SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Marko Haralović, Duje Štolfa

Metode procjene i poboljšanja pravednosti dubokih klasifikatora

Zagreb, 2025.

Ovaj rad izrađen je na Fakultetu elektrotehnike i računarstva, Zavodu za komunikacijske i svemirske tehnologije pod vodstvom izv. prof. dr. sc. Daria Bojanjca i predan je na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2024./2025.

Sadržaj

1. Uvod

Neovisno o tehnološkim sustavima, nepravedno postupanje prema društvenim skupinama oduvijek je bilo prisutno u različitim aspektima ljudskog djelovanja, i u prošlosti i danas. Trenutno se ulažu značajni napori u aktivno smanjivanje diskriminacije i svake vrste pristranosti koja bi mogla do nje dovesti. Kako bismo mogli procijeniti pravednost i smanjiti pristranosti modela, korisno je utvrditi na koje će skupine ljudi utjecati proizvod ili proces koji razvijamo te koje bi nenamjerne posljedice ili štete za njih mogle nastati. Ovisno o vrsti nanesene štete moći će se odrediti i prikladan način za njeno ublažavanje ili uklanjanje. Ovakav pristup pravednosti poznat je kao pravednost među skupinama (engl. *group fairness*), a obilježja na temelju kojih definiramo te skupine nazivaju se osjetljivim značajkama (npr. spol, rasa, prihod).

U medicinskom su području ove štete i rizici vrlo značajni te je iznimno važno osigurati pouzdan i pravedan rad novih tehnoloških rješenja. Analizu pravednosti ćemo stoga provesti i ilustrirati na primjeru razvoja modela za klasifikaciju melanoma iz dermatoskopskih podataka.

1.1. Detekcija melanoma

Melanom je zloćudni tumor uzrokovan nekontroliranom diobom stanica koje proizvode pigment, tzv. melanocita. Te se stanice uglavnom nalaze u epidermi, najudaljenijem sloju kože, pa se većina slučajeva melanoma odnosi na rak kože, dok se rjeđe pojavljuje na mrežnici oka ili sluznicama [1]. Maligni melanociti postaju opasni kada se počnu širiti u dublje slojeve kože, odakle mogu ući u krvotok i limfni sustav te brzo metastazirati u bilo koji dio tijela. Zbog toga je melanom odgovoran za većinu smrtnih slučajeva uzrokovanih rakom kože, iako nije najčešći oblik tog raka [2]. Ovo čini rano otkrivanje melanoma ključnim čimbenikom za potpuno ozdravljenje pacijenta.

Dijagnostički proces za melanom može uključivati velik broj kliničkih pregleda, no konačnu odluku donosi dermatolog, koji mora procijeniti hoće li napraviti biopsiju određene lezije ili ne [1]. Pregled se obavlja vizualno, najčešće uz pomoć dermatoskopa: uređaja koji povećava leziju, smanjuje refleksije te povećava vidljivost određenih značajki lezije i okolne kože. Kriteriji kojima se liječnici koriste pri procjeni sumnje na zloćudnost lezije razvijali su se tijekom vremena te se kreću od nekoliko jednostavnih pravila, koja mogu primjenjivati i sami pacijenti, do analitičkog prepoznavanja složenih struktura lezije i analize pacijentove kliničke povijesti i rizičnih čimbenika. Neki od često korištenih kriterija za procjenu zloćudnosti uključuju:

- **ABCD(E) znakovi upozorenja** [3]
  - Svaka lezija koja ima nepravilan oblik (A, asimetrija), nepravilne rubove (B, *border*), više boja (C, *color*), promjer veći od 6 mm (D) i mijenja se tijekom vremena (E, *evolution*) potencijalno je zloćudna
  - Jednostavan kriterij, uobičajeno ga koriste nestručne osobe, bez dermatoskopa
  - Ne uzima u obzir značajke tipične za nodularni melanom
- **Popis od sedam kriterija** [4]
  - Procjena promjena u veličini, obliku, boji, osjetu, prisutnosti upale, krustama/krvarenju te promjeru većem od 7 mm
- **Popis od tri kriterija** [5]
  - Lezija je vjerojatno zloćudna ako ispunjava barem dva od tri kriterija
  - Kriteriji su: (1) asimetrija boje i strukture (ne nužno oblika), (2) prisutnost atipične pigmentne mreže (rupice ili linije) ili (3) bilo koja vrsta plavo-bijele boje
  - Za primjenu je potreban dermatoskop
- **Analiza uzoraka** [6]
  - Uključuje identifikaciju globalnih uzoraka i lokalnih značajki te njihovo po-

vezivanje sa specifičnim strukturama i uzorcima tipičnima za zloćudne ili dobroćudne lezije

  - Za promatranje finih struktura lezije potreban je dermatoskop

- **Znak ružnog pačeta** [7, 8]

  - Ističe važnost kliničkog konteksta u procesu dijagnoze

  - Procjena atipičnosti lezije temelji se na usporedbi s ostalim lezijama kod istog pacijenta

  - Ako se neka lezija smatra atipičnom ("ružnim pačetom") za određenog pacijenta, vjerojatnost da je zloćudna značajno raste

S obzirom na to da upućivanje pacijenta specijalistu može potrajati, a da melanom može vrlo brzo napredovati, hitne reakcije liječnika obiteljske medicine mogu napraviti razliku između brzog oporavka nakon manje biopsije i višegodišnjeg liječenja metastatskog tumora. Ovdje pouzdano tehnološko rješenje može odigrati važnu ulogu, pomažući ne samo liječnicima opće prakse, već i dermatolozima u donošenju odluka. Razvoj bilo kojeg tehnološkog sustava namijenjenog uporabi u stvarnim kliničkim uvjetima predstavlja veliku odgovornost, a nužan uvjet za njegovu primjenu jest visoka razina pouzdanosti. Sljedeća poglavlja obrađuju teorijsku osnovu procesa osiguravanja te pouzdanosti pri razvoju sustava za detekciju melanoma.

2. Pravednost modela strojnog učenja

2.1. Primjeri nepravednosti modela

Modeli mogu naštetiti nekoj skupini korisnika na mnogo načina, a nerijetko se istodobno susreću i više različitih vrsta šteta. Neke od skupina u koje najčešće možemo kategorizirati ove štete uključuju sljedeće [9]:

- **Štete u alokaciji**
  - Sustav neadekvatno ili pretjerano dodjeljuje resurse, prilike ili informacije
  - npr. sustav koji sustavno ne odobrava osobne zajmove određenoj skupini ljudi koji bi ih zapravo mogli uredno vratiti
- **Štete u kvaliteti usluge**
  - Učinkovitost sustava razlikuje se među skupinama
  - npr. sustav za prepoznavanje govora koji daje lošije rezultate pri obradi izgovora govornika sa stranim naglaskom
- **Štete stereotipizacije**
  - Sustav održava ili pojačava stereotipe
  - I pozitivni i negativni stereotipi mogu biti štetni
  - Primjer su algoritmi za automatsko dovršavanje upita u internetskim tražilicama koji generiraju sadržaje zasnovane na stereotipima
- **Štete brisanja**

- Sustav se ponaša kao da određena skupina ili neka značajna obilježja osobe ne postoje
- npr. aplikacija o povijesti istraživanja DNA ne sadrži informacije o doprinosu Rosalind Franklin

2.2. Ograničenja pravednosti i metričke mjere nepravednosti

Prepoznavanje potencijalnih šteta i njihove vrste ključan je korak u procjeni pravednosti jer se sva daljnja analiza temelji na tim spoznajama. Ono što je najvažnije za sustave strojnog učenja jest da nam to omogućuje određivanje odgovarajućih mjera pravednosti koje sustav mora zadovoljavati. U slučaju pravednosti među skupinama te mjere zahtijevaju da uspješnost modela bude ujednačena među relevantnim skupinama. Ti se zahtjevi nazivaju paritetna ograničenja (engl. *parity constraints*), a pripadne mjere koje kvantificiraju razinu u kojoj je neko ograničenje zadovoljeno nazivaju se mjerama dispariteta (engl. *disparity metrics*) [9].

Najčešća i najprikladnija paritetna ograničenja za zadatak klasifikacije su demografski paritet (engl. *demographic parity*), jednakost mogućnosti (engl. *equalized odds*) i jednakost prilika (engl. *equal opportunity*). Postizanje demografskog pariteta treba imati najveću važnost kada su u određenoj domeni štete u alokaciji najkritičnija vrsta štete. Međutim, sustav koji zadovoljava demografski paritet i dalje može među skupinama ostvarivati različite točnosti i stope lažno pozitivnih rezultata. Stoga, ako su i štete u kvaliteti usluge i štete u alokaciji važne, potrebno je zadovoljiti jednakost mogućnosti. Jednakost prilika također ukazuje i na štete u alokaciji i na štete u kvaliteti usluge, no blaže je ograničenje od jednakosti mogućnosti [9]. Mjere dispariteta detaljnije se obrađuju u poglavlju 7.1., a definirane su kao ili maksimalna razlika ili minimalni omjer odgovarajućih mjera uspješnosti među skupinama.

2.3. Nepravednost u kontekstu klasifikacije melanoma

Kao što je ranije navedeno, ako se pri klasifikaciji melanoma zloćudna lezija pogrešno dijagnosticira kao dobroćudna, pacijentu će biti uskraćeno liječenje, što bi moglo imati

ozbiljne posljedice. U suprotnom se slučaju pacijent provodi kroz dodatne, nepotrebne medicinske pretrage koje će nakon određenog vremena pokazati da je početna dijagnoza bila lažno pozitivna. Drugim riječima, lažno negativni rezultati dovode do uskraćivanja ispravne dijagnoze i potrebnog liječenja. Imajući to na umu, zaključujemo da su za slučaj klasifikacije melanoma štete u alokaciji ozbiljnije od šteta u kvaliteti usluge te da bi postizanje demografskog pariteta trebalo biti prioritet.

2.4. Dizajn pravednog modela strojnog učenja

Pretpostavimo zadatak klasifikacije melanoma iz dermatoskopskih slika pomoću klasifikatora f. Za svaku sliku I imamo pridruženi par oznaka (C, S), gdje C označava klasu (benigna ili maligna lezija), a S označava boju kože. Cilj nam je dizajnirati sustav koji je pravedan s obzirom na atribut S. To znači da za odabranu ciljnu metriku M model mora postići usporedive rezultate za svaku vrijednost s_i iz skupa mogućih vrijednosti S.

Polazimo od pretpostavke da početni model nije pravedan, tj. da postoji disparitet veći od unaprijed definirane granične vrijednosti α, kojom određujemo značajnu razliku u metrici M između skupina. Primjerice, potpuno je moguće da klasifikator radi jednako dobro na potpuno balansiranom skupu podataka ako je S irelevantna značajka (npr. *je li osoba jela sladoled prošli tjedan*). Zbog toga nema smisla unaprijed uvoditi intervencije pravednosti prije evaluacije modela; potrebno je prvo jasno definirati ciljnu metriku, zatim izmjeriti pristranost, pa tek potom odlučiti o korekciji.

Induktivna pristranost modela strojnog učenja onemogućuje univerzalno pravilo za dizajn pravednog modela. Razumno je započeti s klasifikatorom f koji postiže dobru točnost na klasama C, čime eliminiramo mogućnost da nepravednost proizlazi iz nedostatnog kapaciteta modela. Tek tada pristupamo dizajnu modela koji je pravedan u odnosu na S.

No, u slučaju klasifikacije melanoma, ovaj se postupak dodatno komplicira jer su točnost i pravednost međusobno povezane. Drugim riječima, postupak učenja mora uključivati informaciju o S te osigurati da model nauči klasificirati C neovisno o S. To znači optimizirati model tako da se oslanja na značajke koje generaliziraju izvan specifičnih skupina definiranih atributom S.

2.4.1. α-pravednost

Neka M_s označava vrijednost metrike M na podskupu podataka za koji je $S = s$. Kažemo da je model α-pravedan s obzirom na S ako vrijedi:

$$\max_{s_i, s_j \in S} \left| M_{s_i} - M_{s_j} \right| \leq \alpha \tag{2.1}$$

Drugim riječima, maksimalna apsolutna razlika u vrijednosti metrike M između bilo koja dva podskupa definiranih atributom S ne smije prelaziti unaprijed zadanu granicu α.

3. Metode poboljšanja pravednosti

Da bi se neka metoda strojnog učenja mogla primijeniti na problem iz stvarnog svijeta, potrebno je taj problem svesti (barem) na ulaze i izlaze modela te njihov međusobni odnos. Taj proces modeliranja uvodi pristranosti koje omogućuju modelu da uči, ali istovremeno može ukloniti i važne netehnološke kontekste te uvesti neželjene pristranosti koje rezultiraju nepravednim postupanjem prema određenim populacijskim skupinama [10, 9]. Ova se vrsta nepravednosti najučinkovitije ublažava pažljivim razmatranjem tzv. apstrakcijskih zamki [11] tijekom samog procesa dizajna i njihovim aktivnim izbjegavanjem. Osim algoritamskih pristranosti, značajan uzrok nepravednog djelovanja sustava strojnog učenja su i pristranosti skrivene u podacima korištenima za njihovu evaluaciju ili treniranje [10]. Iako je u nekim slučajevima moguće identificirati tu pristranost te ju izravno i potpuno ukloniti iz podataka, u općem slučaju potrebne su dodatne strategije ublažavanja pristranosti kako bi se poboljšala pravednost. Sljedeća poglavlja daju pregled različitih pristupa uklanjanju nepravednosti.

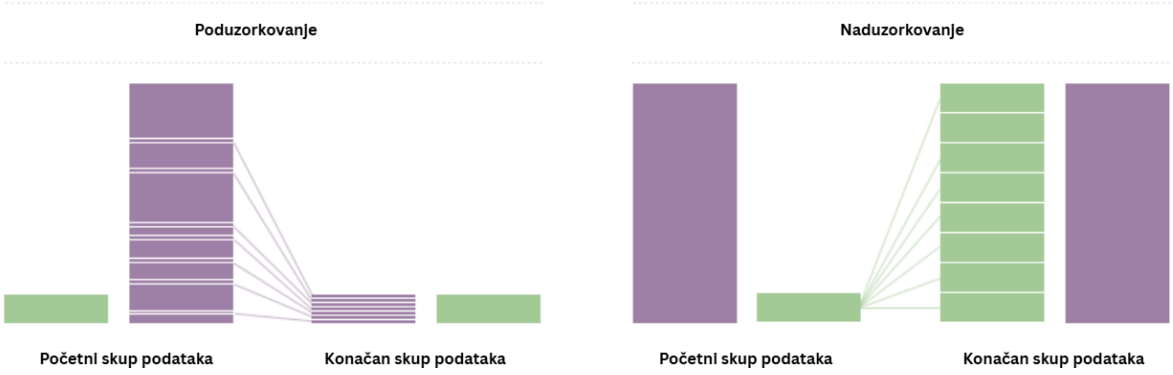
3.1. Augmentiranje i pametno uzorkovanje podataka

Nebalansiran skup podataka inherentno je pristran prema većinskim razredima i može otežati učenje, ali i dovesti do nepravednog rada modela na manjinskim skupinama. Jedan od načina rješavanja ovog problema je umjetno proširivanje manjinskih skupina i njihovo naduzorkovanje tijekom učenja. U tom su procesu ključne augmentacije podataka, jer u skup za učenje unose raznolikost koja je nužna za omogućavanje učenja i izbjegavanje prenaučenosti. Raznolikost korištenog skupa transformacija odredit će prihvatljivu stopu naduzorkovanja, tj. koliko će svaka skupina podataka biti povećana nakon augmentacije.

Postoji više pristupa uravnoteživanju podataka, uključujući poduzorkovanje većinskog razreda, naduzorkovanje manjinskog razreda, korištenje metode SMOTE (*Synthetic Minority Oversampling Technique*) ili generiranje novih uzoraka pomoću GAN-ova (npr. *DermaGAN*). U našem radu koristimo tri različita algoritma za uzorkovanje podataka: uravnoteženo uzorkovanje na razini mini-grupe, nebalansirano uzorkovanje primjera i poduzorkovanje većinskog razreda. Važno je naglasiti da uravnoteženo uzorkovanje na razini mini-grupe osigurava jednaku zastupljenost razreda unutar svake mini-grupe, dok naduzorkovanje ne jamči takvu ravnotežu, pa se primjerice može dogoditi da cijela mini-grupa sadrži samo pozitivne ili samo negativne primjere.

