Obrnuti trio-binning pomoću strojnog učenja

Grbelja, Roko

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 411

OBRNUTI TRIO-BINNING POMOĆU STROJNOG UČENJA

Roko Grbelja

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 411

OBRNUTI TRIO-BINNING POMOĆU STROJNOG UČENJA

Roko Grbelja

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

DIPLOMSKI ZADATAK br. 411

Pristupnik:	**Roko Grbelja (0036515044)**
Studij:	Računarstvo
Profil:	Računarska znanost
Mentor:	doc. dr. sc. Krešimir Križanović

Zadatak: **Obrnuti trio-binning pomoću strojnog učenja**

Opis zadatka:

Ljudski genom je diploidan, što znači da se sastoji od dva haploidna genoma: majčinog i očevog. Prilikom sastavljanja genoma moguće je, osim diploidnog, sastaviti i haploidni genom-kombinaciju majčinog i očevog. Kod sastavljanja diploidnog genoma, podaci dobiveni sekvenciranjem majčinog i očevog genoma mogu se koristiti za razdvajanje haplotipova. Međutim, u slučaju nedovoljne količine podataka, te je podatke moguće iskoristiti kao dodatne podatke za sastavljanje haploidnog genoma. Budući da su majčin i očev genom također diploidni, pri tome je potrebno koristiti samo ispravna očitanja koja odgovaraju djetetovom genomu. Cilj ovog rada je osmisliti metodu koja će iz sekvenciranih podataka majčinog i očevog genoma izdvojiti samo ona očitanja koja odgovaraju djetetovom genomu. Pri tome je potrebno koristiti tehnike strojnog i dubokog učenja. Podatke za učenje i testiranje generirati s pomoću simulatora. Rješenje treba biti napisano kao Jupyter Notebook skripta. Svaki korak postupka iscrpno komentirati.

Rok za predaju rada: 28. lipnja 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 411

OBRNUTI TRIO BINNING POMOĆU STROJNOG UČENJA

Roko Grbelja

Zagreb, Rujan 2024.

SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

DIPLOMSKI ZADATAK br. 411

Pristupnik:	**Roko Grbelja (0036515044)**
Studij:	Računarstvo
Profil:	Računarska znanost
Mentor:	doc. dr. sc. Krešimir Križanović

Zadatak: **Obrnuti trio binning pomoću strojnog učenja**

Opis zadatka:

Ljudski genom je diploidan, što znači da se sastoji od dva haploidna genoma: majčinog i očevog. Prilikom sastavljanja genoma moguće je, osim diploidnog, sastaviti i haploidni genom-kombinaciju majčinog i očevog. Kod sastavljanja diploidnog genoma, podaci dobiveni sekvenciranjem majčinog i očevog genoma mogu se koristiti za razdvajanje haplotipova. Međutim, u slučaju nedovoljne količine podataka, te je podatke moguće iskoristiti kao dodatne podatke za sastavljanje haploidnog genoma. Budući da su majčin i očev genom također diploidni, pri tome je potrebno koristiti samo ispravna očitanja koja odgovaraju djetetovom genomu. Cilj ovog rada je osmisliti metodu koja će iz sekvenciranih podataka majčinog i očevog genoma izdvojiti samo ona očitanja koja odgovaraju djetetovom genomu. Pri tome je potrebno koristiti tehnike strojnog i dubokog učenja. Podatke za učenje i testiranje generirati s pomoću simulatora. Rješenje treba biti napisano kao Jupyter Notebook skripta. Svaki korak postupka iscrpno komentirati.

Rok za predaju rada: 28. lipnja 2024.

Izrada ovog rada plod je 6 godina školovanja, mnogih neprospavanih noći i neiscrpne podrške moje žene Amalije, za koju bih mogao napisati još jedan rad od 38 stranica pun zahvala. Hvala ti.

Sadržaj

Uvod

Human Genome Project (HGP), jedan od najambicioznijih i najvažnijih znanstvenih pothvata u povijesti, započet je 1990. godine s ciljem mapiranja i razumijevanja svih gena ljudskog genoma. Projekt je završen 2003. godine, a njegovi rezultati donijeli su revoluciju u biomedicinskim istraživanjima. Otkrivanje kompletne sekvence ljudske DNK omogućilo je istraživačima da identificiraju gene odgovorne za različite bolesti, unaprijede metode dijagnostike, te razviju ciljane terapije prilagođene individualnim genetskim profilima pacijenata.

Originalni projekt koštao je 2.7 milijarde američkih dolara, a za mapiranje većine ljudskog genoma trebalo je više od dvije godine. Danas se isti ili bolji rezultati mogu dobiti u nekoliko tjedana, za prosječnu cijenu od 600 dolara, a daljnji napredci u tehnologiji obećavaju da bi se ta cijena mogla smanjiti i do 200 dolara po sekvenci (Cara, 2023.).

Iako HGP i dalje stoji kao miljokaz ljudskog postignuća na području genomike, trenutno najkompletniji i najprecizniji referentni ljudski genom ikad proizveden (GRCh38) i dalje ima stotine praznina i rupa na teško odredivim lokacijama na genomu. Telomere-to-Telomere (T2T) consortium je međunarodna kolaboracija znanstvenika i istraživača čiji je cilj sastaviti ljudski genom u potpunosti, od kraja do kraja, uključujući i regije koje su inače teške za sekvenciranje regije kao što su telomeri (krajevi kromosoma) i centromeri (središnji dijelovi kromosoma). U ovu svrhu, koriste se mnoge napredne tehnike sekvenciranja i sastavljanja genoma, a jedna od njih je i trio binning, čija je glavna zadaća pospješiti razdvajanje haplotipova visoke varijacije u alelima kao u ljudskom genomu.

1. Sekvenciranje cijelog genoma

Sekvenciranje cijelog genoma (WGS - eng. whole genome sequencing) jest postupak određivanje slijeda svih, ili gotovo svih, nukleotidnih baza (A, C, G, T) DNK molekule nekog organizma. U čovjeka, taj niz ima približno 3 milijarde nukleotida. Ovaj proces uključuje razbijanje DNK molekule u manje fragmente, sekvenciranje tih fragmenata te, naposljetku, sastavljanje fragmenata natrag u cijeli genom.

1.1. Genom, gen i alel

Prije nego što se upustimo u dublje razmatranje genomike i bioinformatike, važno je razjasniti tri ključna pojma: gen, genom i alel. Ovi se termini ponekad koriste naizmjenično, no postoje razlike koje je bitno razumjeti.

Genom predstavlja cjelokupan genetski materijal nekog organizma, odnosno, sve genetske informacije koje su sadržane u DNK. U suštini, genom je zbir svih genetskih instrukcija koje određuju razvoj i funkcije organizma. Nasuprot tome, gen je specifični segment DNK koji kodira određenu biološku funkciju ili tjelesnu karakteristiku, poput boje očiju ili pigmentacije kože. Drugim riječima, gen je precizno locirano područje na genomu gdje su sekvence nukleotidnih baza organizirane u instrukcije za sintezu proteina koji određuju specifične osobine (Gleichmann, 2020.).

Vrijedi napomenuti da geni koji kodiraju proteine, poznati kao eksoni, čine manje od 2% cjelokupnog ljudskog genoma. Ostatak genoma, koji se sastoji od introna i drugih ne-kodirajućih dijelova, ima niz različitih uloga u regulaciji ekspresije gena i održavanju genomske stabilnosti.

Kada kažemo da netko posjeduje "gen za plave oči", tehnički ispravno bi bilo reći da posjeduje "alel za plave oči". Aleli su različite varijante nukleotidnih sekvenci na određenom lokusu unutar gena. Drugim riječima, aleli predstavljaju varijacije unutar gena. Budući da ljudi nasljeđuju dvije kopije genoma, po jednu od svakog roditelja, ti genomi mogu sadržavati različite alele na istom lokusu. Stoga su ljudi diploidni organizmi, što znači da nasljeđuju dva haplotipa.

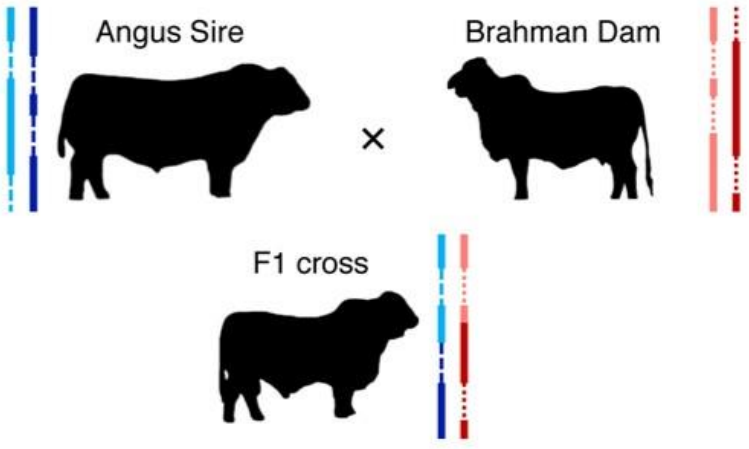
1.2. Heterozigotnost

Heterozigotnost u genetici odnosi se na nasljeđivanje različitih verzija (alela) određenog genetskog markera od svakog biološkog roditelja. Dakle, osoba koja je heterozigotna za taj genetski marker ima dvije različite verzije tog markera.

Uzmimo za primjer osobu smeđih očiju. Osoba je heterozigot za genetski marker koji određuje boju očiju ako ima jedan alel za smeđe oči, a drugi za plave. Ako pak ima par alela za smeđe oči, tada je osoba homozigot za taj genetski marker. Za osobu plavih očiju možemo biti sigurni da je homozigot jer je plava boja očiju recesivno svojstvo te dolazi do izražaja samo ako su oba alela upravo za plave oči.

Aleli nekog gena nalaze se na istom lokusu u oba haplotipa, ali se razlikuje njihov nukleotidni slijed, odnosno sekvenca. Upravo ova činjenica nam je bitna kada nastojimo sastaviti kompletni genom nekog diploidnog organizma, a da pritom točno razdvojimo njegove haplotipove.

Pojava na koju treba obratiti pozornost, a važna je kada govorimo o haplotipovima i njihovom nasljeđivanju, jest genska rekombinacija. Gensku rekombinaciju definiramo kao proces tijekom kojeg dolazi do izmjene genetičkog materijala između homolognih kromosoma tijekom mejoze, odnosno, da aleli prilikom mejoze prelaze s jednog kromosoma na drugi. To znači da prilikom nasljeđivanja kromosoma, dijete ne nasljeđuje samo jedan od dva homologna kromosoma roditelja, već njihovu kombinaciju.



Slika 1: Nasljeđivanje rekombiniranih kromosoma (Koren, i dr., 2018.)

1.3. Sastavljanje genoma

Sastavljanje genoma predstavlja ključni korak u sekvenciranju cijelog genoma, gdje je cilj rekonstruirati puni niz DNA sekvenci organizma iz brojnih manjih, često fragmentiranih sekvenci dobivenih različitim tehnikama sekvenciranja. Sastavljanju genoma može se pristupiti na dva osnovna različita načina: mapiranje na referentni genom i de novo sastavljanje.

