BABES“ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

**Detectarea expresiilor faciale**

Raport practică

Gorea Octavian, Sisteme Distribuite in Internet, 254

2020

**Abstract**

În această lucrare se prezintă diferite abordări prin intermediul cărora se pot face recunoașterea emoțiilor unei persoane din expresiile faciale ale acesteia. Abordările care sunt folosite sunt folosirea unei mașini cu suport vectorial, unde extragerea de caracteristici a avut loc înainte ca datele să fie trimise către aceasta, și folosirea unei rețele neuronale convoluționale.

În cazul mașinii cu suport vectorial, au fost testate 2 metode diferite de extragere a caracteristicilor, și anume Histograma Orientărilor Gradienților (HOG) și Features from accelerated segment test (FAST).

În cazul rețelei neuronale convoluționale, extragerea caracteristicilor desemnate importante are loc de către rețeaua propriu-zisă.

Rezultatele obținute folosind mașina cu suport vectorial (atât cu extragere de caracteristici folosind HOG dar și cu extragere de caracteristici folosind FAST) au fost comparate atât între ele cât și cu rezultatele obținute folosind rețeaua neuronală convoluțională.

În urma unei analize scurte, am decis ca setul de date folosit să fie “Challenges in Representation Learning: Facial Expression Recognition Challenge”, găsit și sub denumirea de FER-2013.

**Cuprins**

[**1.** **Introducere** 4](#_Toc40119789)

[**2.** **Studiul domeniului problemei** 5](#_Toc40119790)

[**3.** **Cercetare și alegerea datelor** 7](#_Toc40119791)

[**3.1.** **Cercetarea lucrăriilor din domeniu** 7](#_Toc40119792)

[**3.1.1.** **Lucrarea 1** 7](#_Toc40119793)

[**3.1.2.** **Lucrarea 2** 9](#_Toc40119794)

[**3.1.3.** **Lucrarea 3** 11](#_Toc40119795)

[**3.1.4.** **Lucrarea 4** 13](#_Toc40119796)

[**3.1.5.** **Comparație a lucrărilor studiate** 14](#_Toc40119797)

[**3.2.** **Alegerea setului de date** 15](#_Toc40119798)

[**4.** **Implementarea programului software** 17](#_Toc40119799)

[**5.** **Concluzii** 22](#_Toc40119800)

[**Bibliografie** 23](#_Toc40119801)

# **Introducere**

Emoțiile sunt factori importanți în viața de zi cu zi, și mai ales în comunicarea dintre oameni. Ele ne ajută să înțelegem intențiile altor oameni. Emoțiile pot să fie extrase atât din contexte scrise, cât și din contexte auditive, vizuale sau chiar din conversații (care poate să fie o combinație dintre cele enumerate anterior, sau chiar toate la un loc). Se poate observa că în general, oamenii iși exprimă emoțiile prin tonalitatea vocii și intermediul expresilor faciale. Expresiile sunt printre cele mai importante canale folosite în comunicarea ditre persoane, și din această cauză, în ultima perioadă de timp a crescut drastic cercetarea emoțiilor exprimate prin intermediul feței.

Putem spune că problema știintifică pe care această lucrare o abordează este una de clasificare, care, din imagini cu fețele unor persoane trebuie să obțină ca și rezultat emoția pe care aceștia o au în acea poză. Problema de clasificare este una care este aplicată pe foarte multe domenii și constă în categorizare, un proces prin care anumite idei sau obiecte sunt recunoscute, diferențiate și înțelese.

Metoda de rezolvare este una foarte simplă, iar pe scurt constă din pregătirea unui set de date (în cazul acesta poze), detectarea caracteristicilor importante din imagini imagine (în cazul acesta vorbim detalii care se găsesc în fața presoanei) și extragerea acestora, introducerea acestora fie în rețeaua neuronală, fie în mașina cu suport vectorial, și obținerea rezultatului clasificării, și anume emoția.

Scopul acestei lucrări este aceea de a crea un sistem capabil de a detecta emoția pe care o are o anumită persoană, la un anumit moment, pe baza unei poze care conține expresia facială a acesteia.

La final se face o comparație între metodele folosite și rezultatele obținute și oferirea unor observații legate de diferite probleme întâmpinate în timpul creării proiectului.

Scopul acestei lucrări este cercetarea unor algoritmi folosiți pentru rezolvarea problemei propuse și crearea unui program software ușor de folosit capabil de a efectua clasificarea oricarei imagini introduse în ea.

Lucrarea aceasta este alcătuită din 5 capitole, după cum urmează:

Capitolul 1, denumit “Introducere” conține o scurtă introducere legată de problema propusă.

Capitolul 2, cu numele “Studiul domeniului problemei” conține un studiu al domeniului problemei.

