Prepoznavanje emocija na osnovu facijalnih ekspresija primenom konvolucionih neuronskih mreža

Anja Dučić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Novi Sad, Srbija
ducic.e215.2024@uns.ac.rs

Anja Lovrić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Novi Sad, Srbija
lovric.e216.2024@uns.ac.rs

***Abstrakt*—Prepoznavanje emocija na osnovu facijalnih ekspresija predstavlja važan problem u oblasti računarske vizije i ima primenu u različitim domenima. U ovom radu upoređene su karakteristike četiri poznate *CNN* arhitekture - *VGG19, InceptionV3, ResNet50, MobileNetV2* i jedne jednostavne konvolucione mreže prilikom rešavanja problema prepoznavanja facijalnih ekspresija. Korišćeni su *transfer learning* i *fine-tuning* pristupi. Naša analiza je pokazala da se na relativno malom i nebalansiranom skupu podataka najbolji rezultati postižu osnovnim i jednostavnim *CNN* mrežama dizajniranim specifično za ovaj zadatak, dok se prilikom upotrebe unapred treniranih modela sa mnogo većim brojem parametara može susresti sa velikim izazovima. Ovaj rad pruža uvid u prednosti i mane upotrebe svakog od modela i daje smernice za izbor modela i njihovih hiperparametara u zavisnosti od slučaja korišćenja.**

Ključne reči—FER, konvolucione neuronske mreže, prepoznavanje emocija, VGG19, InceptionV3, ResNet50, MobileNetV2

I. UVOD

Prepoznavanje emocija na osnovu facijalnih ekspresija predstavlja složen proces kojim se tumače izrazi lica kako bi se omogućio uvid u unutrašnja emocionalna stanja, pri čemu se način izražavanja emocija može značajno razlikovati među pojedincima [1].

Ljudi svoje emocije prenose na različite načine, uključujući facijalne ekspresije, način hodanja, držanje tela, kontakt očima, pa čak i suptilne promene u disanju. Među ovim oblicima neverbalne komunikacije, izrazi lica zauzimaju posebno mesto, jer pružaju najviše informacija o emocijama i namerama osobe čak i kada reči nisu izgovorene [2].

Prepoznavanje emocija je značajno jer omogućava efikasnu socijalnu interakciju i razumevanje tuđeg ponašanja. Posebno je zanimljivo u kontekstu facijalnih ekspresija jer se one mogu analizirati automatski pomoću računarskih sistema i veštačke inteligencije, što otvara mogućnosti za primenu u obrazovanju, zdravstvu, bezbednosti, interakciji čovek-računar i istraživanju ljudskog ponašanja, a takođe omogućava praćenje i procenu reakcija osoba u realnom vremenu [3]. Ukoliko bi računari i elektronski uređaji mogli bolje da razumeju i interpretiraju facijalne ekspresije korisnika, interakcija bi bila lakša i više *user-friendly* [4].

Ovaj zadatak je izazovan zbog velike varijabilnosti u izražavanju pojedinih emocija među ljudima, suptilnih razlika u mimici i prisustva faktora poput osvetljenja, kontrasta i delimičnog zaklanjanja lica [1][3]. Korišćenje *CNN* arhitektura za prepoznavanje emocija na osnovu facijalnih ekspresija zahteva pažljivo podešavanje odabrane arhitekture i hiperparametara, jer duboke mreže mogu biti sklone *overfitting*-u, osetljive na neuravnoteženost podataka i zahtevaju značajne resurse za treniranje i optimizaciju.

U literaturi već postoji značajan broj radova koji koriste konvolucione neuronske mreže za ovaj zadatak. Ipak, ovi pristupi često imaju ograničenja koja se odnose na kvalitet i balansiranost podataka, generalizaciju i prilagodljivost u realnim uslovima. Takođe, uočeno je da modeli često postižu slabije performanse pri prepoznavanju određenih emocija (npr. strah ili iznenađenje) u poređenju sa izraženijim emocijama kao što su sreća i tuga. Pored toga, duboke arhitekture zahtevaju veliku računarsku snagu i memoriju, što može ograničiti brzinu treniranja i eksperimentisanja. Neki modeli su skloni *overfittng*-u na malim ili neuravnoteženim skupovima podataka, dok njihova primena na većim skupovima podataka pokazuje tendenciju ka opadanju performansi [5-7].

U okviru našeg rada koristili smo više arhitektura konvolucionih neuronskih mreža uz *transfer learning* i *fine-tuning* pristupe: *baseline CNN* čija je arhitektura dizajnirana od nule, *VGG19*, *InceptionV3*, *MobileNetV2* i *ResNet50*. Sprovedeno je upoređivanje performansi različitih mreža na istom skupu podakata kako bi se identifikovale njihove prednosti i nedostaci. Poređenje koje je izvršeno je korisno jer omogućava dublje razumevanje odnosa između složenosti modela, kapaciteta za generalizaciju i osetljivosti na karakteristike podataka, što olakšava izbor optimalnog pristupa za realne primene. Naš pristup zasnovan na poznatim *CNN* arhitekturama sa *fine-tuning*-om omogućava modelu da iskoristi prethodno naučene vizuelne karakteristike i prilagodi ih specifičnostima facijalnih ekspresija u korišćenom skupu podataka. Rezultati ovog pristupa pokazuju da složenije mreže mogu biti osetljive na neuravnoteženost podataka, imati slabije performanse pri identifikaciji suptilnijih emocija i potrebu za većim računarskim resursima.

II. POVEZANI RADOVI

U radu [5] predstavljen je unapređeni *VGG19* model za prepoznavanje emocija na osnovu facijalnih ekspresija na slikama. Klasični *VGG19* model poboljšan je dodavanjem *BatchNorm* i *ReLU* slojeva u svaki blok, *Dropout* sloja između poslednjeg konvolucionog i potpuno povezanog sloja, i smanjenjem izlaznog dela na dva potpuno povezana

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

sloja sa *softmax* klasifikatorom. Testiranja na *FER2013* i *CK+* skupovima pokazala su da ovakva arhitektura postiže veću tačnost u poređenju sa osnovnim *VGG19*, ali i sa rešenjima koja su bila najuspešnija na *Kaggle* takmičenju. Unapređeni *VGG19* model je na *FER2013* skupu postigao tačnost od 73.03%, dok je na *CK+* skupu, koji je manji i kontrolisan, dostigao 93.94%. Prednost ovog pristupa je u tome što su relativno jednostavne izmene doprinele značajnom poboljšanju performansi i smanjenju rizika od *overfitting*-a, dok se kao glavni nedostatak navodi da model i dalje ima problema u razlikovanju emocija koje su međusobno slične, poput tuge, gađenja i straha.

U radu [7] predstavljen je model za analizu sentimenta na slikama zasnovan na *InceptionV3* arhitekturi dubokih konvolucionih mreža. Model koristi dodatne *Dense* slojeve, *Dropout* sloj i *softmax* klasifikator za tri klase (pozitivno, negativno, neutralno) kako bi se poboljšala tačnost i smanjio rizik od *overfitting*-a. Testiranja na *CK+*, *FER2013* i *JAFFE* skupovima pokazala su da model postiže značajno bolje rezultate u poređenju sa drugim *transfer learning* pristupima i tradicionalnim mašinskim učenjem, sa tačnošću od 99.57% na *CK+*, 73.09% na *FER2013* i 86% na *JAFFE* skupu. Prednost ovog pristupa je u izuzetno visokoj preciznosti za manje kontrolisane skupove i boljoj robusnosti modela, dok je kao ograničenje istaknuta niža tačnost na većim i raznovrsnijim skupovima podataka.

U radu [3] je istraženo prepoznavanje emocija spram facijalnih ekspresija kada je lice delimično prekriveno maskom, koristeći *transfer learning* pretreniranih dubokih modela (*EfficientNetB0, ResNet50, InceptionV3, Xception, AlexNet*) i klasičnih metoda (*SVM, ANN*). Korišćen je *FER-2013 dataset* sa sedam klasa emocija, uz virtuelnu primenu maski na svako lice. Rezultati pokazuju da *InceptionV3* postiže najbolje performanse, dok veća složenost (npr. *Xception*) smanjuje tačnost. Prednosti u radu jesu efikasno korišćenje pretreniranih modela i visoka tačnost *InceptionV3*. Kao mane navedene su smanjene performanse kod prekrivenog donjeg dela lica i teškoće u prepoznavanju suptilnih emocija.