Mapiranje na referentni genom odnosi se na proces u kojem se postojeća referentna sekvenca koristi kao okvir prema kojem se novodobivene sekvence poravnavaju. Ova metoda omogućuje brzo sastavljanje i često se koristi kod dobro istraženih organizama s visokokvalitetnim referentnim genomima. Međutim, ova metoda može propustiti nove varijante koje nisu prisutne u referentnom genomu.

De novo sastavljanje koristimo kada nemamo pristup referentnoj sekvenci po kojoj bi poravnavali dobivena očitanja, već cijelu sekvencu moramo sastavljati „od nule“. Ova metoda koristi algoritme koji prepoznaju preklapanje među kratkim sekvencama kako bi ih spojili u duže kontinuirane sekvence poznate kao kontizi (contigs). Glavni izazovi u de novo sastavljanju uključuju ponavljajuće sekvence i heterozigotnost, što može otežati preciznu rekonstrukciju genoma (Koren, i dr., 2018.).

1.3.1. Problem ponavljajućih sekvenci

Problem ponavljajućih sekvenci javlja se kad se u genomu ponavljaju duge sekvence (dulje od duljine očitanja/fragmenta). Očitanja koja sadrže dijelove tih sekvenci teško je smjestiti na pravo mjesto u genomu pa kao rezultat dobivamo niz dugih sekvenci (kontiga) isprekidanih prazninama ili ponavljanjima. Napredak tehnologije sekvenciranja dugih očitanja djelomično rješava taj problem, ali sastavljanje i dalje rezultira umjetnim genomom koji je mozaik oba haplotipa, a u stvarnosti ne odražava nijedan od njih. Ovakvo spajanje haplotipova može uzrokovati pojavu lažnih varijanti koje nisu prisutne ni u jednom haplotipu, što može dovesti do pogrešaka u analizi i anotaciji. Idealno, pri sastavljanju želimo par jasno razdvojenih haplotipova.

1.3.2. Problem visoke heterozigotnosti

Pri sastavljanju haplotipova, česta je pojava da alel s jednog haplotipa završi na suprotnom. Standardnim tehnikama preklapanja fragmenata, teško je odrediti točno je li

očitanje koje pokriva lokus na kojem se nalazi određeni alel došlo s majčinog ili očevog genoma. Ovaj problem je izraženiji što je heterozigotnost organizma veća, odnosno što mu se više alela razlikuje.

Metoda trio binninga se pokazala kao izvrstan pristup rješavanju ovoga problema a temelji se na razdvajanju haplotipova prije samog sastavljanja koristeći očitanja dobivena od oca i majke jedinke koju promatramo. S. Koren i drugi demonstrirali su 2018. efikasnost ovog pristupa na de novo sastavljanju genoma goveda a pokazali su i da je kvaliteta haplotip-razdvojenog genoma dobivenog koristeći trio binning veća, što je veća heterozigotnost organizma.

2. Trio binning i obrnuti trio binning

Ustanovili smo da je trio binning dobar pristup sastavljanju genoma razdvojenih haplotipova, te da rješava neke od poznatih problema pri sastavljanju diploidnih genoma. Razjasnimo sada metodiku trio binninga kod sastavljanja genoma djeteta te kako nam ona može pomoći da na sličan način upotrijebimo očitanja s djetetovog genoma kako bi pospješili sastavljanje roditeljskih genoma.

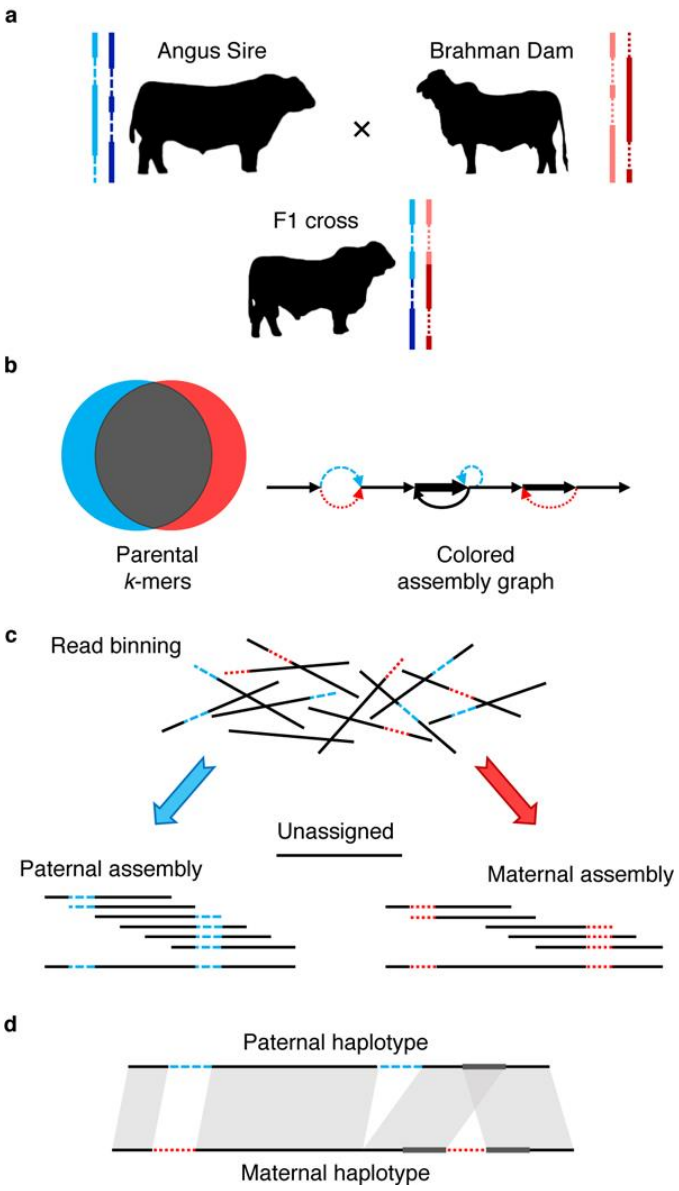
2.1. Trio binning

Trio binning metoda zahtjeva da imamo kvalitetna kratka očitanja s oba roditeljska genoma kako bi razvrstali duga očitanja djetetova genoma u dva bin-a, jedan za svaki haplotip. Očitanja iz odgovarajućih binova tada možemo koristiti u sastavljanju svakog haplotipa zasebno. Trio binning možemo podijeliti na sljedeće osnovne korake:

1. **Sekvenciranje roditelja i potomka:**
    - Pribavljamo očitanja oba roditelja koristeći tehnologiju za kratka očitanja (npr. Illumina). Ova očitanja su kratka i relativno precizna te ih možemo koristiti za identifikaciju jednonukleotidnih polimorfizama ili SNP-eva (Single Nucleotide Polymorphisms) te ostalih jedinstvenih varijanti kako bi razlikovali roditeljske genome.
    - Pribavljamo očitanja potomka koristeći tehnologiju za duga očitanja (npr. PacBio ili OxfordNanopore). Ova su očitanja manje precizna ali puno dulja, te daju veći kontekst za razrješavanje kompleksnih regija genoma.
2. **Detekcija jedinstvenih varijanti:**
    - Poravnavamo kratka očitanja s referentnim genomima roditelja kako bi detektirali SNP-eve i/ili jedinstvene k-mere kako bismo mogli razlikovati roditeljske haplotipove
3. **Binning:**
    - Klasificiramo duga očitanja ovisno o tome koje jedinstvene varijante pronađemo u njima – očinski bin, majčinski bin te nerazrješivi bin, gdje idu očitanja kojima ne možemo sa sigurnošću pripisati haplotip nijednog roditelja

4. Sastavljanje

- Sastavljamo svaki haplotip posebno, koristeći duga očitanja iz odgovarajućih bin-ova
- Rekonstruiramo diploidni genom koristeći sastavljene haplotipove



Slika 2: Trio binning haplotipova goveda (Koren, i dr., 2018.)

2.2. Obrnuti trio binning

Obrnuti trio binning je koncept koji proširuje metodiku trio binninga kako bi se fokus prebacio sa sastavljanja dječjeg genoma pomoću roditeljskih očitanja na sastavljanje genoma roditelja pomoću očitanja s djetetova genoma. Sam pojam nije službeno definiran, no u kontekstu ovog diplomskog rada, obrnutim trio binningom nazivat ćemo metodu u kojoj koristimo kromosomski i haplotipski razdvojen dječji genom kako bi roditeljska očitanja mogli razdvojiti na kromosome kojima pripadaju.

Osnovni koraci ove metode mogu se podijeliti na:

1. **Sekvenciranje roditelja i potomka -** Pribavljamo očitanja roditelja i potomka koristeći tehnologiju za duga očitanja.
2. **Treniranje modela strojnog učenja nad očitanjima potomka –** Pri treniranju modela, pretpostavljamo da su nam dostupne informacije točno s kojeg haplotipa i kromosoma dolazi očitanje djeteta.
3. **Binning –** Koristimo istrenirani model kako bi podijelili roditeljska očitanja u binove koji odgovaraju pojedinim kromosomima.
4. **Sastavljanje –** Koristimo informaciju o kromosomima kojima roditeljska očitanja pripadaju kako bi efikasnije sastavili kromosomski razdvojen roditeljski genom

U ovom diplomskom radu, naglasak je na 2. i 3. koraku postupka, gdje se nastoji razviti robustan i pouzdan model strojnog učenja koji će razdvajati roditeljska očitanja u binove.

3. Podaci

U svrhu izrade diplomskog rada, pripremljen je potpuno sekvenciran i sastavljen referentni diploidni genom čovjeka. Svaki haplotip nalazi se u posebnoj FASTA datoteci koja sadrži sekvence 23 kromosoma. Očitanja potrebna za obrnuti trio binning simulirana su korištenjem alata PBSIM2 (Ono, Asai, & Hamada, 2020) te su pohranjena u odgovarajuće FASTQ datoteke

3.1. Tipovi podataka

3.1.1. FASTA format

FASTA format je standardni format za pohranjivanje nukleotidnih (DNK, RNK) i proteinskih sekvenci. Razvijen je za jednostavno čitanje i razmjenu sekvenci između različitih bioinformatičkih alata. Svaka sekvenca u FASTA formatu sastoji se od dva dijela: zaglavlja (header) i same sekvence.

1. **Linija zaglavlja:** Svaka sekvenca u FASTA datoteci počinje s linijom zaglavlja koja započinje znakom ">" nakon kojeg slijedi identifikator sekvence i opcionalno opis.
2. **Sekvenca:** Nakon linije zaglavlja, sama sekvenca se piše u standardnim jednoslovnim kodovima. U nukleotidnim sekvencama, ta slova su A, C, G, T i ponekad U (za RNA sekvence). U proteinskim sekvencama, svih 20 standardnih aminokiselina predstavljeno je njihovim odgovarajućim kodovima (npr. A za Alanin, R za Arginin itd.). Sekvenca može zauzimati više linija, ali ne bi trebala sadržavati razmake, brojeve ili druge simbole.
3. **Druga sekvenca:** FASTA datoteka može sadržavati više od jedne datoteke. Nakon zadnje linije prijašnje sekvence slijedi prazna linija, zatim linija zaglavlja druge sekvence, a ispod nje slijedi druga sekvenca

Niže je dan primjer FASTA datoteke koja sadrži sekvencu ljudskog citokroma c.

```
>seq1 Human cytochrome c
ATGGCCCCGGTAA
CCCGGTTTAAAGG
TTAACCCGGGAAA
…
```

3.1.2. FASTQ format

FASTQ format je proširenje FASTA formata koji je također široko rasprostranjen u području bioinformatike. Najčešće se koristi za pohranjivanje velikog broja očitanja i njihovih kvalitativnih ocjena.

