

UNIVERSITATEA BABEȘ-BOLYAI

Facultatea de Științe Economice și Gestiunea Afacerilor

Sisteme de Asistare a Deciziilor Economice

Lucrare de disertație

Absolvent,

Oana Luisa **ROMAN**

Coordonator științific,

Lect. univ. dr. Darie **MOLDOVAN**

2021

UNIVERSITATEA BABEȘ-BOLYAI

Facultatea de Științe Economice și Gestiunea Afacerilor

Sisteme de Asistare a Deciziilor Economice

Lucrare de disertație

Identificarea emoțiilor pe baza expresiilor faciale

Absolvent,

Oana Luisa **ROMAN**

Coordonator științific,

Lect. univ. dr. Darie **MOLDOVAN**

2021

Rezumat

Detectarea emoției în expresia feței a devenit o nevoie datorită faptului că se poate aplica în diferite arii, precum: medicină, psihologie, și bineînțeles, inteligență artificială, unde ar putea ajuta la dezvoltarea colaborării între om-calculator, și chiar și comunicării între om-robot. Acest subiect reprezintă o problemă atât solicitantă, din punct de vedere al performanței și sistemelor sau configurărilor de care are nevoie, cât și interesantă în viziunea computerizată, astfel încât au fost efectuate un număr destul de elaborat de lucrări cu privire la această temă. Obiectivul acestei cercetări este de a dezvolta un sistem de recunoaștere a emoției pe baza expresiei faciale utilizând o rețea neuronală convoluțională. Această abordare permite clasificarea a șapte emoții de bază constând în: furie, fericire, supărare, dezgust, frică, surprindere, și o stare neutră.

Cuprins

Abrevieri.....	iv
Lista tabelelor și figurilor.....	v
Introducere	1
1. Studiu Bibliografic	2
2. Metodologie	11
2.1 Setul de date.....	11
2.1.1 <i>FER2013</i>	11
2.1.2 <i>Set de date compus</i>	12
2.2 Arhitectura sistemului și a rețelelor neuronale convoluționale.....	14
2.3 Implementarea sistemului	18
3. Rezultate	23
Concluzii	28
Bibliografie	29

Abrevieri

<i>CK+</i> ,	set de date extins Cohn-Kanade
<i>CNN</i> ,	rețea neuronală convoluțională
<i>FER</i> ,	recunoașterea expresiei faciale
<i>JAFPE</i> ,	expresia facială feminină japoneză
<i>KDEF</i> ,	Karolinska Directed Emotional Faces
<i>LBP</i> ,	model binar local
<i>ReLU</i> ,	unitate liniară rectificată
<i>VGG16</i> ,	arhitectură CNN cu 16 straturi numită după grupul de geometrie vizuală (Visual Geometry Group) de la Oxford

Lista tabelelor și figurilor

Tabele:

Tabel 1. Arhitectura rețelelor neuronale convoluționale.....	17
Tabel 2. Parametrii folosiți pentru funcția de "ImageDataGenerator".....	19
Tabel 3. Clasificarea performanțelor rețelelor obținute pe setul de date FER2013	24
Tabel 4. Clasament privind rezultatele obținute pe seturi de date compuse	27

Figuri:

Figura 1. Diagrama model CNN propusă pentru recunoașterea emoțiilor faciale (Ozdemir, M. A., Elagoz, B., Alaybeyoglu, A., Sadighzadeh, R., & Akan, A., 2019)	3
Figura 2. Arhitectura modelului propus (Li, J., Jin, K., Zhou, D., Kubota, N., & Ju, Z., 2020).....	5
Figura 3. Modelul VGG16 (Hussain, S. A., & Al Balushi, A. S. A., 2020)	6
Figură 4. Robotul NAO - emoțiile și reprezentarea lor în culori (Melinte, D. O., & Vladareanu, L., 2020)	10
Figura 5. Ilustrarea numărului de imagini din fiecare tip de emoție	11
Figura 6. Imagini aleatorii din setul de date reprezentând fiecare emoție (Kaggle - FER2013 data set)	12
Figura 7. Imagini aleatorii reprezentând fiecare emoție din setul de date compus (Kaggle - Emotion- compilation data set)	13
Figură 8. Diagrama fluxului de evenimente.....	15
Figură 9. Exemplu de predicție prin încărcare de fotografie folosind aplicația web	20
Figură 10. Exemplu de caracteristici "Haar Cascade" (OpenCv docs).....	21
Figura 11. Capturi de predicții în timp real prin camera web	22
Figura 12. Progresul acurateței model cu 7 straturi de convoluție și 2 milioane de parametrii	23
Figura 13. Clasament competiție Kaggle (Kaggle - Facial Expression Recognition Challenge - Leaderboard)	25
Figură 14. Matricele de confuzie a modelelor antrenate pe setul de date compus.....	26

Introducere

Expresiile faciale oferă informații importante despre emoțiile unei persoane. Înțelegerea corectă a expresiilor faciale este una dintre sarcinile provocatoare pentru relațiile interumane. (Rani, J., & Garg, K., 2014)

Detectarea automată a emoțiilor folosind recunoașterea expresiilor faciale este acum un domeniu principal de interes în diferite domenii, cum ar fi informatica, medicina și psihologia. Îmbunătățirea abilității de recunoaștere a expresiei faciale este necesară pentru ca un agent inteligent să comunice cu omul ca parte a colaborării mașină-om, precum și cu roboții ca parte a interacțiunii robot-robot. Întrucât cercetările privind recunoașterea expresiei faciale se desfășoară de ani de zile, progresele cercetărilor pe această temă sunt laudabile. Fluctuația ratei de recunoaștere printre clase este una dintre problemele majorității cercetărilor, deoarece acestea au o rată de recunoaștere mai mică pentru a detecta emoții precum dezgustul și frica.

Obiectivul principal al acestei lucrări este de a realiza un sistem care să ajute la recunoașterea celor 7 emoții de bază, astfel se vor descrie două rețele neuronale convoluționale care vor fi folosite în cadrul acestuia. În plus, s-a dezvoltat și o aplicație web cu scopul de a crea un mediu prin care modelele să poate fi testate. Aceasta oferă 2 modalități prin care să se identifice emoțiile: pe baza încărcării unei fotografii sau în timp real, prin camera web. Acuratețea la care ajung cele două rețele, după încercarea a mai multor arhitecturi și abordări, este între 80-82%.

Trebuie menționat faptul că acest proces este unul destul de complex până și pentru oameni. Noi, ca indivizi, reușim să învățăm să recunoaștem aceste emoții de-a lungul timpului, dar un factor important este să cunoaștem și persoana. Pentru a face o diferență între aceste stări trebuie să fim foarte atenți la anumite caracteristici specifice care se regăsesc, în mare parte, la nivelul ochilor, nasului și gurii. Aceste particularități pot fi însă comune în cazul anumitor emoții, iar atunci trebuie privită imaginea de ansamblu.

Capitolele acestei lucrări constau din: un capitol bibliografic care cuprinde diferite lucrări conexe privind recunoașterea expresiei faciale, o prezentare generală a metodologiei utilizate în această cercetare, colectarea și preprocesarea datelor, modul în care a fost implementat sistemul propus, rezultatele obținute, concluzia și obiectivele viitoare.

1. Studiu Bibliografic

În cadrul acestui capitol, se vor prezenta câteva dintre lucrările și articolele de specialitate care urmăresc recunoașterea expresiilor faciale folosind diverse modele, algoritmi, tehnici și seturi de date.

În lucrarea redactată de (Ozdemir, M. A., Elagoz, B., Alaybeyoglu, A., Sadighzadeh, R., & Akan, A., 2019) este propusă o rețea neuronală convoluțională bazată pe arhitectura LeNet pentru recunoașterea a 7 tipuri de expresii faciale: fericire, supărare, furie, dezgust, frică, surprindere și o stare neutră.

Au fost folosite 3 seturi de date: JAFFE – care conține 213 imagini luate de la 10 modele feminine japoneze; KDEF – alcătuit din 4900 de fotografii de la 35 de bărbați și 35 de femei; setul lor de date personalizat cu 140 de imagini de la o singură femeie și un singur bărbat, fiecare expresie facială fiind exprimată de 10 ori de către fiecare participant. Toate seturile de date conțin imagini pentru fiecare dintre cele 7 tipuri de expresii faciale care se vor a fi recunoscute.

Detecția feței s-a făcut cu ajutorul bibliotecii "Haar Cascade" creându-se dreptunghiuri în jurul feței, decupând astfel toate imaginile și ajungând așadar la aceeași dimensiune, 64x64, și transformându-le în imagini gri. Cu arhitectura CNN propusă, se urmărește educarea valorilor pixelilor din regiunea dreptunghiulară, care conțin expresii faciale, creându-se interogări rapide cu ajutorul modelului de rețea neuronală artificială.

În figura de mai jos se poate observa structura rețelei neuronale propusă. Rețeaua imită structura LeNet utilizată în clasificarea expresiei feței în format 2D și include două straturi convoluționale, două straturi max-pooling și un strat complet conectat.