Capitolul 3, și anume “Cercetare și pregătirea datelor” conține studiul unor lucrări care încearcă să rezolve aceeași problemă ca și cea din această lucrare. De asemenea în acel capitol se face și alegerea setului ce date ce va fi folosit aici.

Capitolul 4, “Implementarea programului software” se vorbește despre aplicația software creată.

Capitolul 5, “Concluzie” se descriu rezultatele obținute.

# **Studiul domeniului problemei**

Pentru a vorbi de recunoașterea expresiilor faciale, prima dată trebuie să vorbim despre emoții. În primul rând trebuie dată definița emoției. Există mai multe definiții diferite pentru ceea ce este emoția, în cazul acesta am ales definiția în care aceasta este o “Reacție afectivă de intensitate mijlocie și de durată relativ scurtă, însoțită adesea de modificări în activitățile organismului, oglindind atitudinea individului față de realitate” (DEX ’09 (2009)).

În lucrarea [1], autorii au decis că o emoție este una de bază atunci când 2 criterii sunt îndeplinite, și anume, primul este că emoția trebuie să fie discretă, adică este foarte ușor ca ele să fie diferențiate una de alta. Al doilea criteriu este că emoțiile acestea au evoluat prin adaptarea noastră la mediul încojurător.

Emoțiile pot să fie interpretate astfel:

* Pe categorii – se presupune că acestea sunt împărțite pe clase. În cazul acesta sunt doar câteva seturi finite de emoții umane.
* Dimensional – se prepune că acestea există într-un spectru și nu pot fi definite cu exactitate. În cazul acesta unele modele definesc ca existând 2 sau 3 dimensiuni.

Folosirea unui model pe categorii ar duce la crearea unui clasificator. Ceea ce va fi folosit în cazul aceasta.

Cu toate acestea, nu există o regulă anume legată de care emoții să fie clasificate ca emoții de bază, cu diferiți cercetători având fiecare lista lor. Cea mai populară listă este numită “The Big Six”, și a fost folosită de *Ekman P.* în studiul recunoașterii universal a emoților din expresii faciale.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Emoții de bază** | | | | | | |
| “The Big Six” | | | | | | Altele |
| Fericire | Tristețe | Furie | Dezgust | Frică | Uimire | Dispreț |

Tabelul 1. Lista cu emoții de bază creată de Ekman și Cordaro

Procesul de funcționare a recunoașterii expresiilor faciale constă în folosirea unor tehnologii care folosesc diferiți marcatori biometrici pentru a detecta emoțiile din fețele oamenilor. Pentru rezolvarea acestei probleme se folosesc diferiți algoritmi care detectează fețele, apoi, extrag expresiile faciale și într-un final recunosc emoția. Acest lucru este făcut prin analizarea pozelor sau videoclipurilor.

În general se urmăresc 3 pași:

* Detectarea feței – localizarea feței în cadrul imaginii sau a videoclipului;
* Detectarea reperelor faciale – extragerea informațiilor despre reperele faciale. Acestea pot să fie formele pe care le au diferite componente ale feței sau textura pielii într-o anuită zonă;
* Clasificarea propriu-zisă – analiza rezultatelor de la pasul anterior și clasificarea în categorii prestabilite.

Procesul de funcționare a recunoașterii expresiilor faciale constă în folosirea unor tehnologii care folosesc diferiți marcatori biometrici pentru a detecta emoțiile din fețele oamenilor. Pentru rezolvarea acestei probleme se folosesc diferiți algoritmi care detectează fețele, apoi, extrag expresiile faciale și într-un final recunosc emoția. Acest lucru este făcut prin analizarea pozelor sau videoclipurilor.

Programele software care sunt capabile de rezolvarea acestei probleme pot oferii informații importante despre părerea pe care o au anumiți indivizi despre un anumit produs, o anumită firmă, etc...

Exemple de domenii în care se utilizează recunoașterea în timp real a emoțiilor:

* Testarea jocurilor video – jocurile video sunt create pentru o anumită audiență. De aceea fiecare joc urmărește să invoce diferite emoții și comportamente de la jucători. De aceea există o fază de testare, în care un grup de persoane se joacă jocul, iar creatorii jocului urmăresc reacțiile lor pentru a obține un feedback. Feedback-ul se poate obține si direct de la jucători, însă aceștia nu pot să exprime toate părerile pe care le au. De asemenea ei pot uita unumite detali, din acest motiv este mult mai eficient să se obțină feedback lor în timp real;
* Domeniul auto – din voința de a face vehiculele mai sigure si mai personale, producătorii auto adaugă cât mai multe facilități autovehiculelor lor. În cazul acesta se poate urmării expresia lui pentru a detecta dacă acestuia îi se va face rău;
* Cercetare de piață – companiile efectuează periodic cercetări de piață de mult timp. Detectarea emoțiilor le poate ușura munca exponențial, și să le ofere părerile pe care clienții le au despre anumite produse;
* Interviuri – problema de a determina dacă personalitatea unui candidat este bună pentru locul de muncă oferit este dificilă. Este foarte greu să identifici ceea ce candidatul urmărește cu adevărat, iar în acest caz inteligența artificială poate să intervină pentru a le analiza trăsăturile de personalitate.