Svaka sekvenca u FASTQ datoteci predstavljena je zapisom koji se sastoji od četiri linije:

- **Linija 1: Zaglavlje** - Počinje s '@' nakon čega slijedi identifikator i opcionalni opis, slično zaglavlju u formatu FASTA.
- **Linija 2: Linija sekvence** – Sadrži sekvencu.
- **Linija 3: Razdjelnik** - Počinje znakom '+' i može opcionalno biti praćen istim identifikatorom sekvence i opisom kao u prvoj liniji.
- **Linija 4: Linija kvalitativnih ocjena** - Sastoji se od ASCII znakova koji predstavljaju kvalitativne ocjene za svaku bazu sekvence. Kodiranje ovih ocjena može varirati, ali Phred+33 (koristi ASCII 33 do 75 za predstavljanje kvalitativnih ocjena od 0 do 42) i Phred+64 (koristi ASCII 64 do 104) bili su uobičajeni, iako je Phred+33 sada standard u većini aplikacija.

Niže je dan primjer zapisa jedne sekvence u FASTQ datoteci.

```
@SEQ_ID_1
GATTACA
+
!''*(((
```

3.2. Generiranje očitanja

3.2.1. Stroj za sekvenciranje

Sekvenceri, poznati i kao strojevi za sekvenciranje DNA, sofisticirani su alati koji se koriste za određivanje točnog slijeda nukleotida u molekuli DNA. Ti strojevi rade tako da prvo izvuku i pripreme DNA iz uzorka, a zatim je sekvenciraju raznim kemijskim i fizičkim procesima koji čitaju sljedove baza. Osnovno načelo uključuje fragmentaciju DNA na

upravljive dijelove, koji se zatim umnožavaju i obrađuju kako bi se otkrio slijed baza putem signala koje detektira i interpretira softver stroja.

Trenutno stanje tehnologije sekvenciranja uglavnom je podijeljeno između tehnologija sekvenciranja kratkih i dugih očitanja. Tehnologije sekvenciranja kratkih očitanja, poput onih koje nudi Illumina, proizvode očitanja obično dugih između 100 i 600 baza. One su izuzetno točne, isplative i sposobne brzo obraditi velike količine DNA, što ih čini prikladnim za primjene kao što su sekvenciranje cijelog genoma, profiliranje izražavanja gena i identifikacija SNP-ova. S druge strane, tehnologije sekvenciranja dugih očitanja, kao što su one koje razvijaju Pacific Biosciences (PacBio) i Oxford Nanopore, mogu proizvesti očitanja duga desetke tisuća pa čak i milijune baza. Ta dugotrajna očitanja su ključna za primjene koje zahtijevaju premošćivanje velikih genomskih regija ili složenih struktura, kao što su ekspanzije ponavljanja ili strukturalne varijante, nudeći sveobuhvatniji pogled na genomsku arhitekturu i veću sposobnost za sastavljanje složenih genoma. Svaka tehnologija ima svoje kompromise u pogledu točnosti, duljine očitanja, propusnosti i cijene, oblikujući njihovu primjenu u različitim područjima genomskih istraživanja i kliničke dijagnostike.

3.2.2. PBSIM alat za simuliranje očitanja

Očitanja korištena u ovom radu nisu dobivena pomoću stroja za sekvenciranje, već su simulirana pomoću alat za simuliranje očitanja PBSIM (PacBio Simulator), točnije PBSIM2. Originalno je razvijen u svrhu proučavanja karakteristika pogrešaka koje rade sekvenceri dugih očitanja te se temelji na skrivenim Markovljevim modelima kako bi se generirala očitanja slična onima koji rade stvarni strojevi za sekvenciranje dugih očitanja (Ono, Asai, & Hamada, 2020).

3.2.2.1 Implementacija alata

Detaljna implementacija alata nije od interesa u ovom diplomskom radu, no generativni proces može se ugrubo sumirati na sljedeći način:

1. Odredi se duljina očitanja prema zadanoj distribuciji duljine.
2. Odredi se preciznost očitanja prema zadanoj distribuciji preciznosti.
3. Generiraju se kvalitativne ocjene svakog nukleotida koristeći generativni model na temelju određene preciznosti očitanja i odabrane kemije.

4. Nasumično se određuje pozicija na referentnoj sekvenci te se uzima očitanje zadane duljine.
5. Uvode se pogreške u očitanje na temelju generiranih kvalitativnih ocjena i zadanog omjera pogrešaka brisanja, umetanja i zamjene.

3.2.2.2 Korištenje alata

PBSIM2 poziva se preko komandne linije pri čemu mu se definiraju referentna sekvenca, model (odnosno kemija koju model simulira) te svi ostali parametri potrebni za generiranje očitanja. Primjer jednog takvog poziva dan je niže.

```
pbsim --prefix maternal
--id-prefix M
--depth 5
--length-min 100
--length-max 10000
--difference-ratio 6:50:54
--seed 0
--hmm_model P6C4.model
--length-mean 5000
--length-sd 1000
--accuracy-mean 0.99
I002C_Maternal.fasta
```

Prikazanom komandom alatu zadajemo minimalnu i maksimalnu duljinu očitanja koja može generirati, kao i srednju vrijednost i standardnu devijaciju duljine. Također zadajemo srednju preciznost svakog nukleotida. Od interesa nam je i argument *depth* koji se odnosi na dubinu pokrivenosti. Dubina pokrivenosti 5 govori nam da se pri sekvenciranju, svaki nukleotid našao u 5 jedinstvenih očitanja, odnosno da je, kada bi uzeli u obzir sva dobivena očitanja, cijeli genom sekvenciran 5 puta. Što je veći ovaj parametar, to je bolja pokrivenost genoma očitanjima što rezultira boljom analizom i sastavljanjem, međutim utječe i na računarsku i memorijsku složenost postupka.

3.2.2.3 Rezultat generiranja

Kada pokrenemo alat za neki referentni genom, npr. majčinski haplotip `I002C_Maternal.fasta`, kao rezultat dobivamo 3 datoteke za svaku sekvencu (kromosom) koja je u njoj zapisana. Za prvu sekvencu to su:

1. maternal_0001.ref – Sadrži kopiju referentne sekvence prvog kromosoma
2. maternal_0001.fastq – Generirana očitanja i njihove kvalitativne ocjene s prvog kromosoma
3. maternal_0001.maf – MAF datoteka koja sadrži dodatne informacije o poravnanju očitanja s prvog kromosoma iz FASTQ datoteke s referencom prvog kromosoma.

Nakon završetka generacije, alat također ispisuje i statistiku o simuliranim očitanjima za svaku sekvencu.

4. Korišteni alati

Rješenje je napisano u programskom jeziku Python, čija je ekspresivnost i jednostavnost pisanja koda pogodna kada je potrebno raditi s velikim količinama podataka. Korištena je verzija 3.9, a za upravljanje Python okruženjima korišten je alata Anaconda.

Za pisanje i pokretanje koda, definiranje varijabli, testiranje i provedbu eksperimenata korištene su Jupyter bilježnice, čija interaktivnost i preglednost olakšava cjelokupni proces razvoja kompleksnih rješenja poput onih u području bioinformatike i strojnog učenja.

Pri izradi rješenja korištene su sljedeće biblioteke:

- Numpy – ključna biblioteka koja proširuje mogućnosti rada s nizovima
- Pandas – biblioteka za učitavanje i spremanje podataka te upravljanje podacima
- Biopython – biblioteka za rad s podacima iz bioinformatike, kao što su sekvence genoma, očitanja, i sl.
- Tensorflow – korišten za definiciju modela strojnog učenja i njihovo treniranje
- Scikit-learn – biblioteka koja obuhvaća razne operacije koje se često koriste u projektima strojnog učenja, kao što su definicija modela, upravljanje skupovima za učenje, računanje metrike, i sl.
- Matpplotlib – biblioteka za vizualizaciju podataka

Pri treniranju modela korišteni su Nvidia CUDA Toolkit i CUDA Deep Neural Network (cuDNN) koji nam omogućavaju da modele treniramo na grafičkoj kartici, što znatno ubrzava cijeli postupak.

5. Rješenje

Cilj ovog rada jest razviti postupak koji će na temelju potpuno sekvenciranog i sastavljenog diploidnog genoma djeteta podijeliti očitanja s majčinskog i očinskog genoma u bin-ove koji odgovaraju pojedinom kromosomu. Svrha postupka jest prije samog sastavljanja genoma dobiti znanje o pripadnosti pojedinog očitanja kromosomu, oslanjajući se na pretpostavku da je distribucija k-mera očitanja i izvornog genoma dovoljno dobar pokazatelj te pripadnosti.

5.1. Priprema podataka

5.1.1. Simuliranje mutiranog genoma

Na raspolaganju nam je haplotip-razriješeni genom djeteta. Genom svakog roditelja simuliramo tako da odgovarajući haplotip mutiramo sa stopom mutacije 0.01%. Dakle svaka 10000. baza zamijenjena je za jednu od ostale 3 baze. Mutirane pozicije izabrane su nasumično iz uniformne distribucije. Ono što na ovaj način pokušavamo simulirati jest različitost alela koje je dijete naslijedilo od roditelja (onih koji su rekombinacijom završili na homolognom kromosomu koji je roditelj prenio djetetu) i onih koje dijete nije naslijedilo. Stvarne razlike između roditeljeva genoma i dijela genoma koje je dijete naslijedilo dakako nisu potpuno nasumične, pogotovo kada uzmemo u obzir rekombinaciju gena, no kako za klasifikaciju ne koristimo samu sekvencu već distribuciju k-mera, rješenje ima dozu otpornosti na lokaciju razlika među genomima. Dakako, postoji i određena doza osjetljivosti na samu kvalitetu očitanja koja je generirao simulator, za koja znamo da su u prosjeku precizna 96%.

5.1.2. Simuliranje očitanja

Za potrebe izrade i testiranje rješenja, alatom pbsim generirana su očitanja s 20., 21. i 22. kromosoma. Ovi kromosomi su odabrani jer su najkraći u ljudskom genomu, a računalni resursi su ograničeni, pa su prigodni za testiranje predloženog rješenja. Generirano je ukupno 65325 očitanja za majčinski haplotip, 65325 za mutirani majčinski haplotip, 65763 za očinski haplotip te 65763 za mutirani očinski haplotip. Srednja duljina očitanja je približno

50000 baza, a srednja preciznost očitanja je 96%. Bitno je napomenuti da je točnost od 96% vrlo mala kada govorimo o očitanjima genoma, pogotovo kada uzmemo u obzir da smo za mutaciju genoma koristili pogrešku od samo 0.1%. Oslanjamo se na sposobnost alata pbsim2 da generira očitanja s greškama karakterističnima za strojeve za sekvenciranje koji koriste tehnologiju PacBio te da bi se proizvedeno rješenje stoga moglo lako translatirati i na stvarna očitanja.