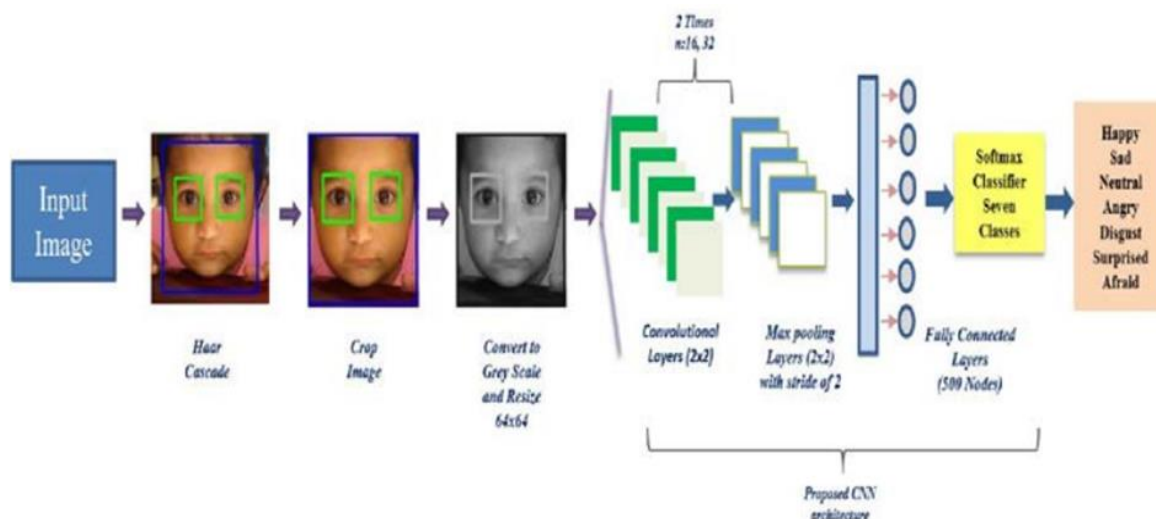


Figura 1. Diagrama model CNN propusă pentru recunoașterea emoțiilor faciale (Ozdemir, M. A., Elagoz, B., Alaybeyoglu, A., Sadighzadeh, R., & Akan, A., 2019)

În formarea rețelei, dimensiunea datelor de test a fost de 25%. Dimensiunea lotului ("batch size") a fost de 32 și numărul de epoci a fost de 500. După instruirea arhitecturii CNN propuse, modelul a fost testat în timp real. În primul rând, fețele umane au fost detectate folosind biblioteca "Haar Cascade" în 30 de imagini pe secundă de la o cameră web. După detectarea imaginilor, acestea au fost trimise modelului și clasele de care aparțin au fost interogate. Ca urmare a predicțiilor, posibilitatea de apartenență la o clasă anume era afișată pe un ecran separat, iar tipul de emoție care avea predicția mai mare era suprascris pe cadrul format de "Haar Cascade". Pentru instruirea arhitecturii LeNet CNN prezentate au fost folosite bibliotecile Keras și TensorFlow. Conform rezultatelor, modelul a ajuns la acuratețea de 96.43%, acuratețea pe datele de validare fiind de 91.81%.

În concluzie, lucrarea propune o metodă de cost și funcționalitate redusă bazată pe arhitectura LeNet CNN. Autorii au ajuns la concluzia că utilizarea unui set de date personalizat a oferit o validare și precizie mai mare decât instruirea doar pe seturile de date existente.

O abordare puțin mai diferită putem regăsi în lucrarea publicată de (Li, J., Jin, K., Zhou, D., Kubota, N., & Ju, Z., 2020) care propune o rețea neuronală convoluțională care folosește un mecanism de atenție, abreviat ACNN, pentru recunoașterea automată a expresiei faciale. Arhitectura este formată din patru module: de extragere a caracteristicilor, de atenție, de reconstrucție și de clasificare. LBP oferă informații despre textura imaginii și apoi prinde mișcările mici ale feței, care pot îmbunătăți performanța rețelei, în timp ce mecanismul de

atenție poate face ca rețeaua să acorde mai mult interes caracteristicilor utile. Pentru îmbunătățirea rezultatelor recunoașterii expresiilor faciale se combină caracteristicile LBP cu cele ale convoluțiilor.

Metoda propusă a fost testată pe cinci seturi de date de expresie facială: CK+, JAFFE, FER2013, Oulu-CASIA și NCUE (Nanchang University Facial Expression) – un set de date colectat de către autori care conține cele 7 expresii faciale de bază de la 35 de studenți absolvenți, 6 femei și 29 de bărbați, cu ajutorul unui senzor Kinect de la Microsoft. Pentru fiecare tip de imagine există 245 de secvențe, fiecare conținând 110 imagini rezultând astfel un total de 26.950 de imagini.

Arhitectura modelului începe de la primul modul de extragere a caracteristicilor compus din două fluxuri de procesare CNN separate. Pentru a preveni ca rețeaua să fie prea complexă, filtrele de convoluție de dimensiuni mici, 3x3, sunt folosite în toate straturile. Ulterior, se fuzionează caracteristicile F1 extrase din imaginile brute cu caracteristicile F2 extrase din LBP și apoi adăugate caracteristicile fuzionate F3 la modulul de atenție.

Acest modul funcționează prin îmbunătățirea caracteristicilor utile și face ca rețeaua să se concentreze mai mult pe aceste caracteristici care sunt vitale pentru recunoașterea expresiei. În acest fel, rețeaua poate recunoaște diferite expresii mai eficient. Apoi, modulul de reconstrucție ajustează ”harta” atenției pentru a crea o hartă de caracteristici îmbunătățită pentru modulul de clasificare. În cele din urmă, straturile complet conectate folosind funcția ”Softmax” sunt utilizate pentru clasificare.

Figura 2. prezintă arhitectura modelului descris anterior. Se poate observa modalitatea prin care modulele comunică între ele și cum, respectiv când, are loc fuzionarea caracteristicilor, trecerea prin modulul de reconstrucție și până la procesul ce ține de clasificare. De altfel, este oferită și o imagine de ansamblu a modelului de reconstrucție. În partea stângă jos, este o legendă care ajută la perceperea mai bună a arhitecturii, și astfel, avem reprezentate straturile convoluționale cu albastru, straturile de activare cu funcția ”ReLU” cu roz-crem, stratul cu funcția ”Sigmoid” cu verde, iar straturile complet conectate cu mov.

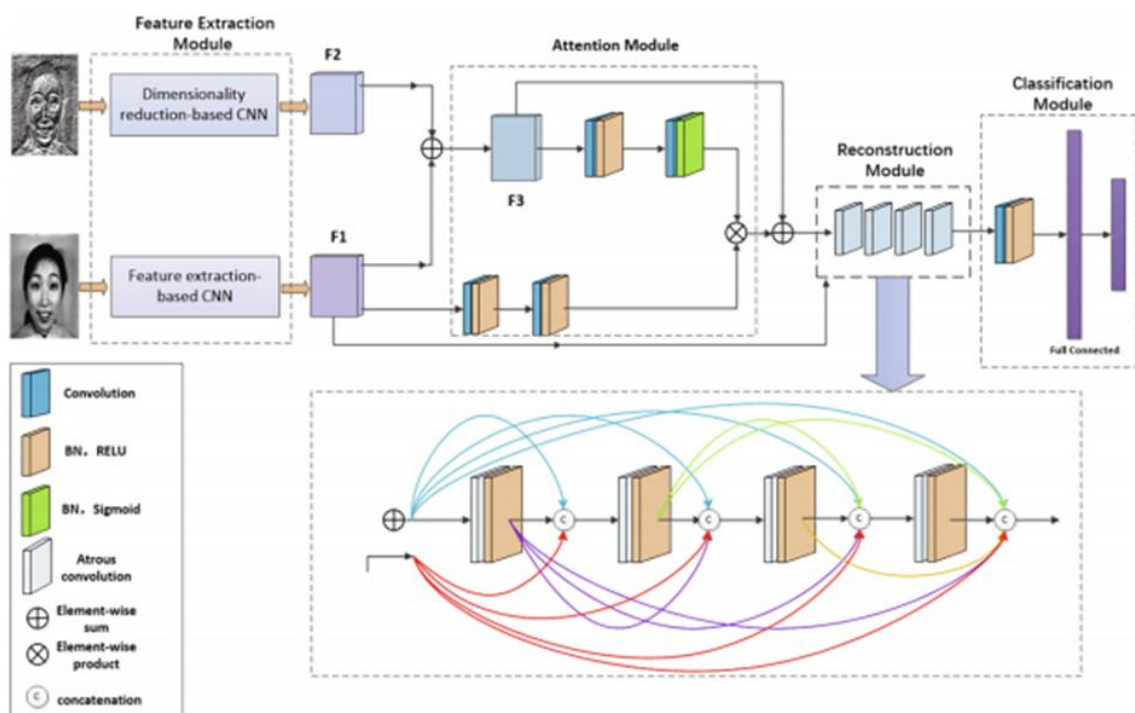


Figura 2. Arhitectura modelului propus (Li, J., Jin, K., Zhou, D., Kubota, N., & Ju, Z., 2020)

Rezultatele obținute sunt împărțite și comparate cu alte lucrări, în funcție de setul de date folosit. De exemplu pe setul de date de la FER2013, modelul propus obține o rată de recunoaștere de 75.82%, în timp ce pe setul de date de la CK+ se obține o rată de 98.68%. Pe setul de date de la JAFFE se obține 98.52%, pe datele de la Oulu-CASIA de 94.63%, iar pe setul lor de date rata recunoașteri emoției pe baza expresiei faciale este de 94.33%. Aceste rezultate arată faptul că modelul propus este superior față de multe alte metode existente pentru aceste seturi de date. În viitor, autorii își propun să îmbunătățească arhitectura astfel încât aceasta să fie potrivită și pentru video-uri sau imagini 3D.

O diferită abordare care folosește detectarea feței prin intermediul camerei web, o regăsim în procesul secvențial al lucrării scrise de (Hussain, S. A., & Al Balushi, A. S. A., 2020), definit în trei faze diferite.

În prima fază, după cum spuneam, este detectată fața umană prin intermediul camerei web, urmând apoi să se analizeze captura pe baza caracteristicilor și a bazei de date utilizate cu suportul modelului rețelei neuronale convoluționale realizat cu ajutorul bibliotecii Keras.

În ultima fază se clasifică expresia facială ca fiind fericită, furioasă, neutră, tristă, dezgustată, sau surprinsă.

Detectarea feței în timp real și delimitarea acesteia printr-o casetă se realizează folosind biblioteca "Haar Cascade", capturile create se salvează apoi într-o bază de date pentru recunoașterea facială. Modelul CNN folosește VGG16 pentru a se potrivi cu fața din baza de date și a face o legătură cu numele feței detectate. În cele din urmă, fața umană recunoscută este clasificată pe baza expresiei în timp real.

Lucrarea prezentată implică histograma modelului binar local pentru a converti imaginea capturată în vector binar. Această procesare ajută la detectarea feței folosind algoritmul "Viola Jones". Acești vectori sunt adăugați pentru a forma un model de arhitectură de rețea pentru clasificarea expresiei faciale utilizând modelul VGG16 CNN descris în figura de mai jos.

