# Nepristrasna detekcija melanoma

Katarina Krstin, Jovan Vučković
Softversko Inženjerstvo i Informacione Tehnologije, Univerzitet u Novom Sadu, Fakultet tehničkih nauka
krstin.sv57.2021@uns.ac.rs, vuckovic.sv64.2021@uns.ac.rs

## I. Problem

Melanom, maligna kožna lezija, je jedan od najagresivnijih oblika raka kože. Njegovo rano otkrivanje je ključno za poboljšanje lečenja kod pacijenata [1]. Standardni dijagnostički postupak je dermatoskopija, koja omogućava detaljniji uvid u strukture kožnih lezija [2]. Ona se služi dermatoskopom koji pravi slike lezija visokog kvaliteta za dalju analizu. Rast broja pacijenata i složenost vizuelne procene ukazuju na potrebu za efikasnijim dijagnostičkim pristupima [1].

Dijagnoza melanoma u kliničkoj praksi se zansiva na ABCDE kriterijumima – asimetrija (engl. *asymmetry*), ivice (engl. *borders*), boja (engl. *color*), prečnik (engl. *diameter*) i evolucija, promene tokom vremena (engl. *evolution*). Njihova primena zahteva vreme i iskustvo dermatologa [3]. Automatizovani sistemi zasnovani na veštačkoj inteligenciji efikasno analiziraju velike skupove dermoskopskih slika [1]. Automatski primenjuju dijagnostičke smernice i izdvajaju podatke od kliničkog značaja [2]. Time se smanjuje broj propuštenih slučajeva, dermatolozi se rasterećuju i omogućuje se fokus na složenije dijagnostičke izazove.

Ovakvi sistemi su predmet brojnih istraživanja, ali ostaje pitanje njihove nepristrasnosti – da li podjednako pouzdano prepoznaju melanome kod pacijenata sa različitim tonovima kože. Nepristrasni modeli doprineli bi ravnopravnijem pristupu zdravstvenoj zaštiti i smanjenju rizika od pogrešne dijagnoze kod ranjivih grupa pacijenata. Nedostatak uravnoteženih skupova podataka, u kojima su zastupljeni svi tonovi kože, otežava izgradnju modela koji imaju jednaku tačnost za sve grupe [1]. Potrebno je razvijati tehnike koje unapređuju nepristrasnost, kako bi modeli zadržali visoke performanse bez favorizovanja određenih populacija.

## II. Teorijske Osnove

Opisani problem rešavamo primenom metoda dubokog učenja. Akcenat se stavlja na nepristrasnost modela. Nepristrasnost u veštačkoj inteligenciji (engl. *fairness in AI*) odnosi se na razvoj algoritama koji ostvaruju ujednačene performanse u različitim demografskim i fiziološkim grupama [4]. U slučaju detekcije melanoma, pristrasni modeli mogu da imaju višu tačnost za pacijente sa svetlijim tonovima kože, dok kod tamnijih tonova ostvaruju značajno slabije rezultate [5]. Takva razlika direktno utiče na pouzdanost i kvalitet dijagnostičkog procesa, što može rezultovati neujednačenim kliničkim ishodima među pacijentima različitih tonova kože.

Dijagnostika melanoma odvija se u različitim uslovima. Klinički faktori potiču od osobina pacijenata i lezija. Ton kože i varijacije u boji lezija direktno utiču na pouzdanost dijagnostike [5]. Neravnomerna distribucija boje lezija povećava rizik od pogrešne klasifikacije. Tehnički faktori potiču od načina prikupljanja i obrade podataka, oni uključuju neravnotežu u dostupnim dataset-ovima, pri čemu većina sadrži slike osoba svetlije kože [1]. Takva pojava uzrokuje pristrasnost modela. Razumevanje ovih faktora pokazuje da problem nije samo u algoritmu. Ono ukazuje da poboljšanja moraju da uključe i kliničku praksu i izgradnju uravnoteženih skupova podataka.