5.1.3. Računanje distribucije k-mera

Jedan od problema na koji u bioinformatici nailazimo, pogotovo ako želimo koristiti metode strojnog učenja, jest nejednaka dimenzionalnost podataka. Kada radimo s očitanjima, svako od njih je različite duljine. Pristup rješenju ovog problema u ovom diplomskom radu je da umjesto sekvence očitanja, koristimo distribucije k-mera za svako očitanje.

K-meri su kratki podnizovi dulje sekvence duljine k i osnovni su koncepti u bioinformatici i genomici jer omogućuju brzo pretraživanje, poravnanje i analizu sekvenci. Duljina k najčešće varira u rasponu od 15 do 31 za kraće genome poput bakterijskih pa sve do 63 za veće genome poput ljudskih, no većinom ovisi o specifičnoj analizi i algoritmu. Njihova konkretna primjena u ovom radu jest da iskoristimo frekvenciju pojavljivanja svakog mogućeg k-mera duljine k u određenoj sekvenci za izgradnju vektora značajki koji bi se primjenjivali u treniranju modela strojnog učenja.

Ideja iza rješenja jest da pomoću distribucije k-mera očitanja s roditeljskih genoma raspodijelimo u bin-ove koji odgovaraju pojedinom kromosomu, slično kao što trio binning razdvaja očitanja djeteta u bin-ove koji odgovaraju pojedinom haplotipu. Roditeljskim očitanjima, međutim, ne možemo razdvojiti haplotipove jer nemamo informaciju o tome s kojeg je homolognog kromosoma pojedino očitanje došlo, odnosno ne znamo je li naslijeđeno od bake ili djeda

Prvo, dakle, moramo iz učitanih sekvenci dobiti spomenute distribucije i načiniti distribucijski vektor.

Distribucijski vektor daje nam informaciju o udjelu pojedinih k-mera u sekvenci. Uzmimo za primjer sljedeći niz:

A C C G C T A A C C G T

Uzmimo za k=3. Broj pojavljivanja 3-mera je sljedeći:

```
{'ACC': 2, 'CCG': 2, 'AAC': 1, 'CGC': 1, 'CGT': 1, 'CTA': 1,
'GCT': 1, 'TAA': 1, 'AAA': 0, …}
```

Skaliramo li rezultate tako da podijelimo sve dobivene brojeve pojavljivanja sa ukupnim brojem 3-mera, dobivamo sljedeću distribuciju 3-mera (koja je ujedno i normalizirana, što je korisno za većinu metoda strojnog učenje):

```
{'ACC': 0.2, 'CCG': 0.2, 'AAC': 0.1, 'CGC': 0.1, 'CGT': 0.1,
'CTA': 0.1, 'GCT': 0.1, 'TAA': 0.1, 'AAA': 0.0, …}
```

Konačni distribucijski vektor konstruiramo tako da na nulti indeks stavimo udio 3-mera 'AAA', na prvi indeks 'AAC', na drugi 'AAG' i tako redom do 'TTT'. Ovime osiguravamo konzistenciju vektora značajki, odnosno da je svaka značajka (svaka kombinacija nukleotida) uvijek na istom mjestu kroz sva očitanja.

Valja napomenuti kako memorijski zahtjevi ovakve reprezentacije očitanja, kao i vremenska složenost njihovog izračunavanja raste s povećanjem željene duljine k-mera k. Dimenzionalnost izračunatih distribucijskih vektora raste sa složenošću $\boldsymbol{O(4^k)}$ dok vrijeme izračunavanja vektora za neko očitanje duljine n raste sa složenošću $\boldsymbol{O(n*k)}$. Sveukupno trajanje računanja distribucijskih vektora za svih 262176 generiranih očitanja je 2 sata i 12 minuta ako koristimo duljinu k-mera 5, a ukupno memorijsko zauzeće distribucija zapisanih u .txt datoteku je 5.36GB.

5.2. Binning očitanja po izvornom kromosomu

5.2.1. Kosinusna sličnost

Prije upotrebe bilo kakvih modela strojnog učenja, za uspoređivanje vektora distribucije k-mera korištena je kosinusna sličnost. Kosinusna sličnost je mjera koja kvantificira sličnost između dva vektora u nekom višedimenzionalnom prostoru. Računa se

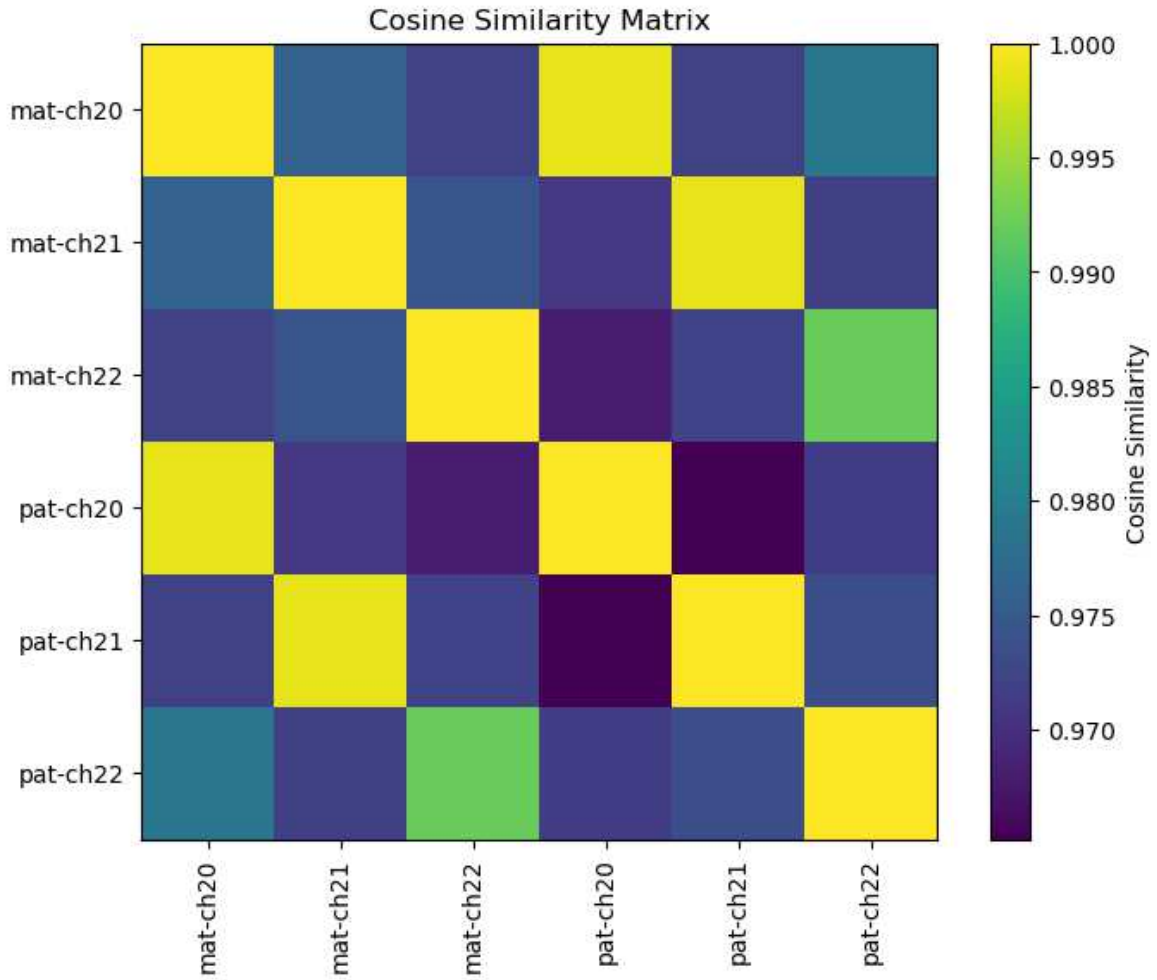
kao kosinus kuta između ta dva vektora. Vektori koji su usmjereni u istom smjeru imaju kosinusnu sličnost od 1, vektori pod pravim kutom (ortogonalni) imaju kosinusnu sličnost od 0, a vektori koji su usmjereni u suprotnom smjeru imaju kosinusnu sličnost od -1. Računa se po formuli:

$$\cos \mathrm{sim}(A, B) = \frac{A * B}{\left||A|\right| * ||B||}$$

Gdje:

- $A * B$ predstavlja skalarni produkt vektora A i B
- $\left||A|\right|, ||B||$ su euklidske norme (duljine) vektora A i B

Prije pokušaja označavanja kojem kromosomu pripadaju očitanja, ispitana je sličnost samih kromosoma. Izračunate su distribucije k-mera za cjelovitu sekvencu svakog kromosoma te je pomoću njih dobivena matrica sličnosti. Ona je korisna jer nam govori za koje kromosome možemo očekivati veću stopu pogreške klasifikacije. Što su sličnije sekvence, za očekivati je da će i očitanja s tih sekvenci biti teže diferencirati.



Slika 3: Matrica sličnosti 20., 21. i 22. kromosoma očinskog i majčinskog haplotipa

Imajući na umu da su djetetov i roditeljski genom i dalje 99.99% slični, te da možemo pretpostaviti da većinu razlika uvodi sama pogreška stroja za sekvenciranje, odnosno simulatora očitanja, očekujemo da je sličnost roditeljskih očitanja i dječjeg genoma zanemarivo različita od sličnosti dječjih očitanja i dječjeg genoma. Klasifikacija na temelju kosinusne sličnosti služi kao baseline za ostatak modela koje testiramo.

5.2.2. Logistička regresija

Prvi model strojnog učenja koji je korišten za klasifikaciju očitanja jest logistička regresija. Logistička regresija je jednostavan i transparentan model koji ima sposobnost multinomijalne (višeklasne) klasifikacije, a daje nam i informaciju o vjerojatnosti pripadnosti primjera svakoj od klasa.

Logistička regresija je linearni model čiji se krajnji rezultat transformira softmax aktivacijskom funkcijom čija je formula dana s:

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

Izlazi softmax funkcije sumiraju se u 1, odakle dolazi i statistička reprezentacija modela koja može biti vrlo korisna kada evaluiramo performanse modela.

