u kostima. Ne osjeća pretjerani umor i nije prisutan zastoj u rastu. Napravljene su razne pretrage. Ustanovljeno je da pacijentica ima uvećanu slezenu i povišeni IGG. Jetra je uredne veličine. Pacijentica navodi kako nije primijetila nastajanje modrica. Pije 10 godina dijagnosticirana joj je anemija za koju trenutno uzima propisanu terapiju. Majka pacijentice boluje od Parkinsonove bolesti.

< 2/2 >



Na temelju navedenog specijalističkog nalaza, tablica s kriterijima i odgovarajućim informacijama o tome jesu li pronađeni može izgledati ovako:

Kriterij	Pronađen (0 = ne, 1 = da)
Splenomegalija (uvećana slezena)	1
Trombocitopenija	0
Bolovi u kostima	1
Anemija	1
Hepatomegalija (uvećana jetra)	0
Gamapatija (povišeni IGG)	1
Žuči kamenci	0
Modrice	0
Zastoj u rastu	0
Astenija (umor)	0
Parkinsonova bolest u obitelji	1

Slika 6.3 Kako koristiti Chat GPT-3 za pronalazak kriterija u specijalističkom nalazu

Na slici (Slika 6.4) prikazano je kako Meta Llama 3.1, generativni model, iz umjetno generiranog specijalističkog nalaza prikazanog na slici (Slika 6.2) može zaključiti koji su kriteriji prisutni kod pacijenta. Rezultati su potpuno jednaki rezultatima koje je generirao Chat GPT-4 nisu ga smetale pravopisne pogreške i sam je znao koje su karakteristike astenije i gamapatije.

Chat with

Meta Llama 3.1 405B

NEW



Na osnovi danog specijalističkog nalaza napravi tablicu u kojoj jedan stupac sadrži kriterij, a drugi u obliku 0 i 1 informaciju je li taj kriterij pronađen u danom nalazu. Kriteriji su splenomegalija, trombocitopenija, bolovi u kostima, anemija, hepatomegalija, gamapatija, žuči kamenci, modrice, zastoj u rastu, astenija i Parkinsonova bolest u obiteljskoj anamnezi. Simptomi određenih kriterija opisani su u nastavku. ANAMNEZA Pacijentica se žali na bolove u kostima. Ne osijeća pretjerani umor i nije prisutan zastoj u rastu. Napravljene su razne pretrage. Ustanovljeno je da pacijentica ima uvećau slezenu i povišeni IGG. Jetra je uredne veličine. Pacijentica navodi kako nije primijetila nastajanje modrica.Pije 10 godina dijagnosticirana joj je anemija za koju trenutno uzima propisanu terapiju. Majka pacijentice boluje od Parkinsonove bolesti.



Tablica sa kriterijima i informacijom o prisutnosti svakog kriterija u danom nalazu:

| Kriterij | Prisutan (1) / Nije prisutan (0) |

| --- | --- |

| Splenomegalija | 1 |

| Trombocitopenija | 0 |

| Bolovi u kostima | 1 |

| Anemija | 1 |

| Hepatomegalija | 0 |

| Gamapatija | 1 |

| Žuči kamenci | 0 |

| Modrice | 0 |

| Zastoj u rastu | 0 |

| Astenija | 0 |

| Parkinsonova bolest u obiteljskoj anamnezi | 1 |

Slika 6.4 Kako koristiti Meta Llama 3.1 za pronalazak kriterija u specijalističkom nalazu

Iz gore prikazanih rezultata možemo zaključiti kako generativni QA modeli mogu s dosta dobrom točnošću u vrlo malom vremenu ispravno izvući potrebne informacije iz danog teksta. Zbog toga je ovo jedan od mogućih načina kako unaprijediti rad razvijanog modela. Plan je u budućnosti koristiti ovakve tehnologije za još uspješnije dijagnosticiranje kako Gaucherove tako i drugih rijetkih bolesti.

Bitno je napomenuti da, kako bismo bili u potpunosti sigurni pouzdanost i točnost ovih modela, potrebno je napraviti dodatno i ekstenzivno istraživanje na većem broju primjera.