Cilj mašinskog učenja kod ovog problema je razviti model koji precizno detektuje maligne lezije i postiže jednaku tačnost za sve tonove kože [5]. Tonovi kože sa slika u *dataset-u* su prethodno označeni prema *Monk Skin Tone* (MST) skali. Ona je standardna desetostruka skala razvijena od strane *Google-a*. Kvantifikuje nijanse kože od svetlih do tamnih. To omogućava evaluaciju i treniranje modela sa ciljem postizanja jednako preciznih predikcija za sve grupe tonova kože [1]. Performanse modela ocenjuju se metrikama kao što su tačnost (engl. *accuracy*), AUC-ROC (engl. *Area Under the Receiver Operating Characteristic Curve*), odaziv (engl. *recall*) i F-mera (engl. *F1-score*), koje pokazuju koliko dobro model razlikuje maligne i benigni slučajeve [2]. Nepristrasnost modela meri se standardnom devijacijom predikcija po grupama tonova kože (engl. *Fairness parity)* i *Fairlearn* metrike, kao što je *demographic parity difference*, da bi se proverilo da li model favorizuje neku grupu [4].

$$Fairness\ parity\ = \sigma\left(\mu_1, \mu_2, \ldots, \mu_G\right)$$

G - broj grupa po tonovima kože (10 po *MST* skali)
$\mu_G$ - prosečna predikcija modela za grupu G

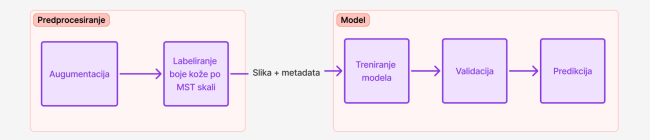
Model mora biti otporan na šum, koji nastaje zbog promena osvetljenja, rezolucije slika ili prisustva artefakata na slici. Interpretabilnost je važna za razumevanje odluka modela i poverenje dermatologa.

Korišćen je *CNN* model zasnovan na *EfficientNetB0* arhitekturi. *EfficientNet* je *state-of-the-art* arhitektura koja postiže visoku tačnost uz manji broj parametara i brže treniranje u poređenju sa klasičnim modelima, poput *ResNet-a* i *VGG-a* [2]. Arhitektura koristi uniformno skaliranje (engl. *compound scaling)*, što omogućava istovremeno povećanje širine, dubine i rezolucije modela, omogućavajući detaljniju analizu lezija bez gubitka brzine obrade. Za poboljšanje generalizacije korišćen je transfer znanja sa modela treniranog na *ImageNet* skupu. *FairSkinToneLoss* dodatno je uveden kako bi se model trenirao da predviđa jednako precizno za sve *MST* grupe tonova kože [5]. *FairSkinToneLoss* funkcija je modifikovana funkcija gubitka koja kažnjava model za razlike u predikcijama po *MST* grupama. Ovaj pristup omogućava da model zadovolji zahteve kliničke preciznosti, nepristrasnosti i robusnosti na varijacije u kvalitetu i osvetljenju slika [5].

## III. Rešenje

Cilj rada je razvoj modela za nepristrasnu detekciju melanoma nezavisno od tona kože. Kao ulaz očekuju se fotografije kožnih lezija uz opcione metapodatke o pacijentu. Izlaz sistema je binarna klasifikacija — benigno ili maligno — uz obezbeđivanje uravnoteženih performansi po svim grupama tonova kože. Na ovaj način, model smanjuje pristrasnost, što ga čini primenjivim u kliničkim i

istraživačkim okruženjima gde je inkluzivnost od ključne važnosti.




Slika 1: Struktura Rešenja

Slika 1 ilustruje korake u rešavanju problema detekcije melanoma.

*A. Pretprocesiranje*

Za izradu rada korišćen je skup podataka *ISIC Challange 2020* podeljen u skup za trening i skup za test. Podaci su nebalansirani, samo 1,7% podataka je maligno, takodje ima malo slika sa najtamnijim tipom kože. U fazi pretprocesiranja, metapodaci sa string vrednostima su kodirani metodom one-hot enkodiranja kako bi se izbeglo uvođenje veštačkog ordinalnog odnosa između kategorija. Sve slike su učitane u RGB formatu i uniformno skalirane na 244×244 piksela radi standardizacije ulaza [6]. Kako bi se obezbedila robusnost modela i smanjilo overfitovanje, primenjena je augmentacija pomoću biblioteke *Albumentations*, uključujući rotacije, translacije, promene kontrasta i osvetljenja, zamućenja i geometrijske transformacije. Svaka maligna slika augmentovana je više puta kako bi se postigla bolja balansiranost skupa [7]. Tonovi kože su labelirani pomoću *MST* skale. Ton kože sa slike se određivao na osnovu skale i *Delta E 2000* algoritma. Boja sa slike uzeta je pomoću super-piksela. Izdvojena je regija kože, boja je uprosečena srednjom vrednošću tona kože i prosleđena algoritmu *Delta E 2000* za označavanje konkretne labele [8].