3.1.1. Uravnoteženo uzorkovanje na razini mini-grupe vs. poduzorkovanje

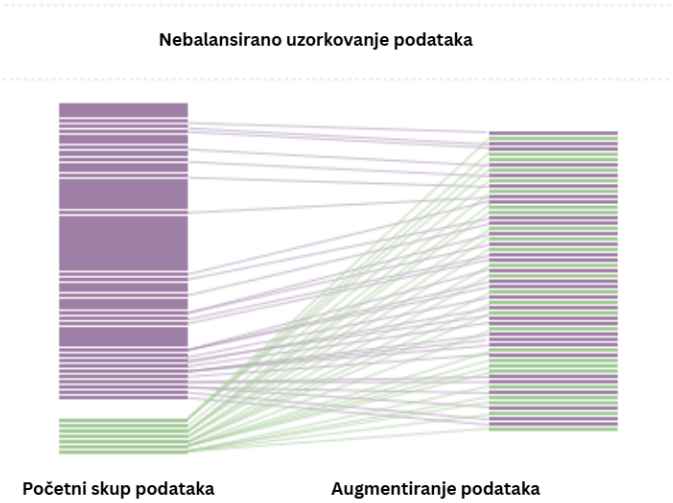
Algoritam uravnoteženog uzorkovanja na razini mini-grupe (engl. *balanced batch sampler*) nasumično udvostručava primjere manjinskog razreda (maligne lezije). S obzirom na to da se već primjenjuju tehnike augmentacije za ublažavanje neravnoteže u korist benignog razreda, jasno je da ovaj pristup može dovesti do prenaučenosti. Suprotno tome, poduzorkovanjem se informacije gube, ponajprije zato što unaprijed nije poznato koje su benigne slike teže za klasifikaciju. Nasumično izbacivanje primjera iz skupa može pogoršati izvedbu modela, osobito jer je potrebno odbaciti velik dio skupa kako bi se broj benignih primjera približio broju malignih. Ova dva pristupa ilustrirana su na slici 3.1.



Slika 3.1. Vizualizacija pod- i naduzorkovanja.

3.1.2. Nebalansirano uzorkovanje primjera

Kako bismo ublažili utjecaj izrazito neuravnoteženih veličina razreda u podatkovnom skupu ISIC 2020, eksperimentirali smo s pametnijom strategijom naduzorkovanja koristeći `ImbalancedDatasetSampler` iz biblioteke `torchsampler`. Ovakvo uzorkovanje dinamički uravnotežuje razdiobe razreda tijekom treniranja, bez izmjene ili dupliciranja podataka. Osim toga, automatski procjenjuje težine uzorkovanja na temelju frekvencija razreda. Na taj način izbjegava se potreba za stvaranjem unaprijed definiranog uravnoteženog skupa podataka, a u kombinaciji s augmentacijama može pomoći u ublažavanju prenaučenosti povećanjem raznolikosti primjera. Slika 3.2. prikazuje algoritam nebalansiranog uzorkovanja.



Slika 3.2. Vizualizacija nebalansiranog uzorkovanja podataka.

3.1.3. Segmentacija lezija (metoda *skin-former*)

U slučaju pravedne detekcije melanoma s obzirom na ton kože, korisna tehnika augmentacije bila bi ona koja transformira sliku iz većinske skupine tona kože u manjinsku. Za ovo postići odabrali smo jednostavnu transformaciju boje u prostoru boja CIELAB, kako je predloženo u [12], te ovu tehniku nazivamo *skin-former*. Ovaj pristup prilagođava vrijednosti svakog piksela koji sadrži kožu (ne sadrži leziju) prema jednadžbama 3.2 i 3.3, na temelju procijenjene vrijednosti pojedinca $\mathrm{ITA}_{\mathrm{original}}$ i željene vrijednosti $\mathrm{ITA}_{\mathrm{target}}$. Kako bismo odredili koje je piksele potrebno transformirati, generirali smo segmentacijske maske lezija korištenjem metode opisane u [13], koja je detaljnije objašnjena u poglavlju 4.1.1. Tijekom učenja ovu transformaciju primjenjujemo na najzastupljenije

skupine tonova kože, pri čemu se nasumično odabire ciljana vrijednost $\mathrm{ITA}_{\mathrm{target}}$ koja odgovara manjinskim skupinama. Transformacija se primjenjuje nasumično, s unaprijed definiranom vjerojatnošću za svaku skupinu tonova kože. Rezultati transformacije prikazani su u Dodatku B

$$\Delta\mathrm{ITA} = \mathrm{ITA}_{\mathrm{original}} - \mathrm{ITA}_{\mathrm{target}} \tag{3.1}$$

$$b = b + (\Delta\mathrm{ITA} \cdot 0.5) \tag{3.2}$$

$$L = L - (\Delta\mathrm{ITA} \cdot 0.12) \tag{3.3}$$

3.2. Modeliranje funkcije gubitka i teških primjera

3.2.1. Dinamičko otežavanje primjera inverznom zastupljenošću razreda (engl. *inverse-frequency weighting*)

Učinke nebalansiranosti podataka moguće je ublažiti i modificiranjem funkcija gubitka koje se koriste tako da se manje zastupljene skupine podataka naglase tijekom učenja. Budući da se unakrsna entropija može izraziti kao:

$$\mathrm{CE} = -\sum_{c=1}^{C} \frac{N_c}{N} \ln P_c. \tag{3.4}$$

prvo rješenje koje se nameće jest ponderirati je tako da se poništi distribucija podataka za treniranje $\frac{N_c}{N}$. Tehnika poznata kao *inverse frequency weighting* prilagođava doprinose klasa tako da se primijene težine $w_c = \frac{N}{N_c} \cdot \alpha$ na svaku klasu. Drugim riječima, množenjem jednadžbe 3.4 inverznom frekvencijom klase, skaliranom faktorom α (koji nazivamo korekcijskim koeficijentom i predstavlja hiperparametar treniranja), dobivamo:

$$\mathrm{CE}_{\mathrm{IFW}} = -\sum_{c=1}^{C} w_c \cdot \frac{N_c}{N} \ln P_c = -\sum_{c=1}^{C} \alpha \ln P_c. \tag{3.5}$$

3.2.2. Unakrsna entropija temeljena na odzivu

Kako je istaknuto u [14], povećanje težina manjinskih razreda može dovesti do značajnog rasta broja lažno pozitivnih rezultata. Razlog tome je što je standardna unakrsna entropija dizajnirana tako da uzima u obzir samo aposteriornu vjerojatnost točnog raz-

reda, zanemarujući distribuciju preko netočnih razreda. To je čini suboptimalnom za naš zadatak, gdje očekujemo izraženu neuravnoteženost pozitivnih i negativnih primjera. Primjerice, dodjeljivanje 50 puta veće težine malignim primjerima u odnosu na benigne može, kako teoretski tako i eksperimentalno, dovesti do problema, jer rezultira naglim porastom lažno pozitivnih rezultata. To često rezultira odzivom od 100 %, ali uz vrlo nisku preciznost.

Kako bi se to adresiralo, [14] predlaže prilagodbu težina gubitka na temelju odziva ostvarenog u prethodnoj epohi učenja. Ideja je započeti treniranje s unakrsnom entropijom ponderiranom inverznom frekvencijom te postupno smanjivati težinu dodijeljenu malignoj klasi kako se njezin odziv približava 100 %. Time se omogućava nastavak učenja i nakon postizanja savršenog odziva, osiguravajući protok gradijenta kroz mrežu, posebno radi poboljšanja preciznosti. U našem je slučaju prioritet maksimizirati odziv za maligne primjere, prihvaćajući rizik niže preciznosti, jer tehnike za ublažavanje neuravnoteženosti razreda služe upravo balansiranju takvih kompromisa. Točna jednadžba glasi:

$$w_{c,t}^{R} = \frac{N}{N_c}\left(1 - R_{c,t}\right) + \epsilon. \tag{3.6}$$

3.2.3. Fokalni gubitak (engl. *focal loss*)

Fokalni gubitak, predložen u [15], modificirana je verzija standardne funkcije gubitka unakrsne entropije, osmišljena kako bi smanjila doprinos dobro klasificiranih primjera te usmjerila učenje na teške, pogrešno klasificirane primjere.

Autori navode kako primjeri koje je jednostavno klasificirati dominiraju gubitkom i gradijentima tijekom treniranja, što može zaustaviti proces učenja, osobito kada postoji velika neuravnoteženost klasa. Iako inverzno ponderiranje po učestalosti može adresirati problem neuravnoteženosti klasa, a metode poput OHEM-a mogu se kombinirati s njim za fokusiranje na teške primjere, takve strategije i dalje mogu dovesti do pristranosti prema pogrešnoj klasifikaciji pozitivnih primjera, bez izričitog adresiranja porasta lažno pozitivnih rezultata. Fokalni gubitak to rješava dinamičkim skaliranjem gubitka i smanjenjem doprinosa lakih negativnih primjera, djelujući tako kao mehanizam uravnoteženja koji naglašava teške negativne primjere tijekom treniranja.

Fokalni gubitak definiran je kao

$$\mathrm{FL}(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t), \tag{3.7}$$

gdje:

- p_t je aposteriorna vjerojatnost ispravnog razreda, izlaz modela
- $\alpha \in [0, 1]$ je faktor uravnoteženja koji kompenzira neuravnoteženost razreda
- $\gamma \geq 0$ je parametar fokusiranja koji kontrolira stopu kojom se doprinos lakih primjera smanjuje

Član $(1 - p_t)^{\gamma}$ predstavlja modulacijski faktor koji smanjuje doprinos gubitku od strane dobro klasificiranih primjera (tj. onih s visokim p_t).

Dodatno, autori su eksperimentirali s uvođenjem apriornih vjerojatnosti razreda izravno u model, čime se osigurava da procijenjena vjerojatnost p za rijetke klase ostane niska prilikom inicijalizacije. Time se u ranim fazama učenja fokus preusmjerava prema nedovoljno zastupljenim (manjinskim) klasama.

3.2.4. Dinamičko biranje teških primjera (engl. *online hard example mining* - OHEM)

Online hard example mining [16] tehnika je balansiranja podataka koja je izvorno osmišljena za učenje dubokih neuronskih mreža u zadacima prepoznavanja objekata. Temelji se na principu razlikovanja primjera koje je lako i teško klasificirati, s ciljem odabira i prioritiziranja potonjih tijekom učenja. Cilj OHEM-a je optimizirati izvedbu modela fokusiranjem treniranja na teže primjere, umjesto da se svi primjeri tretiraju jednako. Postupak učenja uključuje izračunavanje gubitka za svaki primjer unutar mini-grupe, identifikaciju najtežih primjera (onih s najvećim vrijednostima gubitka) te njihovo naglašavanje u sljedećim koracima učenja. Takvi teški primjeri značajnije doprinose ažuriranju gradijenata jer njihove veće vrijednosti gubitka rezultiraju većim gradijentima, potičući model da više uči iz njih.

3.3. Treniranje osjetljivo na domenu (engl. *domain discriminative training*)

Postići grupnu pravednosti znači osigurati ravnomjernu uspješnost klasifikatora kroz sve skupine populacije definirane osjetljivim značajkama. Korisno je te skupine promatrati kao različite domene ulaza u model. Pravedan bi model u tom slučaju bio onaj čiji izlazi ne pokazuju korelacije s razredima ili domenama. Pristup *fairness through awareness* to postiže ugrađivanjem informacija o domeni u model tijekom učenja i eksplicitnim naknadnim uklanjanjem te pristranosti. Informacije o osjetljivim značajkama obično se ugrađuju u model tako da ga se uči razlikovati i ciljne razrede i domene, tj. učimo ga klasificirati primjere u CD razreda umjesto samo u C. Iako se te ugrađene informacije potom mogu ukloniti na više načina, fokusirat ćemo se na jedan pristup koji to radi tako da u odluku klasifikatora ugrađuje apriorno znanje o razdiobi oznaka u ispitnom skupu.

Prilagodbom višerazrednog probabilističkog klasifikatora $f : \mathcal{X} \to \mathcal{Y}$ apriornom vjerojatnosti razreda (engl. *class-prior adaptation*) [17] iz razdiobe razreda u skupu za učenje $\boldsymbol{\rho}$ na raziobu razreda u ispitnom skupu $\boldsymbol{\pi}$ dobijamo prilagođeni klasifikator $g : \mathcal{X} \to \mathcal{Y}$:

$$g(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} g_y(x) \quad \text{za} \quad g_y(x) = \frac{f_y(x)\pi_y}{\rho_y}. \tag{3.8}$$

U slučaju CD-razrednog klasifikatora koji na izlazu daje daje združenu aposteriornu vjerojatnost $P_{train}(y, d \mid x)$, klasifikator prilagođen na razdiobu oznaka u ispitnom skupu $P_{test}(y, d)$ je:

$$P_{\text{test}}(y, d \mid x) = \frac{P_{\text{train}}(y, d \mid x) P_{\text{test}}(y, d)}{P_{\text{train}}(y, d)}. \tag{3.9}$$

Može se pokazati [17, 18] da je prilagođeni klasifikator Bayesov optimalan ako pretpostavimo da su objekti unutar svake skupine (y, d) u ispitnom skupu vizualno slični onim objektima na kojima je klasifikator naučen:

$$P_{\text{test}}(x \mid y, d) = P_{\text{train}}(x \mid y, d). \tag{3.10}$$

Distribucije razreda na skupu za učenje mogu se izračunati kao frekvencije razreda unu-

tar svake domene:

$$P_{\text{train}}(y,d) = \begin{pmatrix} \rho_{1,1} & \rho_{2,1} & \dots & \rho_{C,D} \end{pmatrix} = \begin{pmatrix} \frac{N_{1,1}}{N_{d=1}} & \frac{N_{2,1}}{N_{d=1}} & \dots & \frac{N_{C,D}}{N_{d=D}} \end{pmatrix}. \tag{3.11}$$

Iako je razdioba razreda na testnom skupu $P_{test}(y,d)$ načelno nepoznata, jednadžbu 3.9 možemo pojednostaviti pod pretpostavkom da su d i y na testnom skupu nekorelirani te da je $P_{test}(y)$ uniformno distribuirana:

$$P_{\text{test}}(y,d) = P_{\text{test}}(d \mid y) P_{\text{test}}(y) \propto 1 \tag{3.12}$$

$$\Rightarrow \quad P_{\text{test}}(y,d \mid x) \propto \frac{P_{\text{train}}(y,d \mid x)}{P_{\text{train}}(y,d)} \tag{3.13}$$

Izlazi po razredima C prilagođenog CD-razrednog klasifikatora mogu se dobiti marginalizacijom po domenama D:

$$\begin{aligned} \hat{y} &= \underset{y}{\operatorname{argmax}}\, P_{\text{test}}(y \mid x) \\ &= \underset{y}{\operatorname{argmax}} \sum_{d} P_{\text{test}}(y,d \mid x). \end{aligned} \tag{3.14}$$

3.4. Treniranje neovisno o domeni (engl. *domain independent training*)

Autori rada [18] kao jedan način postizanja ujednačene uspješnosti klasifikatora predlažu učenje neovisnih C-razrednih klasifikatora po domenama D s jednom zajedničkom reprezentacijom značajki. Ovaj je pristup predloženo kao moguće poboljšanje treniranja osjetljivog na domenu u slučajevima kada su granice odluke između razreda dovoljno slične među domenama da ih nije potrebno posebno zasebno učiti. Korištenjem zajedničke reprezentacije značajki osigurano je da su svi dostupni podaci za treniranje korišteni za učenje te reprezentacije. Međutim, važno je napomenuti da pojedinačni klasifikatori vide samo podskup podataka za treniranje koji pripada njihovoj domeni, što može otežati proces učenja.