7. Model za detekciju

U ovom poglavlju bit će opisane tehničke značajke razvijenog algoritma. Bit će opisan znanstveni rad na bazi kojeg je razvijen algoritam, svi kriteriji i parametri koji upućuju na prisutnost bolesti, alati koji su korišteni za razvoj te će biti prikazan razvijeni algoritam i njegove specifikacije.

7.1. Pregled rada po kojemu je napravljen algoritam

Razvoj algoritma za detekciju Gaucherove bolesti temeljio se na članku "Using the Gaucher Earlier Diagnosis Consensus (GED-C) Delphi Score in a Real-World Dataset" autora S. Revel-Vilk G. Chodick i suradnika [7]. Ovaj članak pruža detaljne smjernice za rano i ispravno dijagnosticiranje Gaucherove bolesti, rijetkog autosomno recesivnog stanja, koristeći GED-C Delphi bodovni sustav, koji je razvijen kroz konsenzus 22 stručnjaka u području Gaucherove bolesti. Glavni cilj navedenog rada bio je evaluirati GED-C bodovni sustav na nekolicini pacijenata kod kojih je potvrđena Gaucherova bolest. Istraživanje je provedeno na elektronskim znanstvenim podacima Izraelske zdravstvene ustanove, koja sadrži podatke o 3 milijuna pacijenata.

Ciljevi članka su razviti bodovni sustav koji će pomoći u ranom dijagnosticiranju Gaucherove bolesti. U radu su, kako je već navedeno koristili konsenzus stručnjaka o tome koje kliničke karakteristike i simptome treba promatrati za ranu dijagnozu Gaucherove bolesti. Na osnovu njihovih prijedloga razvijen je bodovni sustav koji svakom simptomu dodjeljuje određeni broj bodova ovisno o njegovoj značajnosti za dijagnozu.

Razvijeni bodovni sustav je testiran na stvarnim podacima i pokazao se učinkovitim u prepoznavanju pacijenata koji zahtijevaju daljnje testiranje.

Na bazi opisanog rada s manjim izmjenama razvijena je baza simptoma koje je potrebno promatrati te je definirano bodovanje. Svi promatrani simptomi i njihove karakteristike opisani su u nastavku.

7.2. Kriteriji

U tablici (Tablica 7.1) prikazani su svi kriteriji promatrani za donošenje odluke je li neka osoba kandidat za daljnje testiranje.

Tablica 7.1 Promatrani kriteriji

KriterijID	Kriterij
1	spleenomegalija
2	trombocitopenija
3	bol u kostima, avaskularne nekroze, frakture
4	anemija
5	hiperferitinemija
6	hepatomegalija
7	gamapatija
8	žučni kamenci
9	krvarenje, modrice, koagulopatija
10	leukopenija
11	niska mineralna gustoća kostiju
12	zastoj u rastu
13	astenija
14	dislipidemija
15	povišene razine angiotenzin-konvertirajućeg enzima
16	Parkinsonova bolest u obiteljskoj anamnezi

U tablici (Tablica 7.2) prikazani su kriteriji sa svojim pripadajućim MKB-10 dijagnozama. Ako pacijent ima određenu dijagnozu njezin kriterij bit će brojan u konačnom broju zadovoljenih uvjeta za daljnje testiranje za Gaucherovu bolest.

Tablica 7.2 MKB-10 šifre

Kriterij	MKB-10 šifre
spleenomegalija	B27; B27.9, B27.1, B27.8, B27.9, R16.1
trombocitopenija	D69.4, D69.5, D69.6
bol u kostima, avaskularne nekroze, frakture	M87.0, M87.9
anemija	D64.8, D64.9
hepatomegalija	R16.0, R16.2
gamapatija	D89.0, D89.2, D47.2
žučni kamenci	K80.2, K80.5, K80.8, K82.9
krvarenje, modrice, koagulopatija	D68.9
niska mineralna gustoća kostiju	M81.0, M81.8, M81.9
zastoj u rastu	R62, R62.0, R62.8, R62.9
astenija	R53, R53.0
dislipidemija	E78.6
Parkinsonova bolest u obiteljskoj anamnezi	G20, G21*, G25*

U tablici (Tablica 7.3) prikazani su kriteriji sa svojim pripadajućim ključnim riječima. Ključne riječi traže se u bilješkama liječnika i ako je pronađena pripadajuća ključna riječ njen kriterij bit će brojan kao zadovoljeni uvjet.