Prilikom optimizacije, (multinomijalna) logistička regresija minimizira gubitak unakrsne entropije, koja je dana formulom:

$$CE = -\sum_{i}^{C} t_i \log(p_i)$$

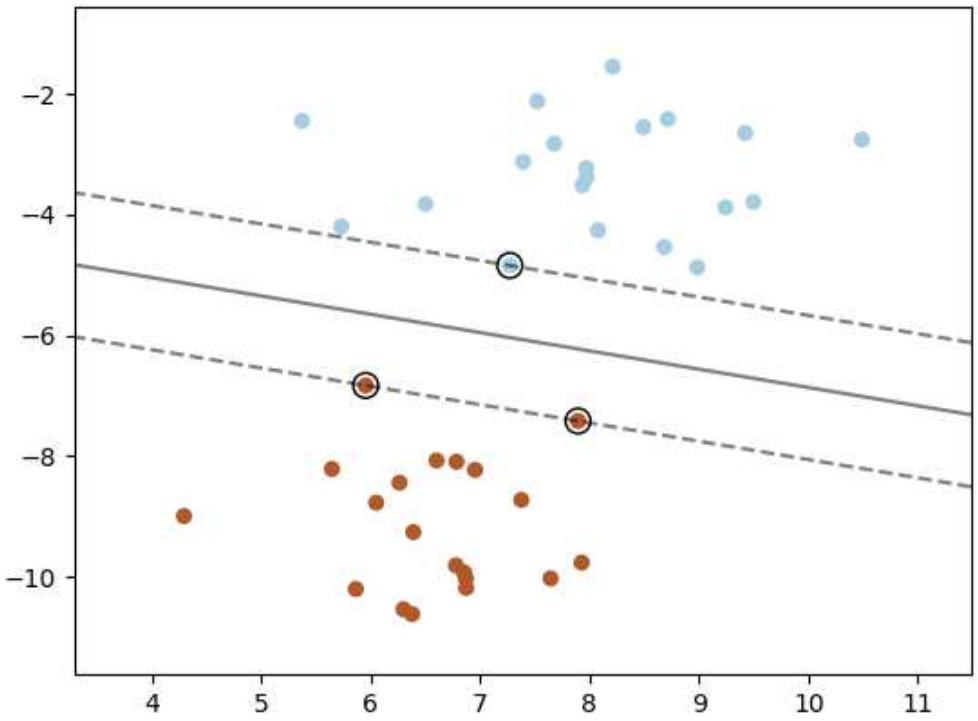
gdje je C broj klasa, t_i oznaka i-te klase, a p_i vjerojatnost i-te klase dobivena softmax funckijom.

Prilikom treniranja modela, pretragom po rešetci, optimizirat ćemo hiperparametre broja komponenti PCA i algoritma optimizacije (solver) koji se koristi pri treniranju. Isprobat ćemo *lbfgs* (limited-memory BFGS), *saga* (Stochastic Average Gradient Accelerated) i *newton-cg* (Newtonova metoda + konjugirani gradijent). Svi solveri podržani su kao opcije koje možemo koristiti prilikom treniranja modela pomoću biblioteke scikit-learn.

5.2.3. Stroj potpornih vektora

Sljedeći model koji je testiran je stroj potpornih vektora (SVM). SVM je nadzirani algoritam strojnog učenja, a zadaća mu je pronaći optimalnu granicu ili hiperravninu koja najbolje razdvaja podatke u različite klase.

SVM traži hiperravninu koja maksimizira marginu, odnosno udaljenost između najužih točaka različitih klasa (poznatih kao potporni vektori). Što je margina veća, to je model robusniji i bolje generalizira na nove podatke. U slučaju linearnog razdvajanja, SVM će odabrati ravninu koja najbolje dijeli podatke.



Slika 4: SVM nalazi jednadžbu pravca naznačenog s punom crtom (u 2d slučaju – pravac, u više dimenzija – hiperravnina).

Za nelinearno razdvojive podatke, SVM koristi tzv. jezgreni trik ("kernel trick") koji preslikava podatke u višu dimenziju gdje se oni mogu linearno razdvojiti. Najčešće korištene jezgre (kerneli) su linearna, polinomijalna i RBF (radijalna bazna funkcija).

SVM je vrlo učinkovit na velikim skupovima podataka i poznat je po tome što daje dobre rezultate čak i kada podaci nisu savršeno linearno razdvojivi.

Hiperparametri koje ugađamo su jezgra, broj komponenti u PCA analizi i inverzni faktor regularizacije C.

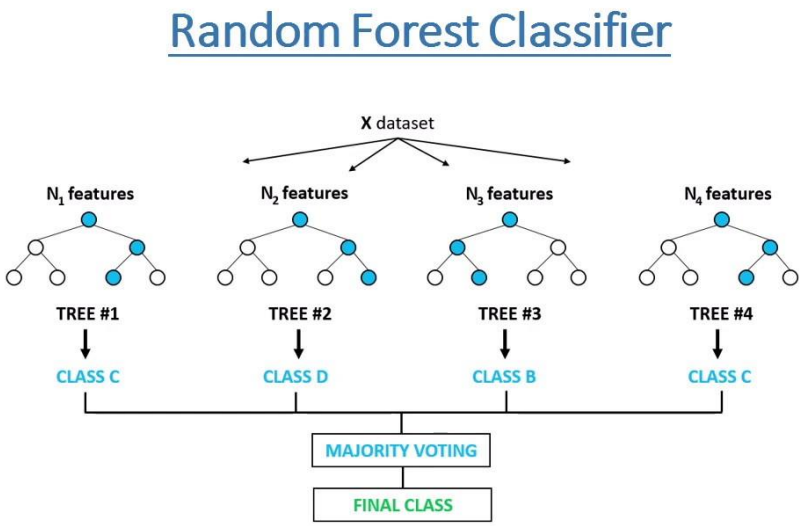
5.2.4. Modeli slučajnih šuma

Model slučajnih šuma je ansambl metoda strojnog učenja, a temelji se na kombiniranju više stabala odluke kako bi se poboljšala točnost predikcija i smanjila varijabilnost (overfitting) pojedinačnih stabala.

1. **Izgradnja više stabala odluke** - Model slučajnih šuma gradi veliki broj stabala odluke. Svako stablo trenira se na različitom podskupu podataka dobivenom

tehnikom bootstrap uzorkovanja, gdje se podaci nasumično uzorkuju s ponavljanjem (tako da neka očitanja mogu biti uključena više puta, dok neka možda neće biti odabrana).

2. **Slučajni odabir značajki** - Pri svakom dijeljenju čvora unutar stabla, model slučajno odabire podskup značajki iz skupa svih značajki i koristi ih za pronalaženje najbolje podjele. Ovo dodaje dodatnu slučajnost u model, smanjuje korelaciju između stabala i pomaže poboljšati generalizaciju modela.
3. **Agregiranje rezultata (glasanjem) -** Za klasifikacijske zadatke, svaki stablo u šumi daje svoju predikciju a konačna predikcija modela određuje se glasanjem većine (najčešće predviđena klasa pobjeđuje).



Slika 5: RF klasifikator iterativno i slučajno dijeli skup značajki na više klasifikatora (stabala), koji donose odluku o neovisno jedan o drugome.

Hiperparametri koje ugađamo su:

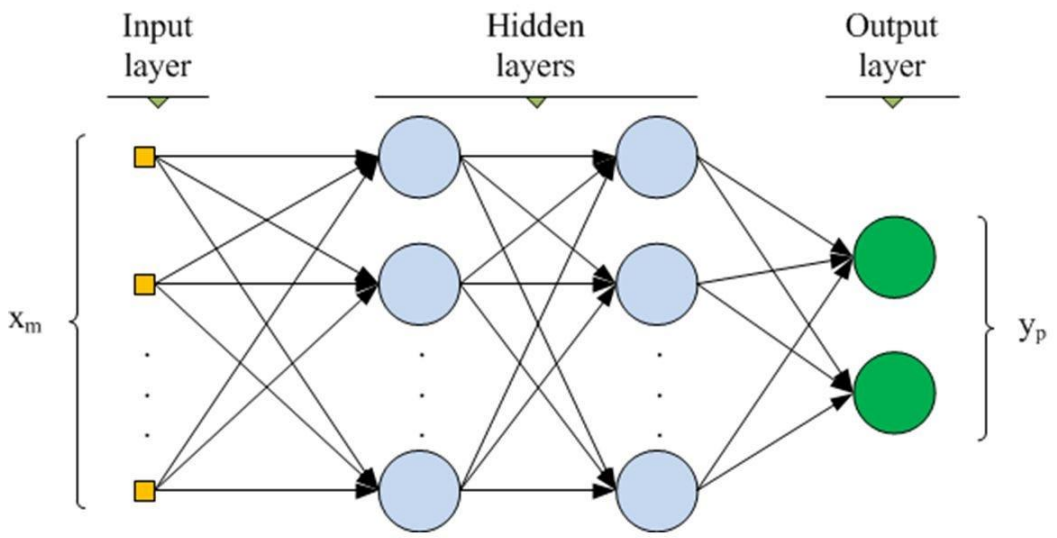
- Broj procjenitelja (*n_estimators*) - Određuje koliko će stabala biti izgrađeno u modelu. Veća vrijednost obično poboljšava performanse modela jer agregira predikcije iz većeg broja stabala, smanjujući varijaciju.
- Maksimalna dubina stabla (*max_depth*) - Kontrolira koliko duboko svako stablo može rasti. Plića stabla generaliziraju bolje i mogu smanjiti pretreniranost, dok dublja stabla omogućuju modelu da preciznije uči iz podataka, ali mogu previše prilagoditi model na treniranim podacima.

- Minimalan broj uzoraka za podjelu čvora (*min_samples*) - Minimalan broj uzoraka koji mora biti prisutan u svakom krajnjem čvoru stabla. Manje vrijednosti omogućuju vrlo specifična razdvajanja, dok veće vrijednosti sprječavaju da stablo postane previše prilagođeno podacima.
- Broj značajki koje se uzimaju u obzir pri traženju najbolje podjele u svakom čvoru. (*max_features*) - Manje značajki po čvoru uvodi više slučajnosti, što može smanjiti korelaciju između stabala i poboljšati generalizaciju modela. Često korištene vrijednosti su *sqrt* (kvadratni korijen ukupnog broja značajki) za klasifikacijske probleme i *log2* za dodatnu slučajnost.

5.2.5. Neuronska mreža

Zadnji model strojnog učenja koji je testiran jest duboka neuronska mreža. Duboke neuronske mreže su složeni modeli strojnog učenja vrlo visokog kapaciteta i sa sposobnošću modeliranja vrlo složenih, nelinearnih odnosa među podacima. Naročito su korisne kada radimo pri radu s velikim skupovima visokodimenzionalnih podataka. Također, neuronske mreže automatski uče hijerarhijsku reprezentaciju podataka, gdje se sirovi podaci, kroz slojeve neuronske mreže, transformiraju u korisne značajke. Ove karakteristike su iznimno pogodne za rad s podacima kakve imamo pred sobom.

Kada govorimo o dubokim neuronskim mrežama, govorimo o mrežama koje, osim ulaznog i izlaznog sloja, imaju i skrivene slojeve. Upravo u skrivenim slojevima mreže događa se učenje odnosa među podacima, odnosno distribucijama k-mera, koje želimo iskoristiti kako bi klasificirali očitanja.



Slika 6: Duboka neuronska mreža (DNN) sastoji se od ulaznog sloja, izlaznog sloja, i jednog ili više skrivenih slojeva.