Izvođenje zaključivanja (engl. *inference*) na ansamblu klasifikatora sa zajedničkom reprezentacijom značajki može se provesti na više načina. Slijedeći [18], zbrajamo iz-

laze softmaksa $s(y, d, x)$ svakog pojedinačnog klasifikatora. Na ovaj smo način efektivno uprosječili granice odluke po razredima:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \sum_{d} s(y, d, x). \tag{3.15}$$

4. Skup podataka ISIC 2020 Challenge

Skup podataka ISIC 2020 Challenge [8] javno je dostupan skup dermatoskopskih slika s popratnim i opširnim metapodacima. Skup uključuje 33 126 anotiranih primjera za učenje i 10 982 ispitne slike bez anotacija. Ove slike prikazuju kožne lezije više od 2 000 pacijenata, pri čemu su zastupljeni i maligni i benigni slučajevi. Skup podataka bio je sastavljen za potrebe natjecanja SIIM-ISIC Melanoma Classification Challenge, koje je održano na platformi Kaggle tijekom ljeta 2020. godine.

4.1. Procjena boje kože iz dermatoskopskih slika

Boja kože rijetko je anotirana u javno dostupnim dermatološkim skupovima podataka, a samo nekolicina skupova sadrži oznake etničke pripadnosti ispitanika ili stvarne boje kože [19]. Osim toga, budući da je boju kože teško mjeriti i kvantificirati, istraživači se nisu usuglasili oko toga koja bi ljestvica trebala biti standard. Primjerice, u kozmetičkoj industriji pokazala se potreba za sitnozrnatom ljestvicom poput L'Oréalove karte boje kože [20, 21]. S druge strane, medicinski su istraživači trebali ljestvicu koja odražava učinke ultraljubičastog zračenja na različite tonove kože, pa su razvili Fitzpatrickovu ljestvicu s šest fototipova kože [22] temeljenu na vjerojatnosti nastanka opeklina ili tamnjenja kože pri izlaganju UV-zračenju. Iako je ova ljestvica u potpunosti subjektivna i izrazito pristrana prema svjetlijim tonovima kože zbog svoje izvorne primjene, često se koristi u istraživanjima, čak i kada ono nije povezano s reaktivnosti kože na sunce. Zbog navedenog, odlučili smo koristiti *individual topology angle* (ITA) kao objektivniju i kontinuiranu mjeru boje kože.

4.1.1. Procjena vrijednosti ITA

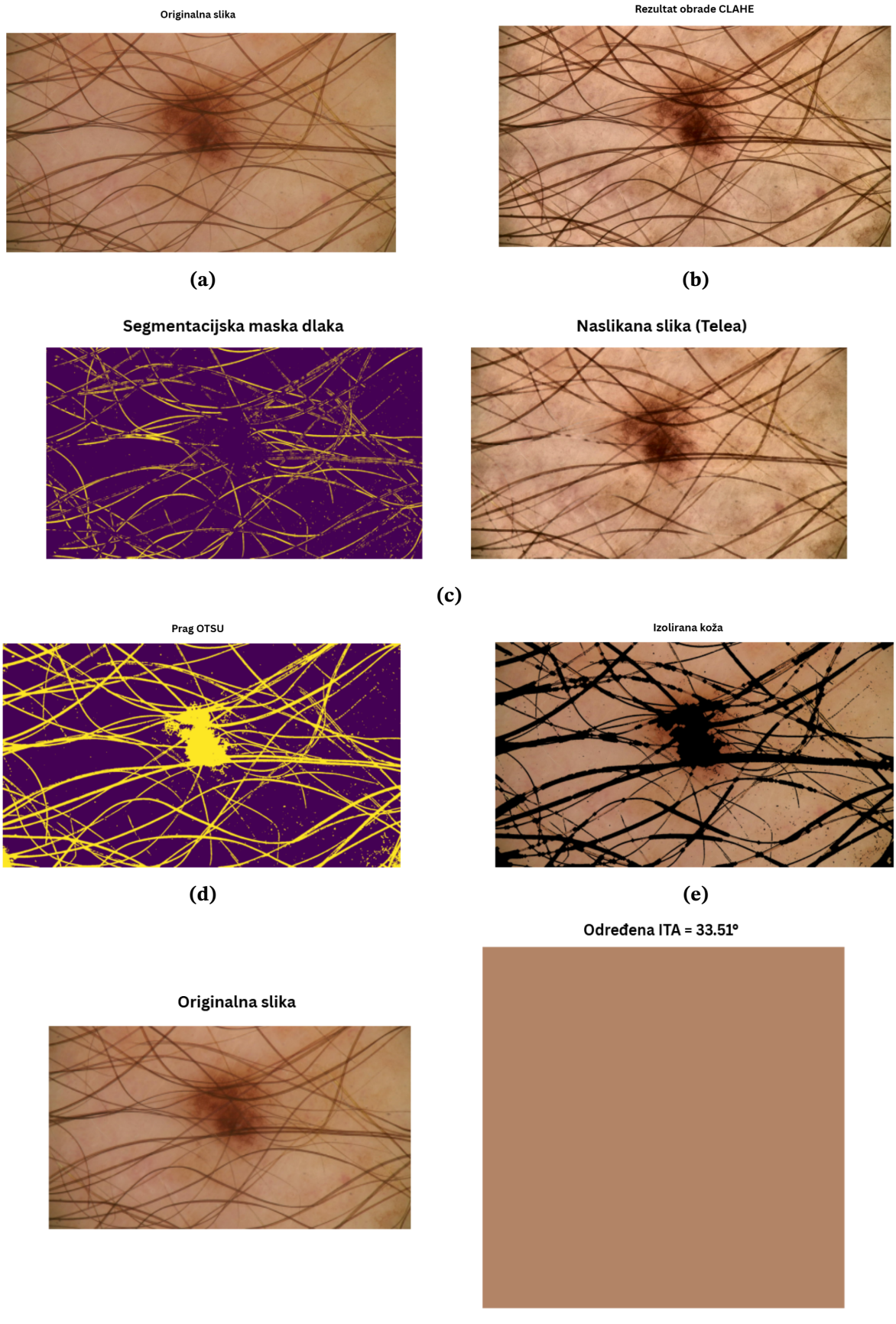
Prvi korak u pripremi i analizi podataka bio je procijeniti vrijednost ITA kože ispitanika za svaku sliku u podatkovnom skupu. Slijedeći postupak opisan u [13], prvo se maskiraju svi pikseli koji ne sadrže kožu (npr. dlačice, lezije), a zatim se preostali pikseli grupiraju kako bi se odredila dominantna boja kože (najbrojnija grupa) na slici. Ako se dobivena boja prikaže u prostoru boja CIELAB, ITA se može procijeniti iz vrijednosti komponente svjetline L i plavožute komponente b. Detaljan postupak prikazan je na slici 4.1. Ukratko, segmentacija se provodi metodama iz obrade slika i uključuje povećavanje kontrasta metodom CLAHE, uklanjanje dlačica pomoću DullRazor algoritma [23] te morfološko proširivanje radi uklanjanja pigmentacija na koži.

Kontinuirane oznake boje kože poput vrijednosti ITA omogućuju testiranje korelacija s kontinuiranim izlazima modela (npr. aposteriornim vjerojatnostima malignog razreda) te općenito sadrže više informacija od diskretnih tipova tonova kože. Ipak, zadatak kojim se bavimo je klasifikacijski, a ciljne varijable su diskretne (benigno, maligno), pa će većina metoda procjene pravednosti zahtijevati grupiranje podataka prema boji kože. Zbog toga diskretiziramo vrijednosti ITA u šest skupina predloženih u [24], prikazanih u tablici 4.1. Kao što ističe [25], a što smo i ranije natuknuli, važno je naglasiti da ove skupine nisu ekvivalentne Fitzpatrickovim fototipovima, koji su subjektivniji i temeljeni na reaktivnosti kože na Sunčevu svjetlost.

Tablica 4.1. Skupine boje kože prema pragovima vrijednosti ITA iz [24]

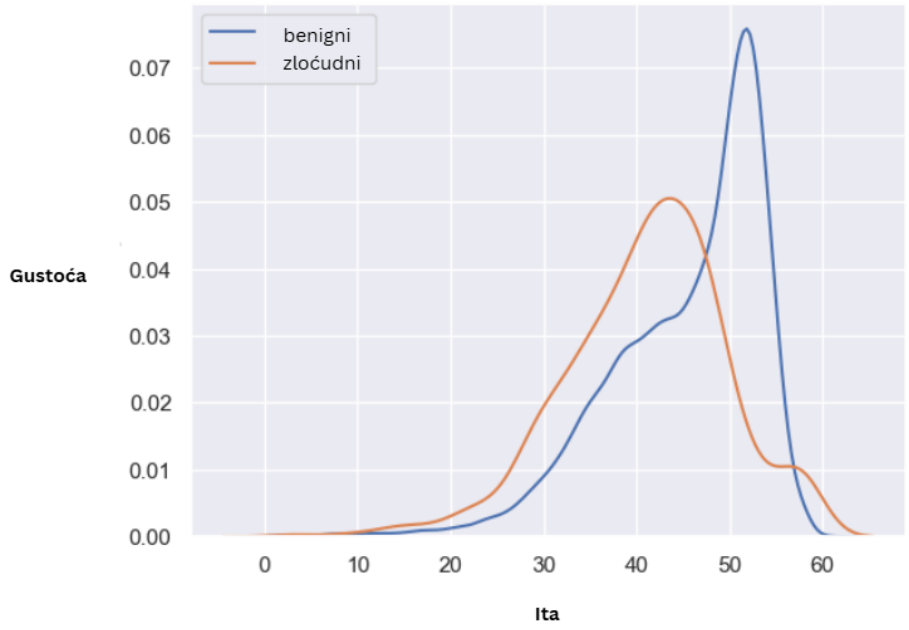
Skupina boje kože	**Raspon vrijednosti ITA**
Vrlo svijetla	> 55
Svijetla	$\langle 41, 55]$
Srednja	$\langle 28, 41]$
Preplanula	$\langle 10, 28]$
Smeđa	$\langle -30.10]$
Tamna	≤ -30

Na slici 4.2. prikazane su procijenjene razdiobe izračunatih vrijednosti ITA za maligne i benigne primjere u dostupnom skupu podataka (sjedinjeni skupovi za učenje i validaciju iz skupa ISIC2020). Očito je da je skup podataka pristran prema svjetlijim tonovima kože, osobito u benignim slučajevima. Distribucija malignih ITA vrijednosti nešto je šira, a srednja je vrijednost blago pomaknuta prema tamnijim tonovima kože.



Slika 4.1. Pojedinačni koraci i rezultati postupka segmentacije kože i procjene boje.

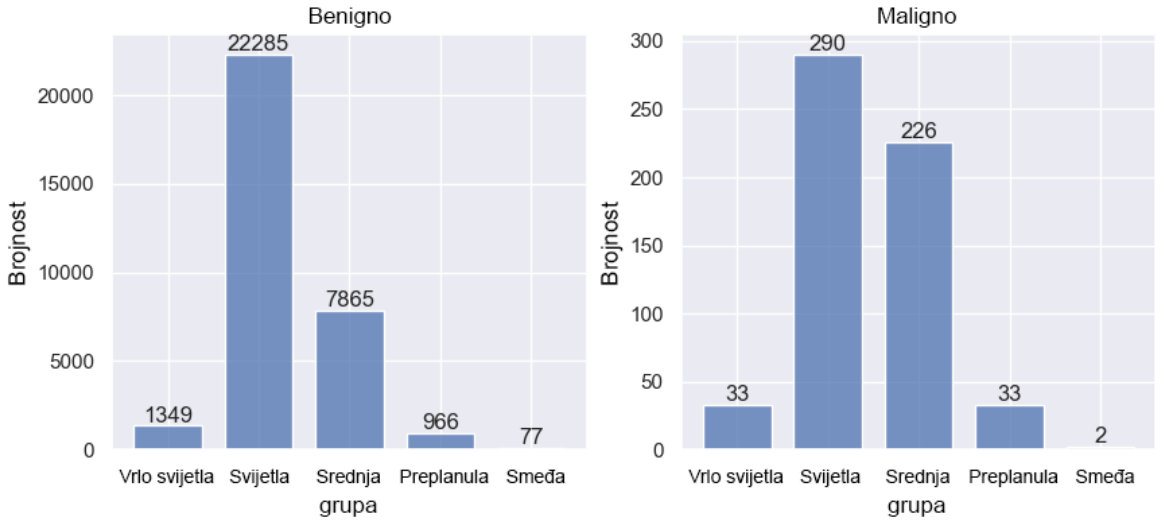
Promatrajući distribucije boje kože prema diskretnim ITA skupinama (slika 4.3.), vidimo da naš skup podataka nema nijedan primjer u tamnoj skupini te da postoje samo dva maligna primjera u smeđoj skupini. Ovi histogrami pokazuju da je, osim što je neuravnotežen prema boji kože, naš skup izrazito neuravnotežen i prema razredima (benigna/maligna). Vidimo da u najznačajnijem slučaju maligne lezije čine samo 1.3% primjera u skupini srednje boje kože. To nije iznenađujuće, budući da se pozitivni primjeri rjeđe pojavljuju pri prikupljanju dermatoskopskih i medicinskih podataka. Dakle, podatkovni skup neuravnotežen je u dva ortogonalna smjera: prema boji kože i prema malignosti, što je važno uzeti u obzir pri odabiru odgovarajućih modela i metoda za klasifikaciju.



Slika 4.2. Distribucije vrijednosti *individual topology angle* po razredima u dostupnim podacima.

4.1.2. Provjera pogrešaka u procjeni

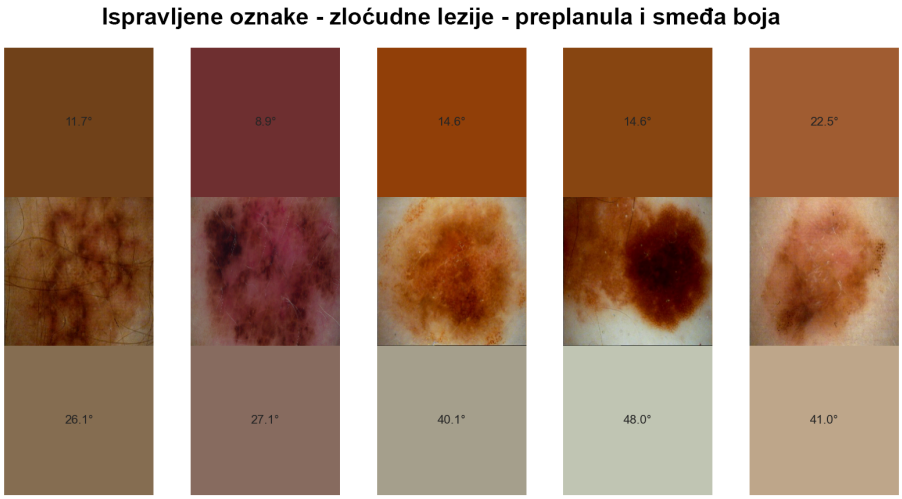
Budući da provodimo automatsku procjenu boje kože, želimo ručno provjeriti pogreške u rezultatima te osigurati da procijenjeni tonovi kože odgovaraju očekivanima za svaku skupinu. Dodatak A prikazuje 64 slike iz svake skupine boje kože, zajedno s procijenjenim ITA vrijednostima i odgovarajućim dominantnim bojama kože. Kod svjetlijih tonova i manjih lezija procjena je uglavnom točna te ispravci nisu potrebni. Primijetili smo da su mnoge slike koje su klasificirane kao smeđe snažno izrezane i prikazuju samo leziju, pri čemu je okolna koža jedva vidljiva. U nekim slučajevima koža je zapravo mnogo svjetlija od predviđene dominantne boje, što znači da je procjena bila pogrešna.



Slika 4.3. Broj slika u svakoj skupini boje kože za benigne i maligne primjere.

Usredotočujući se samo na maligne primjere u smeđoj i preplanuloj skupini (slika 4.4.), pronađeno je pet pogrešno označenih primjera (označeni crvenom bojom). Ove oznake ispravljene su određivanjem prosječne RGB-vrijednosti unutar ručno odabranog područja veličine 5×5 piksela koje sadrži samo kože. Te su RGB-vrijednosti zatim pretvorene u CIELAB prostor kako bi se mogla izračunati odgovarajuća vrijednost ITA. Ispravljene oznake tona kože uspoređene su s automatski procijenjenima na slici 4.5.