Imaginea primită inițial are forma de $224 \times 224 \times 3$, unde 224 reprezintă dimensiunea imaginii, iar 3 numărul de canale. Inițial se găsesc 3 canale, RGB (roșu, verde și albastru), dar după ce aceasta trece prin 2 straturi de convoluție cu funcția de activare "ReLU", dimensiunea devine $224 \times 224 \times 64$. Continuând prin straturile de max-pooling, și cele de convoluție dimensiunea imaginii ajunge la $7 \times 7 \times 512$, ceea ce înseamnă că numărul de canale a crescut considerabil, astfel se pot extrage mai multe informații. În final, straturile complet conectate cu 4096 de noduri și cu ajutorul funcției "Softmax", pot să producă 1000 de rezultate de predicție.

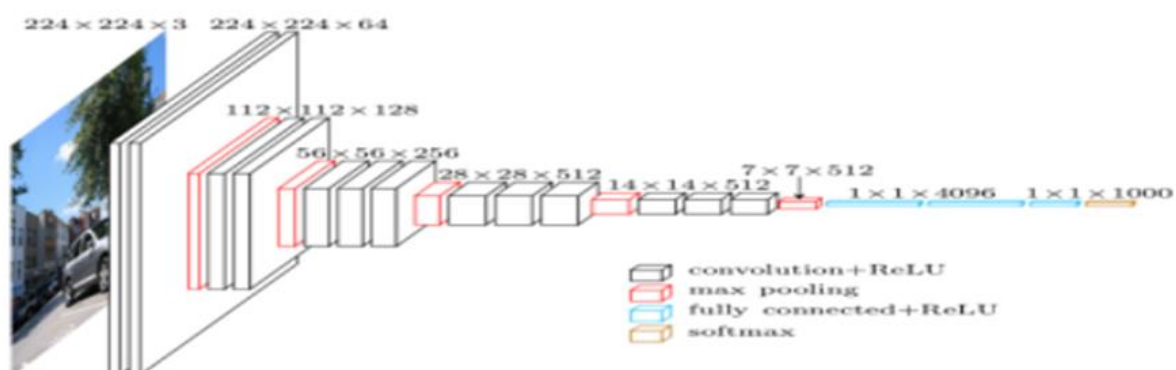


Figura 3. Modelul VGG16 (Hussain, S. A., & Al Balushi, A. S. A., 2020)

Setul de date KDEF folosit pentru acest model propus are 4900 de imagini, dintre care au fost eliminate 1999 de imagini deoarece erau în postură laterală, astfel setul rămâne la 2901 de imagini, împărțindu-se 70% pentru antrenament și 30% pentru testare. Modelul propus ajunge la o performanță de 88% și poate fi folosit și în timp real.

(Sajjad, M., Zahir, S., Ullah, A., Akhtar, Z., & Muhammad, K., 2019) prezintă o lucrare în care se analizează comportamentul uman folosind expresii faciale, luând în considerare unele seriale faimoase. Mai întâi, detectarea feței se face, de asemenea, folosind algoritmul ”Viola-Jones”, iar cu ajutorul algoritmului Kanade-Lucas-Tomasi se urmărește cu precizie fața recunoscută pe tot parcursul videoclipului.

Recunoașterea facială se face folosind caracteristicile histogramei de gradienti orientați (HOG) împreună cu clasificatorul SVM (Support Vector Machine). Urmând apoi ca expresiile faciale să fie recunoscute folosind o rețea neuronală convoluțională. Utilizând HOG imaginile sunt reprezentate în grile formate din celule de tipul $M \times N$. Fiecare celulă păstrează informații despre orientările marginilor în ”coșuri” (”bins”), acesta reprezentând o serie de orientări de gradient.

Dimensiunea mai mică a blocului și a celulei extrage o reprezentare mai robustă a feței și captează modele de forme mici ale ochilor, nasului și gurii. Așadar, în cadrul acestei lucrări, imaginile fețelor sunt reprezentate la dimensiunea de 128×128 , cu dimensiunea celulei de 8×8 și cea a blocului de 2×2 , cu 18 ”coșuri” pe celulă. Această configurație extrage caracteristici precise care sunt suficient de puternice pentru a reprezenta o față pentru recunoaștere.

Datorită faptului că setul de date folosit, KDEF FER, conține o cantitate mică de date, autorii au aplicat abordări de mărire a datelor, folosind biblioteca ”Augmenter”, pentru ca setul creat să fie suficient de capabil pentru instruirea modelului CNN. Datele sunt împărțite în date de validare 20%, de test 20% și de antrenament 60%.

Modelul CNN este instruit pentru 100 de epoci și a atins o validare de 80,5% pentru epoca finală fără aplicarea tehnicii de mărire a datelor. După aplicarea acesteia, precizia de validare a ajuns la 94%. Setul de testare a atins 93,39% precizie, mai mare decât modelele anterioare de pe setul de date folosit, datorită nucleelor de dimensiuni reduse aplicate pe imagini cu dimensiunea 128×128 . Acest lucru permite modelului să învețe diferite modele minuscule regăsite într-o imagine. Precizia claselor de dezgust, fericire, tristețe și surpriză este mai mare de 90%, iar emoțiile de supărare și neutru rămân sub acuratețea de 90%.

În ceea ce privește recunoașterea emoțiilor în serialele TV s-au ales câte 10 episoade din serialele "Friends" și "Extras" și s-au descărcat în format mp4, la o rezoluție scăzută de 360x480 și 30 de cadre pe secundă. Roluri de la seriale sunt împărțite în două categorii, principalii actori fiind recunoscuți după nume, iar ceilalți sunt considerați ca "alții".

De exemplu caracterul numit "Chandler" este recunoscut cu o acuratețe de 88%, iar acuratețea medie ajunge la 89.95%.

(Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I., 2018) examinează performanța a două abordări diferite de deep learning asupra aceluiași set de date, FER2013, pentru recunoașterea expresiei faciale.

Prima arhitectură "Inception", denumită GoogLeNet, conține o rețea alcătuită din 22 de straturi cu parametrii, per total fiind de fapt aproape 100 de straturi. Există părți ale rețelei care sunt executate în paralel, numite "module de început". Nouă astfel de module sunt utilizate în rețea, până când se ajunge la ultimele straturi, unde are loc procesul de clasificare. Contrar celor mai multe arhitecturi de învățare profundă, GoogLeNet nu folosește straturi complet conectate, ci un "average pool", astfel vor fi folosiți mult mai puțini parametrii. O comparație față de următoarea arhitectură este faptul că aceasta utilizează de 12 ori mai puțini parametrii.

Arhitectura prezentată inițial într-o competiție în 2012, AlexNet, este denumită după creatorul său Alex Krizhevsky. În acest caz, rețeaua este formată din 8 straturi complet conectate, iar acest lucru face ca numărul parametrilor să crească considerabil. De exemplu, doar pe primul strat există 105705600 de parametrii, făcând ca supra-adaptarea ("overfitting") să fie inevitabilă.

Analizele acestui studiu s-au împărțit în trei etape. Mai întâi, performanța rețelelor a fost examinată în vederea recunoașterii existenței unui conținut emoțional într-o expresie facială, urmând mai apoi să fie observată performanța în ceea ce privește recunoașterea emoției. În a treia etapă, cele două metode au fost instruite atât pe date emoționale, cât și pe cele neutre. Acuratețea arhitecturilor nu a fost foarte diferită, în ultima etapă, AlexNet fiind cu 1% mai performant decât GoogLeNet, ajungând la o performanță de aproximativ 85%.

Interacțiunea dintre oameni și un robot NAO utilizând rețele neuronale convoluționale este prezentată în lucrarea concepută de (Melinte, D. O., & Vladareanu, L., 2020). Scopul acestei

cercetări este de a dezvolta un canal complet de comunicare între un om și un robot NAO, adică un robot umanoid autonom și programabil dezvoltat de o companie din Franța. Astfel, lucrarea se concentrează pe îmbunătățirea performanței diferitelor tipuri de CNN, în ceea ce privește acuratețea, generalizarea și viteza de inferență, utilizând de exemplu: mai multe metode de optimizare, extinderea bazei de date FER2013 cu imagini din alte baze de date (CK+, JAFFE, KDEF).

Astfel, se prezintă folosirea a două rețele optimizate, una pentru recunoașterea feței și alta pentru recunoașterea expresiei faciale. Modelele CNN utilizate pentru recunoașterea expresiei faciale, VGG, ResNet și InceptionV3, au fost pre-antrenate pe baza de date ImageNet și pot recunoaște obiecte din 1000 de clase, nu neapărat care să aibă legătură cu fețe sau oameni.

Abordarea este deosebită în comparație cu lucrările descrise anterior, folosindu-se învățarea prin transfer și ”fine-tuning”, după cum urmează: în prima fază, straturile complet conectate ale modelului pre-antrenat au fost înlocuite cu două straturi noi inițializate aleatoriu, capabile să clasifice imaginile de intrare în funcție de setul de date și clasele aferente. În timpul acestui antrenament de încălzire, toate straturile convoluționale au fost ”înghețate”, permițând gradientului să se propage înapoi numai prin noile straturi complet conectate. În cea de-a doua etapă, ultimele straturi ale rețelelor convoluționale nu au fost ”înghețate” pentru a permite gradientului să se propage înapoi prin aceste straturi, dar cu o rată de învățare mult mai mică, învățându-se astfel reprezentări la un nivel înalt.

Întregul sistem care să reprezinte acest canal de comunicare a fost implementat pe robotul NAO și a fost împărțit în: captarea imaginii de către robot prin camera acestuia, modelul de recunoaștere a feței, modelul de recunoaștere a emoțiilor faciale și expresia facială a robotului, adică răspunsul său la identificarea clasei. LED-urile pentru ochii robotului sunt colorate în RGB, această caracteristică fiind utilizată pentru asocierea unei emoții cu o culoare: fericit – verde, furios – roșu, trist – albastru, dezgust – galben, neutru – negru, surpriză – alb și frică – portocaliu. Intensitatea culorii va fi ajustată în funcție de probabilitatea de detectare a emoțiilor. Figura de mai jos ilustrează culorile robotului NAO în concordanță cu cele 7 emoții.