*B. Treniranje modela*

Za treniranje modela korišćen je *EfficientNet* kao osnovna arhitektura [9]. Upotrebljen je unapred trenirani model na *ImageNet* skupu, pri čemu je uklonjen klasifikacioni sloj i zamenjen globalnim prosečnim *pooling-om* [10]. Na taj način dobijene su kompaktne reprezentacije slika. One su integrisane sa meta-podacima. Meta-podaci se obrađuju kroz mrežu sa dva potpuno povezana sloja, *Swish* aktivacijom i *batch* normalizacijom [4]. Izlazi iz obe grane – slike i meta-podataka – spajaju se i šalju u završni sloj za binarnu klasifikaciju [10].

Korišćena je prilagođena implementacija *Swish* funkcije sa ručno definisanim gradijentom, kako bi se poboljšao protok gradijenata i stabilnost učenja [12]. Tokom treniranja, model je koristio pet *dropout* slojeva sa 50% verovatnoće isključivanja neurona [14]. Za svaku propagaciju izlaz se računa kroz sve *dropout* slojeve i proseči, čime se dobija konačna predikcija [15]. Meta-podaci, ako su prisutni, prolaze kroz zasebnu mrežu sa *dropout* slojem od 30% između *dense* slojeva [6].

Funkcija gubitka, *FairSkinToneLoss,* zasniva se na binarnoj unakrsnoj entropiji (*BCE)*, proširenoj kaznenim faktorom za neuravnotežene predikcije po grupama tonova kože [11]. Ona se definiše kao:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1 - y_i)log(1 - \hat{y}_i)] + \alpha \cdot \sigma_{MST}(\hat{y})$$

Gde je:

$y_i \in \{0, 1\}$ - stvarna labela (benigno/maligno)

$\hat{y}_i$ - predikcija modela

$N$ - broj primera u *batch-u*

$\sigma_{MST}(\hat{y})$ - standardna devijacija prosečnih predikcija po *MST* grupama tonova kože

$\alpha$ - težina *fairness* kazne

Ovaj faktor kazne podstiče model da smanji razlike u predikcijama između tonova kože, što doprinosi uravnoteženijoj klasifikaciji melanoma i povećava pouzdanost modela za sve tipove kože [4].

## IV. Rezultati

Cilj evaluacije bio je da se ispita da li predloženi model može da postigne visoku tačnost u detekciji melanoma i da istovremeno obezbedi ravnomerne performanse za sve grupe tonova kože prema *MST* skali. Poželjni ishodi evaluacije bili su $AUC \geq 0.85$ i minimalne razlike u performansama između grupa manja od 5%. Neželjeni ishodi su niski odziv (engl. *recall*) za maligne slučajeve ili izražene razlike u performansama između svetlijih i tamnijih tonova kože.

Da bismo utvrdili da li je naše rešenje primenljivo u praksi, model je evaluiran nad unapred podeljenim *ISIC 2020* skupom podataka, koji sadrži odvojene skupove za treniranje i testiranje. Procedura evaluacije obuhvatila je primenu modela na test skupu, uz izračunavanje standardnih metrika performansi: tačnost, odziv (engl. *recall*), F1 mera i *AUC-ROC*. Pored toga, evaluirane su i mere nepristrasnosti korišćenjem *Fairlearn* biblioteke – standardna devijacija predikcija po grupama tonova kože i *demographic parity difference*, pri čemu su kožni tonovi klasifikovani pomoću *Delta E 2000* algoritma i mapirani na *MST* skalu (10 grupa).

Model je postigao ukupnu tačnost od 85,9%, sa *AUC-ROC* od 0,83 i F1 merom od 0,78. Recall za maligne slučajeve iznosio je 0,74, što pokazuje da model prepoznaje pozitivne slučajeve, ali sa značajnim prostorom za poboljšanje. Rezultati po grupama tonova kože prikazani su u Tabeli I.