Rad [4] poredi modele *VGG16*, *DenseNet*, *ResNet50* i *GoogLeNet* za zadatke prepoznavanja lica na *FER2013* skupu podataka. *DenseNet* i *ResNet50* pokazuju najbržu početnu konvergenciju i ranu stabilizaciju, što omogućava postizanje dobrih rezultata uz manje epoha. Model *ResNet50* postigao je najveću tačnost (69.46%). *DenseNet* i *VGG16* ostvaruju visoku tačnost za prepoznavanje emocija sreća (oko 88%) i iznenađenje (81%), dok *GoogLeNet* i *DenseNet* imaju kompaktnije modele, što poboljšava efikasnost i zahtev za memorijom. S druge strane, *GoogLeNet* pokazuje najnižu tačnost, a *VGG16* ima najduže vreme treniranja i najveću veličinu modela (512 MB), što otežava primenu u praksi. Prepoznavanje emocija gađenje i strah je problematično kod svih modela. Na kraju treniranja svi modeli ostvaruju sličnu tačnost koja se kreće u rasponu od 65% do 70%, što ukazuje na mogućnost dodatnog unapređenja i bolje generalizacije na različitim skupovima podataka.

Rad [6] se bavi prepoznavanjem facijalnih emocija sa video snimaka u realnom vremenu. Za klasifikaciju je korišćen *Deep Convolutional Neural Network (DCNN)* treniran od nule, dok su pored njega testirani i pretrenirani modeli *EfficientNet*, *ResNet* i *VGGNet*. *DCNN* je postigao 82.56% tačnosti na trening setu i 65.68% na validacionom setu, nadmašujući pretrenirani pristup. Model je kombinovan sa *Haar* klasifikatorom lica i uspešno prepoznaje emocije poput sreće, tuge, straha, ljutnje, gađenja, iznenađenja i neutralnog stanja u realnom vremenu. Prednost *DCNN* modela je što uči direktno iz *dataset*-a i omogućava efikasnu obradu i instant predikciju emocija, što ga čini pogodnim za praktične primene u interakciji čovek-računar, marketingu i dijagnostici mentalnog zdravlja.

U radu [8] je izvršeno upoređivanje rezultata prepoznavanja facijalnih ekspresija dobijenih upotrebom tri verzije *ResNet* konvolucione mreže – *ResNet18*, *ResNet34* i *ResNet50*. Predloženo rešenje koristi poznate i pretrenirane *ResNet* arhitekture i tehniku *transfer learning*. Poslednji sloj je prilagođen klasifikaciji na *FER 2013* skupu podataka, odnosno klasifikaciji slika u 7 kategorija. Dodatno je izvršen i *fine-tuning* odmrzavanjem pretposlednjeg sloja *layer4* što je doprinelo drastičnom porastu tačnosti modela. Najsloženija mreža *ResNet50* je postigla i najbolje rezultate što se tiče metrika preciznosti, odziva i *F1* mere, ali je zahtevala i najduže vreme treniranja. Jedan od zaključaka koji iznose autori rada je da su jednostavnije *ResNet* arhitekture pokazale veću tačnost i efikasnost na relativno malom skupu podataka. Takođe, autori ukazuju na velike razlike u preciznosti i odzivu za različite klase, što se može donekle poboljšati ciljanom augmentacijom i daljim usavršavanjem modela.

Rad [9] predlaže *EmoNet* arhitekturu koja koristi pretreniranu konvolucionu mrežu *MobileNet* za prepoznavanje facijalnih ekspresija. *MobileNet* je obučena na velikom skupu podataka za različite zadatke, kao što je klasifikacija slika i predstavlja laganu arhutekturu koja se može efikasno koristiti na mobilnim telefonima. Sadrži konvolucione slojeve za ekstrakciju karakteristika ulaznih slika, ali primenjuje *depthwise separable* konvoluciju kako bi se smanjila računska složenost. Akcenat *EmoNet* arhitekture je na tome da se obezbedi tačno prepoznavanje facijalnih ekspresija na hardverski ograničenim uređajima. Primenom *transfer learning* tehnike nad *MobileNet* modelom u sklopu predloženog rešenja postignuta je tačnost od 99.17% na *FER 2013* skupu podataka, a 97.62% na *CK+* skupu podataka, što je bolji rezultat u odnosu na prethodna rešenja sa kojim se poredilo.

Upotrebu *CNN* mreža za rešavanje problema prepoznavanja emocija spram facijalnih ekspresija istražuje i rad [10]. *MobileNet* je upotrebljen za detekciju lica na slikama, dok je *ResNet* korišćen za prepoznavanje emocije. Rad koristi četiri skupa podataka: *Tufts*, *RWTH*, *RAF* i *FER2013*. Na slikama iz *FER2013* skupa uočena je sličnost između pojedinih klasa i minimalne varijacije u izrazima lica. Kao rešenje ovog problema predložena je strategija *Divide-and-Conquer* koja podrazumeva grupisanje emocija u podgrupe koje čine slične klase. Nakon treniranja *CNN* mreže na podgrupama, model predviđa i konačnu emociju za svaku sliku unutar podgrupe. Pokazalo se da ova strategija donosi povećanje tačnosti i bolje generalizovanje. *CNN* je optimizovan upotrebom *learning rate*–a od 0.001, *Adam* optimizatora, prolaskom kroz 10000 epoha i *L1* regulacijom. Postignuta tačnost na *FER2013* skupu podataka je 77.83%.

Rad [11] predlaže metodu za sumarizaciju filmova koja se oslanja na prepoznavanje emocija spram facijalnih ekspresija. Metod podeli film na manje snimke, označi koji snimci su značajniji i na njima primenjuje tehnike za prepoznavanje emocija. Poređene mreže su *MobileNet*,

SqueezNet, *AlexNet*, *GoogleNet* i *ResNet50*. Primenom *transfer learning*-a *ResNet50* mreža sa ulaznim slikama veličine 128x128 piksela je pokazala ubedljivo najbolju tačnost od 93.65%, dok je najskromniji rezultat imala *MobileNet* sa tačnošću od 40.65%. *ResNet50* model sa veličinom slike od 224x224 piksela, na kojoj je i treniran na *ImageNet* skupu, postigao je manju tačnost od 64.83% zbog toga što su težine za slike od 128x128 učene od nule i bolje prilagođene problemu.

III. METODOLOGIJA

U ovoj sekciji opisane su korišćene arhitekture modela i eksperimentalna postavka. Prvo su detaljno predstavljene arhitekture – *Baseline CNN*, *VGG19* [12], *InceptionV3* [13], *RetNet50* [14], *MobileNetV2* [15]– koje su primenjene u zadatku prepoznavanja emocija sa lica. Zatim je data eksperimentalna postavka koja obuhvata opis skupa podataka, korake predobrade podataka, kao i korišćene hiperparametre i proceduru treniranja.

A. Modeli

U ovom radu, koristili smo modele, izuzev *baseline CNN*, sa unapred treniranim težinama na *ImageNet* skupu podataka. Ovo omogućava da modeli iskoriste već naučene osnovne karakteristike iz slika, što ubrzava trening i poboljšava rezultate na našem skupu podataka. U Tabeli I je predstavljen ukupan broj parametara za svaki model, kao i broj parametara koji je treniran.