Din exemplele de mai sus se poate observa că principala utilizare este obținerea de feedback.

Pe scurt avantajele folosirii algoritmilor pentru a obține emoțiile oamenilor, comparativ cu metodele tradiționale sunt:

Pe scurt acestă problemă duce la o economisire a timpului și resurselor epuizate pentru a obține aceste rezultate, comparativ cu metodele tradiționale:

* Rezultate superiore
* Timpul și efortul sunt reduse
* Costuri mici

# **Cercetare și alegerea datelor**

## **Cercetarea lucrăriilor din domeniu**

### **Lucrarea 1**

Prima lucrare parcursă a fost “Automatic emotion recognition through facial expression analysis in merged images based on an artificial neural network.”[2].

*Introducere*

În această lucrare se urmărește rezolvarea problemei de recunoaștere a emoțiilor prin expresii ale feței. Metoda folosită recunoaște 6 emoții de bază ale expresilor feței folosind o rețea neuronală artificială (ANN) în cadrul pozelor care sunt îmbinate. Rezultatul acestui sistem este emoția recunoscută.

Avantajul folosiri pozelor îmbinate este selectarea informațiilor relevante doar în zonele care se presupune că sunt afectate de modificări faciale. Acest lucru duce la reducerea timpului de antrenare a rețelei, performanță crescută a ratei de clasificare și o convergență rapidă a erorii Back-Propagation.

*Aspecte cheie*

Problema adresată în această lucrare este detectarea informațiilor din emoții. În general se urmărește adunatea datelor de la utilizatori sau mediul înconjurător (care pot fi extrase de senzori), obținerea informațiilor din aceste date și într-un final clasificarea informației pentru analizare. Există mai multe tipuri de detecție, printre care se pot enumera: recunoaștere vocală, procesarea limbajului natural, recunoaștere facială (scopul acestei lucrări), etc.

Întotdeauna procesul de detectare trebuie să înceapă cu detectarea zonei în care apare fața, extragerea și urmărirea înformațiilor relevante, și într-un final clasificarea expresiei. Doar apoi, prin intermediul acestor informații se poate trece la analiza expresilor.

*Implementare*

Lucrarea este alcătuită din 2 componente importante:

* Tratamentul imaginii care conține fața;
* Rețeaua neuronală artificială a cărui algoritm recunoaște expresiile faciale.

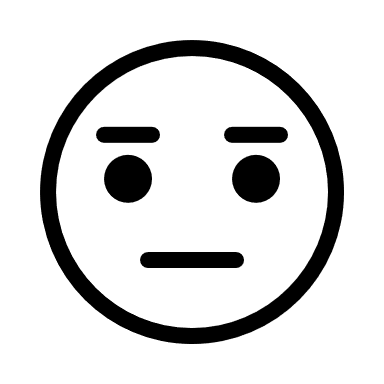
Acest algoritm funcționează în ciclu, și începe prin introducerea unei poze noi care este transformată într-o poză nouă (poză îmbinată) care este pregătită pentru a fi analizată de rețea.

Rețeau a fost antrenată anterior cu serii de poze care formau mulțimea de antrenament cu 6 expresii diferite (6 pentru fiecare persoană), care și ele au primit același tratament de îmbinare.

Într-un final după ce este găsit grupul căruia îi aparține poza, sistemul specifică starea emoțională a acesteia.

Emoția recunoscută

Imaginea introdusă



Tratamentul imaginii:

* Extragerea caracteristicilor faciale (ochi, gură)
* Îmbinarea pozei
* Preprocesarea imaginii (redimensionare și aplicare filtru alb/negru)

Rețeaua neuronală Back-Propagation

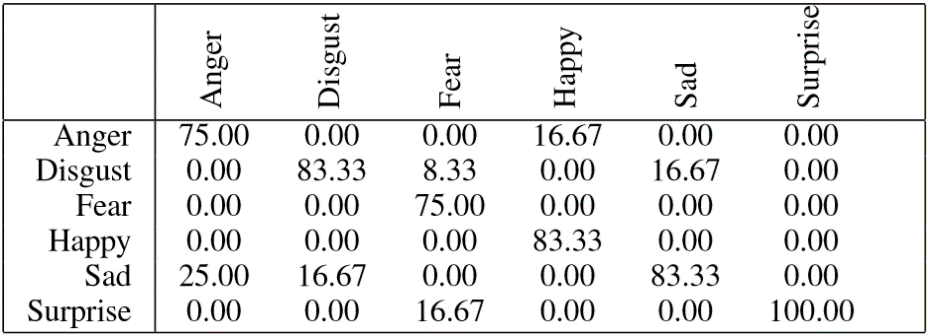
Fig 1. Ciclul de detectare a emoției faciale