Figură 4. Robotul NAO - emoțiile și reprezentarea lor în culori (Melinte, D. O., & Vladareanu, L., 2020)

Din punct de vedere al performanțelor obținute, rețeaua cea mai performantă a reieșit ResNet, cu o precizie de 90,14%, în timp ce VGG a ajuns la o acuratețe de 87%, iar InceptionV3 la 81%. Autorii doresc pe viitor să implice fuziunea altor intrări, cum ar fi: modele audio, imagini în infraroșu, informații de profunzime din modele de fețe 3D; care vor oferi informații suplimentare și vor spori în continuare acuratețea procesului, iar modelul de interacțiune cu robotul umanoid va fi dezvoltat în continuare pentru a fi aplicat în scopuri medicale, în principal pentru îmbunătățirea comunicării și a comportamentului copiilor cu tulburare a spectrului autismului.

2. Metodologie

Acest capitol descrie analiza teoretică a metodelor folosite pentru crearea modelelor și sistemului, arhitectura și implementarea acestora, cât și prezentarea setului de date utilizat și a interfeței grafice dezvoltate cu scopul de a oferi un mediu de testare în timp real.

2.1 Setul de date

2.1.1 FER2013

În prima fază a cercetării s-a ales setul de date FER2013 (Kaggle - FER2013 data set) care conține imagini de dimensiunea de 48x48 de pixeli în tonuri de gri. Fotografiile de fețe care să exprime cele 7 emoții au fost înregistrate automat astfel încât fața să fie mai mult sau mai puțin centrată și să ocupe aproximativ același spațiu, ca și dimensiune, în fiecare poză, dar și să ilustreze variabilități de iluminare, vârstă, poziție, intensitatea expresiei.

Setul este împărțit în date de instruire, care include aproximativ 29.000 de exemple, și date de testare în număr de aproximativ 7000 de poze. Imaginile aferente fiecărei emoții în parte sunt grupate în dosare separate, atât pentru setul de instruire, cât și pentru cel de testare. Un dezavantaj ar putea fi faptul că numărul de fotografii nu este același pentru fiecare tip de emoție.

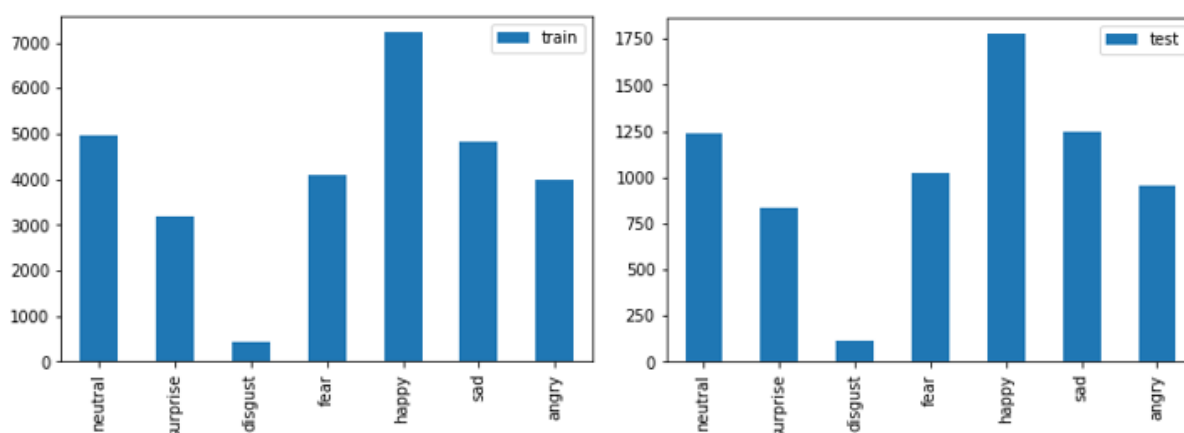


Figura 5. Ilustrarea numărului de imagini din fiecare tip de emoție

După cum putem observa pe baza graficelor din Figura 5., atât în cazul datelor de instruire – reprezentate în diagrama din partea stângă, cât și în cele pentru testare – diagrama din partea dreaptă, cele mai multe imagini ilustrează starea de fericire ("happy"), fiind în număr de 7000 de poze în datele de antrenare și respectiv, 1750 în cele de testare. De asemenea, se poate constata că starea de dezgust ("disgust") este reflectată de cele mai puține ori, sub 1000 de imagini pentru antrenament și sub 250 pentru testare. Restul emoțiilor fiind reprezentate la un nivel destul de egal în ceea ce privește numărul imaginilor. Acest lucru poate crea un dezavantaj destul de mare în procesul de învățare al modelului a acestui tip de emoție, după cum menționam și în partea de introducere.

Figura de mai jos ilustrează câte un exemplu de fotografie aleasă în mod aleatoriu, din fiecare tip de emoție din setul de date. Astfel, putem vedea că persoanele din imagini sunt reprezentate în diferite unghiuri, deci nu sunt neapărat toate fotografiile făcute din față, și în plus, unele pot chiar să conțină și alte obiecte sau înfățișări, cum ar fi de exemplu batista sau mâna, în cazul stării de supărare.



Figura 6. Imagini aleatorii din setul de date reprezentând fiecare emoție (Kaggle - FER2013 data set)

2.1.2 Set de date compus

După crearea, învățarea și testarea modelelor pe primul set de date, s-a mai ales un set de date care este alcătuit din trei seturi de date diferite, printre care și FER2013. Celelalte două seturi de date fiind CK+ și KDEF. Motivul alegerii încă unui set de date, chiar dacă acesta cuprinde și setul folosit inițial, a fost de a vedea dacă mărimea și diversitatea din cadrul imaginilor poate fi un factor care să ajute la performanța modelelor.

KDEF cuprinde în total 4900 de imagini ale expresiilor faciale umane, de la 70 de persoane, 35 de femei și 35 de bărbați, care afișează cele 7 emoții de bază, privite din 5 unghiuri diferite. Criteriile de selecție a indivizilor a fost ca aceștia să aibă vârsta cuprinsă între 20 și 30 de ani, fără barbă, mustață, cercei sau ochelari de vedere, și de preferință fără machiaj.

CK+ constă din fotografii obținute prin înregistrarea a 210 adulți folosind 2 camere Panasonic AG-7500 sincronizate hardware. Participanții au avut vârsta cuprinsă între 18 și 50 de ani și au fost de diferite etnii. Aceștia au efectuat 23 de fețe, iar fiecare afișaj a început și s-a încheiat cu o față neutră.

Setul extins este împărțit în date de instruire, date de testare și validare. Am ales să combinăm datele de validare cu cele datele de testare pentru a crește numărul acestora, dat fiind faptul că și pe setul de date FER2013 am avut doar date de instruire și testare. Acestea sunt de asemenea împărțite în fișiere diferite pentru fiecare emoție în parte. Așadar, ajungem la un număr de 31293 de imagini pentru datele de validare și de 10436 pentru datele de testare. Acest set de date combinat poate fi descărcat tot de pe Kaggle, de la adresa (Kaggle - Emotion-compilation data set).

Prin urmare, și cu acest set de date există anumite diferențe între numărul fotografiilor care exprimă o emoție, de exemplu, în acest caz imaginile care exprimă starea neutră sunt mult mai numeroase decât toate celelalte, ajungând până peste 10.000 de imagini pentru partea de antrenament și peste 3500 pentru testare. Cu toate că scopul folosirii acestui set combinat a fost pentru a avea un număr mai crescut de imagini care să prezinte starea de dezgust, numărul acestora rămâne în continuare sub 1000 de fotografii în setul de instruire, respectiv sub 250 în cazul setului pentru testare.

În Figura 7. sunt afișate imagini reprezentând expresia facială a fiecărei emoții alese în mod aleatoriu din al doilea set de date. Se pot observa de altfel, diferențele în ceea ce constau cadrele, culorile și dimensiunile fotografiilor față de imaginile din primul set de date. După descrierea setului de date KDEF, ar părea că imaginile care ilustrează starea de furie și dezgust sunt din cadrul acestuia.

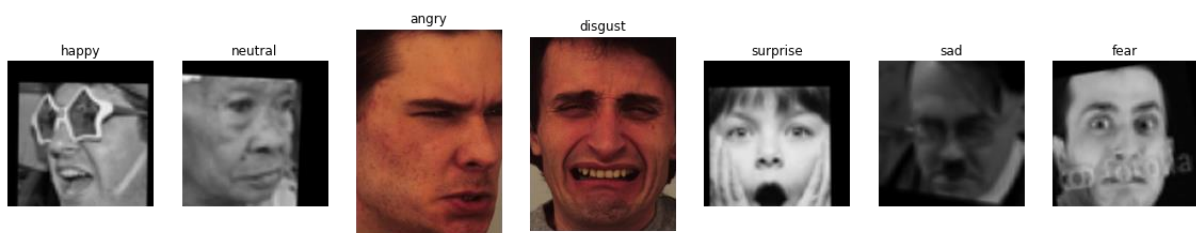


Figura 7. Imagini aleatorii reprezentând fiecare emoție din setul de date compus (Kaggle - Emotion-compilation data set)

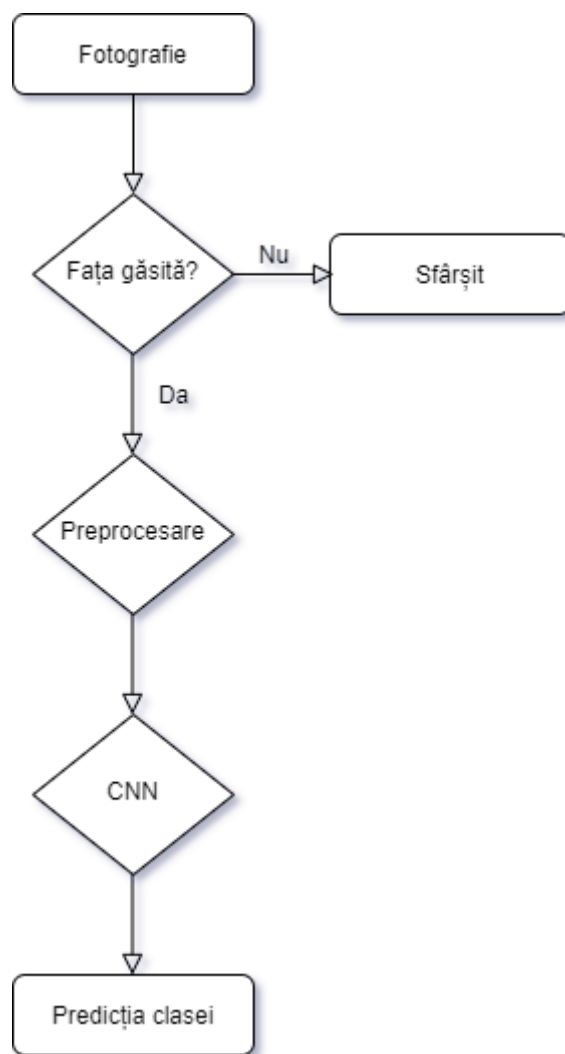
2.2 Arhitectura sistemului și a rețelelor neuronale convoluționale

Sistemul general dezvoltat în această lucrare este considerat format din două părți: o parte care reprezintă crearea și antrenarea rețelelor, și o parte care constă dintr-o aplicație web, dezvoltată cu scopul de a testa modelele atât în timp real, cât și prin încărcarea de fotografii.