Slika 4.4. Svi maligni primjeri označeni kao preplanuli ili smeđi. Slike s netočno procijenjenom bojom kože označene su crvenim pravokutnikom.



Slika 4.5. Maligni primjeri (srednji red) iz smeđe i preplanule skupine s pogrešno procijenjenim vrijednostima ITA (gornji red) i ručno ispravljenim oznakama (donji red).

4.2. Artefakti na slikama

Artefakti poput kirurških oznaka, dlaka i zatamnjenih rubova jasno su vidljivi na slici A1. Uz to, po cijelom se podatkovnom skupu mogu uočiti i fizička ravnala, mjehurići zraka u nanesenoj imerzijskoj tekućini te mjerila iz dermatoskopa. Ovakvi su artefakti široko rasprostranjeni u svim dermatoskopskim skupovima podataka i njihov utjecaj na uspješnost modela dobro je istražena [26]. Važno je biti svjestan njihove prisutnosti jer vizualni artefakti mogu lako uvesti nepoželjne pristranosti u modele i smanjiti njihovu sposobnost generalizacije (primjerice, ako model nauči povezivati prisutnost ljubičastih oznaka s malignom klasom).

5. Modeliranje

U ovom poglavlju raspravljamo o odabiru arhitektura modela i načinu njihova treniranja. Detaljno razmatramo modele, postupke treniranja modela i linearne evaluacije (engl. *linear probing*) odabranih modela, kao i način na koji smo uveli treniranje uz pomoć segmentacije radi podrške procesu finog ugađanja (engl. *fine-tuning*).

5.1. Arhitekture modela

Metode analize, procjene i povećanja pravednosti u strojnom učenju ilustriramo na zadatku binarne klasifikacije. Cilj tog zadatka odrediti je kojoj od dvije semantičke kategorije dana slika pripada te joj ovisno o tome dodijeliti oznaku. U posljednjem desetljeću duboke arhitekture pokazale su nadmoć nad plitkim modelima u izlučivanju semantičkih značajki iz strukturiranih podataka. Konvolucijski modeli posebno su uspješni na slikovnim podacima, ali im konkuriraju i neke transformerske arhitekture. U sljedećim poglavljima opisano je nekoliko dubokih arhitektura za klasifikaciju slika koje su razmatrane u eksperimentalnom dijelu ovog rada. Ovo uključuje i arhitekture koji su isključivo konvolucijske ili transformerske, ali i one koje kombiniraju ta dva pristupa.

5.1.1. ConvNeXt

Arhitektura modela ConvNeXt [27] kombinira principe rada uobičajenih konvolucijskih umjetnih neuronskih mreža s principima rada transformerskih arhitektura kao što je Vision Transformer (ViT). Ovakav model zadržava učinkovitost i induktivne pristranosti konvolucijskih mreža (npr. ekvivarijantnost translacije), dok uvodi suvremena poboljšanja poput velikih jezgri, normalizacije na razini sloja i invertiranih uskih grla (engl. *inverted bottlenecks*). Modeli ConvNeXt nadzirano su trenirani na velikom podatkovnom skupu ImageNet-21K, a neki su dodatno fino ugođeni na skupu ImageNet-1K. Najbolji

model postigao je točnost od 87,8% `acc@1` na skupu ImageNet-1K.

Odabrali smo ga kao dobru početnu točku iz nekoliko razloga: induktivne pristranosti, mogućnosti prijenosa znanja (engl. *transfer learning*), dobre performanse na skupovima srednje veličine te dostupnost radova koji ga koriste u obradi medicinskih slika ([28], [29]).

5.1.2. ConvNeXtV2

Prirodno je bilo ispitati i uspješnost ConvNeXtV2 arhitekture [30] kao nadograđene i poboljšane inačice izvornog modela ConvNeXt. Ova verzija koristi nenadzirano treniranje tehnikom maskiranog modeliranja slika (engl. *masked image modeling*), nadahnutom maskiranim autokoderima, gdje je cilj modela rekonstruirati izbrisane dijelove slike. Ovakvim se učenjem poboljšava kvaliteta naučenih značajki, što je vidljivo i u rezultatima tog rada. Ostvarivši točnost od 88.9 % na skupu ImageNet-1k, model ConvNeXtV2 nadmašio je svog prethodnika i postigao tadašnje stanje tehnike.

5.1.3. EfficientNetV2

EfficientNetV2 [31] skup je konvolucijskih neuronskih mreža koji nadograđuje originalnu arhitekturu EfficientNet uvođenjem bržeg treniranja, bolje točnosti i bolje skalabilnosti. Kombinira složeno skaliranje modela (engl. *compound scaling*) s progresivnim strategijama učenja te koristi Fused-MBConv blokove u ranim slojevima radi brzine, dok u dubljim slojevima zadržava MBConv blokove zbog njihove reprezentacijske snage. Model je postigao točnost od 88,5% na skupu ImageNet-1K.

Odluku da istražimo EfficientNetV2 kao jednu od glavnih arhitektura potaknula je njegova široka uporaba u rješenjima natjecanja ISIC 2020. Velik dio najbolje rangiranih rješenja koristio je jednu ili više varijanti EfficientNeta, često u sklopu ansambla, što pokazuje njegovu praktičnu učinkovitost u klasifikaciji lezija kože. Ova raširena uporaba i dobre performanse u domeni učinile su ga logičnim izborom za istraživanje.

5.1.4. DINOv2

DINOv2 [32] je samonadzirani vizualni transformer treniran korištenjem okvira DINO (engl. *Self-Distillation with No Labels*). Model uči vizualne reprezentacije bez označenih

podataka koristeći arhitekturu učitelj–učenik, pri čemu obje mreže obrađuju više pogleda iste slike. Mreža učenik trenira se tako da njezini izlazi odgovaraju izlazima mreže učitelja, koja se ažurira eksponencijalnim pomičnim prosjekom. Ova metoda potiče model na učenje semantički bogatih značajki.

DINOv2 smatra se temeljnim modelom (engl. *foundation model*) u računalnom vidu. Budući da je treniran na velikom skupu neoznačenih podataka, značajke dobivene iz zamrznutog DINOv2 modela mogu se koristiti za linearno ispitivanje, pri čemu se na njegovu okosnicu dodaje klasifikacijska glava koja se trenira na izdvojenim značajkama za izvođenje klasifikacijskog zadatka. Ideja je da su značajke DINOv2 dovoljno bogate da klasifikacijska glava postigne dobre rezultate bez treniranja same okosnice.

Vrijedi napomenuti da su autori u ablacijskoj analizi zaključili kako model ne pokazuje jasne obrasce pristranosti prema spolu, boji kože ili dobi. Na slici 5.1. iz izvornog rada vidimo kako izgledaju prve glavne komponente tih značajki:

Slika 5.1. Vizualizacija prvih nekoliko PCA komponenti značajki DINOv2 modela (iz izvornog rada DINOv2)

U našem slučaju koristimo DINOv2 u postavci linearnog ispitivanja. Okosnica je zamrznuta, a trenirali smo samo klasifikacijsku glavu s značajno manjim brojem parametara na značajkama dobivenih okosnicom. Budući da je većina naših modela imala sklonost prenaučavanja zbog izrazite nebalansiranosti klasa, vjerovali smo da bi ovakav pristup s fiksiranom jezgrom mogao biti posebno koristan za postizanje bolje sposobnosti generalizacije na našem skupu podataka.

6. Implementacijski detalji

Za ovaj projekt koristili smo Python 3.10, sustav `conda` za postavljanje okruženja i upravljanje paketima, te `PyTorch`[1] kao glavnu biblioteku za duboko učenje. Biblioteku `submitit`[2] koristili smo za slanje poslova na `Slurm`, `Docker`[3] za kontejnerizaciju koda, a `tmux`[4] za pokretanje koda u odvojenim terminalima i izbjegavanje prekida SSH sesija. `GitHub` je služio za kontrolu verzija i suradnju, dok je `Kaggle`[5] korišten za početno istraživanje modela.

Ostale važne biblioteke uključuju `kneed`[6] za detekciju točke „koljena" funkcije (korisno za ITA-predikciju), `fairlearn`[7] za izračun metrika pravednosti, `torchsampler`[8] za učinkovito naduzorkovanje, `timm`[9] za izradu modela, `seaborn`[10] i `matplotlib`[11] za vizualizaciju te `tensorboardX`[12] za zapisivanje eksperimenata. Izvoz u `ONNX`[13] i `TorchScript`[14] je podržan i opisan u kasnijem poglavlju.

Također smo koristili `HuggingFace`[15] za implementaciju našeg modela i koda, čime je model postao dostupan za korištenje i zaključivanje putem biblioteke Hugging Face Transformers.

[1] `https://pytorch.org`
[2] `https://github.com/facebookincubator/submitit`
[3] `https://www.docker.com`
[4] `https://github.com/tmux/tmux`
[5] `https://www.kaggle.com`
[6] `https://github.com/arvkevi/kneed`
[7] `https://fairlearn.org`
[8] `https://github.com/ufoym/imbalanced-dataset-sampler`
[9] `https://github.com/huggingface/pytorch-image-models`
[10] `https://seaborn.pydata.org`
[11] `https://matplotlib.org`
[12] `https://github.com/lanpa/tensorboardX`
[13] `https://onnx.ai`
[14] `https://pytorch.org/docs/stable/jit.html`
[15] `https://huggingface.co/`

6.1. Modularni dizajn

Glavni princip kojeg smo se morali držati u razvojnom dijelu rada bila je osigurati održivost i nadogradivost koda pomoću modularnosti. Bilo je potrebno brzo iterirati u razvoju i isprobavati što više različitih postavki eksperimenata, što je zahtijevalo oblikovanje koda koji je jednostavan za održavanje i još jednostavniji za nadogradnju. U sljedećim odjeljcima prolazimo kroz glavne komponente razvijenog sustava, fokusirajući se na rad prilagođenog DataLoadera, kako smo omogućili nadogradnju potpuno novih arhitektura uz minimalne promjene koda, te kako upravljamo različitim algoritmima uzorkovanja, optimizatorima, funkcijama gubitka, težinama gubitka, izračunom metrika i sličnim parametrima.

6.1.1. Dizajn DataLoadera

`LocalISICDataset` je modularna implementacija razreda `Dataset` iz biblioteke PyTorch, posebno prilagođena skupu podataka ISIC 2020. Odgovarajući *dataloader* podržava strukturu skupa podataka temeljenu na direktorijima sa zasebnim mapama za `train/test`, `benign/malignant` slike i pripadajuće segmentacijske maske lezija. Maske se koriste samo ako želimo izrezati kožu i učiti isključivo na maskiranim slikama. U slučajevima kada je dostupan `skin_color_csv`, za svaki se primjer dodatno vraćaju i vrijednost ITA, oznaka skupine boje kože i oznaka iz Fitzpatrickove skale.

Strategije augmentiranja, odabira prostora boja i naduzorkovanja podataka ključne su za poboljšanje uspješnosti naučenih modela i osiguravanje njihovog pouzdanog rada. Za maligne primjere moguće je selektivno primijeniti korisnički definirane augmentacije (npr. rotacije, promjene svjetline). *Dataloader* podržava uzorkovanje s omjerom naduzorkovanja izvedenim iz broja augmentacija, što je korišteno za umjetno povećanje broja pozitivnih primjera. Podaci se mogu učitavati u prostorima boja RGB i CIELAB te se dodatno može provesti i transformacija tona kože *skin-former* opisana u poglavlju 3.1.3. Ako je `segment_out_skin=True`, *dataloader* koristi segmentacijske oznake lezija kako bi zamaskirao piksele koji ne sadrže leziju.

Svaki poziv metode `__getitem__` vraća trojku:

```
(image: Tensor, label: int, group: int)
```

gdje `group` označava ton kože pacijenta i koristi se za analizu pravednosti te za treniranje osjetljivo na domenu (grupu).

6.1.2. Dizajn klasifikatora

Klasifikacijski model strukturiran je kao modularni omotač oko okosnice kao ekstraktora značajki, omogućujući fleksibilnost u prebacivanju između različitih arhitektura modela. Svaku okosnicu slijedi klasifikacijska glava čime je cijeli model prilagođen za binarnu klasifikaciju melanoma. Kako bismo podržali treniranje osjetljivo na domenu, klasifikator se može opcionalno proširiti da proizvodi izlaze dimenzija `broj_razreda` × `broj_domena`. Klasifikator prihvaća parametre za kontrolu korištenja prednaučenih težina (uključujući i ImageNet-1K i ImageNet-22K) te zamrzavanja okosnice za potrebe linearne evaluacije. Svaki model vraća ili klasifikacijski token (npr. u ViT/DINOv2) ili objedinjeni vektor značajki (npr. u CNN-ovima), koji se zatim prosljeđuje potpuno povezanoj glavi `nn.Linear`. Glava se dinamički prilagođava broju ciljnih klasa.

Podržane okosnice uključuju arhitekture ConvNeXt, ConvNeXtV2, EfficientNetV2 i DINOv2. Za dodati novu okosnicu potrebno je definirati prilagođenu funkciju za stvaranje modela (`create_X_model`), a potom proširiti logiku odabira modela kako bi uključivala tu dodatnu okosnicu. Postupak odabira okosnice može se sažeti pseudokodom 1

Algorithm 1 Odabir okosnice u modulu `MelanomaClassifier`

function MELANOMACLASSIFIER(*model_name*, *num_classes*, *pretrained*, *in_22k*, *freeze*)
  if *model_name* sadrži "convnext_" **then**
    Učitaj ConvNeXt backbone
    Zamijeni glavu s `Linear(num_features, num_classes)`
  else if *model_name* sadrži "efficientnet" **then**
    Učitaj EfficientNetV2 s navedenim parametrima
  else if *model_name* sadrži "convnextv2" **then**
    Učitaj ConvNeXtV2 s navedenim parametrima
  else if *model_name* sadrži "dinov2" **then**
    Učitaj DINOv2 backbone
    Dodaj `Linear(num_features, num_classes)` na CLS izlaz
  else
    Baci grešku: Nepodržani model
  end if
  return Omotač s okosnicom
end function

6.1.3. Dizajn funkcije gubitka

Definiramo nekoliko prilagođenih funkcija gubitka: `OhemCrossEntropy`, `RecallCrossEntropy`, `DomainIndependentLoss`, `DomainDiscriminativeLoss` i `FocalLoss`, uz podršku za `InverseFrequencyWeighting`. Odabir funkcije gubitka i uključivanje dinamičkog otežavanja primjera kontrolira se putem argumenata naredbenog retka, što omogućuje jednostavno provođenje eksperimenata s različitim ciljevima. Ovaj modularni dizajn također omogućuje jednostavno korištenje metoda za evaluaciju i treniranje neovisno ili osjetljivo na domenu. Za dodavanje nove funkcije gubitka potrebno je samo dodati razred koji nasljeđuje `nn.Module` i ima metodu `forward` te proširiti odgovarajući argument naredbenog retka.

6.1.4. Odabir mjera uspješnosti i optimizatora

Izračun metrika razlikuje se ovisno o tome trenira li se standardni binarni klasifikator ili klasifikator osjetljiv na domenu. U potonjem slučaju, dimenzija izlaza iz modela je `broj_razreda` × `broj_domena`, pa se ti izlazi prije evaluacije dodatno moraju svesti na vektor veličine `broj_razreda`. Metrike se mogu jednostavno proširiti kako bi podržale obje paradigme učenja, a definiramo nekoliko funkcija za obradu predikcija. Optimizator se stvara pomoću obrasca tvornice, preuzetog i prilagođenog iz službenog ConvNeXt repozitorija[16].

6.2. Verzioniranje koda

Koristili smo git i GitHub za verzioniranje i upravljanje kodom. Kako bismo održali dosljedan stil koda, konfigurirali smo `pre-commit` hookove za automatsko sortiranje bibiloteka i formatiranje koda koristeći `isort` i `black`. Ovi alati uključeni su kao opcionalne vanjske bibiliteke. Također smo definirali `pyproject.toml` datoteku za centralizaciju pravila formatiranja za oba alata, koja se automatski primjenjuju pri pokretanju alata putem naredbenog retka ili `pre-commit` hooka.