Tablica 7.3 Ključne riječi

Kriterij	Ključne riječi
spleenomegalija	Povećana slezena Uvećana slezena Splenomegalija Splenektomija
trombocitopenija	trombocitopenija
bol u kostima, avaskularne nekroze, frakture	Bol u kostima Idiopatska aseptična nekroza kosti Osteonekroza
anemija	anemija
hepatomegalija	Povećana jetra Uvećana jetra Hepatomegalija
gamapatija	Povišeni IGG Gamaglobulinemija Gamapatija
žučni kamenci	Žučni kamenac Kolecistektomija
krvarenje, modrice, koagulopatija	Modrice Produženi APTV Povišeni APTV Krvarenje Koagulopatija
niska mineralna gustoća kostiju	Osteoporoza Bolest kostiju Prijelom kostiju Prijelom kuka Slomljena kost Ozljeda
zastoj u rastu	Zastoj u rastu Usporen rast
astenija	Umor Slabost
povišene razine angiotenzin-konvertirajućeg enzima	Povišeni ACE
Parkinsonova bolest u obiteljskoj anamnezi	Parkinson u obitelji

U tablici (Tablica 7.4) prikazani su svi laboratorijski uvjeti potrebni za odrediti je li određeni kriterij zadovoljen ili nije. Prikazani uvjeti uspoređuju se s provedenim laboratorijskim pretragama.

Tablica 7.4 Laboratorijski uvjeti

Kriterij	Laboratorijski uvjeti
trombocitopenija	1-1010080 - K - Trombociti <= 140
anemija	1-1010020 K - Hemoglobin - ispod referentne vrijednosti
leukopenija	1-1010100 K - Leukociti - snižena referentna vrijednost
povišene razine angiotenzin-konvertirajućeg enzima	1-2080040 S - HDL-kolesterol - sniženi HDL

7.3. Korišteni alati

Model je napravljen koristeći SQL Server Management Studio. SQL Server Management Studio za upravljanje relacijskim bazama podataka je sustav koji omogućava izvođenje složenih SQL upita te analizu velikih skupova podataka.

7.4. Implementacija

Model je zapravo SQL procedura koja nizom SQL upita u bazu pohranjuje broj kriterija koje zadovoljava određeni pacijent. Trebalo je dobro razmisliti i razraditi sustav kako brojati kriterije samo jednom. Moguć je pronalazak više različitih ključnih riječi, nekoliko dijagnoza i snižene laboratorijske vrijednosti za isti kriterij. To je riješeno tako da se za svaki pronađeni pokazatelj sprema samo id kriterija kojemu pripadaju.

Proceduru možemo podijeliti na tri glavna djela. U jednom SQL upitom pretražujemo sve slobodne tekstove u bazi te radimo provjeru nalazi li se i u jednom od njih neka od traženih ključnih riječi. Ako postoje, u privremenu tablicu potrebno je pohraniti id pacijenta i šifru kriterija kojeg opisuje pronađena ključna riječ, dijagnoza ili stršeci laboratorijski nalaz. Kod pretrage ključnih riječi bilo je potrebno paziti na velika i mala slova te mogućnost da određene riječi budu sadržane u drugim riječima, kao što su naprimjer umor i tumor. Problem velikih i malih slova riješen je tako što su sva slova pretvorena u mala slova koristeći funkciju LOWER(). Problem pronalaska ključnih riječi u dijelu drugih dužih riječi riješen je dodavanjem razmaka na početku svih traženih pojmova. U drugom djelu

pretražuju se dijagnoze. Za svakog pacijenta s aktivnom dijagnozom koja je karakteristična za GB, u privremenu tablicu sprema se id kriterija kojeg opisuje uspostavljena dijagnoza. Treći dio radi na istom principu, ali ne gleda postojanje ključnih riječi i prisutnost određenih dijagnoza već promatra rezultate provedenih pretrag - jesu li rezultati unutar rizičnog intervala.