Structura aplicației web este simplistă și compusă din:

- directorul "models" care conține toate modelele și clasificatorul folosit pentru detectarea feței;
- directorul "templates" unde se găsesc patru șabloane în format ".html" care prezintă paginile dezvoltate;
- un director "uploads" în care se stochează fotografiile încărcate;
- patru fișiere de script care includ funcționalitatea, respectiv funcțiile și importurile care sunt necesare pentru rularea aplicației, mai exact pentru realizarea predicției emoției în ambele modalități.

Pe lângă sistemul general descris mai sus, se consideră tot un sistem, modul în care se vor realiza predicțiile în ceea ce privește emoția clasată din expresia facială regăsită în datele de intrare. Diagrama fluxului de evenimente al acestui sistem este ilustrată în figura următoare. După cum se poate observa, la primul pas, sistemul preia o imagine și detectează fața din fotografie cu ajutorul unui clasificator. În continuare, dacă fața este găsită, imaginea este trimisă spre preprocesare, în caz contrar se va afișa mesajul "Nu s-a detectat o față." și procesul se încheie. În cele din urmă, imaginea augmentată este introdusă în CNN pentru a face o predicție despre clasa recunoscută, reprezentând emoția.



Figură 8. Diagrama fluxului de evenimente

Pe parcursul acestei cercetări au fost create și testate diferite arhitecturi de rețele neuronale convoluționale, cu scopul de a găsi unul care să rezulte cea mai bună performanță. În principal, diferențele dintre acestea constau în numărul straturilor, filtrelor și nodurilor de convoluție, straturilor complet conectate, straturilor de dropout, parametrilor.

În cele ce urmează, se vor prezenta două modele secvențiale încercate care au obținut, comparativ cu celelalte, cele mai bune performanțe, și care, de altfel, vor fi folosite și în aplicația dezvoltată cu scopul de a putea testa aceste modele atât în timp real, cât și prin încărcarea de fotografii.

Primul model care este folosit pentru a clasifica expresiile faciale conține câte două straturi de convoluție, care sunt urmate de câte un strat de max-pooling, de mărimea 2x2. Acest strat de

max-pooling calculează valoarea maximă pentru fiecare zonă din "harta" caracteristicilor. Fiecare strat conține un număr de noduri care crește exponențial, începând cu 16, care reprezintă 2^4 și continuând până la ultimul strat de convoluție, având 512 noduri, care este de asemenea singurul care nu are pereche. Prin urmare, în total sunt 7 straturi de convoluție cu dimensiunea nucleului de 3×3 și numărul de noduri diferit.

Funcția de activare utilizată în stratul de convoluție este "ReLU", aceasta reprezentând o funcție liniară care va emite direct intrarea dacă este pozitivă, iar în caz contrar va returna 0. A devenit o funcție implicită de activare pentru multe tipuri de rețele neuronale, deoarece un model care o folosește este mai ușor de antrenat și de cele mai multe ori obține performanțe mai bune. Acestea fiind motivele pentru care este folosită și în acest model.

După cum spuneam și la descrierea setului de date folosit, imaginile furnizate către model sunt de dimensiunea 48×48 . Forma de intrare a modelului fiind astfel $(48, 48, 1)$, unde 1 se referă la numărul de canale existente în imaginile de intrare deoarece acestea sunt în tonuri de gri.

Urmează apoi, un strat complet conectat care constă din 512 noduri. La fel ca și la straturile de convoluție, funcția de activare "ReLU" a fost aplicată și pentru stratul ascuns. Iar după acest strat, a fost inserat un strat "dropout" cu valoarea de 0,5, ceea ce înseamnă că în mod aleator se dezactivează 50% dintre noduri din stratul ascuns cu scopul de a evita supra-adaptarea. În cele din urmă, stratul de ieșire al modelului este format din 7 noduri, deoarece avem cele 7 clase de emoții. Ca și funcție de activare în stratul de ieșire a fost folosită funcția "Softmax". După toate aceste configurări au reușit un număr total de parametri de 2,772,74, care de asemenea reprezintă și numărul de parametri antrenabili.

Optimizatorul de model folosit este "Adam", cu o rată de învățare în mod implicită de 0.001. Această metodă se consideră a fi "eficientă din punct de vedere al calcului, are puține cerințe de memorie, invariantă pentru redimensionarea diagonală a gradientilor și este foarte potrivită pentru problemele mari în ceea ce privește datele sau parametrii". (Keras API reference / Optimizers / Adam)

În plus, se mai folosește o funcție de pierdere care este utilizată în sarcini de clasificare multi-clasă, numită "loss='categorical_crossentropy'". Este potrivită pentru acest model deoarece o imagine poate aparține doar unei categorii de emoții, iar modelul trebuie să decidă din care.

Scopul celui de-al doilea model a fost ca numărul de parametri să fie crescut, astfel încât să se testeze dacă această diferență ajută sau nu la performanța modelului. Așadar, acest model are un număr total de 32,115,463 parametri, care de asemenea reprezintă și parametri antrenabili. Modelul este creat din patru straturi de convoluție, grupate câte două și urmate de câte un strat de max-pooling de 2x2. Numărul de filtre folosite sunt, încep de la 32, până la 256, iar dimensiunea nucleului este de 3x3. Urmează apoi un strat complet conectat care conține 1024 de noduri și un strat "dropout" cu valoare de 0,5. De asemenea, au fost folosite aceleași funcții de activare, " ReLu" și "Softmax", cât și funcția de optimizare "Adam".

O privire de ansamblu asupra arhitecturilor celor două rețele neuronale convoluționale care au fost concepute pentru această lucrare este prezentată în tabelul de mai jos, Model 1 reprezentând primul model prezentat, iar Model 2 cel de-al doilea model prezentat.

Tabel 1. Arhitectura rețelelor neuronale convoluționale

Configurări	Detalii – Model 1	Detalii – Model 2
I strat de convoluție	16 filtre, 3x3, ReLu	32 filtre, 3x3, ReLu
Al II-lea strat de convoluție	32 filtre	64 filtre
I strat de max-pooling	2x2	2x2
Al III-lea strat de convoluție	64 filtre	128 filtre
Al IV-lea strat de convoluție	128 filtre	256 filtre
Al II-lea strat de max-pooling	2x2	2x2
Al V-lea strat de convoluție	128 filtre	-
Al VI-lea strat de convoluție	256 filtre	-
Al III-lea strat de max-pooling	2x2	-
Al VII-lea strat de convoluție	512 filtre	-
Al IV-lea strat de max-pooling	2x2	-
I strat conectat	512 noduri, ReLu	1024 noduri, ReLu
I strat de "dropout"	0.5	0.5
Stratul de ieșire	7 clase, SoftMax	7 clase, SoftMax
Funcția de optimizare	Adam	Adam
Număr parametri	2,772,743	32,115,463

În plus, pentru optimizarea modelelor au fost folosite câteva funcții de "callback" de la Keras, cu ajutorul cărora se poate vizualiza performanța antrenamentului modelului, se poate preveni supra-adaptarea. Acest set de funcții se aplică la anumite etape ale procedurii de instruire.

În cazul ambelor rețele neuronale convoluționale s-au aplicat următoarele funcții de callback:

- "ModelCheckpoint" – pentru a salva modelul ca un fișier cu scopul de punct de control (în format "h5") după fiecare epocă reușită. Aceste fișiere vor fi folosite apoi în aplicația web și pe baza lor se vor face predicțiile.
- "EarlyStopping" – pentru a reduce supra-adaptarea prin încheierea procesului de instruire dacă nu se vede niciun progres de învățare. S-a ales monitorizarea valorii de pierdere și permitem antrenamentului să continue să ruleze 3 epoci fără modificări înainte de oprire.
- "ReduceLROnPlateau" – asemănătoare cu ce descrisă mai sus, dar cu scopul de a reduce rata de învățare atunci când valoarea nu se schimbă după un număr de 6 epoci, valoare setată pentru aceste modele. De exemplu, "*Epoch 00086: ReduceLROnPlateau reducing learning rate to 0.000100000000474974513.*" arată că după epoca numărul 86 rata de învățare a fost actualizată.

2.3 Implementarea sistemului

Considerăm sistemul ca fiind alcătuit din două părți: o parte care s-a ocupat de crearea și antrenarea rețelelor neuronale convoluționale, și a doua parte care constă în aplicația web dezvoltată cu scopul de a testa modelul.

Bibliotecile necesare pentru această cercetare sunt Keras, TensorFlow, numpy, OpenCV și matplotlib. TensorFlow a fost folosit ca și parte de backend, în timp ce Keras a oferit funcțiile de activare, optimizatorul, straturile convoluționale, etc., iar matplotlib a fost folosit pentru a genera matricile de confuzie și anumite grafice. Toate acestea au fost importate și instalate în ambele părți ale sistemului.

Prima parte a fost realizată folosind "Google Colab", un mediu de dezvoltare web pentru Python oferit de Google, care de altfel oferă și posibilitatea de a folosi GPU, unități de procesare grafică, ajutând astfel ca antrenarea modelului să se efectueze mult mai rapid. Această parte a sistemului conține operațiile necesare pentru încărcarea și vizualizarea datelor, preprocesarea acestora, crearea modelelor și antrenarea lor, și obținerea rezultatelor.