Imaginea procesată este trimisă către rețeaua neuronală care folosește un algoritm Back-Propagation cu arhitectură Feed-Forward. Acesta conține 1200 neuroni în stratul de intrare (reprezintă pixelii imaginii 40 x 30). Pentru funcția de activare s-a folosit funcția sigmoidală . S-a folosit un singur strat ascuns, iar pentru numărul de neuroni pe stratul ascuns s-a folosit , unde L este numărul de straturi ascunse, este numărul de neuroni de intrare și este numărul de sample-uri de intrare. În final stratul de ieșire conține 6 neuroni, unul pentru fiecare emoție ce poate să fie rezultată.

*Teste și rezultate*

Pentru testare au folosit atât poze din propria bază de date, cât și poze din baza de date Cohn-Kanade (CK).

Prima testare a fost efectuată folosind pozele din propria bază de date unde mulțimea de antrenare conține un grup de 36 poze de la 6 indivizii, iar cea de testare 72 de imagini noi de la 12 indivizi. Toate imaginile au fost procesate corespunzător înainte să fie introduse în rețeaua neuronală. Pentru calcularea eficienței s-a folosit o matrice de confuzie. Per total acuratețea este de 83.3%, iar fiecare emoție are o acuratețe de peste 75%.



Tabelul 2. Tabelul cu matricea de confuzie a sistemului folosind imaginile îmbinate[3]

A doua testare a fost folosind același test, dar cu zona de la gură și ochi separate. Rezultatele sunt următoarele:

* Folosind doar ochii:
  + Acuratețe generală: 80.6%
  + Acuratețe per emoție de peste 66.67%
* Folosind doar gura:
  + Acuratețe generală: 68%
  + Acuratețe per emoție de peste 58.33%

Ultimate testare a fost cu setul de date din baza CK. Din 97 de indivizi au fost selectați 40, iar pentru fiecare individ sunt 6 poze, deci per total 240 de poze. Antrenarea a avut loc cu 15 indivizi (90 poze), iar testarea cu 25 indivizi (150 poze). Per total acuratețea este de 84%, iar fiecare emoție are o acuratețe de peste 76%.

Rezultatele arată că rețelele neuronale au un potențial mare în recunoașterea emoțiilor pe baza imaginilor. Se recomandă ca pe viitor să se folosească rețele neuronale dinamice pentru rezolvarea acestor probleme.

### **Lucrarea 2**

A doua lucrare parcursă a fost “Facial expression recognition using a hybrid CNN–SIFT aggregator.”[3].

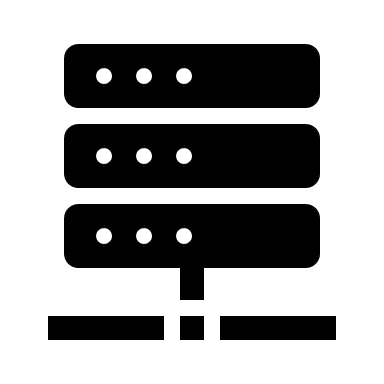
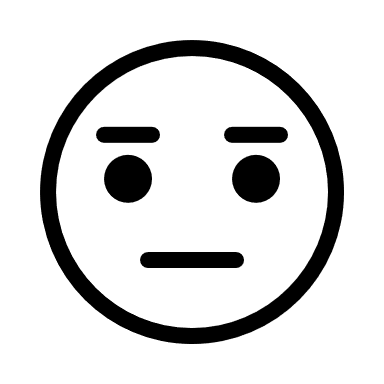
*Introducere*

Pentru această lucrare se urmărește combinarea folosirii unei rețele neuronale convoluționale împreună cu SIFT. Imaginea crudă trece prin straturile rețelei convoluționale înainte să se combine cu caracteristici SIFT sau SIFT dense. Atât modelul tradițional cât și cel dens sunt combinate cu modelul CNN. Ambele modele sunt antrenate și testate separat. Față de alte lucrări, în cazul acesta, combinarea v-a avea loc în timpul fazelor de antrenare și testare.