TABELA I. BROJ PARAMETARA MODELA

Model	Ukupan broj parametara	Broj treniranih parametara
Baseline CNN (*RGB* ulaz)	2,053,335	684,359
Baseline CNN (*grayscale* ulaz)	2,051,607	683,783
VGG19	48,609,630	9,506,055
InceptionV3	56,181,822	11,371,783
ResNet50	53,537,687	14,967,815
MobileNetV2	4,510,679	1,043,911

1) Baseline CNN: Za potrebe poređenja performansi složenijih arhitektura, implementirali smo *baseline CNN* model. Ovaj model je jednostavna duboka konvoluciona mreža koja služi kao referentna tačka za ocenu efikasnosti *fine-tuning*-a složenijih *CNN* modela. *CNN* je izabran zbog svoje sposobnosti da prepozna suptilne obrasce i automatski ekstraktuje složene karakteristike direktno iz vizuelnih podataka, što ga čini pogodnim za prepoznavanje facijalnih izraza lica. Implementirali smo dve varijante ovog modela: *RGB* verziju i *grayscale* verziju, kako bismo ispitali uticaj boje slike na performanse modela. Ulazne slike za ove modele imaju dimenzije 48x48 piksela. Oba modela imaju sličnu arhitekturu: tri konvoluciona sloja sa *ReLU* aktivacijom, svaki praćen *MaxPooling* slojem. Nakon konvolucionih slojeva, izlaz se pretvara u jednodimenzionalni vektor pomoću *Flatten* sloja, što omogućava povezivanje sa potpuno povezanim slojevima. Sledeći *Dense* sloj sa 128 neurona i *ReLU* aktivacijom obrađuje izdvojene karakteristike, dok *Dropout* slojevi smanjuju rizik od *overfitting*-a, a *BatchNormalization* sloj ubrzava i stabilizuje treniranje. Na kraju, *Dense* sloj sa *softmax* aktivacijom klasifikuje slike u neku od sedam vrsta emocija: ljutnja, gađenje, strah, sreća, tuga, iznenađenje i neutralno.

2) VGG19: VGG19 je duboka konvoluciona neuronska mreža koja sadrži 16 konvolucionih slojeva sa *ReLU* aktivacijama, *MaxPooling* slojevima i 3 potpuno povezana sloja. Arhitektura prati jednostavan i ponavljajući obrazac, što je čini lakšom za razumevanje i implementaciju. Zbog dubine strukture i velikog broja parametara, *VGG19* može efikasno izdvajati i kombinovati jednostavne vizuelne obrasce u složenije karakteristike što poboljšava prepoznavanje detalja na slikama.

Ulazne slike za ovaj model *resize*-ovane su na dimenzije 96x96 piksela, budući da originalne dimenzije slika (48x48 piksela) izazivaju pogoršanje performansi modela. S druge strane, slike prevelikih dimenzija gube na kvalitetu, a model ih obrađuje presporo i zahtevaju mnogo više računarske snage i memorije, čineći trening neefikasnim.

U ovom radu, uklonjena su tri originalna potpuno povezana sloja sa vrha mreže koja su inače korišćena za klasifikaciju na 1000 *ImageNet* klasa. Umesto njih, izlaz iz *VGG19* osnove prolazi kroz *GlobalAveragePooling2D* sloj, koji pretvara prostornu mapu karakteristika u vektor fiksne dužine, čime se zadržavaju ključne karakteristike, a smanjuje se broj parametara. Nakon toga, *Dense* sloj sa 128 neurona i *ReLU* aktivacijom uči reprezentacije specifične za naš skup podataka. Primenjene su preporuke rada [5]: dodavanje *BatchNormalization* sloja nakon *Dense* sloja omogućava stabilizaciju i ubrzavanje treninga tako što normalizuje aktivirane vrednosti, dok *Dropout* sloj služi za regularizaciju i smanjenje rizika od *overfitting*-a. Na kraju, *Dense* sloj sa *softmax* aktivacijom klasifikuje slike u jednu od sedam prethodno pomenutih emocija.

Fine-tuning je sproveden otključavanjem poslednjih slojeva *VGG19*, počevši od *block5_conv1* jer ovi slojevi obrađuju specifične, karakteristike višeg nivoa, koje je potrebno prilagoditi našem *dataset*-u.

3) InceptionV3: Mreža koristi *stem* blokove, *inception* module i *reduction* blokove da pripreme ulaz, izdvoje karakteristike i smanje dimenzije i broj parametara [16]. Ovaj model je izabran zbog svojih inovativnih *inception* modula, koji omogućavaju efikasno izdvajanje karakteristika slike kroz paralelne konvolucione slojeve različitih veličina kernela. Ovo smanjuje računarske zahteve i čini je efikasnom i fleksibilnom arhitekturom za prepoznavanje emocija.

Ulazne slike za ovaj model *resize*-ovane su na dimenzije 130x130 piksela, jer *resizing* slika na još veće dimenzije smanjuje njihov kvalitet i time negativno utiče na performanse modela.

U ovom radu, uklonjen je originalni klasifikacioni sloj sa vrha mreže i modifikovan prema potrebama našeg rada, ali utemeljen na preporukama rada [3]. Umesto njega, izlaz iz *InceptionV3* osnove prolazi kroz *GlobalAveragePooling2D* sloj, koji kompresuje prostornu mapu karakteristika u vektor fiksne veličine i smanjuje broj parametara. Nakon toga, *Dense* sloj sa 128 neurona i *ReLU* aktivacijom uči reprezentacije specifične za *dataset*, dok *BatchNormalization* i *Dropout* slojevi stabilizuju treniranje i smanjuju rizik od *overfitting*-a. Na kraju, *Dense* sloj sa *softmax* aktivacijom klasifikuje slike u jednu od sedam prethodno pomenutih emocija.

Fine-tuning je sproveden otključavanjem poslednjih slojeva *InceptionV3*, počevši od sloja *conv2d_268*, jer ovi slojevi obrađuju karakteristike višeg nivoa koje je potrebno prilagoditi našem *dataset*-u.

4) ResNet50: ResNet predstavlja duboku konvolucionu mrežu koja uvodi *residual* blokove čime se postiže efikasnije treniranje dubokih neuronskih mreža i rešava se jedan od problema dubokih mreža, problem nestajućeg gradijenta. U ovom radu, izabrana je arhitektura sa 50 slojeva kako bi se mogla koristiti gotova implementacija modela u *Keras* biblioteci [17], ali i jer se prema eksperimentima iz rada [8] mogu očekivati bolji odziv i preciznost u odnosu na *ResNet* mreže sa manje slojeva.

Ulaznim slikama je prilagođena veličina povećavanjem na 128x128 piksela. Model je naučen na slikama veličine 224x224 piksela i filteri su optimizovani za slike većih dimenzija od naših originalnih slika (48x48) i zato gubi sposobnost prepoznavanja na toliko malim ulazima. Izabrana vrednost predstavlja kompromis između računarske efikasnosti i sposobnosti modela da uči relevantne obrasce u podacima. Eksperimenti na većim slikama su značajno povećavali brzinu obučavanja i potrošnju memorije.

Korišćen je model originalno ima 1000 izlaza. U našem slučaju, poslednji sloj sa *softmax* aktivacionom funkcijom određuje pripadnost jednoj od sedam klasa. Dodatno je izvršen i *fine-tuning* treniranjem sloja *layer4* kako je predloženo u radu [8]. Ovaj blok čine četiri konvoluciona sloja koja se nalaze pre *average pooling* i poslednjeg izlaznog sloja mreže i utvrđeno je da se tačnost može drastično povećati omogućavanjem da *layer4* uči na specifičnom skupu podataka.

5) MobileNetV2: Ovaj model predstavlja unapređenu arhitekturu tradicionalnih residualnih mreža koja omogućuje efikasno korišćenje na uređajima sa ograničenim resursima poput mobilnog telefona. Jedna od najvećih razlika u odnosu na standardne residualne mreže jeste to što se na ulazu i izlazu koriste tanji slojevi, dok se mreža širi na srednjem sloju. Smanjenje gubitaka informacija se postiže tako što se u srednjem sloju koristi linearna konvolucija bez aktivacije. Složenost mreže i broj parametara se smanjuje upotrebom *depthwise separable* konvolucije [15].

Ulaz u model predstavljaju slike dimenzija 96x96 piksela. Ova rezolucija je izabrana eksperimentalnim putem. Model je prilagođen radu sa slikama manjih dimenzija, a povećavanje veličine u odnosu na originalnu (48x48 piksela) je doprinelo poboljšavanju performansi.