Inițial datele au fost încărcate în Google Drive sub forma a două arhive, una care reprezintă datele de antrenare, iar cealaltă datele de testare, putând apoi să fie extrase într-un fișier

temporar în Google Colab. După extragerea acestora, au fost create câteva figuri care să ilustreze anumite aspecte despre date, precum:

- Numărul imaginilor care reprezintă fiecare emoție, din ambele categorii de date, reprezentat sub forma unui tabel. De exemplu, există 4097 de fotografii care să exprime frica în datele de antrenare și doar 436 care să exprime dezgustul. În ceea ce privește datele de testare, se regăsesc 1233 pentru surprindere și 958 pentru furie.
- Diferențele dintre acestea în grafice de bare, ca cele expuse la secțiunea "2.1 Setul de date".
- Câte o fotografie din fiecare tip de emoție aleasă în mod aleatoriu.

Keras API facilitează preprocesarea datelor prin introducerea funcției "ImageDataGenerator" care prin mai multe operații poate fi aplicată pe setul de date existent pentru a genera mai multe date noi, cu scopul de a îmbunătăți învățarea profundă a modelului în ceea ce privește clasificarea imaginilor. Parametrii folosiți pentru această funcție sunt mărirea imaginilor, întoarcerea lor în mod orizontal și redimensionare. Tabelul de mai jos reprezintă parametrii folosiți cu valorile aferente.

Tabel 2. Parametrii folosiți pentru funcția de "ImageDataGenerator"

TIPUL DE PARAMETRU	VALOAREA
Mărirea imaginilor – "zoom_range"	0.3
Întoarcerea orizontală – "horizontal_flip"	True
Redimensionare – "rescale"	1./255

Urmează apoi să se implementeze, pe rând, fiecare model descris în subcapitolul precedent. După definirea acestora, se realizează și un rezumat care să reflecte toate straturile folosite și se calculează astfel numărul de parametri. Apoi, după declararea funcțiilor de "callback", începe antrenarea modelelor și obținerea rezultatelor.

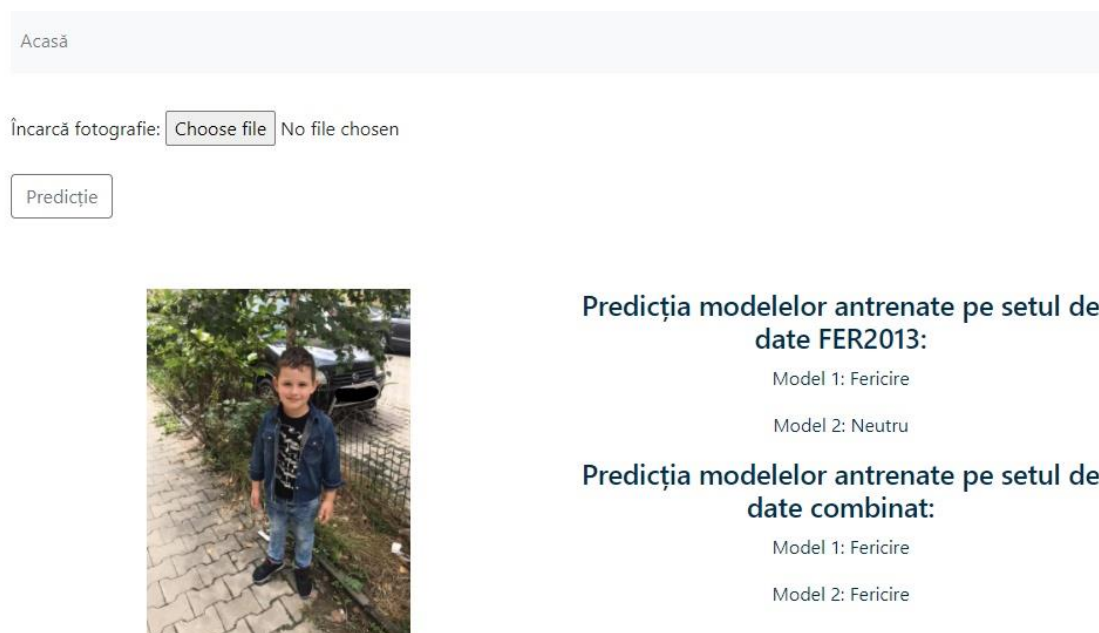
Pentru a putea folosi modelele în aplicația web fără a mai fi nevoie să le antrenăm din nou, la finalizarea acestora, fiecare se salvează sub forma unui fișier cu extensia ".h5" care cuprinde toate informațiile necesare: arhitectura, detalii despre compilare, "greutățile" modelului.

Pe baza rezultatelor, se creează mai întâi, un grafic care să reflecte creșterea acurateței modelului, atât în cazul datelor de antrenare, cât și pe cele de testare, în concordanță cu numărul de epoci. Urmând apoi să se compună matricea de confuzie care să ajute la vizualizarea

performanțelor rețelelor, și mai exact să se poată observa nivelul predicției fiecărei emoții în parte.

Aplicația web este scrisă, de asemenea, în limbajul de programare Python, folosindu-se framework-ul web, Flask. Aceasta oferă o interfață grafică creată cu ajutorul tehnologiilor HTML și CSS. Interfața este destul de simplă, pe pagina de "Acasă" sunt prezentate cele două metode de testare, acestea putând fiind accesate prin apăsarea unui buton.

În cazul alegerii de a încărca o fotografie din calculatorul personal, imaginea este afișată în partea dreapta, iar în stânga se vor afișa predicțiile celor două rețele antrenate atât pe setul de date FER2013, cât și pe cel extins, așadar, vor fi prezentate patru clasificări.



Acasă

Încarcă fotografie: Choose file No file chosen

Predicție

Predicția modelelor antrenate pe setul de date FER2013:

Model 1: Fericire

Model 2: Neutru

Predicția modelelor antrenate pe setul de date combinat:

Model 1: Fericire

Model 2: Fericire

Figură 9. Exemplu de predicție prin încărcare de fotografie folosind aplicația web

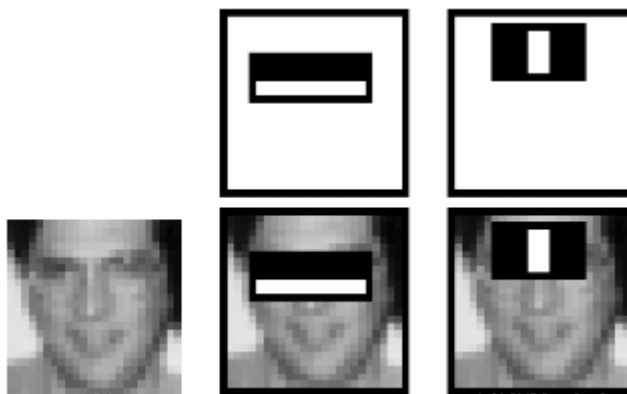
În figura de mai sus este ilustrat un exemplu de predicție a modelelor și afișarea rezultatelor acestora asupra unei fotografii încărcate din calculator. După cum se poate vedea, ambele modelele, modelul cu 7 straturi de convoluție și 2 milioane de parametrii reprezentat prin

”Model 1”, și respectiv, modelul cu 4 straturi de convoluție și 32 milioane de parametrii simbolizat prin ”Model 2”, în cazul antrenării pe setul de date combinat, clasifică emoția ca fiind cea de ”Fericire” reprezentând o predicție corectă. Modelul cu 4 straturi de convoluție și 32 milioane de parametrii antrenat pe setul de date FER2013 are ca rezultat clasa ”Neutru”.

Pentru realizarea clasificărilor, imaginea va fi mai întâi convertită în tonuri de gri, acest lucru este necesar deoarece rețelele au fost antrenate pe astfel de fotografii. Apoi, se va detecta fața cu ajutorul clasificatorului ”Haar Cascade Frontal Face” (GitHub - OpenCV - Haar Cascade), oferit de librăria OpenCV, sub forma unui fișier XML, care poate fi găsit pe GitHub.

Teoretic, acesta reprezintă un algoritm care folosește anumite caracteristici de detectare a muchiilor sau liniilor care au fost propuse de către Paul Viola și Michael Jones în lucrarea ”Rapid object detection using a boosted cascade of simple features” (Viola, P., & Jones, M., 2001). Este o abordare bazată pe învățarea automată, în care se antrenează un model pe o mulțime de imagini pozitive, în cazul celui menționat, acestea reprezentând imagini ale fețelor, și imagini negative, imagini fără fețe, fiind extrase anumite caracteristici. Pe baza acestora, clasificatorul învață să detecteze o față.

Fiecare caracteristică reprezintă o singură valoare obținută prin scăderea sumei pixelilor sub dreptunghiul alb din suma pixelilor de sub dreptunghiul negru. Un exemplu poate fi observat în figura de mai jos. Rândul de sus prezintă două caracteristici bune, prima pare să se concentreze asupra proprietății că regiunea ochilor este adesea mai întunecată decât cea a nasului și obrazilor, iar a doua se bazează că ochii sunt mai întunecați decât puntea nasului. (OpenCv docs)



Figură 10. Exemplu de caracteristici ”Haar Cascade” (OpenCv docs)

Cu ajutorul funcției ”detectMultiScale” existentă în clasificator se va face detectarea feței pe imaginea primită. Dacă nu a fost găsită niciuna atunci se va returna fals și se va afișa mesajul ”Nu s-a găsit o față.”, dar dacă aceasta a fost găsită, imaginea va fi redimensionată în 48x48 pixeli și se va continua cu predicția modelelor și afișarea rezultatelor.

În ceea ce privește testarea în timp real, în cadrul interfeței, va apărea un ecran care va reflecta imaginea primită prin intermediul camerei web. Se vor efectua capturi la fiecare 2 secunde, iar acestea vor fi preprocesate în același mod ca și în cazul fotografiilor încărcate sau atunci când rețeaua a fost instruită. În cadrul acestei modalități, clasificarea se va face însă doar de către rețeaua care a atins cea mai bună acuratețe, din această cercetare.

Clasa detectată va fi afișată pe ecran în timp real, deasupra unui dreptunghi albastru care evidențiază zona feței detectate, precum exemplele din figura următoare, unde se ilustrează 6 din cele 7 emoții: furie în prima pictogramă, urmată de fericire, supărare, frică, surprindere și starea neutră. De asemenea se exemplifică și cazul în care modelul nu detectează o față umană și se afișează un mesaj aferent.