[16] `https://github.com/facebookresearch/ConvNeXt`

6.3. Treniranje i infrastruktura

Treniranje je provedeno na udaljenom GPU poslužitelju kojem smo pristupali putem protokola SSH. Svaka sesija treniranja pokretana je pomoću *shell* skripte koja je prosljeđivala argumente naredbenog retka našoj glavnoj skripti za učenje. Kako bismo osigurali neprekinuto izvođenje, koristili smo `tmux`, terminalski multiplekser koji omogućuje odvajanje i ponovno spajanje terminalskih sesija. To nam je omogućilo pokretanje dugih sekvenci eksperimenata i njihovo praćenje prema potrebi.

Za praćenje napretka treniranja implementirali smo prilagođeni sustav automatskog zapisivanja koji bilježi:

- Vrijeme treniranja po mini-grupi i po epohi
- Brzinu učenja i minimalnu brzinu učenja
- Vrijednosti gubitka i točnost po klasi
- *Weight decay* i potrošnju memorije

Argumenti naredbenog retka također su se zapisivali, a sustav je podržavao spremanje modela nakon svake epohe, kao i spremanje najboljeg modela prema ciljnoj metrici kao što je F1-mjera ili odziv za malignu klasu. Rezultati evaluacije za svaku epohu zapisivali su u datoteku `training.log`, što nam je omogućilo praćenje uspješnosti modela kroz sva izvođenja.

Također smo podržavali treniranje s mješovitom preciznošću koristeći PyTorchov paket za automatsku mješovitu preciznost (AMP), koje se može uključiti pomoću oznake `-use_amp true`. Omogućena je i podrška za distribuirano treniranje na više GPU kartica pomoću modula `torch.distributed.launch`, a proširili smo ga i podrškom za više čvorova na sustavu SLURM pomoću biblioteke `submitit`, kako što je i predviđeno u službenom ConvNeXt repozitoriju.

Eksperimenti su provedeni na skupu od 8 NVIDIA L4 GPU-ova, svaki s 24 GB memorije. Za svako izvođenje spremali smo sljedeće izlaze:

- `events.out.tfevents.*` (TensorBoard logovi)

- `config.json` (konfiguracija naredbenog retka)
- `training.log` (metrike po epohi)
- `checkpoint_epoch_X.pth` i `best_checkpoint.pth`

6.4. Podrška za ONNX i TorchScript

Kako bismo osigurali kompatibilnost s različitim implementacijskim okruženjima i okvirima za zaključivanje, kao i posluživanje u aplikacijama, podržavamo izvoz modela u formate TorchScript[17] i ONNX[18]. TorchScript je reprezentacija PyTorch modela koja se može izvoditi neovisno o Pythonu. Funkcija `convert_to_torchscript` sprema težine modela u formatu TorchScript:

```
def convert_to_torchscript(model, input_tensor, output_path):
    model.eval()
    scripted_model = torch.jit.trace(model, input_tensor)
    scripted_model.save(output_path)
```

Listing 6..1: Spremanje modela u formatu TorchScript

format ONNX (Open Neural Network Exchange) je format otvorenog koda koji je namijenjen namijenjen predstavljanju modela strojnog učenja. Izvoz se vrši pomoću funkcije `export_model_to_onnx`:

```
def export_model_to_onnx(model, input_tensor, output_path):
    torch.onnx.export(
        model, input_tensor, output_path, export_params=True,
           opset_version=11,do_constant_folding=True, input_names
           =['input'], output_names=['output'], dynamic_axes={'
           input': {0:'batch_size'}, 'output': {0: 'batch_size'}
    )
```

Listing 6..2: Spremanje modela u formatu ONNX

[17] https://pytorch.org/docs/stable/jit.html
[18] https://onnx.ai/

Podrška za oba formata omogućuje prenosivost i brzinu tijekom implementacije, neovisno o infrastrukturi za posluživanje.

7. Eksperimenti

U sljedećim poglavljima prikazani su rezultati naših eksperimenata na zadatku klasifikacije melanoma. Prva sekcija obuhvaća implementacijske detalje zajedničke za sve eksperimente. Sljedeće sekcije 7.2.-7.6. prikazuju rezultate i procjenjuju pravednost svakog pristupa analizom metrika na razini grupe (raščlanjenih), skupnih i metrika na razini slike.

7.1. Implementacijski detalji

Naše modele trenirali smo koristeći kosinusnu adaptivnu stopu učenja [33], AdamW optimizator [34], treniranje u punoj preciznosti te ulazne slike rezolucije 224×224. Skup slika s zloćudnim lezijama umjetno smo proširili pomoću nekoliko augmentacijskih tehnika: horizontalnim preslikavanjem, vertikalnim preslikavanjem, nasumičnom rotacijom manjom od 15 stupnjeva te blagim translacijama. Također smo eksperimentirali s metodom *ColorJitter*, no odlučili smo je ne uključiti kao standardnu augmentaciju, iz razloga što smo smatrali kako je takva augmentacija koja mijenja specifične obrasce boje nepogodna za detekciju inače specifičnih obrazaca pigmentacije zloćudnih lezija.

Pri podjeli podataka na skupove za učenje i validaciju morali smo paziti da ne uvedemo dodatnu neuravnoteženost u podatke te da sve skupine kože budu dobro zastupljene u oba podskupa. Kako je analiza podataka iz poglavlja 4. otkrila da je skupina *smeđe boje kože* bila podzastupljena, odlučili smo je spojiti sa skupinom *preplanule boje kože*. Nadalje, poznato je da podatkovni skup ISIC2020 sadrži više primjeraka istih slika, pa je ukupno 425 duplikata uklonjeno prije treniranja. Rezultirajući broj primjera u svakom podskupu naših podataka za odabranu strategiju podjele prikazan je u tablici 7.1.

Rezultate eksperimenata izvještavamo dvojako; prikazujemo standardne klasifika-

Tablica 7.1. Broj slika u korištenim skupovima za učenje i validaciju za različite skupine tonova kože.

Skupina boje kože	Oznaka lezije	Skup za učenje	Skup za validaciju	Omjer
0 Vrlo svijetla	benigna	1072	268	4:1
	maligna	16	17	1:1
1 Svijetla	benigna	17515	4379	4:1
	maligna	231	58	4:1
2 Srednja	benigna	62774	1569	4:1
	maligna	181	46	4:1
3 Preplanula + Smeđa	benigna	834	209	4:1
	maligna	16	16	1:1

cijke metrike našeg modela, ali i metrike pravednosti, najčešće u obliku metričkih razlika ovisno o boji kože. Tablica 7.2. prikazuje kako su ove nestandardne metrike izračunate. Odlučili smo se za metričke razlike koje se računaju kao maksimalna razlika između određenih raščlanjenih (po skupinama) mjera uspješnosti. To znači da bi ove metričke razlike bile blizu 0 za model koji je pravedan, odnosno koji ima ujednačenu uspješnost među svim definiranim skupinama. Alternativni pristup bio bi korištenje minimalnog omjera između odgovarajućih raščlanjenih metrika.

Tablica 7.2. Mjere uspješnosti klasifikacije i pravednosti.

Metrika	Formula
Balansirana točnost (engl. *Balanced accuracy*)	$\frac{\text{Osjetljivost} + \text{Specifičnost}}{2}$
Stopa odabira (engl. *Selection rate*)	$\frac{TP + FP}{TP + FP + TN + FN}$
Razlika FPR-a (engl. *FPR difference*)	$\max_g \text{FPR}_g - \min_g \text{FPR}_g$
Razlika jednakosti prilika (razlika osjetljivosti/odziva/TPR-a) (engl. *Equal opportunity difference*)	$\max_g \text{Osjetljivost}_g - \min_g \text{Osjetljivost}_g$
Razlika jednakosti mogućnosti (engl. *Equalized odds difference* – EOD)	$\max\{\text{Razlika TPR-a}, \text{Razlika FPR-a}\}$
Razlika demografskog pariteta (engl. *Demographic parity difference* – DPD)	$\max_g \text{StopaOdabira}_g - \min_g \text{StopaOdabira}_g$

7.2. Početni modeli

7.2.1. ConvNeXt-Tiny vs. Efficient-Net M/L

Kao osnovni model odabrali smo model ConvNeXt-Tiny [27] koji je predtrenirian na ImageNet-22k i dodatno ugođen na ImageNet-1k s ulaznom rezolucijom od 224×224. Primijenili smo augmentacije na slikama s zloćudnim lezijama, koristili stopu učenja od 5e-5 te postavili parametar raspada težina (engl. *weight decay*) na 5e-3. Model je treniran 10 epoha koristeći RGB slike i dinamičko otežavanje primjera inverznom zastupljenošću razreda uz standardni gubitak unakrsne entropije. Tablica 7.3. prikazuje rezultate početnih eksperimenata treniranih na podacima u prostorima RGB i CIELAB. Učenje modela na slikama u prostoru boja CIELAB pokazalo je blagi porast osjetljivosti u usporedbi s RGB prostorom; poboljšanje od 5.8 postotnih bodova, s 50.4% na 56.2%. F1 mjera je dosljedno ostajala ispod granice od 30%, a općenito smo primijetili da je veća ukupna osjetljivost povezana s nižim vrijednostima DPD-a. Iz tablice također možemo vidjeti da još uvijek nismo postigli zadovoljavajuću razinu osjetljivosti jer je i dalje ispod 60%. Konfiguracija eksperimenta *Recall CE + CIELAB + IFW* odabrana je kao osnovica za daljnje eksperimente.

Tablica 7.3. Usporedba ConvNeXt-Tiny modela treniranih na RGB i CIELAB ulazima.

Eksperiment	Balansirana točnost	Osjetljivost	FPR	F1	DPD	Razlika osjetljivosti
Recall CE + RGB + IFW	73.8	50.4	2.8	35.8	7.7	31.6
Recall CE + CIELAB + IFW	75.3	56.2	5.6	27.0	12.5	27.5
CIELAB + OHEM	65.4	32.9	2.1	28.6	4.9	42.3

Modeli arhitekture EfficientNetV2 (tablica 7.4.) ostvaruju povećanje osjetljivosti i mjere F1 u odnosu na osnovicu. EfficientNetV2-M postigao je dosad najveću osjetljivost, dok je EfficientNetV2-L imao problema s prenaučenošću na skupu za učenje. Razlog prenaučenosti dodjeljujemo prekapacitiranosti tog modela.

Tablica 7.4. Usporedba uspješnosti između modela EfficientNetV2-M i EfficientNetV2-L.

Eksperiment	Bal. točnost	Osjetljivost	FPR	F1	DPD	Razlika osjetljivosti
Bazni model + EfficientNetV2-M	75.0	56.9	6.9	23.8	14.0	37.5
Bazni model + EfficientNetV2-L	72.7	50.4	5.1	26.0	9.7	54.5

7.2.2. Linearna evaluacija DinoV2 ViT base modela

Eksperimentirali smo s linearnim ispitivanjem treniranjem jednog klasifikacijskog sloja nad predtreniranim zamrznutim DINOv2 modelom kao okosnicom koristeći slike rezolucije 518×518 u prostoru boja RGB. Rezultati su pokazali ili savršenu osjetljivost ili gotovo nikakvu osjetljivost, što ukazuje na ekstremne slučajeve prenaučenosti i podnaučenosti. Ove nedosljednosti signalizirale su nestabilnost u učenju smislenih granica odluke u ovom postavu. Stoga smo odlučili ne nastaviti s ovim pristupom.

Tablica 7.5. Metrike linearne evaluacije DINO base modela na rezoluciji 518×518 (RGB slike). Osj. označava osjetljivost.

Model	Točnost	Osj.	F1	DPD	Osj. skupine 0	Osj. skupine 1	Osj. skupine 3	Osj. skupine 4
Dino ViT-14 Base + IFW unakrsna entropija	24.1	97.1%	5.1%	5.2%	100.0%	94.8%	97.8%	100.0%
Dino ViT-14 Base + naduzorkovannje	75.2	80.3%	11.9%	23.4%	94.1%	70.7%	82.6%	93.8%
Dino ViT-14 Base + poduzorkovanje	95.6	4.4%	4.0%	3.0%	0.0%	7.0%	4.4%	0.0%

7.2.3. ConvNeXt-Tiny + naduzorkovanje

Promatrali smo utjecaj različitih funkcija gubitka pri korištenju strategije naduzorkovanja. Svi modeli trenirani su koristeći ConvNeXt-Tiny arhitekturu na slikama rezolucije 224×224 u prostoru boja CIELAB. Cilj je bio promatrati hoće li i kada model početi pokazivati znakove prenaučenosti.

U svim je slučajevima naduzorkovanje dovelo do prenaučenosti. Model je naučio gotovo savršenu granicu odluke između malignih i benignih slučajeva na skupu za učenje, postižući gotovo savršene rezultate tijekom treniranja. Međutim, naučena granica nije dobro generalizirala na validacijski skup podataka, gdje je uspješnost opala. Ovo ukazuje na to da je model zapamtio naduzorkovane maligne slučajeve umjesto da je naučio općenitije obrasce. Kao rezultat toga, odlučili smo ne nastaviti s rješenjima temeljenima na naduzorkovanju u našem konačnom modelu. Vrijedi spomenuti kako bi ovaj pristup mogao biti valjan u slučaju spajanja više skupova podataka, odnosno prilikom proširenja trenutnog skupa podataka s novim primjerima malignih lezija (kao manjinskog razreda).

Tablica 7.6. Naduzorkovanje s različitim funkcijama gubitka na ConvNeXt-Tiny modelu koristeći 224×224 CIELAB slike. Osj. označava osjetljivost.

Model	Točnost	Osj.	F1	DPD	Osj. skupine 0	Osj. skupine 1	Osj. skupine 3	Osj. skupine 4
ConvNeXt-Tiny + naduzorkovanje	96.0	39.4%	29.0%	21.7%	52.9%	32.8%	54.7%	31.3%
ConvNeXt-Tiny + naduzorkovanje + IFW + recall CE	96.2	40.9%	31.1%	43.7%	76.5%	32.8%	39.1%	37.5%
ConvNeXt-Tiny + naduzorkovanje + OHEM + IFW	96.1	34.3%	26.9%	31.2%	58.8%	27.6%	30.4%	43.8%

7.2.4. Ablacija ulazne rezolucije slika

Usporedba rezolucija u Tablici 7.7. pokazuje da povećanje rezolucije slika ne dovodi uvijek do poboljšanja uspješnosti modela. Za ConvNeXt-Tiny s treniranjem osjetljivim na domenu (boju kože), rezolucija od 224×224 dala je najbolji balans pravednosti i točnosti, postigavši najveću osjetljivost (72.3%) i najmanju razliku demografskog pariteta (26.2 %). Suprotno tome, na višim rezolucijama (448×448 i 672×672) uspješnost se smanjila, s nižom osjetljivošću i većim razlikama među skupinama.

S druge strane, EfficientNetV2-M imao je koristi od veće rezolucije 448×448, gdje je postigao najveću ukupnu osjetljivost (76.6%), iako je mjera F1 ostala niska zbog visokih stopa lažno pozitivnih primjera. Model ConvNeXt-Tiny treniran pomoću strategije *skin-former* pokazao je veću stabilnost kroz različite rezolucije, zadržavajući osjetljivost između 64% i 67% te pokazavši najuravnoteženiju osjetljivost među skupinama na rezoluciji 672×672.

Sveukupno, iako određene arhitekture poput EfficientNeta imaju koristi od većih rezolucija, naši rezultati sugeriraju da 224×224 ostaje najboljim izborom za ConvNeXt-Tiny modele.