Za potrebe skupa podataka u ovom radu, izvršena je zamena poslednjeg sloja kako bi se vršila klasifikacija na sedam emocija. Dodata je i regulacija kako bi se smanjila mogućnost *overfitting*-a, u vidu dva *Dropout* sloja, između kojih je *Dense* sloj sa 128 neurona i *ReLU* aktivacionom funkcijom. Za poboljšavanje rezultata, umesto korišćenja težina naučenih na *ImageNet* skupu podataka za sve slojeve mreže, dodatno je izvršen je i *fine-tuning* na slojevima koji čine *block*_16.

B. *Eksperimentalna postavka*

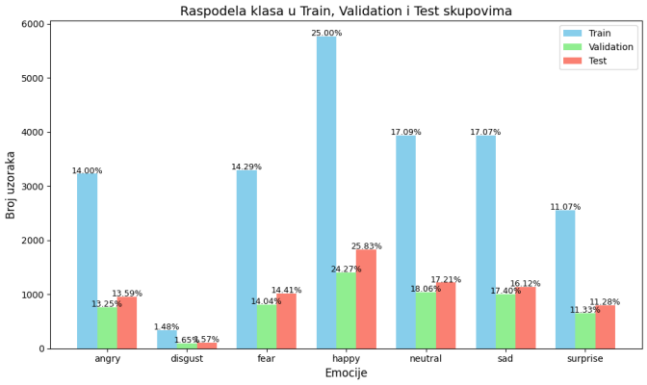
1) Skup podataka: Za potrebe ovog rada korišćen je skup podataka "*Facial Expression Recognition (FER) Dataset*" [18], dostupan na *Kaggle* platformi. Ovaj skup podataka je često korišćen u istraživanju prepoznavanja emocija lica, pružajući veliki broj slika za treniranje modela. Skup sadrži ukupno 35.887 slika lica, sa rezolucijom 48x48 piksela u sivim tonovima, podeljenih u sedam vrsta emocija: sreća, tuga, iznenađenje, strah, gađenje, ljutnja i neutralno. Primer slika i labela je dat na Slici 1.

Slika 1. Primer slika iz skupa podataka.

Organizacija skupa je zasnovana na podfolderima, gde svaki podfolder nosi naziv odgovarajuće emocije i sadrži slike koje joj pripadaju. Originalna podela skupa na *training* (28.821 slika) i *validation* (7.066 slika) set je zadržana, pri čemu je *validation* set korišćen kao *test* set za potrebe evaluacije modela.

Korišćenje javno dostupnog i široko prihvaćenog skupa podataka, kao što je *FER dataset*, omogućava direktno poređenje performansi razvijenog modela sa rezultatima drugih istraživača, što doprinosi objektivnosti i reproduktivnosti rada.

Distribucija klasa u skupu je neravnomerna, sa značajno manjim brojem uzoraka u klasi gađenje. Broj slika u svakoj klasi, kao i procenat u odnosu na ukupan broj slika je prikazan na Slici 2. Ovaj disbalans predstavlja specifičan izazov koji je uzet u obzir prilikom obučavanja modela. Kao što je izloženo u [4] i uočeno tokom izrade našeg eksperimenta, ovaj skup podataka ima određen broj pogrešno labeliranih slika što dodatno utiče na slabiji kvalitet rezultata.



Slika 2. Raspodela klasa u *Train*, *Validation* i *Test* skupovima.

2) Obrada podataka: Originalni skup podataka je učitan i pripremljen za treniranje, validaciju i testiranje. Za potrebe validacije, 20% trening skupa je izdvojeno kao validacioni skup. Ova podela omogućava praćenje performansi modela tokom treniranja i podešavanje hiperparametara. Kako bi se delimično uravnotežio broj primera po klasama, izvršen je *oversampling* za klasu gađenje u trening skupu, čime je ukupan broj trening primera povećan na 32.385. Ulazni podaci su normalizovani korišćenjem *min-max* skaliranja. Normalizacija je primenjena na sve *dataset*-ove (trening,

validacija i test). Normalizacija ulaza u *ResNet50* je izvršena u okviru funkcije *preprocess_input*, iz *Keras* biblioteke, specijalizovane za pretprocesiranje ulaza u ovu mrežu. U slučaju *MobileNetV2* mreže za normalizaciju je takođe korišćena funkcija *preprocess_input* iz *Keras* biblioteke kako bi pikseli bili u opsegu vrednosti od -1 do 1 što predstavlja optimalan ulaz za ovu mrežu. Kako bi modeli koji zahtevaju trokanalne ulaze (*RGB*) mogli da obrade slike, slike se po potrebi mogu konvertovati u *RGB* format. Za potrebe pojedinih modela, slike su dodatno *resize*-ovane na određene dimenzije, korišćenjem *bilinear* ili *bicubic* interpolacije. Takođe je testirana i metoda *nearest neighbor*, ali je previše isticala pojedinačne piksele, što je dovodilo do neprirodnog prikaza i lošijih performansi modela. Na kraju, za povećanje robusnosti modela i smanjenje *overfitting*-a, na trening skup je primenjena augmentacija, koja uključuje horizontalni *flip*, rotacije, translacije, *zoom*, kontrast i osvetljenje, čime model uči invariantne karakteristike lica u različitim uslovima.

3) Treniranje: Trening svih modela je izveden sa hiperparametrima koji su eksperimentalno podešeni. Tokom treniranja korišćene su *callback* funkcije: *EarlyStopping*, *ReduceLROnPlateau* i *ModelCheckpoint*, koje su kontrolisale proces treniranja, sprečavale *overfitting* i osiguravale da evaluacija koristi najbolje trenirani model. *EarlyStopping* je zaustavljao trening ukoliko *val_loss* nije beležio poboljšanje tokom definisanog broja epoha, *ReduceLROnPlateau* je smanjivao *learning rat*e kada *val_loss* stagnira, dok je *ModelCheckpoint* čuvao najbolju verziju modela tokom treniranja, takođe prema *val_loss* metrici.

a) Baseline CNN: Korišćen je *batch size* od 128, početni *learning rate* 5e-4 i ukupno 100 epoha. Pri manjem *batch size*-u (32 i 64), treniranje je bilo nestabilno i oscilacije *val_loss*-a su bile veće, dok je veći *batch size* (256) dodatno smanjivao sposobnost modela da generalizuje. Veći početni *learning rate* izazivao je nagle promene *val_loss*-a i loše predikcije na testnom setu podataka. Za optimizaciju je korišćen *Adam* optimizator koji automatski prilagođava *learning rate* za svaki parametar na osnovu istorije gradijenata i time omogućava brzu i stabilnu konvergenciju. *EarlyStopping* funkcija je omogućila automatski prekid treniranja ukoliko *val_loss* nije beležio poboljšanje tokom 10 epoha, čime se izbegava nepotrebno produžavanje treninga i *overfitting*.

b) VGG19: Za *transfer learning VGG19* modela eksperimentalno su odabrani *batch size* od 128, početni *learning rate* 5e-3 i ukupno 80 epoha, uz *Adam* optimizator. Pri manjem *batch size*-u (64) treniranje je bilo vremenski zahtevno i sa većim oscilacijama *val_loss*-a, dok je veći *batch size* (256) uzrokovao *overfitting* uočen na graficima metrika tokom epoha: *accuracy* i *loss* na trening setu su napredovali, a na validacionom stagnirali, a potom se pogoršali. Početni *learning rate* je eksperimentisan: veći *LR* je izazivao veće oscilacije *training loss*-a što je negativno uticalo na konvergenciju modela, dok manji *LR* bio izrazito vremenski zahtevan. *EarlyStopping* funkcija je omogućila automatski prekid treniranja ukoliko *val_loss* nije beležio poboljšanje tokom 6 epoha.

Nakon toga, sproveden je *fine-tuning* otključavanjem slojeva od *block5_conv1* pa naviše, uz manji *learning rate* (1e-4) i dodatnih 25 epoha. Ovaj pristup omogućava modelu da prilagodi karakteristike višeg nivoa specifične za *dataset*, bez prepisivanja već naučenih osnovnih karakteristika iz *ImageNet* skupa. Tokom *fine-tuninga* najbolji model je sačuvan u 87. epohi, što je značajno poboljšalo performanse u odnosu na inicijalno treniranje zaustavljeno u 20. epohi.