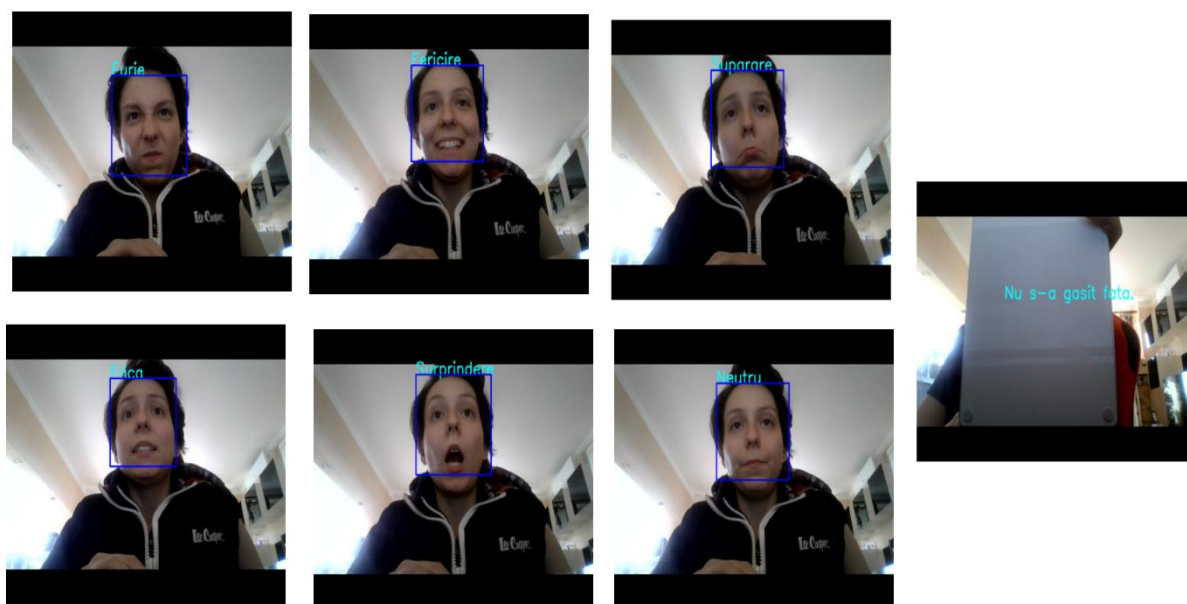


Figura 11. Capturi de predicții în timp real prin camera web

3. Rezultate

În acest capitol se vor prezenta rezultatele obținute și se va observa că diferențele ce țin de acuratețea modelelor sunt destul de variate în funcție de setul de date folosit și de numărul emoțiilor din care este alcătuit.

Indiferent de setul de date folosit, cât și de numărul emoțiilor alese, au fost folosite aceleași rețele neuronale convoluționale descrise în capitolul anterior, la care o să se facă referire prin ”model cu 7 straturi de convoluție și 2 milioane de parametrii”, reprezentând primul model prezentat, și respectiv, ”model cu 4 straturi de convoluție și 32 milioane de parametrii”.

Prin urmare, în ceea ce privește acuratețea modelelor pe primul set de date ales, FER2013, s-a obținut aproximativ același rezultat: 57,2% primul model, și respectiv, 57,38% cel de-al doilea. În urma acestor rezultate, putem spune că, cel puțin în cazul celor două rețele, numărul parametrilor nu reprezintă un avantaj sau dezavantaj asupra performanței.

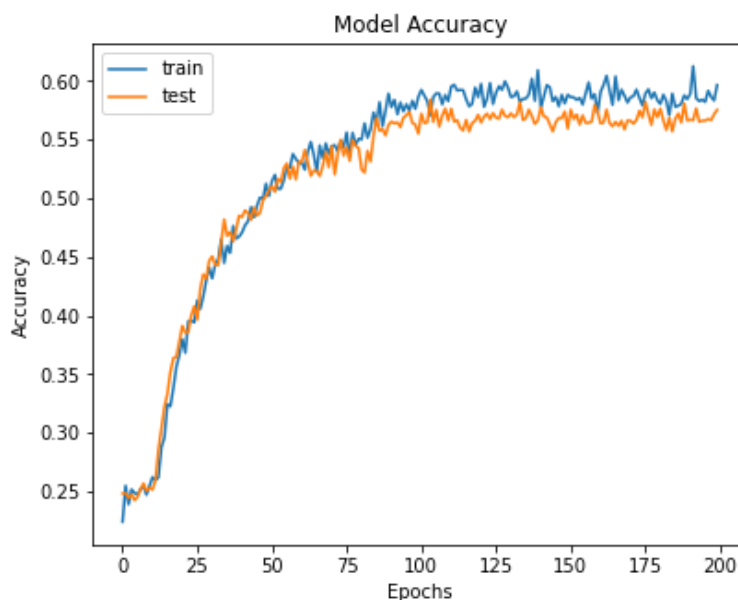


Figura 12. Progresul acurateții model cu 7 straturi de convoluție și 2 milioane de parametrii

Figura de mai sus ilustrează progresul acurateții obținută pe datele de antrenare, înfățișată prin culoarea albastru și de testare, simbolizat cu portocaliu, în cazul primei rețele antrenate. Axa X reprezentând numărul de epoci, iar axa Y figurând rata de recunoaștere. Evoluția crește o dată cu atingerea unui număr mai mare de epoci, reușind să atingă un procent în jur de 57%.

Nu putem spune că acest rezultat este unul performant, având un număr considerabil de alte performanțe cu care să se poată compara. În primul rând, dacă luăm în considerare doar rezultatele menționate în capitolul 1 de studiu bibliografic, se poate observa că rețelele dezvoltate în această lucrare au obținut cea mai mică acuratețe dintre toate celelalte prezentate.

În Tabel 3. se poate vedea o clasificare descendentă din punctul de vedere al acurateței obținute, compusă din rezultatele cercetărilor studiate, care au antrenat modelele tot pe setul de date FER2013, și cele dobândite. Pe primul loc se clasează rețeaua AlexNet, urmată la o mică diferență de GoogLeNet, ambele fiind descrise în aceeași lucrare. Prin urmare, există o deosebire considerabilă între performanțele modelelor prezentate în această lucrare față de celelalte rețele, aproximativ între 18 – 28%.

Tabel 3. Clasificarea performanțelor rețelelor obținute pe setul de date FER2013

Lucrare	Acuratețea obținută
(Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I., 2018) - AlexNet	85%
(Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I., 2018) - GoogLeNet	84%
(Li, J., Jin, K., Zhou, D., Kubota, N., & Ju, Z., 2020)	75.82%
Model cu 4 straturi de convoluție și 32 milioane de parametri	57.38%
Model cu 7 straturi de convoluție și 2 milioane de parametri	57.2 %

În schimb, dacă facem o comparație cu clasamentul realizat în cadrul competiției organizată de Kaggle, în anul 2013, la care au participat 56 de echipe, iar acolo a fost folosit setul de date FER2013, rețelele s-ar afla pe locul 17, modelul cu 4 straturi de convoluție și 32 milioane de parametri, respectiv locul 18, modelul cu 7 straturi de convoluție și 2 milioane de parametri. Echipa câștigătoare a obținut o performanță de 71%, în timp ce cea mai jos clasată de 20%. Restul rezultatelor variind între aceste valori.

Figura 13. ilustrează partea de clasament unde s-ar plasa rețelele dezvoltate în această cercetare, indicat prin linia roșie, iar prin culoarea galbenă s-a pus în evidență locurile și rezultatele între care s-ar clasa acestea. Locul 16 reușind o performanță de 57.59%, iar locul 17 de 56.81%. Considerând aceste rezultate s-ar putea spune că cele 2 modele nu au realizat o acuratețe chiar așa de scăzută, dar în același timp trebuie luat în considerare faptul că această competiție a avut loc acum 8 ani, și de atunci studiile privind acest subiect au avansat într-un mod accelerat. Captura de ecran a fost editată pentru a se putea trasa aceste modificări relatate mai sus.







15	▲ 1	12AngryBird		0.57899
16	▼ 1	shiggles		0.57592
17	▲ 1	multiboost		0.56812
18	▼ 1	Pikqu		0.56199
19	—	Anil Thomas		0.55363
20	—	dova		0.54444

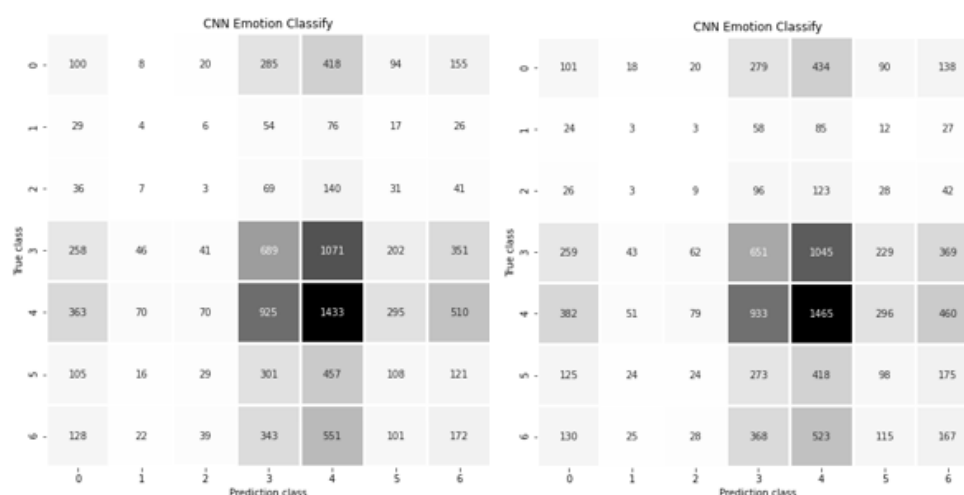
Figura 13. Clasament competiție Kaggle (Kaggle - Facial Expression Recognition Challenge - Leaderboard)

O altă abordare încercată cu scopul de a obține o performanță a rețelelor mai bună, a fost de a scoate din setul de date fotografiile care exprimă emoția de dezgust, în ideea că celelalte emoții ar putea fi mai bine învățate de către rețele. În plus, această stare era redată, atât în cazul datelor de instruire, cât și în cazul celor de testare, de cele mai puține imagini, cu diferențe sesizabile față de celelalte emoții. Prin urmare, nici această modalitate nu a rezultat o performanță mai bună a modelelor, obținându-se o acuratețe cu doar 1% mai performantă în cazul rețelei cu 7 straturi de convoluție și 2 milioane de parametrii, de 58.43% și cu aproape 2% mai scăzută în cazul celei cu 4 straturi de convoluție și 32 milioane de parametrii, de 55.79%.