CNN

Imaginea

introdusă

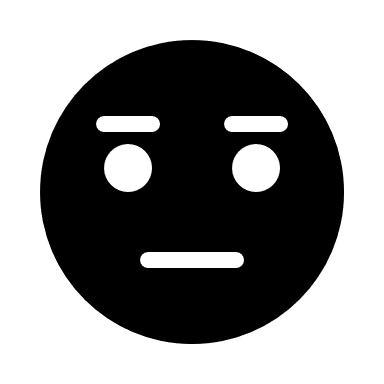
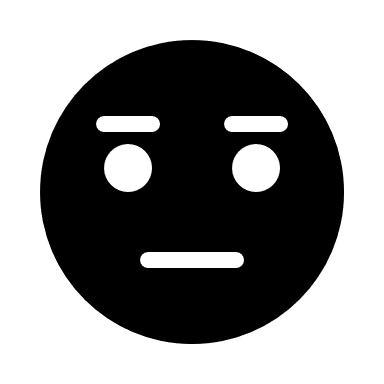


SIFT

Dense SIFT

Extragere puncte importante

Dense Detector



Descriptor

Emoție detectată

Descriptor

K-Means

Emoție detectată

Fig 2. Prezentarea generală a metodelor propuse[3]

*Pre-Procesarea*

Toate imaginile sunt redimensionate la 48x48 pixeli. Fiecare imagine este amplificată de 10 ori cu diferite transformări liniare (flip orizontal, rotație la un unghi aleator, zoom în unul din cele 4 colțuri). Într-un final pozele sunt normalizate.

*Rețeaua convoluțională*

Rețeaua conține 6 straturi convoluționale, 3 straturi Max-Pooling, urmate de 2 straturi dense (2048 neuroni/strat) conectate în întregime. Fiecare dintre staturile Max-Pooling si dense sunt urmate de straturi de dropout. Într-un final vine un strat softmax cu 7 ieșiri.

Ca și funcție de activare s-a optat pentru funcția Leaky ReLU .

*SIFT*

SIFT este folosit pentru a extrage puncte cheie din pozele cu fețe. Descriptorul SIFT este determinat prin partiționarea pozei în pătrate de 4x4. Pentru fiecare din cele 16 pătrate, se obține un vector de lungime 8. Prin unirea tuturor vectorilor se obține unul de lungime 128 pentru fiecare punct cheie. Pentru a folosi descriptorul în clasificare trebuie ca vectorul să aibă dimensiune fixă, pentru acest lucru se folosește K-means, care grupează descriptorii într-un set de clustere. După aceea, punctele cheie importante sunt calculate din numărul de descriptori din fiecare cluster. Vectorul final are o dimensiune de 2048. Vectorii trec printr-un strat conectat în întregime, iar rezultatul este unit cu cel al modelului CNN.

Comparativ, Dense SIFT nu are nevoie de extragere puncte cheie, acesta împarte poza în regiuni egale de pixeli. Astfel iasă 16 regiuni, fiecare descrisă de 128 de caracteristici, deci în total 2048 caracteristici pentru toată poza.

Pentru a crește acuratețea modelului, se face media sumei între rezultatele de la CNN, de la CNN+SIFT și de la CNN+Dense SIFT.

*Testare și rezultate*

Pentru testare se specifică că în cazul setului de date de la FER 2013, modelul a fost antrenat pe 28709 exemple, setul public fiind folosit pentru validare iar setul privat pentru testare. Fiecare rețea a fost antrenată pentru 300 de epoci cu batch-uri de 128 de poze. În cazul setului de date de la CK+, s-a specificat că s-a folosit un model pre-antrenat. S-au folosit doar 309 poze dar doar pentru 20 de epoci pentru a evita overfitting.

Rezultatele sunt și ele foarte bune, în cazul setului de date de la FER 2013, combinația de CNN+SIFT+D-SIFT a fost superior atât peste CNN+SIFT și peste CNN+D-SIFT, cu o acuratețe de 73.4%. La fel a fost și cazul pentru setul de date de la CK+, unde s-a obținut o acuratețe de 99.1%.

În concluzie, agregarea Dense-SIFT și CNN oferă rezultate uimitoare pentru aceste seturi de date.

### **Lucrarea 3**

Cea de a treia lucrare parcursă a fost “ Facial expression recognition using a hybrid CNN–SIFT aggregator.”[4].

*Introducere*

Obiectivul acestei lucrări este clasificarea fețelor oamenilor în una din cele 6 emoții universale sau în cea de a șaptea emoție, și anume cea neutră. În această lucrare se vor testa 3 tipuri de modele, și anume “Decision Tree”, Rețea neuronală Feed-Forward și Rețea neuronală convoluțională.

Arborele de decizieeste o tehnică de învățare supervizată care prezice o valoare pentru input-uri, pe baza unor reguli învățate cu un set de date folosite pentru antrenare. Se poate imagina ca și arbore de decizii masiv care funcționează pe reguli if-then-else.