Tablica 7.7. Usporedba rezolucija među arhitekturama (CIELAB, različite ulazne veličine). Osj. označava osjetljivost

Model	Rezolucija	Točnost	Osj. testa	F1	DPD	Osj. skupine 0	Osj. skupine 1	Osj. skupine 3	Osj. skupine 4
ConvNeXt Tiny osjetljiv na domenu	224x224	89%	72.3%	21.5%	26.2%	88.2%	62.1%	73.9%	87.5%
ConvNeXt Tiny osjetljiv na domenu	448x448	94%	48.2%	25.8%	25.1%	64.7%	39.7%	47.8%	62.5%
ConvNeXt Tiny osjetljiv na domenu	672x672	94%	46.0%	24.9%	30.2%	64.7%	34.5%	52.2%	50.0%
EfficientNetV2 M	224x224	92.4%	56.9%	23.8%	37.5%	82.4%	44.8%	58.7%	68.8%
EfficientNetV2 M	448x448	83.6%	76.6%	16.3%	31.6%	94.1%	67.3%	87.0%	62.5%
ConvNeXt Tiny + skin-former	224x224	90.3%	66.4%	22.3%	36.5%	88.2%	51.7%	69.6%	87.5%
ConvNeXt Tiny + skin-former	448x448	87.6%	67.2%	18.4%	28.9%	64.7%	58.6%	71.7%	87.5%
ConvNeXt Tiny + skin-former	672x672	91.1%	64.2%	23.2%	24.7%	76.5%	51.7%	71.7%	75.0%

7.3. Transformacije boje kože

Promatrali smo uspješnost modela ConvNeXt-Tiny na dvjema vrstama slika: slikama na kojima su modelu bile vidljive samo lezije (koža je bila maskirana) te slikama transformiranim pomoću pristupa *skin-former* [35] (kao što je opisano u poglavlju 3.1.).

Tablica 7.8. Uspješnost modela ConvNeXt-Tiny na 224x224 CIELAB slikama s transformacijama boje kože (*skin-former*) i maskiranjem kože.

Eksperiment	Balansirana točnost	Osjetljivost	Stopa lažnih pozitiva	F1	DPD	Razlika osjetljivosti
Bazni model + skin-former	78.6	66.4	9.2	22.3	17.6	36.5
Bazni model + maskirana koža	67.0	37.2	3.3	25.7	10.1	58.2
OHEM + CIELAB + IFW + skin-former	64.4	31.4	2.6	24.8	5.1	33.8

Transformacija maskiranja kože rezultirala je smanjenjem uspješnosti u odnosu na prethodne rezultate. Vjerujemo da se objašnjenje nalazi u činjenici da smo koristili predtrenirane težine na ImageNetu, a ova promjena distribucije ulaznih podataka bila je preznačajna da joj se model prilagodi. Osim toga, ova transformacija često je dovodila do većeg broja lažno pozitivnih predikcija te do nestabilnog treniranja kada se nije koristila inicijalizacija predtreniranim težinama, budući da su se modeli lako preučenili na skup za treniranje. Zbog toga smo odlučili ne provoditi daljnje eksperimente s ovom transformacijom. Kao rezultat, okrenuli smo se transformaciji *skin-former* za ablacijske studije s ciljem poboljšanja rezultata. Ona se pokazala jednostavnijom za treniranje, postigla je za 9.5 postotnih bodova veću osjetljivost od EfficientNetV2-M modela te je generirala manji broj lažno pozitivnih predikcija za svaku od demografskih skupina.

7.4. Treniranje osjetljivo na domenu

Ovdje prikazujemo tri naša najbolja pokusa koristeći transformaciju *skin-former* i treniranje osjetljivo na domenu. Istražili smo strategije neovisne o domeni i osjetljive na istu, a rezultate prikazujemo u tablici 7.9.

Tablica 7.9. Uspješnost treniranja ovisno i neovisno o domeni na ConvNeXt-Tiny modelima korištenjem 224×224 CIELAB slika s opcionalnom augmentacijom *skin-former*.

Eksperiment	Balansirana točnost	Osjetljivost	Stopa lažno pozitivnih	F1	DPD	Razlika osjetljivosti
Bazni model + treniranje osjetljivo na domenu	80.8	72.3	10.7	21.5	18.9	26.2
Bazni model + treniranje neovisno o domeni	59.3	19.7	1.0	23.5	2.1	12.2
Bazni model + treniranje osjetljivo na domenu + skin-former	80.0	74.5	14.5	17.4	23.8	25.4

Treniranje neovisno o domeni pokazalo je loše rezultate na testnom skupu. Model je dosljedno predviđao većinu slika kao benigne, što je rezultiralo niskom osjetljivošću. Pokušaj eksplicitnog učenja granice odluke između klasa samo je povećao stopu lažno pozitivnih predikcija i smanjio preciznost. Stoga smo odlučili ne nastaviti s ovom procedurom treniranja.

S druge strane, treniranje osjetljivo na domenu dalo je najviše rezultate osjetljivosti u svim pokusima, uz zadovoljavajuću stopu lažnih negativa. Najbolji pokus povećao je osjetljivost za 8.1 postotnih bodova, dosegnuvši ukupnu osjetljivost od 74.5%. Bili smo spremni žrtvovati dio preciznosti, što je rezultiralo padom F1 mjere s početnih 40% na svega 17.4%. S obzirom na omjer neuravnoteženosti 49 prema 1 u našem skupu podataka, smatrali smo ovaj kompromis prihvatljivim, budući da smo mogli tolerirati veći broj lažno pozitivnih predikcija i višu stopu lažno pozitivnih.

Zbog velikog broja provedenih eksperimenata, ovdje prikazujemo samo dio njih. Za potpuni pregled svih glavnih eksperimenata u tabličnom obliku, pogledajte dodatak C

7.5. Evaluacija pravednosti na razini grupe

U tablici 7.10. vidimo da grupe 1 i 2 imaju osjetljivost od samo 45% i 46%, što odgovara stopama lažno negativnih (FNR) od 55% i 54%. Sličan obrazac se pojavljuje u tablici 7.11., gdje osjetljivost za najbrojniju grupu iznosi samo 43%. Tablica 7.14. također pokazuje umjerenu osjetljivost u tim grupama, oko 45%.

ConvNeXt treniran na slikama u prostoru boja RGB (Tablica 7.10.) odbačen je zbog niske osjetljivosti u najbrojnijim grupama i visoke stope lažnih negativa. ConvNeXt treniran na CIELAB slikama (Tablica 7.11.) imao je još veće stope lažnih negativa i nižu osjetljivost u grupi s najboljim rezultatima, i on je isključen iz daljnjeg razmatranja. EfficientNet-M (Tablica 7.14.) pokazao je nešto nižu ukupnu osjetljivost, ali s manjom stopom lažnih negativa, pri čemu je najveća pogreška zabilježena u grupi s najtamnijom bojom kože. Odlučili smo nastaviti s tim modelom. Odlučujući faktor bila je razlika u osjetljivosti među grupama: ConvNeXt-tiny (RGB) imao je razliku osjetljivosti od 31%, ConvNeXt-tiny (CIELAB) 28%, dok je EfficientNet-M imao najnižu razliku od samo 14%, čak bolju od varijante s augmentacijom *skin-former* koja je iznosila 18%.

Tablica 7.10. Uspješnost po grupama: ConvNeXt-tiny + RGB + recall CE

Grupa	F1	Osjetljivost	Točnost	Stopa odabira	Balansirana točnost	FPR	FNR
0	60%	76%	94%	9%	86%	5%	24%
1	31%	45%	97%	3%	71%	2%	55%
2	31%	46%	94%	5%	71%	4%	54%
3	46%	56%	91%	10%	75%	7%	44%

Tablica 7.11. Uspješnost po grupama: ConvNeXt-tiny + CIELAB + recall CE

Grupa	F1	Osjetljivost	Točnost	Stopa odabira	Balansirana točnost	FPR	FNR
0	62%	71%	95%	8%	83%	4%	29%
1	21%	43%	96%	4%	70%	3%	57%
2	24%	63%	88%	12%	76%	11%	37%
3	42%	69%	86%	16%	78%	12%	31%

Među preostalim modelima u tablicama 7.12., 7.13. i 7.14., F1 je gotovo identičan po grupama. Svi modeli pokazuju blagu pristranost u stopi odabira prema podzastupljenim grupama. ConvNeXt-tiny model koji je koristio treniranje osjetljivo na domenu ističe se s najvišom osjetljivošću kroz sve grupe, pri čemu je najniža osjetljivost na razini grupe još uvijek respektabilnih 62% za najbrojniju grupu boje kože. Varijanta s augmentacijom *skin-former* također je postigla dobre rezultate u smislu osjetljivosti.

Tablica 7.12. Uspješnost po grupama: ConvNeXt-tiny + *skin-former*

Grupa	F1	Osjetljivost	Točnost	Stopa odabira	Balansirana točnost	FPR	FNR
0	59%	88%	93%	12%	91%	7%	12%
1	16%	52%	93%	7%	73%	6%	48%
2	19%	70%	83%	18%	77%	16%	30%
3	39%	88%	81%	24%	84%	20%	13%

Sveukupno, model ConvNeXt-Tiny treniran osjetljvo na domenu pokazuje se kao najpravedniji i najučinkovitiji u smislu grupne osjetljivosti, zbog visoke osjetljivosti i najmanje razlike u osjetljivosti između grupa. Jedina blaga zabrinutost odnosi se na varijaciju u stopi odabira među grupama, što se očituje kroz stopu lažnih pozitiva. Ipak, to nije osobito zabrinjavajuće za našu primjenu (detekcija malignih lezija), budući da prioritet stavljamo na smanjenje lažno negativnih rezultata. U tom kontekstu, viša stopa lažnih pozitiva je prihvatljiva, osobito kada pomaže izbjeći propuštanje malignih slučajeva u

Tablica 7.13. Uspješnost po grupama: ConvNeXt-tiny + treniranje osjetljivo na domenu

Grupa	F1	Osjetljivost	Točnost	Stopa odabira	Balansirana točnost	FPR	FNR
0	56%	88%	92%	13%	90%	8%	12%
1	17%	62%	92%	8%	77%	8%	38%
2	19%	74%	82%	20%	78%	18%	26%
3	36%	88%	78%	27%	83%	22%	13%

Tablica 7.14. Uspješnost po grupama: EfficientNet-M

Grupa	F1	Osjetljivost	Točnost	Stopa odabira	Balansirana točnost	FPR	FNR
0	55%	82%	92%	12%	87%	7%	18%
1	18%	45%	95%	5%	70%	5%	55%
2	21%	59%	87%	13%	73%	12%	41%
3	37%	69%	84%	19%	77%	15%	31%

iznimno nebalansiranom skupu podataka.

7.6. Statistička evaluacija pravednosti

Dosad smo razmatrali samo metrike na razini grupa i mjere dispariteta kao načine kvantificiranja pravednosti. Sljedećim eksperimentima želimo ići korak dalje i sagledati pravednost naših modela iz drugačije perspektive, analizirajući njihovu uspješnost na razini pojedinačnih slika. Budući da se bavimo klasifikacijom, koristit ćemo predviđene aposteriorne vjerojatnosti pripadanja primjeru malignoj klasi $P(y = \text{malignant} \mid x)$ kao mjeru uspješnosti na razini slike. Na taj način, pravednost koju statistički testiramo odgovara jednakoj razini pouzdanosti klasifikacije među grupama boje kože. Ovo poglavlje fokusira se na tri modela prikazana u tablici 7.15., koji su pokazali najbolju ukupnu uspješnost i pravednost.

Tablica 7.15. Najuspješniji eksperimenti odabrani za daljnju analizu pravednosti. Bazna je postavka treniranja konfiguracija ConvNeXt-Tiny + Recall CE + CIELAB + IFW.[1]

Alias	Eksperiment
Eksperiment 1	Bazna postavka treniranja + skin-former
Eksperiment 2	Bazna postavka treniranja + treniranje osjetljivo na domenu
Eksperiment 3	Bazna postavka treniranja + EfficientNet-M

[1] Tablica podrazumijeva da u trećem eksperimentu arhitektura EfficientNet-M nadomješta ConvNeXt-Tiny.

7.6.1. Podskup pozitivnih primjera

Želimo testirati ponašaju li se naučeni modeli različito među grupama boje kože na pozitivnim (malignim) primjerima, neovisno o tome jesu li ih modeli ispravno klasificirali ili ne. Na taj način dobivamo cjelokupni pregled pravednosti svakog modela i provjeravamo razlike u razini pouzdanosti klasifikacije na svim malignim primjerima. Tablica 7.16. uspoređuje naša tri najuspješnija modela i prikazuje srednje aposteriorne vjerojatnosti za svaku grupu boje kože te pripadajuće jednostrane ANOVA statistike. Nulta i alternativna hipoteza za ovaj eksperiment glase:

H_0 ... Srednje vjerojatnosti $P(y = \text{malignant} \mid x)$ jednake su za sve maligne primjere u svim grupama boje kože

H_1 ... Najmanje jedna srednja vjerojatnost razlikuje se od ostalih

Tablica 7.16. Srednje aposteriorne vjerojatnosti malignog razreda za pozitivne primjere (slike malignih lezija) u različitim grupama boje kože. Uspoređena su tri različita modela pomoću jednostrane ANOVA statistike i pripadajućih p-vrijednosti.

Grupa boje kože	Eksperiment 1	Eksperiment 2	Eksperiment 3
0	0.838 ± 0.323	0.320 ± 0.280	0.784 ± 0.334
1	0.508 ± 0.410	0.230 ± 0.272	0.435 ± 0.410
2	0.658 ± 0.376	0.251 ± 0.253	0.581 ± 0.404
3	0.735 ± 0.285	0.261 ± 0.223	0.653 ± 0.361
ANOVA	$F = 4.142$ $p = 0.008$	$F = 0.520$ $p = 0.669$	$F = 3.983$ $p = 0.009$

Rezultati analize varijance pokazuju da možemo odbaciti H_0 na razini značajnosti 0.05 u eksperimentima 1 i 3. To ukazuje na prisutnost pristranosti vezane uz boju kože u ta dva modela; preciznije, ti modeli nisu jednako sigurni u svoje odluke među različitim grupama. Promatrajući srednje vrijednosti po grupama, čini se da modeli 1 i 3 daju više prilika (višu sigurnost u odluku) manjinskim grupama boje kože. Međutim, znajući da je grupa 1 bila većinska u skupu za učenje, možemo s pravom pretpostaviti da su rezultati za ostale grupe previše optimistični. Srednje vrijednosti za eksperiment 2 sugeriraju vrlo stabilnu klasifikacijsku pouzdanost među grupama boje kože, unatoč činjenici da je skup za učenje bio izrazito neuravnotežen. Budući da ova analiza uključuje i primjere koji su pogrešno označeni kao benigni, sva tri modela pokazuju visoku varijancu.

7.6.2. Podskup istinito pozitivnih primjera

Promatrajući samo primjere koji su ispravno označeni kao maligni, možemo provjeriti je li naš model bio manje siguran pri dodjeljivanju te oznake primjerima iz različitih grupa boje kože. Rezultati sljedećeg statističkog testa prikazani su u tablici 7.17.

H_0 ... Srednje vjerojatnosti $P(y = \text{malignant} \mid x)$ jednake su za primjere ispravno klasificirane kao pozitivne kroz sve grupe boje kože

H_1 ... Najmanje jedna srednja vjerojatnost razlikuje se od ostalih

Tablica 7.17. Srednje aposteriorne vjerojatnosti malignog razreda primjera ispravno klasificiranih kao malignih za različite grupe boje kože. Tri različita modela uspoređena su korištenjem prikazanih jednosmjernih ANOVA statistika i p-vrijednosti.

Grupa boje kože	**Eksperiment 1**	**Eksperiment 2**	**Eksperiment 3**
0	0.950 ± 0.112	0.650 ± 0.049	0.930 ± 0.120
1	0.879 ± 0.158	0.639 ± 0.047	0.865 ± 0.157
2	0.885 ± 0.148	0.639 ± 0.048	0.903 ± 0.118
3	0.827 ± 0.157	0.627 ± 0.069	0.881 ± 0.134
ANOVA	$F = 1.612$ $p = 0.192$	$F = 0.131$ $p = 0.941$	$F = 0.741$ $p = 0.531$

Kada se uzimaju u obzir samo ispravno pozitivne predikcije, ne pronalazimo statistički značajne dokaze o razlikama u pouzdanosti i u sva tri eksperimenta ne odbacujemo nul-hipotezu. Za razliku od rezultata prethodne analize varijance, ova nije otkrila nikakve pristranosti u eksperimentima 1 i 3. Ovo dodatno pokazuje da, promatrajući samo istinito pozitivne primjere, dobivamo blago pristran i manje općenit prikaz pravednosti našeg modela. Stoga sumnjamo da modeli u eksperimentima 1 i 3 mogu biti pravedni (dosljedno pouzdani) u slučajevima kada su točni, ali ipak pokazuju određenu pristranost. Kao što je i očekivano te ranije napomenuto, varijance u vjerojatnostima znatno su niže nego u tablici 7.16.