Pe baza unei analize asupra matricelor de confuzie a rețelei cu 4 straturi de convoluție și 32 milioane de parametrii, obținute după antrenarea și testarea pe setul de date FER2013 cu toate cele 7 emoții și doar cu cele 6 emoții, lipsind starea de dezgust, nu se poate spune că ar exista totuși o emoție pentru care modelul să ofere o mai bună predicție comparativ cu cele rezultate în cazul antrenării și testării pe toate cele șapte clase. Preciziile sunt destul de apropiate, de exemplu, în ceea ce privește recunoașterea emoției de frică, în cazul antrenării pe toate stările, a fost identificată de 90 de ori, iar în cazul instrucției doar pe cele 6 emoții de 82 de ori.

O mică îmbunătățire există asupra claselor de furie, fericire, stare neutră și supărare, dar diferențele sunt atât de mici încât nu se pot considera a fi semnificative. De exemplu, emoția de supărare a fost identificată de 258 ori în cazul antrenării pe setul de date care conține doar 6 emoții, în timp ce, iar în cazul instrucției pe setul de date cu toate emoțiile s-a detectat de 247 de ori.

Antrenarea rețelelor neuronale convoluționale pe setul de date compus rezultă o mai bună acuratețe obținută din partea ambelor modele: 80% modelul cu 7 straturi de convoluție și 2 milioane de parametri și 82.87% model cu 4 straturi de convoluție și 32 milioane de parametri. Această performanță oferă concluzia că un set de date mai dezvoltat, care să conțină atât un număr mai mare de imagini, cât și cadre din diferite unghiuri, ajută la procesul de învățare a unei CNN.



Figură 14. Matricele de confuzie a modelelor antrenate pe setul de date compus

În figura de mai sus sunt ilustrate matricele de confuzie a celor două modele, model cu 7 straturi de convoluție și 2 milioane de parametri în partea stângă, respectiv model cu 4 straturi de convoluție și 32 milioane de parametri în partea dreaptă, antrenate pe al doilea set de date, unde axa X reprezintă clasa de predicție a modelului și axa Y clasa adevărată. Numerele de la 0-6 reprezentând fiecare emoție în parte, după cum urmează: 0 – furie, 1 – dezgust, 2 – frică, 3 – fericire, 4 – stare neutră, 5 – supărare, 6 – surprindere.

Privind diagonala matricelor, unde se intersectează valorile reprezentative pentru clasa recunoscută împreună cu clasa adevărată, se poate observa că, în cazul ambelor rețele cea mai cunoscută emoție este cea de neutru, fiind zona cea mai închisă la culoare din matrice, cu o valoare în jur de 1400, urmată de fericire, puțin peste 600 de detectări corecte. Starea de dezgust și frică fiind în continuare cele mai puțin identificate, mai puțin de 10 în ambele cazuri.

De asemenea, se poate vedea faptul că starea neutră este adesea încurcată cu starea de fericire, și invers, fiind de mai multe ori identificată emoția de fericire, când clasa adevărată era starea neutră, decât atunci când emoția reprezentată era chiar aceasta. În același mod fiind și în cazul invers, când fericirea era cea ilustrată, dar predicția a fost de stare neutră. De aici s-ar putea

trage concluzia că aceste două emoții diferite sunt reprezentate destul de asemănător în setul de date, și din acest motiv pot fi atât de ușor confundate. Încă o dovadă că setul de date are o importanță extrem de mare în ceea ce privește antrenarea unei rețele neuronale convoluționale.

În lucrările studiate nu s-au găsit cercetări care să antreneze sau să testeze rețelele neuronale convoluționale pe acest set de date compus din FER2013, CK+ și KDEF, dar totuși există anumite rezultate obținute pe un set de date creat din alăturarea mai multor astfel de date, descrise în tabelul de mai jos.

Tabel 4. Clasament privind rezultatele obținute pe seturi de date compuse

Lucrare – set de date – Arhitectura CNN	Acuratețea obținută
(Ozdemir, M. A., Elagoz, B., Alaybeyoglu, A., Sadighzadeh, R., & Akan, A., 2019) – JAFFE, KDEF, set de date personalizat - LeNet	96.43%
(Melinte, D. O., & Vladareanu, L., 2020) - FER2013, CK+, JAFFE, KDEF - ResNet	90.14%
(Melinte, D. O., & Vladareanu, L., 2020) - VGG	87%
Model cu 4 straturi de convoluție și 32 milioane de parametrii	82.87%
(Melinte, D. O., & Vladareanu, L., 2020) – InceptionV3	81%
Model cu 7 straturi de convoluție și 2 milioane de parametrii	80 %

Comparativ cu aceste performanțe, modelul cu 4 straturi de convoluție și 32 milioane de parametrii se clasează pe locul 3, la o diferență de aproape 4% față de locul din față, și respectiv de aproximativ 13% față de primul loc. În timp ce, modelul cu 7 straturi de convoluție și 2 milioane de parametrii se situează ultimul din acest clasament creat. Totuși, diferențele dintre acuratețea obținută în rețelele din lucrările studiate și modele prezentate în această lucrare și antrenate pe acest set de date compus sunt mult mai diminuate față de cele realizate în cazul folosirii setului de date FER2013.

Concluzii

În această lucrare de cercetare, obiectivul a fost să se studieze diferite abordări care să ofere îmbunătățiri în ceea ce privește recunoașterea celor șapte emoții de bază (fericire, supărare, furie, frică, surprindere, dezgust și o stare neutră) pe baza expresiilor faciale a unei persoane cu ajutorul unei rețele neuronale convoluționale.

Am ajuns la concluzia că un set de date de dimensiuni mai mari, care să conțină de altfel și imagini abordate din diferite posturi, de diferite culori și mărimi, ajută la obținerea unei performanțe mai bune a rețelelor, cu o diferență de mai mult de 20%. Prin această abordare s-a ajuns la acuratețea de aproape 83% obținută de modelul cu 4 straturi de convoluție și 32 milioane de parametri, și respectiv 80% în cazul modelului cu 7 straturi de convoluție și 2 milioane de parametri, în timp ce progresul maxim în cazul instruirii pe setul de date mai redus, FER2013, a fost de 57%.

O bază de date care să cuprindă un număr impresionant de imagini reprezentând fiecare emoție, de la indivizi de diferite etnii și vârste, cu cadre din diverse unghiuri și ipostaze, ar avea un mare impact asupra performanțelor modelelor ce urmăresc învățarea profundă pentru recunoașterea emoțiilor. Adăugarea mai multor date în fiecare clasă ajută la obținerea unor rezultate mai precise mai ales datorită faptului că învățarea profundă este o abordare bazată pe date.

În viitor se dorește dezvoltarea rețelelor neuronale convoluționale prezentate în această lucrare astfel încât acestea să realizeze o predicție cât mai bună asupra oricărui tip de emoție, din cele șapte de bază. Scopul acestei optimizări este ca modelele să poată fi folosite în cadrul unei aplicații mobile destinate monitorizării și analizei emoțiilor pe care le au, în special copiii, în timpul petrecut pe telefon. Motivația pentru dezvoltarea acestui sistem a pornit de la anumite statistici cu privire la creșterea depresiei în rândul copiilor sub 12 ani datorită utilizării excesive a telefoanelor și petrecerea a prea puțin timp cu familia sau prietenii în mediul real. Astfel, o asemenea aplicație ar putea ajuta, mai ales, părinții să vadă prin ce stări trece copilul său în funcție de aplicația pe care o folosește și astfel să ia anumite decizii, în cazul în care statisticile oferite prezintă îngrijorări.

Bibliografie

- Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I. (2018). Deep learning approaches for facial emotion recognition: A case study on FER-2013. În V. P. Ioannis Hatzilygeroudis, *Advances in Hybridization of Intelligent Methods*. Springer International Publishing AG.
- GitHub - OpenCV - Haar Cascade. (fără an). Preluat de pe <https://github.com/opencv/opencv/blob/master/data/haarcascades/>
- Hussain, S. A., & Al Balushi, A. S. A. (2020). A real time face emotion classification and recognition using deep learning model. *Journal of Physics: Conference Series*.
- Kaggle - Facial Expression Recognition Challenge - Leaderboard. (fără an). Preluat de pe <https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/leaderboard>
- Kaggle - Emotion-compilation data set. (fără an). Preluat de pe <https://www.kaggle.com/qnkhua/emotion-compilation>
- Kaggle - FER2013 data set. (fără an). Preluat de pe <https://www.kaggle.com/msambare/fer2013>
- Keras API reference / Optimizers / Adam. (fără an). Preluat de pe <https://keras.io/api/optimizers/adam/>
- Li, J., Jin, K., Zhou, D., Kubota, N., & Ju, Z. (2020). Attention mechanism-based CNN for facial expression recognition. *Neurocomputing*.
- Melinte, D. O., & Vladareanu, L. (2020). Facial Expressions Recognition for Human–Robot Interaction Using Deep Convolutional Neural Networks with Rectified Adam Optimizer. *Sensors*, 20(8).
- OpenCv docs. (fără an). Preluat de pe https://docs.opencv.org/3.4/db/d28/tutorial_cascade_classifier.html
- Ozdemir, M. A., Elagoz, B., Alaybeyoglu, A., Sadighzadeh, R., & Akan, A. (2019). Real time emotion recognition from facial expressions using cnn architecture. 2019 Medical Technologies Congress.
- Rani, J., & Garg, K. (2014). Emotion detection using facial expressions-A review. *International Journal of Advanced Research in Computer Science and Software Engineering*.
- Sajjad, M., Zahir, S., Ullah, A., Akhtar, Z., & Muhammad, K. (2019). Human behavior understanding in big multimedia data using CNN based facial expression recognition. *Mobile networks and applications*, 1-11.
- Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition*.