Rețeaua neuronală Feed-Forward este un sistem de algoritmi care încearcă să identifice relații într-un set de date folosind metode care imită funcționarea creierului uman.

Rețeaua neuronală convoluțională este o rețea neuronală care conține straturi de convoluție care este o operație matematică care din 2 funcții crează o a treia funcție. Se folosesc atît straturi conectate în întregime (ca și la o rețea feed-forward) cât și straturi convoluționale care nu sunt conectate către toți neuronii.

Convoluția este foarte eficientă în recunoașterea de imagini și în clasificare comparativ cu o rețea tradițională feed-forward deoarece permite reducerea numărului de parametrii în o rețea și avantajul de concentrare pe anumite zone.

*Setul de date și alegerea modelului*

Setul de date folosit a fost FER-2013, care conține 37,887 poze cu cele 7 expresii: nervos, degustat, înfricoșat, fericit, trist, surprins și neutru. Se specifică ca acest set de date conține un număr imens de poze cu emoții fericite, și un număr foarte mic de poze cu emoții degustate. Acest lucru poate afecta rezultatele finale.

Din cele 37,887 poze, 80% au fost folosite pentru antrenare și 20 % pentru testare. În general se recomandă rația 60-20-20. Fețele identificare erau transformate în poze alb-negru și micșorate la o rezoluție de 48x48 pixeli.

S-au testat în total 4 modele, arborele de decizie și 3 modele de rețele neuronale.

În cazul arborelui de decizie, s-a obținut o acuratețe de 30.84%, o acuratețe foarte mică, astfel că testările au continuat pe modelele de rețele neuronale.

Apoi a urmat rețeaua neuronală feed-forward, unde acesta era formată din 3 straturi, intrare, ascuns și ieșire, după statul de intrare și și cel ascuns urma un dropout de 0.2 pentru a reduce overfitting, funcția de activare folosită a fost ReLU ,iar pe stratul de ieșire a fost de tip softmax. Din păcate acest model a prezis mult prea mult emoția nervos, și a fost abandonat, cu o acuratețe de 17.38%.

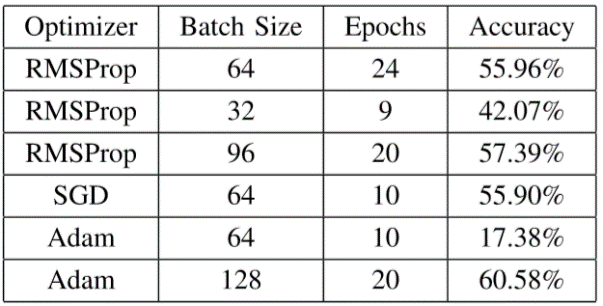
Următoarea rețea a fost una convoluțională simplă, unde aceasta era formată din 2 straturi convoluționale 2D, urmate de un strat max-pooling 2D, care era urmat de 2 straturi complet conectate dense. Din nou funcția de activare folosită a fost ReLU. Rezultatele acestui model au fost asemănătoare cu cele ale modelului anterior doar că predominant prezicerea a fost pentru poze cu expresii fericite, acuratețea fiind de 24.72%.

Într-un final a urmat modelul propus de ei, acesta este asemănător cu modelul propus anterior, dar conține încă 4 straturi convoluționale, iar max-pooling și dropout-ul au fost folosite mai eficient. Antrenarea era setată să aibă loc pentru 100 de epoci, însă după 20 de opoci se oprea, din moment ce nu se mai observau îmbunătățiri. Acuratețea acesteia a ajuns la 55.61%.

*Testarea*

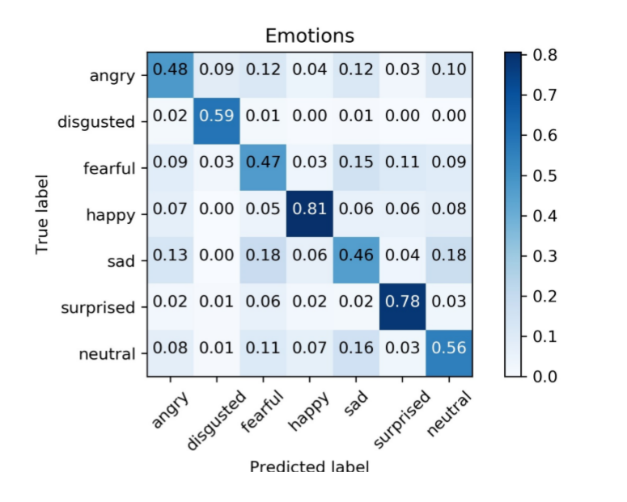
Pentru măsurarea acurateții s-a folosit următoarea formulă: .

S-au efectuat mai multe teste pe modelul final dar cu ajustări ale parametrilor, iar în final cea mai bună acuratețe obținută a fost de 60.58% cu optimizer-ul Adam.