Prije treniranja konačnog modela, pretragom po rešetci optimizirani su hiperparametri broja skrivenih slojeva, broj neurona po skrivenom sloju te metoda optimizacije. Testirani su optimizatori ADAM i RMSPROP. Svi modeli trenirani su koristeći ReLU (Rectified Linear Unit – zglobnica) aktivacijsku funkciju u skrivenim slojevima te softmax funkciju u izlaznom sloju. Učenje je provedeno u 50 epoha koristeći veličinu minigrupe 32.

6. Rezultati

6.1. Podjela skupa podataka

Prilikom treniranja modela korištena je metoda *train_test_split()* biblioteke scikit-learn kako bi brzo i efikasno podijelili skup podataka na skup za treniranje i skup za testiranje. Za treniranje modela i optimizaciju hiperparametara korištena su dječja očitanja, dok se konačna evaluacija svih modela vrši nad očitanjima roditeljskog genoma. Evaluaciju radimo nad majčinskim haplotipom, no postupak je isti i za očinski haplotip.

6.2. Metrika

Pri ocjenjivanju performansi svih testiranih modela koriste se sljedeće metrike:

- Točnost (accuracy) – Udio ispravno klasificiranih primjera u odnosu na ukupan broj primjera.
- Preciznost (precision) – Udio ispravno klasificiranih primjera jedne klase (stvarni pozitivi) u odnosu na sve primjere koje je model predvidio kao pripadnike te klase.
- Odziv (recall) – Udio ispravno klasificiranih primjera jedne klase (stvarni pozitivi) u odnosu na sve stvarne primjere te klase (stvarni pozitivi i lažni negativi).
- F1 metrika – Harmonijska sredina preciznosti i odziva. Kombinira ove dvije metrike kako bi stvorila jedinstvenu mjeru koja balansira preciznost i odziv.

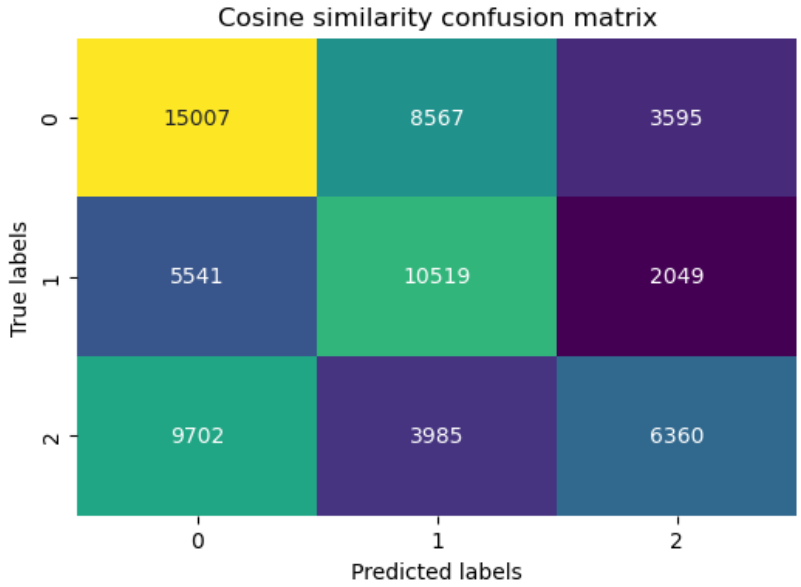
Prilikom optimizacije hiperparametara, odabiremo onaj model koji ima najveću ocjenu točnosti, a samu optimizaciju radimo metodom pretrage po rešetci, gdje treniramo po jedan model za svaku moguću kombinaciju hiperparametara.

Za svaki odabrani model ispisujemo i matrice zabune kako bi dobili uvid u to koji su kromosomi problematični za razlikovanje.

6.3. Performanse modela

6.3.1. Kosinusna sličnost

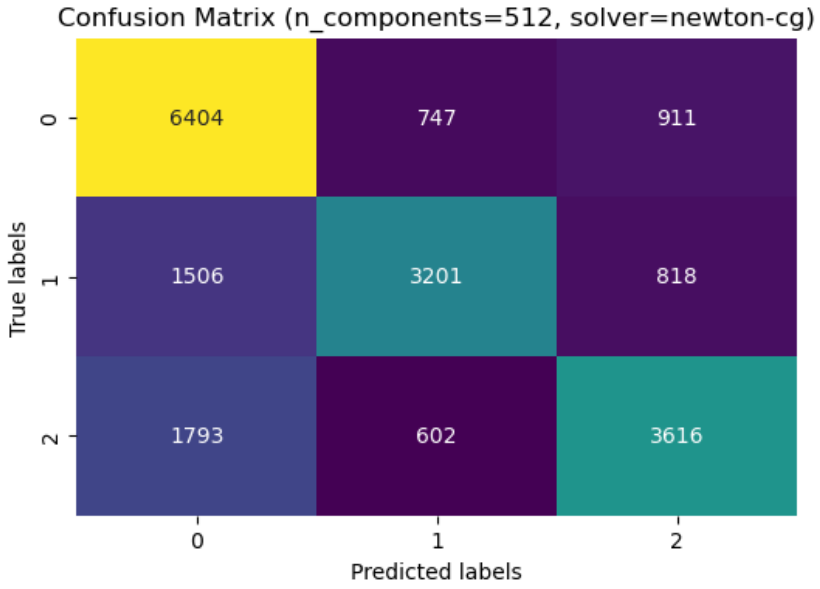
Baseline za performanse modela je klasifikacija očitanja na temelju kosinusne sličnosti pojedinom kromosomu. Nad očitanjima je najprije provedena analiza glavnih komponenti (PCA) kako bi se dimenzionalnost vektora distribucije smanjila na 512 najdiskriminativnijih značajki. Točnost ove klasifikacije je 0.49 na skupu za testiranje, što očigledno nije dovoljno pouzdano za bilo kakvu stvarnu primjenu, ali je bolje od nasumične klasifikacije, čija je točnost na 3 genoma 0.33. Ova informacija nam govori da distribucije k-mera u sebi nose potencijal za pouzdaniju diskriminaciju očitanja.



	precision	recall	f1-score	support
0.0	0.50	0.55	0.52	27169
1.0	0.46	0.58	0.51	18109
2.0	0.53	0.32	0.40	20047
accuracy			0.49	65325
macro avg	0.49	0.48	0.48	65325
weighted avg	0.50	0.49	0.48	65325

6.3.2. Logistička regresija

Za logističku regresiju, najbolje performanse imaju modeli s najviše značajki uključenih u PCA, odnosno 512. U stvarnoj primjeni uključili bi dakako svih 1024 značajki, no radi brzine testiranja, korišten je manji broj značajki. Od svih isprobanih solvera najtočnijim se pokazao *newton-cg* koji postiže točnost od 0.6714 na skupu za testiranje.

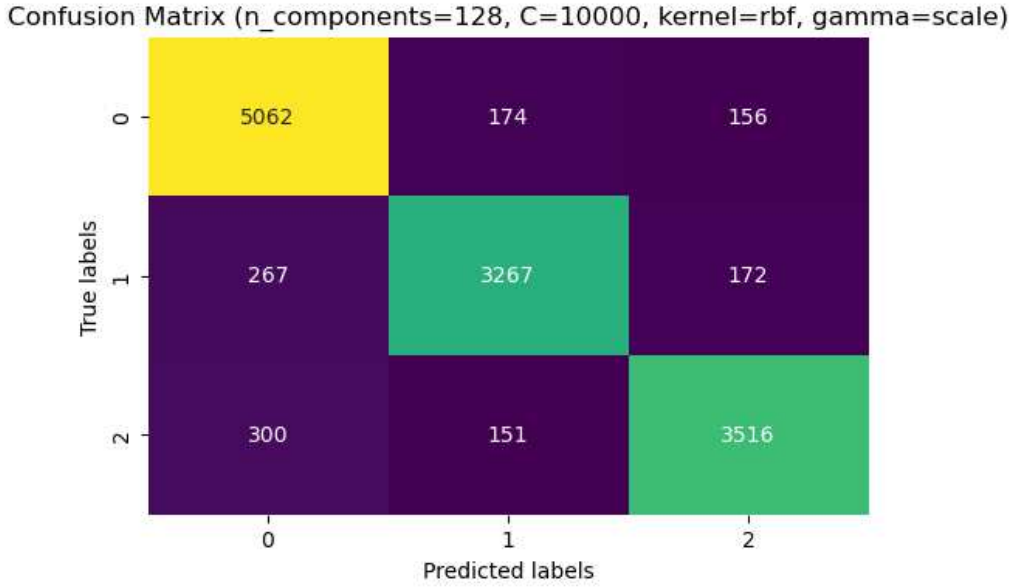



```
n_components = 512, solver = newton-cg
Accuracy: 0.6746
Classification Report:
              precision    recall  f1-score   support

         0.0       0.66      0.79      0.72      8062
         1.0       0.70      0.58      0.64      5525
         2.0       0.68      0.60      0.64      6011

    accuracy                           0.67     19598
   macro avg       0.68      0.66      0.66     19598
weighted avg       0.68      0.67      0.67     19598
```

6.3.3. Stroj potpornih vektora

Stroj potpornih vektora pokazao se kao vrlo efikasan klasifikator očitanja. Pretraživanjem po rešetki ustanovljeno je da veći broj korištenih značajki i smanjivanje utjecaja regularizacije daje bolje performanse modela. Testirana je linearna jezgra i radijalna

bazna jezgra (RBF kernel) koja, očekivano daje značajno bolje rezultate. Za 128 komponenti, i inverzni faktor regularizacije postavljen na 10000, model na skupu za testiranje daje točnost od 0.9066.