8. Zaključak

Cilj ovog istraživanja bio je procijeniti i poboljšati pravednost dubokih modela strojnog učenja na primjeru klasifikacije lezija kože na primjeru iznimno neuravnoteženih skupova dermatoskopskih slika. U tu svrhu koristili smo skup podataka ISIC 2020 za čije smo primjere odredili kontinuirane oznake za boju kože (vrijednost ITA). Pomoću ovih vrijednosti razdvojili smo podatkovni skup u diskretne skupine određene bojom kože ispitanika. Otkrili smo da je skup podataka izrazito neuravnotežen u korist benignih slika te da omjer broja benignih i malignih primjera iznosi približno 49:1. Ova neuravnoteženost otežava učenje jer uvodi značajnu početnu pristranost u korist klasifikacije slika kao benignih. Još se jedna neuravnoteženost u podacima javlja u raspodjeli primjera po boji kože, pri čemu većina slika pripada osobama svjetlije puti.

U ovom smo radu opisali različite pristupe razvoju modela koji uzimaju u obzir ove dvije neuravnoteženosti te izbjegavaju pristranosti o boji kože i o razdiobi oznaka u skupu za učenje. Nakon iscrpnog eksperimentiranja, uspješno smo ublažili pristranost i na razini boje kože i na razini oznaka, postigavši visoki odziv od 72.3 % na našem validacijskom skupu uz zadovoljavajuće mjere točnosti. Pravednost najuspješnijih modela analizirali smo na tri razine: na razini cijelog ispitnog skupa pomoću mjera pravednosti, usporedbom uspješnosti među skupinama temeljenim na boji kože te na razini pojedinačnih slika analizom varijance.

Najuspješniji model koristio je prilagodbu apriornom vjerojatnosti razreda (engl. *prior-shift inference*), oblik treniranja osjetljivog na domenu koji prilagođava model tijekom zaključivanja kako bi ublažio probleme koji nastaju kada se distribucije podataka za treniranje i testiranje značajno razlikuju. Konkretno, prilagodili smo naš model kako bi podržavao klasifikaciju u CD razreda, gdje C označava broj ciljnih razreda, a D broj domena (boja kože). U našem je slučaju to bio klasifikator s osam izlaza (2×4). Pokazali

smo izvedivost te korisnost takvog treniranja u svrhu povećanja pravednosti modela na primjeru klasifikacije dermatoskopskih slika.

8.1. Budući rad

U budućem radu planiramo proširiti naš skup podataka dodatnim dermatoskopskim slikama. Pretpostavljamo da bi takav pristup značajno poboljšao uspješnost i sposobnost generalizacije modela, kao što je potkrijepljeno literaturom [35]. Nadalje, planiramo naučiti zasebni segmentacijski model kako bismo poboljšali izdvajanje lezija od kože. Naša trenutna metoda segmentacije lezija povremeno je davala netočne maske i uvodila artefakte, poput označavanja rubova slike kao dijela lezije.

Konačno, razmatrali smo metode ansambla, ali smo ih odlučili izostaviti zbog malog broja pozitivnih uzoraka, osobito unutar nedovoljno zastupljenih skupina boje kože. Stratificirana k-stuka unakrsna provjera (engl. *k-fold cross-validation*) nije bila izvediva jer bi razdvajanje podataka uz očuvanje zastupljenosti na razini skupina narušilo pravednost i konzistentnost uspješnosti modela. Treniranje ansambla obično ima koristi od neovisnog paralelnog učenja, no u našem bi slučaju različiti modeli naučili slične obrasce, što bi umanjilo doprinose ansambla. Ovom bismo se pristupu ponovno vratili kada bismo značajno umjetno uvećali podatkovni skup ili uključili dodatne skupove podataka (npr. iz drugih godina natjecanja ISIC).

Također, planiramo koristiti i *ArcFaceLoss* s ciljem jasnijeg razdvajanja komponenti svih klasa. *ArcFaceLoss* omogućuje postizanje veće odvojivosti između domena i klasa, čime se nadamo prijeći s grube na detaljnu klasifikaciju. Kod ovog gubitka ideja je izračunati kut θ između težinskog vektora i vektora značajki primjera te povećati marginu između susjednih skupina primjera, čime se povećava diskriminativnost značajki klasifikatora. Hipoteza je da bi ovako usmjeren model mogao postići bolju separabilnost značajki i veću preciznost.

Zahvala

Posebnu zahvalu željeli bismo posvetiti prof. dr. sc. Siniši Šegviću. Prije svega na silnom znanju koje nam je pružio, bilo stručnom ili životnom. Hvala mu na brižnosti, ohrabrivanju i pruženim uvidima u znanost.

Literatura

[1] J. Dinnes, J. J. Deeks, M. J. Grainge, N. Chuchu, L. Ferrante di Ruffano, R. N. Matin, D. R. Thomson, K. Y. Wong, R. B. Aldridge, R. Abbott, M. Fawzy, S. E. Bayliss, Y. Takwoingi, C. Davenport, K. Godfrey, F. M. Walter, i H. C. Williams, "Visual inspection for diagnosing cutaneous melanoma in adults", *Cochrane Database of Systematic Reviews*, sv. 2018, br. 12, prosinac 2018. https://doi.org/10.1002/14651858.cd013194

[2] T. J. Brinker, A. Hekler, A. H. Enk, C. Berking, S. Haferkamp, A. Hauschild, M. Weichenthal, J. Klode, D. Schadendorf, T. Holland-Letz, C. von Kalle, S. Fröhling, B. Schilling, i J. S. Utikal, "Deep neural networks are superior to dermatologists in melanoma image classification", *European Journal of Cancer*, sv. 119, str. 11–17, rujan 2019. https://doi.org/10.1016/j.ejca.2019.05.023

[3] L. Thomas, P. Tranchand, F. Berard, T. Secchi, C. Colin, i G. Moulin, "Semiological value of abcde criteria in the diagnosis of cutaneous pigmented tumors", *Dermatology*, sv. 197, br. 1, str. 11–17, 1998. https://doi.org/10.1159/000017969

[4] R. M. MacKie, "Clinical recognition of early invasive malignant melanoma." *BMJ*, sv. 301, br. 6759, str. 1005–1006, studeni 1990. https://doi.org/10.1136/bmj.301.6759.1005

[5] H. P. Soyer, G. Argenziano, I. Zalaudek, R. Corona, F. Sera, R. Talamini, F. Barbato, A. Baroni, L. Cicale, A. Di Stefani, P. Farro, L. Rossiello, E. Ruocco, i S. Chimenti, "Three-point checklist of dermoscopy", *Dermatology*, sv. 208, br. 1, str. 27–31, 2004. https://doi.org/10.1159/000075042

[6] R. P. Braun, H. S. Rabinovitz, M. Oliviero, A. W. Kopf, i J. H. Saurat,

"Pattern analysis: a two-step procedure for the dermoscopic diagnosis of melanoma", *Clinics in Dermatology*, sv. 20, br. 3, str. 236–239, svibanj 2002. https://doi.org/10.1016/s0738-081x(02)00216-x

[7] C. Gaudy-Marqueste, Y. Wazaefi, Y. Bruneu, R. Triller, L. Thomas, G. Pellacani, J. Malvehy, M.-F. Avril, S. Monestier, M.-A. Richard, B. Fertil, i J.-J. Grob, "Ugly duckling sign as a major factor of efficiency in melanoma detection", *JAMA Dermatology*, sv. 153, br. 4, str. 279, travanj 2017. https://doi.org/10.1001/jamadermatol.2016.5500

[8] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvehy, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, i P. Soyer, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context", *Scientific Data*, sv. 8, str. 34, 2021. https://doi.org/10.1038/s41597-021-00815-z

[9] Fairlearn, "Fairness assessment", https://fairlearn.org/v0.12/user_guide/assessment/index.html, accessed: 2025-04-29.

[10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, i A. Galstyan, "A survey on bias and fairness in machine learning", kolovoz 2019. https://doi.org/10.48550/ARXIV.1908.09635

[11] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, i J. Vertesi, "Fairness and abstraction in sociotechnical systems", u *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19. ACM, siječanj 2019., str. 59–68. https://doi.org/10.1145/3287560.3287598

[12] A. Corbin i O. Marques, "Assessing bias in skin lesion classifiers with contemporary deep learning and post-hoc explainability techniques", *IEEE Access*, sv. 11, str. 78 339–78 352, 2023. https://doi.org/10.1109/access.2023.3289320

[13] M. Benčević, M. Habijan, I. Galić, D. Babin, i A. Pižurica, "Understanding skin color bias in deep learning-based skin lesion segmentation", *Computer*

Methods and Programs in Biomedicine, sv. 245, str. 108044, ožujak 2024. https://doi.org/10.1016/j.cmpb.2024.108044

[14] M. Kačan, M. Ševrović, i S. Šegvić, "Dynamic loss balancing and sequential enhancement for road-safety assessment and traffic scene classification", *IEEE Transactions on Intelligent Transportation Systems*, sv. 25, br. 11, str. 15 628–15 640, 2024. https://doi.org/10.1109/TITS.2024.3456214

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, i P. Dollár, "Focal loss for dense object detection", https://arxiv.org/abs/1708.02002, 2017.

[16] R. G. Abhinav Shrivastava, Abhinav Gupta, "Training region-based object detectors with online hard example mining", https://arxiv.org/abs/1604.03540v1, 2016.

[17] A. Royer i C. H. Lampert, "Classifier adaptation at prediction time", u *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, lipanj 2015., str. 1401–1409. https://doi.org/10.1109/cvpr.2015.7298746

[18] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, i O. Russakovsky, "Towards fairness in visual recognition: Effective strategies for bias mitigation", u *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020., str. 8916–8925. https://doi.org/10.1109/CVPR42600.2020.00894

[19] R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, i J. Zou, "Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review", *JAMA Dermatology*, sv. 157, br. 11, str. 1362, studeni 2021. https://doi.org/10.1001/jamadermatol.2021.3129

[20] L'Oréal Research and Innovation, "Expert in skin and hair types around the world". [Mrežno]. Adresa: https://www.loreal.com/en/articles/science-and-technology/expert-inskin/

[21] L. N. Montoya, J. S. Roberts, i B. S. Hidalgo, "Towards fairness in ai for melanoma detection: Systemic review and recommendations", 2024. https://doi.org/10.48550/ARXIV.2411.12846

[22] T. B. Fitzpatrick, “The validity and practicality of sun-reactive skin types i through vi”, *Archives of Dermatology*, sv. 124, br. 6, str. 869, lipanj 1988. https://doi.org/10.1001/archderm.1988.01670060015008

[23] T. Lee, V. Ng, R. Gallagher, A. Coldman, i D. McLean, “Dullrazor®: A software approach to hair removal from images”, *Computers in Biology and Medicine*, sv. 27, br. 6, str. 533–543, studeni 1997. https://doi.org/10.1016/s0010-4825(97)00020-6

[24] A. Chardon, I. Creotis, i C. Hourseau, “Skin colour typology and suntanning pathways”, *International Journal of Cosmetic Science*, sv. 13, br. 4, str. 191–208, kolovoz 1991. https://doi.org/10.1111/j.1467-2494.1991.tb00561.x

[25] M. Osto, I. H. Hamzavi, H. W. Lim, i I. Kohli, “Individual typology angle and fitzpatrick skin phototypes are not equivalent in photodermatology”, *Photochemistry and Photobiology*, sv. 98, br. 1, str. 127–129, studeni 2021. https://doi.org/10.1111/php.13562

[26] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, i M. H. Yap, “Analysis of the isic image datasets: Usage, benchmarks and recommendations”, *Medical Image Analysis*, sv. 75, str. 102305, siječanj 2022. https://doi.org/10.1016/j.media.2021.102305

[27] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, i S. Xie, “A convnet for the 2020s”, https://arxiv.org/abs/2201.03545, 2022.

[28] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jaeger, i K. Maier-Hein, “Mednext: Transformer-driven scaling of convnets for medical image segmentation”, https://arxiv.org/pdf/2303.09975, 2024.

[29] Y. Liu, Y. Wang, J. Zhang, X. Li, i H. Chen, “Deep learning-based skin lesion classification using dermoscopic images: A comprehensive review”, https://www.sciencedirect.com/science/article/abs/pii/S0010482523004250, 2023.

[30] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, i S. Xie, “Convnext v2: Co-designing and scaling convnets with masked autoencoders”, 2023. https://doi.org/10.48550/ARXIV.2301.00808

[31] Q. V. L. Mingxing Tan, “Efficientnetv2: Smaller models and faster training”, https://arxiv.org/abs/2104.00298, 2021.

[32] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, i P. Bojanowski, “Dinov2: Learning robust visual features without supervision”, 2023. https://doi.org/10.48550/ARXIV.2304.07193

[33] F. H. Ilya Loshchilov, “Sgdr: Stochastic gradient descent with warm restarts”, https://arxiv.org/abs/1704.00109, 2017.

[34] ——, “Decoupled weight decay regularization”, https://arxiv.org/abs/1711.05101, 2017.

[35] A. Corbin i O. Marques, “Assessing bias in skin lesion classifiers with contemporary deep learning and post-hoc explainability techniques”, https://ieeexplore.ieee.org/document/10162202/, 2023.

Sažetak

Metode procjene i poboljšanja pravednosti dubokih klasifikatora

Marko Haralović, Duje Štolfa

Prilikom razvoja sustava baziranih na strojnom učenju potrebno je osigurati da rade pravedno i da neće prouzročiti štetu nijednoj skupini korisnika. Pouzdani tehnološki alati za automatsku detekciju melanoma imaju ključnu ulogu u ranom prepoznavanju bolesti i pravovremenom liječenju. Postojeći dermatoskopski skupovi podataka ograničeni su u pogledu podzastupljenosti osoba s tamnijim tonovima kože, što predstavlja rizik da modeli budu nepravedni spram demografskih skupina.

Analizirali smo različite pristupe za poboljšanje pravednosti: treniranje osjetljivo na ton kože, ciljano uzorkovanje podataka, prilagodbe funkcije gubitka te metode postprocesiranja. Pravednost modela ocjenjivali smo na tri razine: cijelog skupa podataka, pojedinih skupina (prema boji kože) te na razini pojedinačnih slika, koristeći kombinaciju statističkih alata i standardnih klasifikacijskih metrika.

Rezultati su pokazali da većina jednostavnijih pristupa, zbog izrazito neuravnoteženog skupa podataka, ima vrlo nizak broj ispravno prepoznatih malignih lezija te nisu primjenjivi u praksi. Treniranje osjetljivo na boju kože pokazalo se kao najučinkovitiji način povećavanja pravednosti, dok su strategije proširenja podataka temeljene na transformacijama boje kože dodatno pridonijele boljoj ravnoteži između skupina.

Ključne riječi: dermatoskopske slike, pravednost, duboko učenje, klasifikacija melanoma, treniranje osjetljivo na ton kože

Summary

Fairness assessment and unfairness mitigation methods for deep classifiers

Marko Haralović, Duje Štolfa

A key requirement for any machine-learning system is that it performs fairly and doesn't cause harms to any group of users. Reliable technological tools for automatic melanoma detection play a key role in early disease recognition and timely treatment. Existing dermatoscopic datasets are limited in terms of the underrepresentation of people with darker skin tones, which poses a risk that models may be unfair toward demographic groups.