Tabelul 3. Comparația hiperparametrilor[4]

Pentru obținerea matricii de confuzie s-a verificat numărul de poze pentru fiecare emoție prezisă de model, cu numărul de poze pentru fiecare emoție din setul original de date.nt ce nu se mai observau îmbunătățiri. Acuratețea acesteia a ajuns la 55.61%.



Tabelul 4. Matricea de confuzie[4]

*Concluzie*

Acuratețea finală a fost obținută folosind un optimizer Adam cu parametrii modificați. Alte lucrări au obținut rezultate mai bune, însă în cazul acesta rezultatele slabe se pot atribui și folosirea unui singur set de date, unde a fost o discrepanță mare în numărul de poze de la unele emoții. De asemenea aici s-a folosit și un model destul de simplist, scopul fiind arătarea că cu un număr modest de date și un model simplist se pot obține totuți rezultate bune.

### **Lucrarea 4**

Ultima lucrare parcursă a fost “Facial expression recognition with recurrent neural networks.”[5].

*Introducere*

Pentru această lucrare se folosește un model 3D al feței, modelul Candide-3 care conține 79 de parametrii care indică forma componentelor feței, cum ar fi gura, ochii, sprâncenele printr-un set de 116 hărți anatomice. Acești parametri sunt estimați cu ajutorul așa zișilor “learned objective functions”. Aceste caracteristici sunt apoi trimise către rețeaua neuronală recurentă pentru clasificare. Avantajul folosirii rețelelor recurente este faptul că acestea pot obține informații din secvențe de date succesive. Asfel, clasificarea se face pe baza datelor extrase din poze succesive din secvențe succesive care afișează expresii faciale.

Pentru recunoșterea unei expresii faciale, în mod normal se parcurg 3 pași: detectarea feței, extragerea de caracteristici și clasificarea expresiei.

*Implementare*

Pentru a extrage informații de nivel înalt, parametrii modelului trebuie să fie estimați pentru a descrie cel mai bine fața dintr-o poză. Acest lucru este rezolvat de Model Fitting și este adresată printr-o funcție obiectivă care evaluează cât de bine se potrivește parametrizarea modelului pentru o poză. Autorii propun ca funcția să fie învățată pe poze adnotate. Funcția aceasta este deobicei proiectată manual.

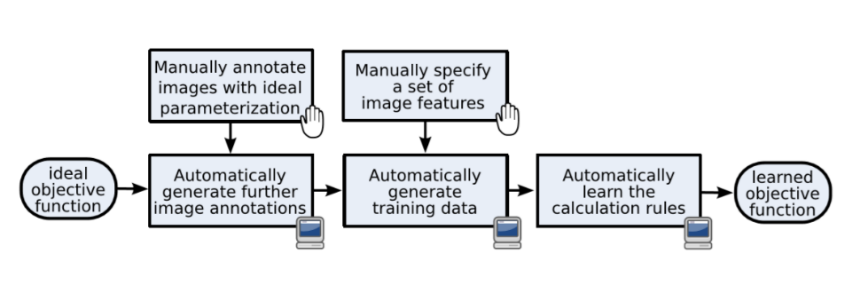


Fig 3. Abordarea prin învățare[5]

Avantajele folosirii unei astfel de funcții este că aceasta selectează doar caracteristicile importante dintre toate cele oferite, ceea ce reduce numărul de calcule necesare, și reduce timpul necesar.

În cazul extragerii caracteristicilor, s-au identificat 6 expresii faciale universale indiferent de cultură sau originea persoanei. Parametrii de deformare descriu constituția feței și poziția celor 116 mape. Acești parametri sunt extrași de la fiecare poză.

Pentru clasificare, se trimite secvența de caracteristici extrase. Sistemul acesta este bazat pe clasificare pe baza de frame-uri, pentru aplicări în timp real.

Din cauză că mașinile vector, arborii de decizie și rețelele neuronale feed-forward sunt proiectate pentru modele statice, s-a optat pentru folosirea unei rețele neuronale recurente care folosește celule LSTM.

*Testare*

În cazul testării s-a testat rețeaua recurentă atât cu celule LSTM uni-direcționale cât și bi-direcționale iar acesta a fost testastă pe setul de date Cohn-Kanade, scopul fiind clasificarea cu success a unor videoclipuri în cele 6 emoții de bază.

Rețeaua neuronală era formată dintr-un strat de intrare cu 35 de neuroni, un strat ascuns cu 150 neuroni, 100 de celule LSTM se aflat pe straturile ascunse de fordward și backward, iar stratul de ieșire avea 6 neuroni, unul pentru fiecare emoție. Pe durata antrenării, fiecare cadru al fiecărui videoclip a fost clasificat separat pe baza expresiei per general din videoclip. După ce antrenarea a avut loc cu success, acuratețea a fost obținută prin însumarea clasificărilor cadrelor și apoi normalizarea pentru a obține o clasificare probabilistică pentru fiecare secvență.