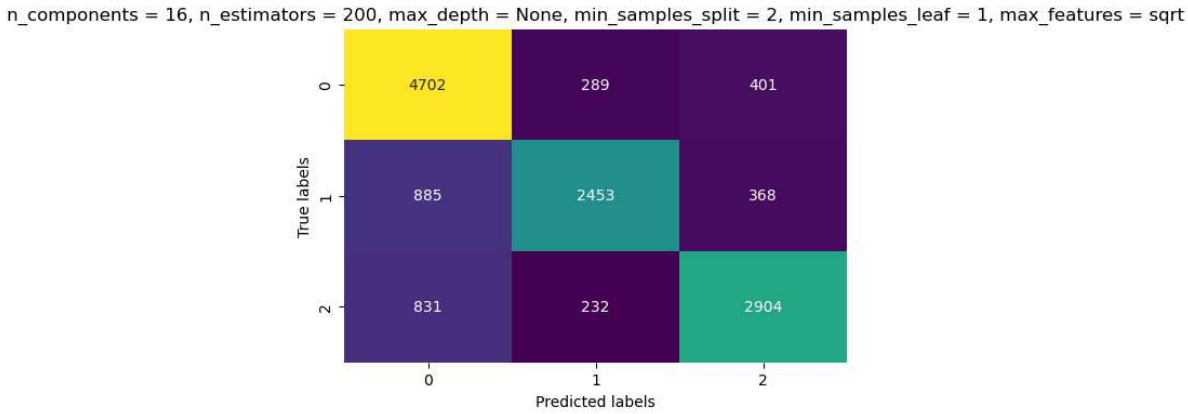



```
n_components = 128, C = 10000, kernel = rbf, gamma = scale
Accuracy: 0.9066
Classification Report:
              precision    recall  f1-score   support

         0.0       0.90      0.94      0.92      5392
         1.0       0.91      0.88      0.90      3706
         2.0       0.91      0.89      0.90      3967

    accuracy                           0.91     13065
   macro avg       0.91      0.90      0.90     13065
weighted avg       0.91      0.91      0.91     13065
```

6.3.4. Modeli slučajnih šuma

Sve modele slučajnih šuma trenirali smo nad 16 komponenti dobivenih PCA analizom radi bržeg treniranja. U daljnjoj evaluaciji, model s optimiranim hiperparametrima ćemo trenirati nad svih 1024 značajki. Najbolju točnost pokazao je model sa sljedećim hiperparametrima:

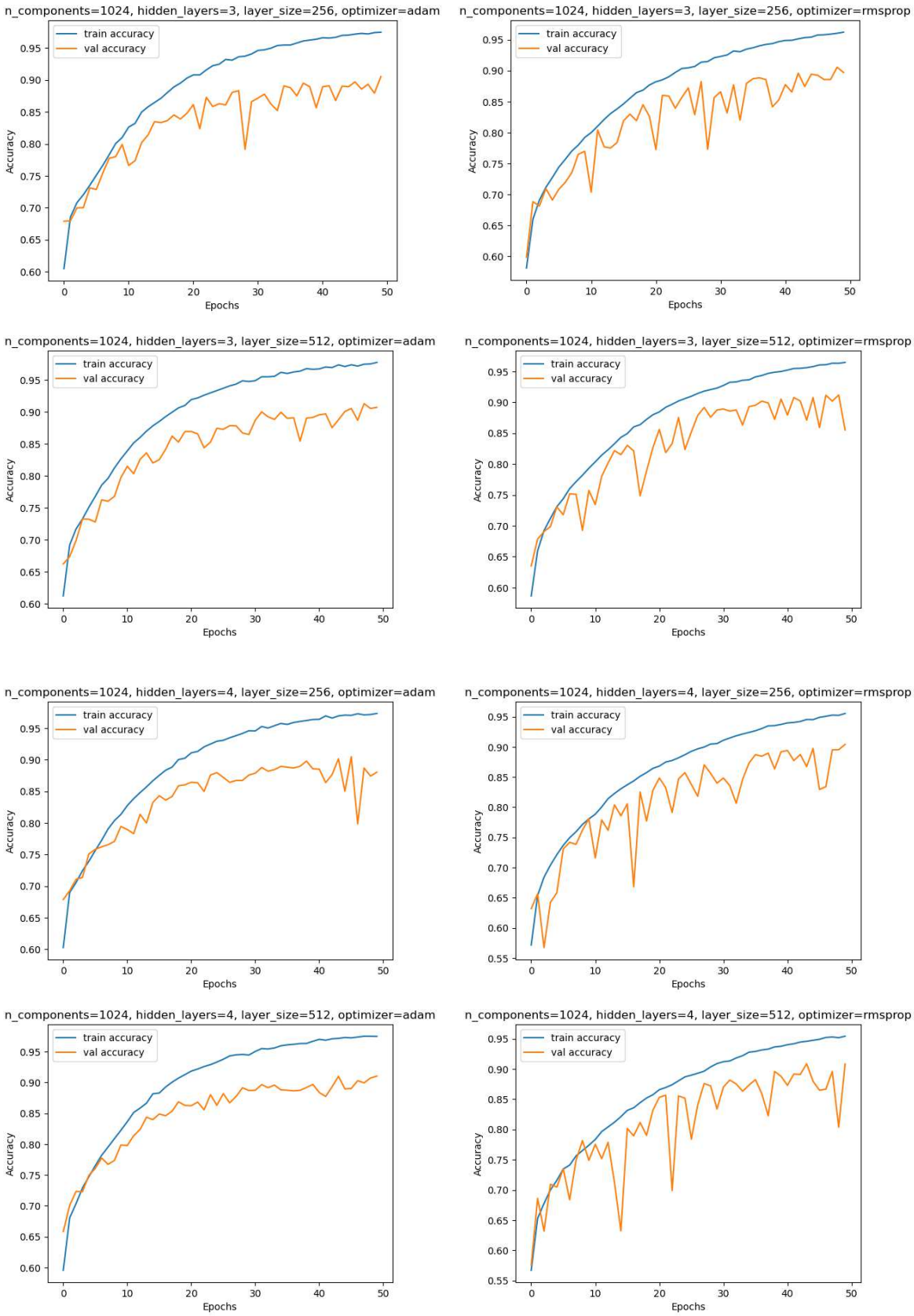
```
n_components': 16, 'n_estimators': 200, 'max_depth': None,
'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'sqrt'
```




```
Accuracy: 0.7699
Classification Report:
              precision    recall  f1-score   support

         0.0       0.73      0.87      0.80      5392
         1.0       0.82      0.66      0.73      3706
         2.0       0.79      0.73      0.76      3967

    accuracy                           0.77     13065
   macro avg       0.78      0.76      0.76     13065
weighted avg       0.78      0.77      0.77     13065
```

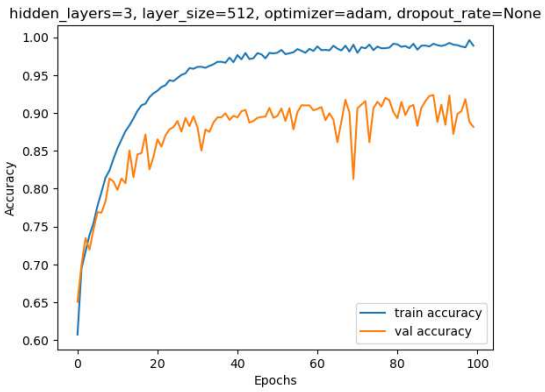
6.3.5. Neuronska mreža

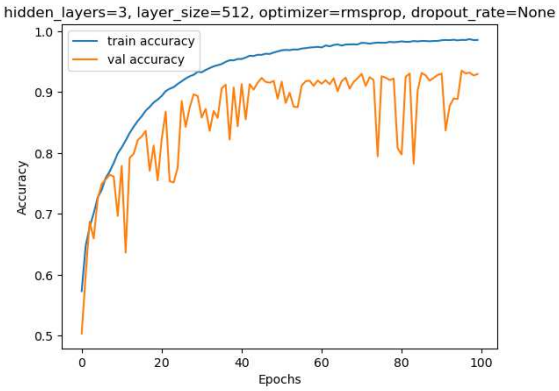
Prilikom optimizacije hiperparametara neuronske mreže, ustanovljeno je da, očekivano, veći broj skrivenih slojeva te veći broj neurona u skrivenim slojevima pridonosi većoj točnosti modela. Ipak, pri završnim epohama treniranja, naročito kod optimizatora RMSPROP, koji pokazuje veće oscilacije funkcije gubitka, počinjemo uočavati naznake pretreniranja.

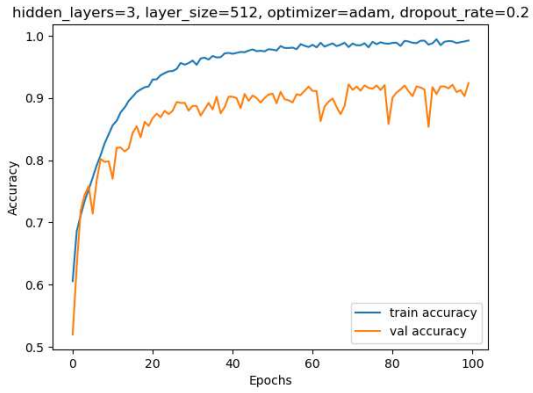


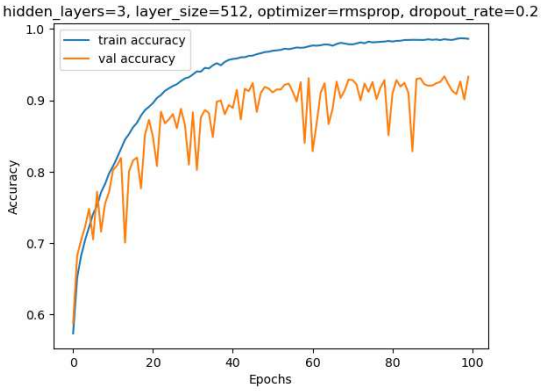
Slika 7: Točnost na skupu za treniranje i na skupu za validaciju tokom treniranja neuronske mreže s različitim hiperparametrima.

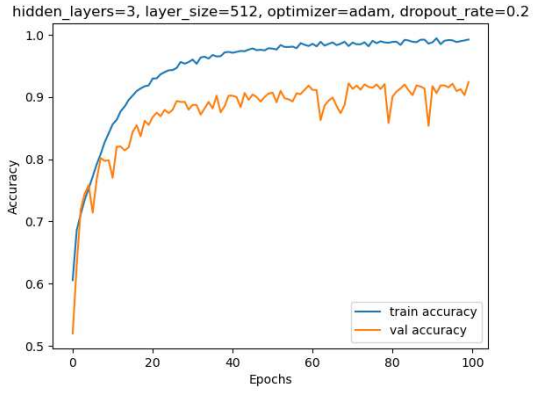
Kod ostalih modela se regularizacija pokazala nepotrebnom, jer svi modeli pokazuju bolje performanse kada značajno smanjimo utjecaj regularizacije, no kod neuronske mreže, taj utjecaj je dodatno ispitan. Ponovljena je pretraga po rešetci za hiperparametre broja skrivenih slojeva, metode optimizacije te stope dropout-a neurona uz fiksnih 512 neurona po skrivenom sloju. Također, povišen je i broj epoha učenja na 100 te veličina mini-grupe na 64.