Pokušaji dodatnog *fine-tuninga* delimično i čitavih 4. i 5. blokova, nisu doveli do poboljšanja; *val_loss* je stagnirao, a ni *accuracy*, *AUC* i *recall* nisu pokazivali poboljšanja, dok su *precision* vrednosti bile razbacane. *Macro F1* skor je bio bolji bez dodatnog *fine-tuninga*, što je ukazivalo na rizik od *overfitting*-a kod dodatnog dotreniranja.

Eksperimenti sa dodatnim ponderisanjem klasa (*class weights*) za strah, iznenađenje i ljutnja klase, kako bi se poboljšao *recall*, doveli su do neznatnog poboljšanja za ove klase, a istovremeno su pogoršali rezultate na drugim klasama i treniranje je postalo znatno sporije. Slična situacija je bila i sa višestrukim augmentacijama za ove klase, koje su povećavale trajanje treniranja, ali nisu značajno poboljšale rezultate. Zbog toga su ovi pristupi odbačeni u konačnoj konfiguraciji modela, koja se bazira na obrađenom skupu podataka, inicijalnoj augmentaciji, *transfer learning*-u i *fine-tuning*-u od *block5_conv1* naviše.

c) InceptionV3: Za *transfer learning InceptionV3* modela eksperimentalno su odabrani *batch size* od 128, početni *learning rate* 5e-4 i ukupno 45 epoha, uz *Adam* optimizator. Pri manjem *batch size*-u (32 i 64) treniranje je bilo generalno nestabilno sa većim oscilacijama *val_loss*-a, dok je veći *batch size* (256) uzrokovao *overfitting* uočen na graficima metrika tokom epoha: *accuracy* i *loss* na trening setu su napredovali, a na validacionom stagnirali. Početni *learning rate* je eksperimentisan: veći *LR* po uzoru na radove [3] i [7] je izazivao veće oscilacije *training loss*-a što je negativno uticalo na konvergenciju modela, dok je odabrana vrednost (5e-4) omogućila efikasniju optimizaciju slojeva. *EarlyStopping* funkcija je zaustavljala trening ukoliko *val_loss* nije beležio poboljšanje tokom 6 epoha.

Nakon toga, sproveden je *fine-tuning* otključavanjem slojeva od *conv2d_268* pa naviše, uz manji *learning rate* (5e-5) i dodatne 22 epohe. Tokom *fine-tuninga* najbolji model je sačuvan u poslednjoj 78. epohi, što je značajno poboljšalo performanse u odnosu na inicijalno treniranje završeno u 45. epohi.

Eksperiment sa dodatnim ponderisanjem klasa za emocije: strah, iznenađenje i ljutnja nije značajno poboljšao performanse modela, a imao je negativan uticaj na druge klase. Takođe, prema preporuci iz rada [7], testirana je zamena jednog *Dense* sloja od 128 neurona sa dva takva sloja uz *BatchNormalization* i *Dropout*, ali konačni *macro F1* skor nije poboljšan, a sam proces treniranja je bio značajno sporiji. Zbog toga je konačna konfiguracija modela za *InceptionV3* zasnovana na obrađenom skupu podataka, inicijalnoj augmentaciji, *transfer learning*-u i *fine-tuning*-u od sloja *conv2d_268* naviše, bez dodatnog ponderisanja i povećanja broja *Dense* slojeva.

d) ResNet50: Prilikom *transfer learning*-a ovog modela, odabran je *batch size* 64 i početni *learning rate* 1e-3. Upotreba većeg *batch size*-a je bila previše računarski zahtevna. Za optimizaciju je korišćen *Adam* optimizator. Funkcija *EarlyStopping* je zaustavljala trening ukoliko

val_loss nije beležio poboljšanje tokom 5 epoha. Trening je zaustavljen posle 13 epoha. Pokušaj korišćenja parametra *class_weight* za rešavanje problema nebalansiranosti podataka nije doneo napredak u rezultatu, stoga se ne nalazi u konačnoj verziji modela. Isto važi i za pokušaj ciljane dodatne augmentacije manjih klasa.

Nakon toga, odmrznuta su četiri bloka poznata kao *layer4* i izvršen je *fine-tuning*. U ovom slučaju, *learning rate* je značajno smanjen na vrednost od 1e-5 kako bi model iskoristio težine naučene u prethodnim epohama. Povećavanje ovog parametra dovodilo je gubljenja težina naučenih tokom prošlih epoha i uništavanja prethodnog znanja. *Fine-tuning* je završen posle 7 epoha.

e) MobileNetV2: Prilikom primene tehnike *transfer learning* na ovaj model korišćen je *batch size* 128, *Adam* optimizator i početni *learning rate* 1e-3. Nebalansiranost podataka je delimično regulisana dodavanjem parametra modela *class_weight* koji tera model da ređe klase bolje prepoznaje tako što više kažnjava greške nad tim klasama. Dodatno podešavanje vrednosti težina je dovodilo do toga da model preferira određene klase. Isprobana je dodatna augmentacija ređih klasa, međutim ovaj metod nije dodatno doprineo poboljšanju rezultata. Vrednosti težina za svaku klasu su utvrđene eksperimentalno. *EarlyStopping* je zaustavljao trening ukoliko *val_loss* nije beležio poboljšanje tokom 10 epoha. Treniranje se završilo posle 76 epoha.

Dalje je vršen i dodatan *fine-tuning* svih slojeva koji čine *block_16* u cilju dodatnog popravljanja *macro f1* mere. *Fine-tuning* ovog bloka je završen posle 25 epoha uz smanjen početni *learning rate* na 5e-5 i ponovnu upotrebu *Adam* optimizatora.

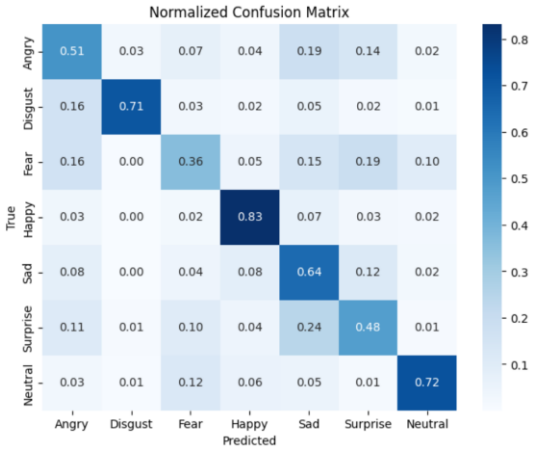
IV. Rezultati i diskusija

U radu je testirano više konvolucionih neuronskih mreža uz *fine-tuning* pristup za prepoznavanje emocija na osnovu facijalnih ekspresija: *VGG19*, *InceptionV3*, *MobileNetV2* i *ResNet50*, kao i *baseline CNN* treniran od nule. Svi modeli evaluirani su na istom testnom skupu podataka, a performanse su merene pomoću metrika: *accuracy*, *precision*, *recall*, *AUC*, *F1*, dok je *loss* korišćen za praćenje napredovanja treniranja i konvergencije modela.

Baseline RGB model postigao je *accuracy* od 61.48%, *precision* od 57.68%, *recall* od 60.61% i *AUC* od 89.32%, dok *macro F1 score* iznosi 58.53%. *F1* skorovi po klasama iznose: ljutnja – 49.79%, gađenje – 56.94%, strah – 40.50%, sreća – 83.68%, tuga – 57.50%, iznenađenje – 48.57% i neutralno – 72.73%. Ove metrike pokazuju da model ostvaruje umerene performanse, sa relativno dobrom sposobnošću razlikovanja klasa (*AUC* 89.32%), ali performanse variraju među pojedinačnim emocijama, što se ogleda u nižem *macro F1* skoru od 58.53% i ipak ukazuje na izazove u prepoznavanju retkih ili suptilnih emocija.