Na slikama (Slika 7.1 i Slika 7.2) prikazana je napravljeni procedura.

```

CREATE OR ALTER PROCEDURE bodoviZaMBO
AS
BEGIN
-- Temporary tablica #labRezultati u kojoj su pohranjeni jmbg,
-- kriterijiID koji zadovoljavaju kriterije za Gaucherovu bolest
SELECT DISTINCT bodovanje.jmbg, bodovanje.KriterijID
INTO #labRezultati
FROM (
  (SELECT DISTINCT rezultatiTablica.jmbg, glnu.KriterijID
  FROM (
    (SELECT gn.PacijentID AS jmbg, gnp.sifra, TRY_CAST(REPLACE(gnp.Rezultat, ',', '.')) AS FLOAT)
    AS rezultat, gnp.status
    FROM [dbo].[GrupiraniNalaz] gn
    JOIN GrupiraniNalazPretraga gnp ON gn.GrupiraniNalazID = gnp.GrupiraniNalazID)
  UNION ALL
  (SELECT jmbg, sifra_pretrage, TRY_CAST(REPLACE(rezultat, ',', '.')) AS FLOAT) AS rezultat, status
  FROM [dbo].[POVIJEST_NALAZI] pn
  JOIN [dbo].[POVIJEST_NALAZI_LAB_REZULTATI] pnl ON pnl.nalaz_id = pn.id
  JOIN MAMM_KATALOG_PRETRAGA_STAVKE m ON m.pretraga_id = pnl.Pretraga_id)
  UNION ALL
  (SELECT jmbg, mamm_sifra_pretrage AS sifra, TRY_CAST(REPLACE(rezultat, ',', '.')) AS FLOAT)
  AS rezultat, urr.komentar AS status
  FROM [dbo].[UPUTNICE_REZULTATI] ur
  JOIN [dbo].[UPUTNICE_REZULTATI_GRUPE] urg ON ur.id = urg.uputnice_rezultati_id
  JOIN UPUTNICE_REZULTATI_REZULTATI_POSTUPKA urr ON urr.uputnice_rezultati_grupe_id = urg.id
  JOIN HKMB_MAMM hm ON hm.hkmb_id = urr.SifraRezultata)
  ) AS rezultatiTablica
  JOIN GAUCHER_LAB_NALAZ_UVJETI glnu ON rezultatiTablica.sifra = glnu.Sifra
  AND (
    (rezultatiTablica.rezultat > glnu.minimum AND rezultatiTablica.rezultat < glnu.maksimum)
    OR rezultatiTablica.status = glnu.Status
  )
  )
  )

UNION ALL
-- Pronalazak aktivnih dijagnoza za pacijente
(SELECT DISTINCT s.jmbg, gdi.KriterijID
FROM SLUCAJ s
JOIN GAUCHER_DIJAGNOZE gdi ON s.glavna_dijagnoza_id = gdi.Sifra
WHERE s.status_slucaja NOT IN ('5', '9'))
UNION ALL
-- Pronalazak ključnih riječi koje se pojavljuju u tekstualnim nalazima pacijenata
(SELECT DISTINCT tekstTablica.jmbg, gkr.KriterijID
FROM (
  SELECT jmbg, LOWER(CONCAT_WS(' ', status, anamneza, terapija, preporuka, dijagnoza_lat, nalaz))
  AS tekst
  FROM [Gaucher_DB-test].[dbo].[PRIMARNA_POVIJEST]
  UNION ALL
  SELECT jmbg, LOWER(nalaz) AS tekst
  FROM [Gaucher_DB-test].[dbo].[POVIJEST_NALAZI]
  ) AS tekstTablica
  JOIN GAUCHER_KLJUCNE_RIJECI gkr ON tekstTablica.tekst LIKE '%' + gkr.KljucnaRijec + '%'
  WHERE EXISTS (
    SELECT 1
    FROM [GAUCHER_KLJUCNE_RIJECI]
    WHERE tekstTablica.tekst LIKE '%' + [GAUCHER_KLJUCNE_RIJECI].KljucnaRijec + '%'
  )
  )
) AS bodovanje;