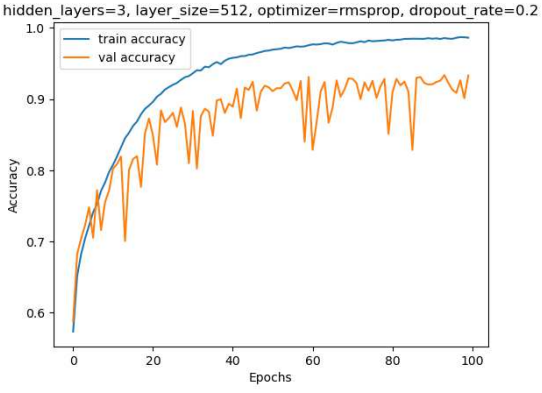


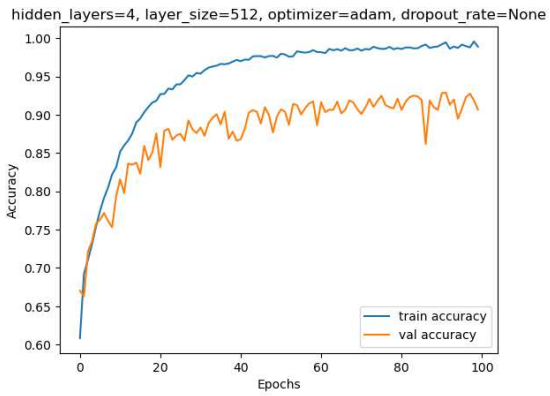


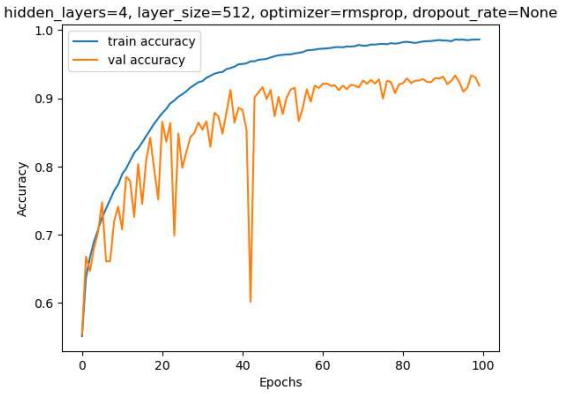


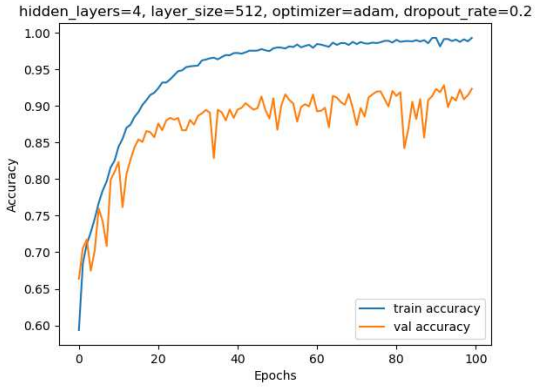


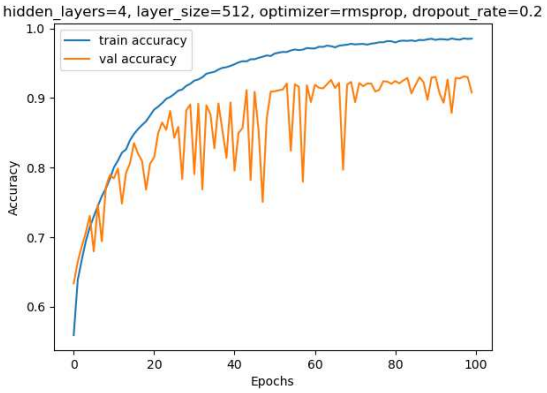


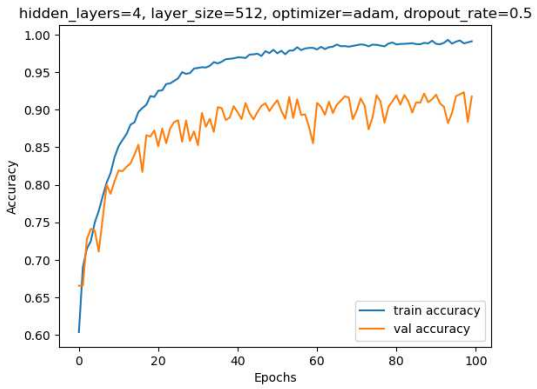


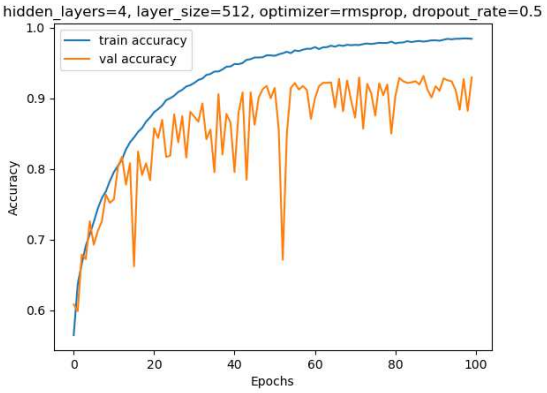












Nakon uvođenja promjena, najbolja točnost, koja iznosi 0.93, dobivena je za model s 3 skrivena sloja veličine 512 neurona, koristeći optimizator RMSPROP uz stopu dropout-a 0.2.

6.4. Evaluacija na roditeljskim očitanjima

Jednom kad imamo optimalne modele, treniramo ih ponovno, koristeći dobivene optimalne hiperparametre, na svih 1024 značajki. Zatim ih evaluiramo nad roditeljskim očitanjima, te uspoređujemo njihove rezultate.

model	broj parametara	točnost
LR	3075	0.6784
SVM	38761	0.9880
RF	~2441550	0.9565
DNN	1051651	0.9715

Tablica 1: Veličine modela i njihova točnost na roditeljskim očitanjima

Svaki od testiranih modela pokazao se jednako točnim i na neviđenim očitanjima roditelja. Ovi rezultati potvrđuju da nedostatak regularizacije ne utječe negativno na generalizaciju modela. Svi modeli još su jednom testirani tako što su istrenirani na očitanjima jednog roditelja (npr. majčinog), a zatim evaluirani na očitanjima drugog (npr. očevog).

model	broj parametara	točnost
LR	3075	0.4536
SVM	38761	0.3590
RF	~2441550	0.5150
DNN	1051651	0.4139

Tablica 2: Točnosti modela na očitanjima drugog roditelja

Iako su očev i majčin haploptip u suštini vrlo slični (na slici 3 možemo vidjeti da su im distribucije više od 99% slične), istrenirani modeli su znatno manje točni u ovako postavljenom eksperimentu. Ako pogledamo točnosti, možemo vidjeti da je SVM, koji je u prijašnjem eksperimentu bio najtočniji, sada najlošiji u klasifikaciji. RF model, koji je bio bolji samo od logističke regresije, koja ionako ima vrlo malen kapacitet u usporedbi s ostalim modelima, sada je najtočniji klasifikator. Njegovu točnost možemo pridijeliti regularizaciji koju sam model nasumičnih šuma ima ugrađen u sebe, a koja je spriječila da se model prenauči nad očitanjima s kromosoma drugog roditelja. SVM, kojem smo značajno smanjili regularizaciju najlošije generalizira.

Zaključak

Postupak obrnutog trio binninga sveli smo na klasifikacijski problem, gdje želimo svakom roditeljskom očitanju dodijeliti klasu koja označava s kojeg kromosoma je očitanje došlo. Znanje o referentnom genomu djeteta koristili smo kako bi istrenirali model strojnog učenja koji bi obavljao taj zadatak.

Distribucija k-mera pokazala se kao dobar način reprezentacije očitanja te pogodan za strojno učenje jer očitanja varijabilne duljine svodi na normalizirani distribucijski vektor dimenzije 4^k. Koristeći ovakvu reprezentaciju djetetovih očitanja trenirali i optimizirali 4 modela strojnog učenja te smo uspjeli postići točnost od 98.80% pomoću stroja potpornih vektora, međutim ovaj model nam ne daje vjerojatnosnu distribuciju klasa. Budući da bi hipotetski sljedeći korak u metodi obrnutog trio binninga bio sastavljanje referenci kromosoma pomoću očitanja klasificiranih istreniranim modelom, vjerojatnost klasa mogli bismo interpretirati kao sigurnost modela u svaku klasu te iskoristiti tu informaciju za odluku hoćemo li očitanje koristiti u sastavljanju ili ne.

Radi prirode podataka s kojima smo radili, ispostavilo se da modeli profitiraju od doze prenaučenosti, ako koristimo model za binning očitanja samo jednog roditelja. Ako razmislimo o potencijalnoj primjeni ovog postupka, gdje je pretpostavka da imamo potpuno sastavljen referentni genom djeteta, s odvojenim haplotipima i kromosomima, prihvatljivo je pretpostaviti i da možemo istrenirati po jedan model za binning očitanja svakog roditelja. Valjalo bi ispitati i kako odabir duljine k-mera utječe na točnost klasifikacije te dublje eksperimentirati s različitim hiperparametrima modela, naročito s različitim metodama i jačinom regularizacije.

Literatura

1. Cara, E. (11. Travanj 2023.). *The Human Genome Project Turns 20: Here's How It Altered the World*. Dohvaćeno iz MIT Biology: https://biology.mit.edu/the-human-genome-project-turns-20-heres-how-it-altered-the-world/#:~:text=The%20original%20project%20cost%20%242.7,over%20a%20two-year%20span.

2. Gleichmann, N. (9. Ožujak 2020.). *Gene vs Allele: Definition, Difference and Comparison*. Dohvaćeno iz Technology Networks: https://www.technologynetworks.com/neuroscience/articles/gene-vs-allele-definition-difference-and-comparison-331835#:~:text=Genes%20are%20chunks%20of%20DNA,is%20known%20as%20their%20genotype.

3. Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., . . . Phillippy, A. M. (22. Listopad 2018.). De novo assembly of haplotype-resolved genomes with trio binning. *Nature biotechnology*.

4. Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (5. Lipanj 2020.). Long-read human genome sequencing and its applications. *Nature Reviews Genetics 21*, str. 597-614.

5. Ono, Y., Asai, K., & Hamada, M. (25. Rujan 2020). PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics, Volume 37, Issue 5*, str. 589-595.

6. Szalai, C. (15. Listopad 2023.). Arguments for and against the whole-genome sequencing of newborns. *American journal of translational research*, str. 6255-6263.

Sažetak

Cilj ovo rada je istražiti primjenu strojnog učenja za klasifikaciju roditeljskih očitanja ovisno o njihovom izvornom kromosomu koristeći dječja očitanja i poznati referentni genom djeteta. Iznesena je pozadina o problemu de novo sekvenciranja ljudskog genoma te kako trio binning pospješuje rezultate sastavljanja dječjeg genoma, a zatim je predloženo rješenje obrnutog trio binninga koji bi na sličan način pospješio sastavljanje roditeljskih genoma. Pomoću distribucije k-mera trenirani su različiti modeli strojnog učenja te uspoređene njihove performanse. Naposljetku, dani su prijedlozi za daljnji razvoj predložene metode.

Abstract

This thesis aims to explore the application of machine learning for classifying parental reads based on their originating chromosome, using child reads and the known reference genome of the child. The background of the de novo human genome sequencing problem is presented, as well as how trio binning improves the assembly of the child's genome. A reverse trio binning solution is proposed to similarly enhance the assembly of parental genomes. Various machine learning models were trained using k-mer distribution, and their performances were compared. Finally, suggestions for further development of the proposed method are provided.

Skraćenice

HGP	*Human Genome Project*	Projekt humanog genoma
T2T	*Telomere-to-Telomere*	Telomere-to-Telomere
WGS	*Whole Genome Sequencing*	Sekvenciranje cijelog genoma
SNP	*Single Nucleotide Polymporphism*	Jednonukleotidni polimorfizam
SVM	*Support Vector Machine*	Stroj potpornih vektora
RF	*Random Forest*	Model slučajnih šuma
RBF	*Radial Basis Function*	Radijalna bazna funkcija
NN	*Neural Network*	Neuronska mreža
DNN	*Deep Neural Network*	Duboka neuronska mreža
ReLU	*Rectified Linear Unit*	Funkcija zglobnice