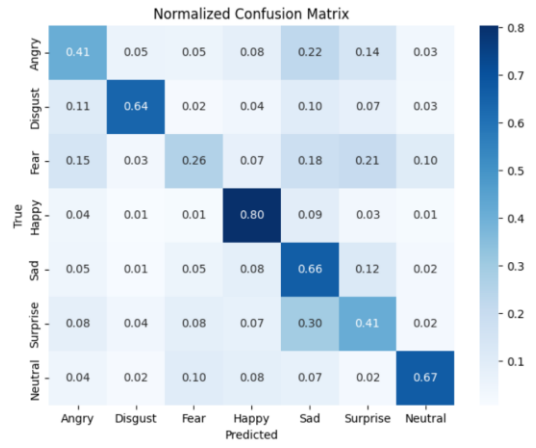
Baseline grayscale model postigao je *accuracy* od 61.65%, *precision* od 59.84%, *recall* od 60.82% i *AUC* od 89.17%, dok *macro F1 score* iznosi 60%. *F1* skorovi po klasama iznose: ljutnja – 50.54%, gađenje – 65.29%, strah – 41.29%, sreća – 83.74%, tuga – 56.61%, iznenađenje – 48.94% i neutralno – 73.56%.

U poređenju sa *RGB* verzijom, *grayscale* model pokazuje blago poboljšanje *macro F1* skora i *precision*-a, što sugeriše da uklanjanje informacija o bojama može doprineti stabilnijem prepoznavanju pojedinih klasa, posebno vizuelno sličnih. I dalje se uočava neravnomerna efikasnost po klasama, ali generalno, *grayscale* ulazi omogućavaju marginalno bolje balansiranje performansi među emocijama i bolju sposobnost modela da razlikuje klase. Oba modela povremeno greše u klasifikaciji određenih emocija zbog sličnih opštih izraza lica: strah se ponekad predviđa kao ljutnja ili iznenađenje, dok se iznenađenje ponekad klasifikuje kao tuga. Matrica konfuzije za *grayscale* model je data na Slici 3.



Slika 3. Matrica konfuzije za *baseline CNN – grayscale*.

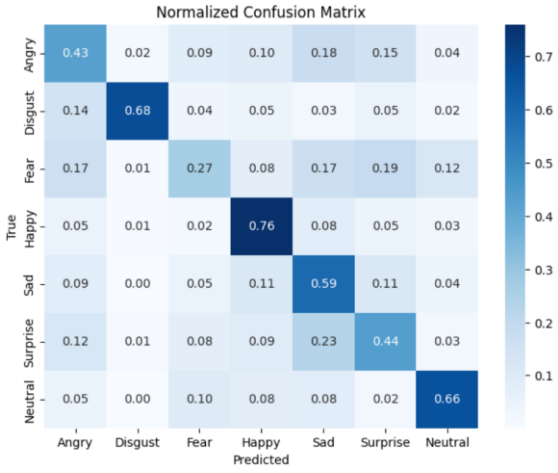
VGG19 model postigao je *accuracy* od 56.74%, *precision* od 52.10%, *recall* od 55.15% i *AUC* od 86.75%, dok *macro F1 score* iznosi 52.05%. *F1* skorovi po klasama iznose: ljutnja – 44.73%, gađenje – 41.40%, strah – 32.96%, sreća – 79.46%, tuga – 53.85%, iznenađenje – 43.24% i neutralno – 68.69%. *Recall* je najniži za klasu strah, slično kao i kod *baseline* modela, što ukazuje na teškoće u pouzdanom prepoznavanju ove emocije. *VGG19* model često pogrešno klasifikuje ljutnju, strah i iznenađenje zbog sličnih vizuelnih karakteristika – ljutnju ponekad predviđa kao tugu, strah kao iznenađenje, a iznenađenje kao tugu. Postoje i druge, manje izražene greške u klasifikaciji, ali ova tri su najizraženija. Celokupna matrica konfuzije je prikazana na Slici 4.



Slika 4. Matrica konfuzije za *VGG19*.

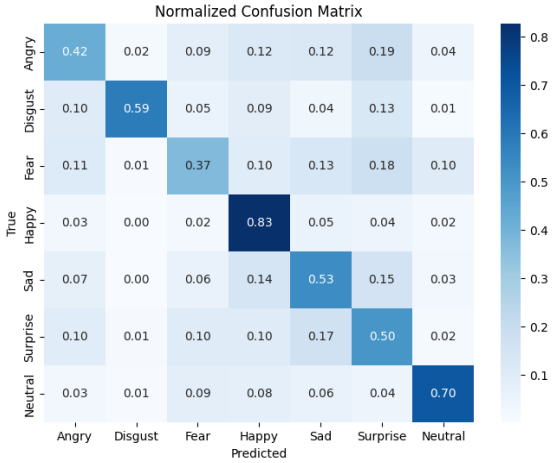
U poređenju sa *baseline RGB* i *grayscale* modelima, *VGG19* ostvaruje niži *macro F1 score* i opštu tačnost, što ukazuje na slabije performanse na ovom *dataset*-u uprkos složenijoj arhitekturi. Posebno se vidi pad u prepoznavanju emocija poput straha i gađenja. Iako model pokazuje solidnu sposobnost prepoznavanja jasno izraženih emocija poput sreće i neutralnosti, rezultat govori da veće i dublje mreže nisu automatski pogodne za male i neuravnotežene skupove podataka, dok *baseline* modeli, posebno *grayscale*, pokazuju stabilnije *macro* metrike i ipak bolju ravnotežu među klasama.

InceptionV3 model postigao je *accuracy* od 55.04%, *precision* od 53.17%, *recall* od 54.58% i *AUC* 85.71%, dok *macro F1 score* iznosi 53.40%. *F1* skorovi za svaku od klasa su sljedeći: ljutnja – 42.32%, gađenje – 61.48%, strah – 32.99%, sreća – 74.97%, tuga – 52.19%, iznenađenje – 44.69% i neutralno – 65.14%. U poređenju sa *baseline RGB* i *grayscale* modelima, *InceptionV3* ostvaruje nešto slabije performanse u svim metrikama, posebno u *macro F1* skoru, što ukazuje na manje ujednačeno prepoznavanje emocija. Kao i kod prethodnih modela, *recall* je najmanji za emociju strah, dok model povremeno meša ljutnju, strah i iznenađenje, sa blažim greškama između ostalih klasa što se vidi u matrici konfuzije datoj na Slici 5. Ove metrike sugerišu da složenija arhitektura *InceptionV3*, u ovom *dataset*-u, nije donela značajnu prednost u odnosu na jednostavnije *baseline* mreže.



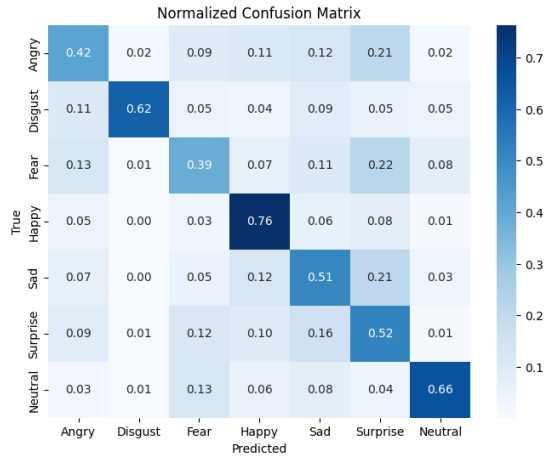
Slika 5. Matrica konfuzije za *InceptionV3*

ResNet50 model je postigao *accuracy* od 58.49%, *precision* od 58.63%, *recall* 56.25%, *AUC* 86.78% i *macro F1 score* od 56.29%. *F1 score* za svaku od klasa je sledeći: ljutnja 45.76%, gađenje 58.56%, strah 42.22%, sreća 77.14%, tuga 52.70%, iznenađenje 47.89% i neutralno 69.80%. Matrica konfuzije na Slici 6 pokazuje da se i kod *ResNet50* modela javlja problem pri prepoznavanju najčešće straha, a zatim i ljutnje. Najčešće greške su razlikovanje ljutnje i iznenađenja, kao i straha i iznenađenja, što predstavlja izazove i za *baseline* model. *Macro F1* nije dostigao vrednost *baseline* modela, kao ni druge posmatrane metrike. Ipak, rezultati *ResNet50* su blago bolji u odnosu na druge posmatrane modele: *VGG19*, *InceptionV3* i *MobileNetV2*.



Slika 6. Matrica konfuzije za *ResNet50*.