```

Slika 7.1 Prvi dio procedure

```

MERGE PacijentiBodovi AS Dest
USING (
    SELECT p.hzzo_id AS mbo,
           COUNT(DISTINCT lr.KriterijID) AS bodovi,
           GETDATE() AS CreatedOn
    FROM PACIJENTI p
    LEFT JOIN #labRezultati lr ON p.jmbg = lr.jmbg
    WHERE p.hzzo_id IS NOT NULL
    GROUP BY p.hzzo_id
) AS src
ON Dest.mbo = src.mbo
WHEN MATCHED THEN
    UPDATE SET Dest.bodovi = src.bodovi,
              Dest.UpdatedOn = GETDATE()
WHEN NOT MATCHED THEN
    INSERT (mbo, bodovi, CreatedOn, UpdatedOn)
    VALUES (src.mbo, src.bodovi, src.CreatedOn, GETDATE());
END;

```

Slika 7.2 Drugi dio procedure

8. Rasprava

U ovom poglavlju dan je kratak osvrt na to što je moglo biti drugačije odrađeno, odnosno što bi to omogućilo da projekt bude kvalitetniji i bolji. Gaucher screening projekt ima veliki potencijal za daljnje unaprjeđenje. U poglavlju 6 već je detaljno opisano kako se ovaj projekt može poboljšati **korištenjem specijalističkih podataka** kao izvorom informacija uz pomoć obrade prirodnog jezika. No, osim toga, postoje još brojni načini kako poboljšati razvijeni model za identifikaciju pacijenata u riziku. Neki od načina da se projekt poboljša su **prevladavanje izazova vezanih uz dostupnost podataka, promjena određenih parametara** nakon dobivenih prvih povratnih informacija o tome kakve su performanse algoritma te **izgradnja novih algoritama** strojnog učenja kada se uspije izgraditi kvalitetna ulazna baza podataka. Važno je napomenuti kako se s vremenom očekuje unaprjeđenje podataka s čime će cijeli sustav postati bolji.

8.1. Izazovi s dostupnošću i obradom podataka

Projekt bi bio bolje izveden da nije bilo izazova s obradom podataka. Nisu se svi podaci mogli obrađivati kako je bilo poželjno, ponajprije zbog odredba propisanih u GDPR-u. Također algoritam bi imao puno bolje performanse kada bi skup podataka bio veći te ne samo iz jednog doma zdravlja nego i iz drugih izvora, kako iz Hrvatske tako i svijeta. Idealna situacija bila bi kada bi početni skup podataka sadržavao informaciju o tome boluje li neka osoba od Gaucherove bolesti ili ne te da ima adekvatan broj osoba s dijagnosticiranom bolesti. Takav skup podataka omogućio bi da algoritam ima veću točnost te bi nam omogućilo korištenje nekih drugih metoda pronalaska pacijenata u riziku.

8.2. Validacija i unaprjeđenje algoritma

Kao što je već navedeno, validacija algoritma bit će provedena kroz izvještaj dobiven od Doma zdravlja Zagreb – Centar. Izvještaj će se okvirno sastojati od podataka o tome koliko je pacijenata algoritam detektirao, koliko je njih poslano na testiranje nakon liječničke provjere te koliko ih zaista ima Gaucherovu bolest. Kasnije, kada ta dijagnoza bude unesena u sustav bit će moguća daljnja obrada podataka i nadogradnja algoritma.

Još jedna velika prednost bila bi kada bi unutar izvještaja bili dostupni detaljniji podatci i informacija tko je točno, a tko netočno klasificiran. Naravno potrebno je provjeriti sa stručnjacima za GDPR koje podatke će uistinu biti moguće koristiti. Informacija o lažno pozitivnim i lažno negativnim rezultatima bila bi korisna za daljnje prilagođavanje parametara algoritma. Također bi bila korisna i informacija od liječnika opće prakse zašto je određenog pacijenta odlučio poslati na daljnje pretrage, a nekog drugog ne.

8.3. Izgradnja novih algoritama

Jedan od najboljih načina za dodatno poboljšati model za identifikaciju pacijenata u riziku je korištenjem različitih algoritama strojnog učenja, koji s velikom točnošću mogu klasificirati razne bolesti. Preduvjet za korištenje takvih algoritama je stvaranje konkretnog podatkovnog skupa koji se odnosi na populaciju nama zanimljivog područja. Idealno bi bilo da se validacijom napravljenog algoritma pronađe dovoljno pacijenata s Gaucherovom bolesti te da se uspije napraviti označeni skup podataka na kojemu bi se mogli trenirati modeli strojnoga učenja.