Tabela I: Rezultati

| *MST* grupa | *Accuracy* | *Recall* | *F1* | *AUC* |
|---|---|---|---|---|
| 1-2 (svetla) | 87,5% | 0,76 | 0,79 | 0,84 |
| 3-5 (srednja) | 85,8% | 0,73 | 0,77 | 0,82 |
| 6-8 (tamna) | 84,1% | 0,71 | 0,74 | 0,80 |
| 9-10 (vrlo tamna) | 83,5% | 0,68 | 0,72 | 0,79 |

Standardna devijacija tačnosti između grupa bila je 1,7%, dok je *demographic parity difference* iznosio 0,11, što ukazuje na vidljive razlike u performansama između grupa, naročito kod tamnijih tonova kože.

Niske performanse modela delom se mogu pripisati neravnomernoj zastupljenosti tonova kože u *ISIC 2020* skupu podataka. Iako je augmentacija primenjena na maligne slučajeve radi balansiranja klasa, tamniji tonovi kože su i dalje slabije zastupljeni, što se odražava u nižoj tačnosti i recall vrednostima za MST grupe 6–10. Ručnim pregledom predikcija uočavamo da model ponekad ne

prepoznaje maligne lezije na tamnijim tonovima kože, dok kod svetlih i srednjih tonova preciznost ostaje znatno viša.

*FairSkinToneLoss* je uveden da bi se smanjila pristrasnost između grupa, vrednosti standardne devijacije i *demographic parity difference* pokazuju da blage razlike i dalje postoje, naročito kod najtamnijih tonova kože. Ove razlike ukazuju na to da model nije u potpunosti uspeo da generalizuje i da je potrebna dodatna optimizacija i veća zastupljenost svih *MST* grupa u trening skupu.

Uprkos ovim ograničenjima, model pokazuje solidne performanse za svetlije i srednje tonove kože, što ga čini pouzdanim u kontekstu ranog prepoznavanja melanoma u tim grupama. Međutim, za kliničku primenu kod tamnijih tonova kože potrebne su dodatne strategije, kao što su prošireni skupovi podataka, naprednije tehnike augmentacije i dodatna fina podešavanja modela.

## Literatura


[1] Laura N Montoya, Jennafer Shae Roberts, and Bel´en S´anchez Hidalgo: *Towards Fairness in AI for Melanoma Detection: Systematic Review and Recommendations*

[2] Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J., & Soyer, H. P. (2021). *A patient-centric dataset of images and metadata for identifying melanomas using clinical context*. Scientific Data, 8, 34.

[3] N. R. Abbasi, H. M. Shaw, D. S. Rigel, R. J. Friedman, W. H. McCarthy, I. Osman, A. W. Kopf, and D. Polsky, "Early diagnosis of cutaneous melanoma: Revisiting the ABCD criteria," vol. 292, no. 22, pp. 2771– 2776.

[4] NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI: *A Survey on Bias and Fairness in Machine Learning*

[5] *Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set*

[6] Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, Alexander Schlaefer *Skin Lesion Classification Using Ensembles of Multi-Resolution EfficientNets with Meta Data*

[7] Suorong Yang, Weikang Xiao, Mengchen Zhang, Suhan Guo, Jian Zhao, Furao Shen, 2022. *Image Data Augmentation for Deep Learning: A Survey*

[8] Suorong Yang, Weikang Xiao, Mengchen Zhang, Suhan Guo, Jian Zhao, Furao Shen *Automatic skin lesion segmentation on dermoscopic images by the means of superpixel merging*

[9] Tan, M. & Le, Q. V., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[10] Deng, J. et al., "ImageNet: A Large-Scale Hierarchical Image Database," *CVPR*, 2009.

[11] Chollet, F., *Deep Learning with Python*, 2nd Edition, Manning Publications, 2021.

[12] Ramachandran, P., Zoph, B., Le, Q.V., "Searching for Activation Functions," *arXiv preprint arXiv:1710.05941*, 2017.

[13] Ioffe, S. & Szegedy, C., "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *ICML*, 2015.

[14] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. Journal of Machine Learning Research, 15, 1929-1958.

[15] Gal, Y., & Ghahramani, Z. (2016). *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. In ICML.