MobileNetV2 je postigao *accuracy* od 56.65%, *precision* 56.63%, *recall* od 55.54% i *AUC* 86.05%. *Macro F1* je dostigao vrednost od 55.87%. Za svaku od klasa, postignuti su sledeći *F1* skorovi: ljutnja 44.89%, gađenje 61.88%, strah 42.24%, sreća 75.16%, tuga 51.11%, iznenađenje 45.67 % i neutralno 70.11%. Postignuti rezultati su skromniji u odnosu na *baseline CNN* što pokazuje da složenija arhitektura nije donela poboljšanje u odnosu na *baseline CNN*, a specifična arhitektura *MobileNet*-a daje prednost brzini u odnosu na mogućnost raspoznavanja sitnijih detalja što dovodi do lošijih rezultata u odnosu na složenije mreže kao što je *ResNet50*. U poređenju sa složenom *VGG19*, postignut je bolji *F1 score*, ali manji *accuracy* i *AUC*. I kod ovog modela matrica konfuzije, prikazana na Slici 7, ukazuje na probleme prilikom prepoznavanja retkih i sličnih emocija. Najproblematičnija emocija je i za ovaj model bila strah. Ostali modeli su veliki broj grešaka pravili u klasifikaciji iznenađenja kao tuge, međutim *MobileNetV2* je napravio manji broj takvih grešaka, dok je najčešći problem bila klasifikacija straha kao iznenađenja.



Slika 7. Matrica konfuzije za *MobileNetV2*.

Analizom svih matrica konfuzije i rezultata predviđanja svake od klase zasebno uočena je izražena neravnoteža. Najbolji rezultati dobijeni su za emociju sreće, budući da su facijalne ekspresije za nju jako izražene i jedinstvene prema [19], što u velikoj meri olakšava modelu. Pored toga, ova emocija ima manje preklapanja sa drugim klasama. Takođe, broj podataka je znatno veći u odnosu na ostale emocije, što omogućava bolje učenje i klasifikaciju. Najslabija

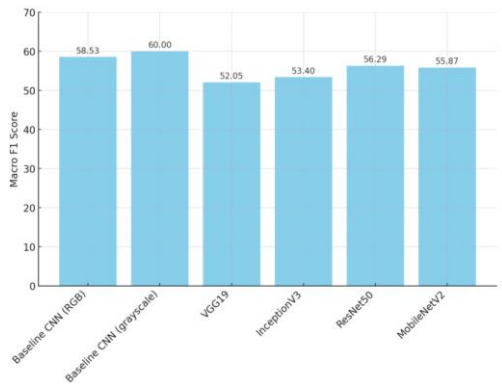
klasifikacija vrši se na slikama sa emocijom straha, jer su prema [19] izražajne karakteristike manje jasno definisane i često se preklapaju sa drugim emocijama poput tuge ili iznenađenja. Dosta pogrešnih klasifikacija je izvršeno između ljutnje, straha i iznenađenja. Osim slične i male količine slika dostupnih za ove emocije, ova pojava se objašnjava i sličnošću ekspresija lica pri izražavanju ovih emocija.

Rezultati svih upoređivanih modela dati su u Tabeli II. *VGG19*, *InceptionV3*, *ResNet50* i *MobileNetV2* pokazuju slične profile performansi. *VGG19* ima marginalno veće *accuracy* i *recall* u odnosu na *InceptionV3*, što znači da malo preciznije prepoznaje prave pozitivne primere i generalno hvata više instanci ispravnih klasa. S druge strane, *InceptionV3* beleži nešto bolji *precision*, što ukazuje da model ređe pogrešno svrstava uzorke u klasu kojoj oni u stvari ne pripadaju. *Macro F1* je malo veći kod *InceptionV3*, što sugeriše da je ukupna ravnoteža između preciznosti i odziva po klasama nešto povoljnija, odnosno da je malo pravedniji prema manje zastupljenim klasama. Slične performanse pokazuje i *MobileNet*. Upoređujući sa radom [11], naši rezultati pokazuju veću tačnost *MobileNet* arhitekture nego u radu, kao i slabije vrednosti metrike *accuracy* u odnosu na *ResNet50*. U našem radu, najveće vrednosti postigao je *ResNet50* koji ima nešto višu preciznost, odziv i *macro F1* što ukazuje da on bolje razlikuje klase, daje manji broj pogrešnih pozitivnih predikcija i ukupno pruža najbolji balans.

TABELA II.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	AUC
Baseline CNN (*RGB* ulaz)	64.48	57.68	60.61	58.53	89.32
Baseline CNN (*grayscale* ulaz)	61.65	59.84	60.82	60.00	89.17
VGG19	56.74	52.10	55.15	52.05	86.75
InceptionV3	55.04	53.17	54.58	53.40	85.71
ResNet50	58.49	56.83	56.25	56.29	86.78
MobileNetV2	56.65	56.63	55.54	55.87	86.05

Radi jasnog upoređivanja modela, na Slici 8 je dat prikaz postignutih *F1* skorova koji na jasan način ističe *baseline* kao najuspešniji model, ali i male razlike između svih modela.



Slika 8. Prikaz *macro F1* skorova

Prednost *VGG19*, *InceptionV3*, *ResNet50* i *MobileNetV2* u odnosu na *baseline* model je to što su ovi modeli trenirani na ogromnom *ImageNet* skupu podataka, tako da *fine-tuning* omogućava da već naučene opšte karakteristike budu iskorišćene za naš specifičan zadatak. Sa druge strane, na našem manjem i neuravnoteženom skupu podataka, složenost modela može dovesti do nešto nestabilnijeg učenja, *overfitting*-a i slabijih performansi.

VGG19 model, sa svojom složenijom arhitekturom u odnosu na *baseline* modele i velikim brojem parametara, ima prednost u sposobnosti da uči detaljnije i apstraktnije vizuelne reprezentacije lica, što je potencijalno korisno na većim i raznovrsnijim skupovima podataka. Univerzalnost *fine-tune*-ovanog *VGG19* modela ogleda se u tome što nije ograničen samo na jedan skup podataka. Jednom adaptiran, može lakše da se koristi za druge zadatke u domenu prepoznavanja emocija. Veći broj parametara takođe znači znatno duže vreme treniranja u poređenju sa *baseline* modelima – *VGG19* je otprilike dvadesetak puta po epohi sporiji od *baseline* modela. Međutim, model i dalje pokazuje solidnu sposobnost prepoznavanja izraženih emocija i generalno ujednačenih klasa, naročito zahvaljujući višeslojnoj arhitekturi koja daje prethodno pomenutu mogućnost *fine-tuning*-a.

Prednost *InceptionV3* modela jeste u njegovoj modularnosti i sposobnosti da istovremeno uči karakteristike različitih razmera kroz *inception* module, što potencijalno može poboljšati prepoznavanje složenih obrazaca na većim i raznovrsnijim skupovima podataka. Model je znatno sporiji u treniranju u poređenju sa *baseline* mrežom – otprilike deset puta po epohi, iako je znatno brži od *VGG19*. Ipak, i dalje pokazuje solidnu sposobnost prepoznavanja češćih i izraženijih emocija, a modularna arhitektura omogućava fleksibilnost za *fine-tuning* i adaptaciju na specifične zadatke.

ResNet50 predstavlja mrežu sa velikim brojem parametara i zbog toga je i očekivan lošiji rezultat na malom skupu podataka kao što je naš, kao što je pomenuto u [8]. Male dimenzije sliku su gotovo onemogućavale da ova mreža uči, dok je povećavanje dimenzija slika značajno popravilo rezultate. Pretpostavka je da se ovaj rezultat može lako prevazići dodatnim povećavanjem veličine na 224x224 piksela. Prednost *ResNet50* je to što ona zbog specifične arhitekture ima sposobnost da prepozna i najsuptilnije detalje, ali problem nastaje prilikom praktične upotrebe na našem skupu podataka zbog nebalansiranosti i malog broja primera. U odnosu na *MobileNetV2*, brzina treniranja je dosta veća i zahteva više resursa.