Neki od algoritama koji bi mogli biti uspješni u dijagnosticiranu Gaucherove bolesti su algoritam slučajne šume (engl. Random Forest), naivni Bayes (engl. Naive Bayes), algoritam k-najbližih susjeda (engl. k-Nearest Neighbors – KNN) te brojni drugi. Uspješnost različitih algoritma za dijagnosticiranje rijetkih bolesti istražena je i prikazana u radu [36].

Zaključak

Razvoj metodologije i algoritma za detekciju Gaucherove bolesti, opisani u ovom radu, imaju važnu ulogu u unaprjeđenju dijagnosticiranja rijetkih bolesti. Diplomski rad napravljen je u sklopu projekta koji je uključivao sve relevantne dionike. Projekt realiziran u suradnji s globalnom biofarmaceutskom tvrtkom Takeda, MCS grupom te Domom zdravlja Zagreb – Centar pokazuje kako moderna tehnologija i njen neprestani razvoj mogu značajno unaprijediti dijagnosticiranje i liječenje raznih medicinskih stanja. Metodologija i algoritam osmišljeni su kako bi se korištenjem brojnih medicinskih podataka liječnicima pruži alat za brže, preciznije i jednostavnije identificiranje osoba koji su kandidati za daljnje testiranje za Gaucherovu bolest.

U ovom radu dana je motivacija, prikazana je metodologija te određeni konkretni koraci za razvoj modela za identifikaciju pacijenata koji imaju Gaucherovu bolest. Iako za vrijeme izrade i pisanja diplomskog rada nije bilo moguće verificirati i valorizirati algoritam, razrađeni su i prikazani daljnji koraci koji se očekuju nakon početka upotrebe razvijenog algoritma. Također, jasno su prikazane mogućnosti unaprjeđenja.

Uvođenje naprednih tehnologija u zdravstvenu domenu donosi brojne prednosti. Između ostalog, smanjuje se vrijeme potrebno za postavljanje ispravne dijagnoze, što je kritično za veliki broj bolesti kod kojih pravovremena dijagnoza može značajno poboljšati ishode liječenja. Algoritmi poput ovdje prikazanoga, mogu pomoći prepoznati pacijente koji potencijalno ne bi bili identificirani kroz standardne dijagnostičke procese.

Projekt u sklopu kojeg je nastao ovaj diplomski rad, ne samo da pridonosi digitalizaciji hrvatskog zdravstvenog sustava, već postavlja i temelje za poboljšanje dijagnostičkih procesa i liječenje raznih bolesti. Korištenjem modernih tehnologija i alata, moguće je značajno poboljšati zdravstvenu skrb. Iz svega navedenog očito je kako spomenuti projekt i ovaj diplomski rad imaju potencijal značajno unaprijediti zdravstvo i postati primjer za slične inicijative u budućnosti.