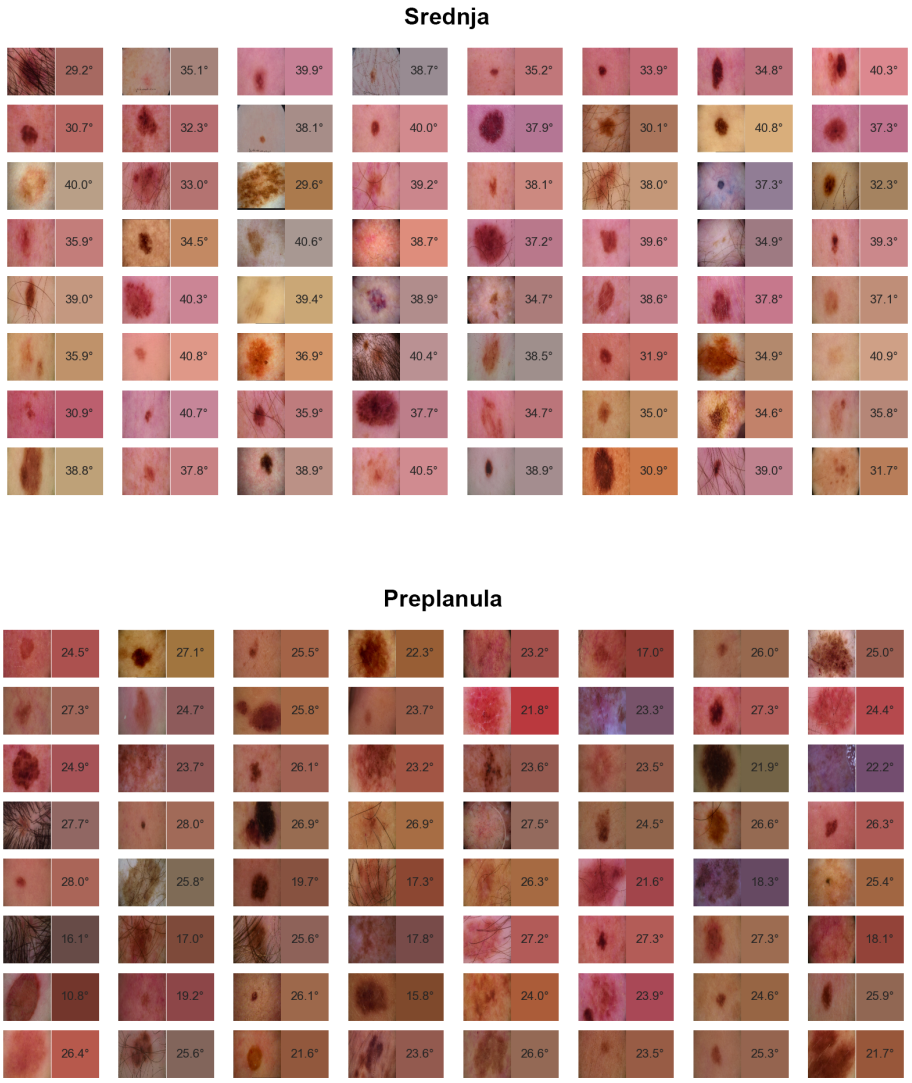
We analyzed different strategies for reducing bias: domain independent and discriminative training, targeted class sampling, loss function adjustments, and post-processing methods. We evaluated model fairness at three levels: the entire dataset, individual groups (by skin color), and individual images, using a combination of statistical tools and classification metrics.

The results showed that most simpler approaches, due to the highly imbalanced dataset, have a very low number of correctly detected malignant lesions and are therefore not applicable in practice. Domain discriminative approaches proved to be the most effective way of increasing fairness, while data augmentation strategies based on skin color transformations further contributed to better balance across groups.

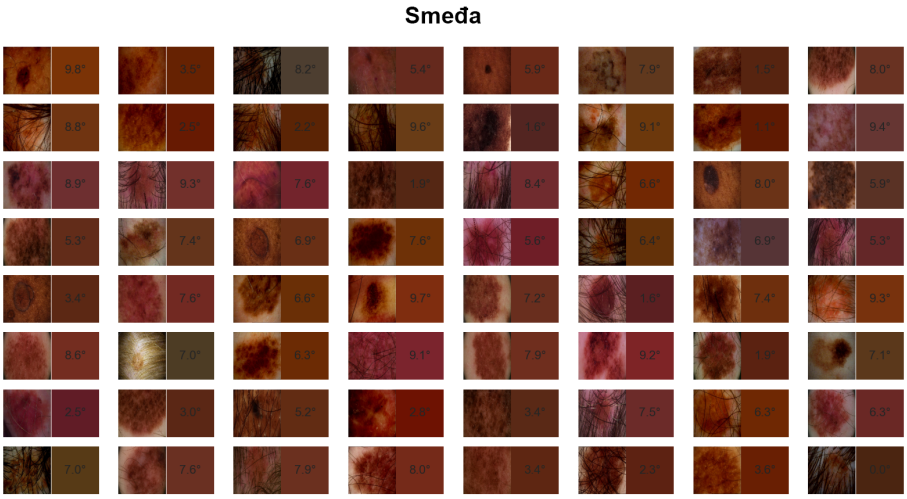
Keywords: dermatological imaging, fairness, deep learning, melanoma classification, skin-tone aware training

Privitak A: Procjena vrijednosti ITA

Slika A1. Slike s procijenjenim ITA vrijednostima i dominantnim bojama kože. Skupine: vrlo svijetla i svijetla.

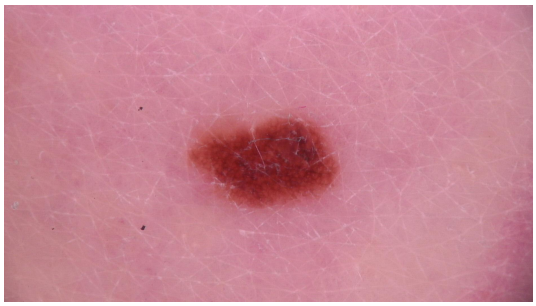


Slika A2. Slike s procijenjenim ITA vrijednostima i dominantnim bojama kože. Skupine: srednje svijetla i preplanula.

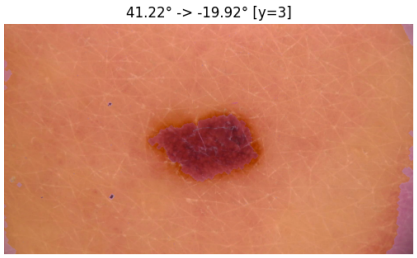


Slika A3. Slike s procijenjenim ITA vrijednostima i dominantnim bojama kože. Skupina: smeđa.

Privitak B: Vizualizacija augmentacije *skin-former*

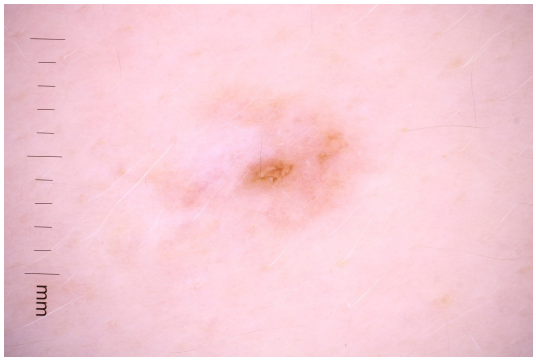


(a) Izvorna slika

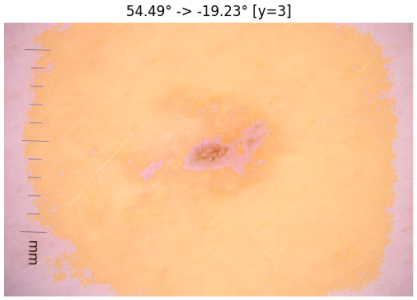


(b) Slika nakon skin-former transformacije. ITA smanjen za 61,14°

Slika B1. Usporedba prije i poslije skin-former augmentacije za ISIC_0052212.



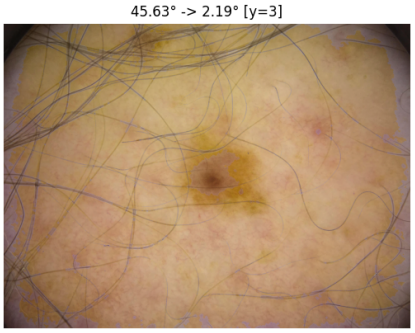
(a) Izvorna slika; vidljiv je artefakt ravnala.



(b) Slika nakon skin-former transformacije. ITA smanjen za 73,72°. Tamniji rubovi slike stvorili su artefakte u segmentacijskim maskama.

Slika B2. Usporedba prije i poslije skin-former augmentacije za ISIC_0075914.

(a) Izvorna slika



(b) Slika nakon skin-former transformacije. ITA smanjen za 43,42°. Dlaka je uspješno maskirana.

Slika B3. Usporedba prije i poslije skin-former augmentacije za ISIC_0078712.

Privitak C: Rezultati odabranih eksperimenata

Tablica C1. Prikaz ključnih metrika za svaki od eksperimenata

Eksperiment	Točnost	Uravnotežena točnost	F1	Odziv	Stopa odabira	DPD	Jedn. prig.
ConvNeXt Tiny + IFW + CE (RGB)	0.971	0.578	0.197	0.168	0.0148	0.0439	0.2888
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu (RGB)	0.962	0.738	0.358	0.504	0.0379	0.0770	0.3164
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu	0.937	0.753	0.270	0.562	0.0661	0.1250	0.2748
ConvNeXt Tiny + IFW + OHEM	0.966	0.654	0.286	0.328	0.0271	0.0491	0.4229
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu 50 epoha	0.951	0.693	0.266	0.423	0.0456	0.1116	0.5366
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu + *skin-former*	0.903	0.786	0.222	0.664	0.1038	0.1755	0.3651
ConvNeXt Tiny + IFW + OHEM + *skin-former*	0.960	0.644	0.248	0.314	0.0320	0.0509	0.3382
ConvNeXt Tiny + IFW + osjetljivo na domenu	0.890	0.808	0.215	0.723	0.1193	0.1886	0.2617
ConvNeXt Tiny + IFW + naduzorkovanje	0.945	0.726	0.275	0.496	0.0546	0.0949	0.4442
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu + neovisno o domeni	0.973	0.593	0.235	0.197	0.0142	0.0210	0.1217
ConvNeXt Tiny + fokalni gubitak	0.970	0.642	0.292	0.299	0.0219	0.0457	0.5184
EfficientNetV2-M + IFW + unakrsna entropija t.n. odzivu	0.924	0.750	0.238	0.569	0.0789	0.1395	0.3753
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu + segmentacija kože	0.955	0.670	0.257	0.372	0.0396	0.1008	0.5822
EfficientNetV2-L + IFW + unakrsna entropija t.n. odzivu	0.940	0.727	0.260	0.504	0.0600	0.0965	0.5446
ConvNeXt-Tiny + naduzorkovanje	0.960	0.683	0.290	0.394	0.0360	0.0608	0.2169
ConvNeXt-Tiny + naduzorkovanje + IFW + unakrsna entropija t.n. odzivu	0.962	0.691	0.311	0.409	0.0340	0.0762	0.4371
ConvNeXt-Tiny + naduzorkovanje + OHEM + IFW	0.961	0.659	0.269	0.343	0.0323	0.0735	0.3124
Dino ViT-14 Base + IFW unakrsna entropija	0.241	0.598	0.051	0.971	0.7783	0.2297	0.0517
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu + osjetljivo na domenu + *skin-former*	0.853	0.800	0.174	0.745	0.1576	0.2379	0.2543
ConvNeXt Tiny osjetljivo na domenu + IFW + unakrsna entropija t.n. odzivu + 448x448	0.942	0.717	0.258	0.482	0.0571	0.1023	0.2505
EfficientNetV2-M + IFW + unakrsna entropija t.n. odzivu + 448x448	0.836	0.802	0.163	0.766	0.1756	0.2113	0.3162
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu + *skin-former* + 448x448	0.876	0.776	0.184	0.672	0.1315	0.1300	0.2888
ConvNeXt Tiny osjetljivo na domenu + IFW + unakrsna entropija t.n. odzivu + 672x672	0.942	0.706	0.249	0.460	0.0562	0.0708	0.3022
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu + *skin-former* + 672x672	0.911	0.780	0.232	0.642	0.0948	0.1360	0.2475

Tablica C2. Metrike povezane s pravednosti kroz eksperimente

Eksperiment	FNR	FPR	FN pogreška	FP pogreška	Razl. FNR	Razl. FPR	Razl. toč.	Razl. urav. toč.	Razl. F1	Razl. odziva
ConvNeXt Tiny + IFW + CE (RGB)	0.832	0.012	0.0626	0.0026	0.2888	0.0212	0.0508	0.1343	0.3286	0.2888
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu (RGB)	0.496	0.028	0.0837	0.0040	0.3164	0.0473	0.0667	0.1512	0.2988	0.3164
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu	0.438	0.056	0.0831	0.0056	0.2748	0.0901	0.0965	0.1359	0.4008	0.2748
ConvNeXt Tiny + IFW + OHEM	0.672	0.021	0.0786	0.0035	0.4229	0.0208	0.0330	0.2040	0.4199	0.4229
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu 50 epoha	0.577	0.038	0.0827	0.0046	0.5366	0.0633	0.0613	0.2367	0.3802	0.5366
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu + *skin-former*	0.336	0.092	0.0791	0.0071	0.3651	0.1331	0.1226	0.1786	0.4234	0.3651
ConvNeXt Tiny + IFW + OHEM + *skin-former*	0.686	0.026	0.0777	0.0039	0.3382	0.0237	0.0647	0.1701	0.3251	0.3382
ConvNeXt Tiny + IFW + osjetljivo na domenu	0.277	0.107	0.0750	0.0075	0.2617	0.1495	0.1384	0.1275	0.3857	0.2617
ConvNeXt Tiny + IFW + naduzorkovanje	0.504	0.045	0.0837	0.0051	0.4442	0.0558	0.0819	0.1944	0.3119	0.4442
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu + neovisno o domeni	0.803	0.010	0.0666	0.0025	0.1217	0.0065	0.0520	0.0576	0.1941	0.1217
ConvNeXt Tiny + fokalni gubitak	0.701	0.016	0.0767	0.0031	0.5184	0.0160	0.0584	0.2618	0.5059	0.5184
EfficientNetV2-M + IFW + unakrsna entropija t.n. odzivu	0.431	0.068	0.0829	0.0062	0.3753	0.1068	0.1115	0.1735	0.3678	0.3753
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu + segmentacija kože	0.628	0.033	0.0810	0.0043	0.5822	0.0893	0.1207	0.2732	0.3772	0.5822
EfficientNetV2-L + IFW + unakrsna entropija t.n. odzivu	0.496	0.051	0.0837	0.0054	0.5446	0.0502	0.0542	0.2520	0.3773	0.5446
ConvNeXt-Tiny + naduzorkovanje	0.606	0.028	0.0818	0.0041	0.2169	0.0366	0.0565	0.0996	0.2045	0.2169
ConvNeXt-Tiny + naduzorkovanje + IFW + unakrsna entropija t.n. odzivu	0.591	0.026	0.0823	0.0039	0.4371	0.0394	0.0712	0.1996	0.3342	0.4371
ConvNeXt-Tiny + naduzorkovanje + OHEM + IFW	0.657	0.026	0.0795	0.0039	0.3124	0.0505	0.0765	0.1458	0.3235	0.3124
Dino ViT-14 Base + IFW unakrsna entropija	0.029	0.774	0.0282	0.0102	0.0517	0.2433	0.2182	0.1216	0.1255	0.0517
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu + osjetljivo na domenu + *skin-former*	0.255	0.145	0.0730	0.0086	0.2543	0.2037	0.1878	0.0807	0.2742	0.2543
ConvNeXt Tiny osjetljivo na domenu + IFW + unakrsna entropija t.n. odzivu + 448x448	0.518	0.048	0.0837	0.0052	0.2505	0.0701	0.0818	0.1167	0.3281	0.2505
EfficientNetV2-M + IFW + unakrsna entropija t.n. odzivu + 448x448	0.234	0.163	0.0709	0.0090	0.3162	0.2027	0.1948	0.2355	0.3114	0.3162
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu + *skin-former* + 448x448	0.328	0.120	0.0786	0.0079	0.2888	0.1093	0.1074	0.1010	0.2783	0.2888
ConvNeXt Tiny osjetljivo na domenu + IFW + unakrsna entropija t.n. odzivu + 672x672	0.540	0.048	0.0835	0.0052	0.3022	0.0450	0.0667	0.1413	0.3429	0.3022
ConvNeXt Tiny + IFW + unakrsna entropija t.n. odzivu + *skin-former* + 672x672	0.358	0.083	0.0803	0.0068	0.2475	0.0999	0.1009	0.1192	0.3662	0.2475