MobileNetV2 arhitektura ima mali broj parametara u odnosu na ostale poznate *CNN* arhitekture korišćene u ovom radu što omogućava znatno brže treniranje ovog modela i smanjuje sklonost model ka *overfit*-ovanju. Zbog toga, ovaj model je ima potencijal da daje bolje rezultate na malim skupovima podataka i na slikama manjih rezolucija. Prednost upotrebe ovog modela jeste brzina, efikasnost i mogućnost upotrebe slika manjih dimenzija, dok je mana to što arhitektura nije osmišljena tako da hvata finije detalje što može biti presudno za slučaj raspoznavanja emocija sa lica.

Baseline grayscale model ostvaruje najbolje performanse uprkos jednostavnijoj arhitekturi iz nekoliko razloga. Prvo, manja i plića mreža bolje odgovara veličini i karakteristikama našeg *dataset*-a; složeniji modeli poput

često imaju previše parametara u odnosu na količinu podataka, što ih čini sklonim *overfitting*-u i otežava generalizaciju. Drugo, korišćenje *grayscale* slika uklanja informacije o boji koje nisu presudne za prepoznavanje facijalnih ekspresija, čime se smanjuje „vizuelni šum“ i model se može fokusirati na ključne oblike i teksture lica. Pokazali smo da ovaj model postiže nešto viši *macro f1* u odnosu na istu arhitekturu, ali kada ulazne slike imaju tri kanala – *RGB*. Treće, *baseline* model je treniran od nule direktno na našem *dataset*-u, što znači da je potpuno prilagođen specifičnostima tih podataka, dok pretrenirani modeli mogu biti previše vezani za vizuelne obrasce sa *ImageNet*-a, koji se razlikuju od facijalnih ekspresija i tako gube efikasnost. Dalje, jednostavnija arhitektura može bolje balansirati između učenja suptilnih detalja i očuvanja stabilnosti tokom treniranja, što se reflektuje u višem *macro F1* skoru, pokazatelju bolje ujednačenosti performansi između klasa, uključujući i retke ili teže prepoznatljive emocije poput straha ili iznenađenja. Takođe, najefikasniji je model po brzini. Ipak, ima i određena ograničenja. Kao manji i jednostavniji model treniran od nule, njegova sposobnost da uči kompleksnije reprezentacije je ograničena, što znači da bi mogao imati slabiju generalizaciju na većim i raznovrsnijim skupovima podataka. Iako generalno daje stabilnije rezultate od složenijih modela, i dalje ima izražene poteškoće sa određenim klasama kao što su strah, ljutnja i iznenađenje. Konačno, celokupan rezultat dobijen u ovom radu sugeriše da veće i dublje mreže nisu automatski superiorne za male i neuravnotežene skupove podataka.

V. Zaključak

U ovom radu ispitane su različite arhitekture konvolucionih neuronskih mreža za rešavanje problema prepoznavanja emocija spram izraza lica na slikama. Eksperimentalni rezultati su pokazali da je za izabrani skup podataka efikasnije koristiti jednostavnije *CNN* arhitekture nego unapred trenirane duboke mreže sa velikim brojem parametara. Najbolje se pokazala *baseline CNN* dizajnirana specijalno za rešavanje ovog problema. Utvrđeno je da je za sve pomenute modele i dalje izazovno razlikovati svaku od sedam emocija iz *FER* skupa.

Buduća proširenja ovog projekta bi mogla obuhvatati ispitivanje i upotrebu naprednijih tehnika za povećavanje balansiranosti podataka, kao i dodavanje još skupova podataka čime bi se mogli drastično poboljšati rezultati na složenijim unapred treniranim mrežama. Za primenu prepoznavanja facijalnih emocija u realnim slučajevima, potrebno je iskoristiti predložene arhitekture nad podacima koji stižu u realnom vremenu sa veb-kamere. Da bi se to postiglo, zaključujemo da je zbog bolje generalizacije bolje koristiti neku od unapred treniranih modela iz rada, ali uz povećanje količine podataka za *fine-tuning*. Korišćenje *MobileNet* bi omogućilo i efikasno prepoznavanje emocija sa kamera na mobilnim telefonima.

U zavisnosti od potrebe, arhitekture ispitane u ovom radu se, uz već napomenuta potencijalna poboljšanja u budućnosti, mogu koristiti u realnim slučajevima zarad poboljšanja interakcije čovek-računar, u obrazovanju, zdravstvenoj zaštiti, robotici.

Reference

[1] A. S. Vyas, H. B. Prajapati, and V. K. Dabhi, “Survey on Face Expression Recognition using CNN,” IEEE Xplore, Mar. 01, 2019. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8728330.

[2] U. Hess, “Nonverbal Communication,” Encyclopedia of Mental Health, vol. 3, no. 2, pp. 208–218, 2016, doi: https://doi.org/10.1016/b978-0-12-397045-9.00218-4.

[3] A. Agarwal and S. Susan, “Emotion Recognition from Masked Faces using Inception-v3,” pp. 1–6, Mar. 2023, doi: https://doi.org/10.1109/rait57693.2023.10126777.

[4] C. Qian, J. Alexandre, and S. J. Fong, “Analysis of deep learning algorithms for emotion classification based on facial expression recognition,” Proceedings of the 2024 8th International Conference on Big Data and Internet of Things, Sep. 2024, doi: https://doi.org/10.1145/3697355.3697382.

[5] L. Bi, S. Tang, and C. Li, “A Facial Expression Recognition Method Based on Improved VGG19 Model,” International Journal of Advanced Computer Science and Applications, vol. 15, no. 7, 2024, doi: https://doi.org/10.14569/ijacsa.2024.0150725.

[6] D. Bhagat, A. Vakil, Rajeev Kumar Gupta, and A. Kumar, “Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN),” Procedia computer science, vol. 235, pp. 2079–2089, Jan. 2024, doi: https://doi.org/10.1016/j.procs.2024.04.197.

[7] G. Meena, K. K. Mohbey, and S. Kumar, “Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach,” International Journal of Information Management Data Insights, vol. 3, no. 1, p. 100174, Apr. 2023, doi: https://doi.org/10.1016/j.jjimei.2023.100174.

[8] H. Sheng and M. Lau, “Optimising Real-Time Facial Expression Recognition with ResNet Architectures,” Journal of Machine Intelligence and Data Science, vol. 5, 2024, doi: https://doi.org/10.11159/jmids.2024.005.

[9] A. Livingston, “FACIAL EMOTIONAL RECOGNITION USING MOBILENET BASED TRANSFER LEARNING,” EPRA International Journal of Research & Development (IJRD), vol. 8, no. 7, pp. 1–1, Jul. 2023, Accessed: Oct. 05, 2025. [Online]. Available: https://eprajournals.com/IJSR/article/11017.

[10] D.-H. Lee and J.-H. Yoo, “CNN Learning Strategy for Recognizing Facial Expressions,” IEEE access, vol. 11, pp. 70865–70872, Jan. 2023, doi: https://doi.org/10.1109/access.2023.3294099.

[11] I. Ul Haq, A. Ullah, K. Muhammad, M. Y. Lee, and S. W. Baik, “Personalized Movie Summarization Using Deep CNN-Assisted Facial Expression Recognition,” Complexity, vol. 2019, pp. 1–10, May 2019, doi: https://doi.org/10.1155/2019/3581419.

[12] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv.org, Apr. 10, 2015. https://arxiv.org/abs/1409.1556.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” arXiv.org, 2015 https:/arxiv.org/abs/1512.00567

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 2015. Available: https://arxiv.org/pdf/1512.03385.

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” arXiv.org, 2018. https://arxiv.org/abs/1801.04381.

[16] A. BRITAL, “Inception V3 CNN Architecture Explained .,” Medium, Oct. 24, 2021. https://medium.com/@AnasBrital98/inception-v3-cnn-architecture-explained-691cfb7bba08

[17] F. Chollet et al., “Keras,” GitHub repository, 2015. [Online]. Available: https://keras.io.

[18] J. Oheix, "Facial Expression Recognition Dataset," 2017. [Online]. Available: https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset/data.

[19] E. G. Krumhuber, L. I. Skora, H. C. H. Hill, and K. Lander, “The role of facial movements in emotion recognition,” Nature Reviews Psychology, Mar. 2023, doi: https://doi.org/10.1038/s44159-023-00172-1