Literatura

- [1] *Gaucher Screening in Croatia*, Project deescription, October 2023., Takeda
- [2] M. Merkler, I. Šimić, I. Pećin, et al., *Gaucherova bolest – smjernice za dijagnozu i liječenje odraslih bolesnika*, Liječnički vjesnik, Vol. 136 No. 5-6, 2014.
- [3] Takeda, *Tko smo mi, o nama* <https://www.takeda.com/hr-hr/tko-smo-mi/o-nama>
- [4] *O MCS Grupi*, <https://mcs.hr/o-nama/>
- [5] Dom zdravlja Zagreb – Centar, *O nama*, <https://dzz-centar.hr/o-nama/>
- [6] Fakultet elektrotehnike i računarstva
https://hr.wikipedia.org/wiki/Fakultet_elektrotehnike_i_ra%C4%8Dunarstva
- [7] Revel-Vilk S.; Chodick G.; Shalev V.; Lotan R.; Zarakowska K.; Gadir N. *Using the Gaucher Earlier Diagnosis Consensus (GED-C) Delphi Score in a Real-World Dataset*. Int. J. Transl. Med. 2022, 2, 506-514.
- [8] Revel-Vilk, S.; Szer, J.; Zimran, A. *Gaucher disease and related lysosomal storage diseases*. In Williams Hematology, 10th ed.; Kaushansky, K., Lichtman, M., Prchal, J., Levi, M., Burns, L.J., Linch, D., Eds.; McGraw-Hill: New York, NY, USA, 2021.
- [9] Stolk, P. , Willemen, M.J. & Leufkens, H.G. *Rare essentials: drugs for rare diseases as essential medicines*. Bull. World Health Organ. 84, 745–751 (2006).
- [10] Dawkins HJS, Draghia-Akli R, Lasko P, Lau LPL, Jonker AH, Cutillo CM, Rath A, Boycott KM, Baynam G, Lochmüller H, Kaufmann P, Le Cam Y, Hivert V, Austin CP; *International Rare Diseases Research Consortium (IRDiRC)*. *Progress in Rare Diseases Research 2010-2016: An IRDiRC Perspective*. Clin Transl Sci. 2018 Jan;11(1):11-20, Epub 2017 Oct 23. PMID: 28796411; PMCID: PMC5759730.
- [11] Tambuyzer, E. *Rare diseases, orphan drugs and their regulation: questions and misconceptions*. Nat. Rev. Drug Discov. 9, 921–929 (2010)
- [12] *Rijetke bolesti*, HRVATSKI SAVEZ ZA RIJETKE BOLESTI, <https://rijetke-bolesti.com/rijetke-bolesti/>
- [13] Ellen Sidransky, MD; Chief Editor: Maria Descartes, MD, *Gaucher Disease*, Medscape, Oct 05, 2023.
- [14] Atul Mehta, *Epidemiology and natural history of Gaucher's disease*, European Journal of Internal Medicine, Volume 17, Supplement, 2006, Pages S2-S5, ISSN 0953-6205
- [15] InalJDHFKJk
- [16] Stone WL, Basit H, Master SR. *Gaucher Disease*. StatPearls. 2023 Jan.
- [17] Gary, S. E., Ryan, E., Steward, A. M., & Sidransky, E. (2018). *Recent advances in the diagnosis and management of Gaucher disease*. Expert Review of Endocrinology & Metabolism, 13(2), 107–118

- [18] Stirnemann J, Belmatoug N, Camou F, Serratrice C, Froissart R, Caillaud C, Levade T, Astudillo L, Serratrice J, Brassier A, et al. *A Review of Gaucher Disease Pathophysiology, Clinical Presentation and Treatments*. International Journal of Molecular Sciences. 2017; 18(2):441.
- [19] EduBridge, <https://www.edubridgeindia.com/blog/wp-content/uploads/2022/02/Data-Analyst-Tool-Blog-.jpg>
- [20] R. Sheldon, *What is a data set?*, TechTarget, <https://www.techtarget.com/whatis/definition/data-set>
- [21] *Raspodjeljena pohrana velike količine strukturiranih podataka*, Raspodijeljena obrada velike količine podataka, FER, 21.3.2016
- [22] Praveen, Shagufta & Chandra, Umesh. (2020). *Influence of Structured, Semi-Structured, Unstructured data on various data models*. International Journal of Scientific and Engineering Research. 8. 67-69.
- [23] E. de Haan, M. Padigar, S. El Kihal, R. Kübler, J. E. Wieringa, *Unstructured data research in business: Toward a structured approach*, *Journal of Business Research*, Volume 177, 2024, 114655, ISSN 0148-2963
- [24] *Healthcare Unstructured Data*, truenorth, <https://www.truenorthitg.com/healthcare-unstructured-data/>
- [25] *baza podataka*. Hrvatska enciklopedija, mrežno izdanje. Leksikografski zavod Miroslav Krleža, 2013. – 2024. Pristupljeno 12.8.2024.
- [26] *Osnove baza podataka*, Access za Microsoft 365 Access 2021
- [27] Ming-Syan Chen, Jiawei Han and P. S. Yu, "Data mining: an overview from a database perspective," in IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883, Dec. 1996, doi: 10.1109/69.553155.
- [28] *Elektronički zdravstveni zapisi i umjetna inteligencija: utakmica napravljena na nebu*, shaip, 14.09.2021.
- [29] Šimundić, Ana-Maria, *laboratorijski informacijski sustavi // Medicinska informatika / Kern, Josipa ; Petrovečki, Mladen (ur.)*. Zagreb: Medicinska naklada, 2009. str. 89-100
- [30] *LIS – Laboratorijski informacijski sustav*, POLIKLINIKA ANALIZA LAB <https://www.poliklinika-analizalab.hr/o-nama/lis-laboratorijski-informacijski-sustav/>
- [31] *Radiološki informacijski sustav (RIS i PACS)*, Sve što vas zanima o radiologiji <https://radiologija-hr.blogspot.com/2007/11/radioloki-informacijski-sustav-ris-i.html>
- [32] Adam Bohr, Kaveh Memarzadeh, Chapter 2 - *The rise of artificial intelligence in healthcare applications*, Artificial Intelligence in Healthcare, Academic Press, 2020
- [33] D. Lopez, *Your Guide to Natural Language Processing (NLP)*, Data Science Central, 17.01.2019.
- [34] *Which Industries Are Making the Most of NLP Applications?*, RAI <https://revolveai.com/nlp-applications-in-different-industries/>
- [35] *What are the best NLP models for question answering?*, Powered by AI and the LinkedIn community, <https://www.linkedin.com/advice/0/what-best-nlp-models-question-answering-skills-machine-learning-rzbgf>

- [36] P. Roman-Naranjo, A.M. Parra-Perez, J.A. Lopez-Escamez, *A systematic review on machine learning approaches in the diagnosis and prognosis of rare genetic diseases*, Journal of Biomedical Informatics, Volume 143, 2023, 104429, ISSN 1532-0464

Sažetak

Razvoj algoritma za detekciju Gaucherove bolesti

Rijetke bolesti, poput Gaucherove bolesti, pogađaju do 5 osoba na 10.000 ljudi i često su opasne po život. Njihovo dijagnosticiranje traje prosječno 5 godina što uvelike smanjuje šansu za uspješno liječenje. U Hrvatskoj je Gaucherova bolest dijagnosticirana u samo 34 % slučajeva. Razlog tome su nedostatak znanja i stručnjaka vezanih za ovu bolest te visoka cijena potrebnih pretraga. U sklopu ovog rada razvijen je algoritam koji, na temelju raznih podataka o pacijentu, pomaže liječnicima brže uspostaviti ispravnu dijagnozu. Ovime se povećavaju šanse za pravovremeni početak liječenja što će poboljšati kvalitetu života oboljelih i unaprijediti digitalizaciju zdravstva u Hrvatskoj.

Ključne riječi: relacijska baza podataka, Microsoft SQL Server, rijetke bolesti, Gaucherova bolest, klasifikacijski algoritam, digitalizacija zdravstva

Summary

Development of an algorithm for the detection of Gaucher disease

Rare diseases, such as Gaucher disease, affect up to 5 in 10 000 people and are often life-threatening. Their diagnosis takes an average of 5 years, which greatly reduces the chance of successful treatment. In Croatia, Gaucher's disease is diagnosed in only 34 % of expected cases. The reason for this is the lack of knowledge and experts related to this disease and the high cost of the tests needed establish the diagnosis. As part of this work, an algorithm was developed that, based on various patient data, helps doctors establish the correct diagnosis more quickly. This increases the chances for a timely start of treatment, which will improve the quality of life of patients and advance the digitalization of health care in Croatia.

Keywords: relational database, Microsoft SQL Server, rare diseases, Gaucher's disease, classification algorithm, digitalization of health care

Skraćenice

GDPR	<i>General Data Protection Regulation</i>	Opća uredba o zaštiti podataka
GB	<i>Gaucherova bolest</i>	Gaucherova bolest
SQL	<i>Structured Query Language</i>	Strukturni upitni jezik
AI	<i>Artificial Intelligence</i>	Umjetna inteligencija
ML	<i>Machine Learning</i>	Strojno učenje
NLP	<i>Natural Language Processing</i>	Obrada prirodnog jezika