*Rezultate*

Rețeaua bi-direcțională a avut o performanță mai bună, însă rețeaua uni-direcțională oferă potențial pentru folosirea în timp real. În ambele cazuri seturi de antrenare a fost învățat foarte rapid de rețea, performanța fiind mai bună dacă s-ar fi folosit un set de date mai mare.

### **Comparație a lucrărilor studiate**

După cum se poate observa, fiecare din aceste lucrări aduce ceva nou pe masă. Prima lucrare a adus o acuratețe extrem de bună, a doua lucrare a demonstrat că combinarea unei rețele neuronale CNN cu alte metode poate aduce rezultate mai bune decât folosind doar rețeaua CNN. A treia lucrare a avut o rețea convoluțională simplistă si a demonstrat că setul de date este foarte important pentru a obține o acuratețe bună. Ultima lucrare a inovat folosind modele 3D.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Lucrarea 1** | **Lucrarea 2** | **Lucrarea 3** | **Lucrarea 4** |
| **Date pe care s-a lucrat** | Serii de poze care conțin 6 expresii diferite (6 pentru fiecare persoană) | Seturi de poze de pe CK și FER, preprocesate. | Serii de poze care conțin 7 expresii diferite | Candide-3 + funcție pentru face model fitting |
| **Algoritmi AI folosiți** | Feed-Forward Back-Propagation Artificial Neural Network | Rețea neuronală convoluțională + SIFT/Dense SIFT | Rețea neuronală convoluțională | Learning Objective Function + Rețea neuronală recurentă |
| **Măsurarea calității detecției** | Procentual, pe baza unui “Confusion Matrix” | Procentual, comparație între fiecare model | Procentual, succes/număr total | Mean Error Rate |
| **Rezultate concrete** | Acuratețe de peste 83%, rezultate promițătoare | Combinarea CNN cu SIFT/Dense SIFT are rezultate uimitoare | Rezultate bune, însă folosirea mai multor seturi de date ar fi dus la rezultate mai bune | Algoritmul funcționează corespunzător pentru secvențe de date. |

Tabelul 5. Comparația lucrărilor studiate

## **Alegerea setului de date**

Printre cele mai importante aspecte, în general, când vine vorba de folosirea unei rețele neronale este folosirea unui set de date adecvat. Un set de date bun trebuie să conțină un număr mare de date, să nu conțină date lipsă sau date inutile, este structurat într-un fel în care este ușor să se umble cu el și este ușor de înțeles.

Pentru recunoașterea emoțiilor din expresiile faciale, am luat în considerare următoarele seturi de date:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Baza de date** | **Imagini** | **Subiecte** | **Expresii** | **Locație** |
| CK+ | 593 sevențe | 123 | 6 + 2 | [Click](http://www.consortium.ri.cmu.edu/ckagree/) |
| MMI | 740 IMG, 2900 VID | 25 | 6 + 1 neutră | [Click](https://mmifacedb.eu/) |
| JAFFE | 213 IMG | 10 | 6 + 1 neutră | [Click](https://zenodo.org/record/3451524) |
| TFD | 112,234 IMG | ? | 6 + 1 neutră | [E-mail](mailto:josh@mplab.ucsd.edu) |
| FER-2013 | 35,887 IMG | ? | 6 + 1 neutră | [Click](https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/overview) |
| AFEW 7.0 | 1,809 VID | ? | 6 + 1 neutră | [Click](https://sites.google.com/site/emotiwchallenge/) |
| SFEW 2.0 | 1,766 IMG | ? | 6 + 1 neutră | [Click](https://cs.anu.edu.au/few/emotiw2015.html) |
| Multi-PIE | 755,370 IMG | 337 | Zâmbet, surprins, strâmbat, degustat, țipat, neutru | [Click](http://www.flintbox.com/public/project/4742/) |
| BU-3DFE | 2,500 IMG | 100 | 6 + 1 neutră | [Click](http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html) |
| Oulu-CASIA | 2,880 secvențe | 80 | 6 | [Click](http://www.cse.oulu.fi/CMV/Downloads/Oulu-CASIA) |
| RaFD | 1,608 IMG | 67 | 6 + 2 | [Click](http://www.socsci.ru.nl:8180/RaFD2/RaFD) |
| KDEF | 4,900 IMG | 70 | 6 + 1 neutră | [Click](https://www.emotionlab.se/kdef/) |
| AffectNet | 450,000 Label IMG | ? | 6 + 1 neutră | [Click](http://mohammadmahoor.com/databases-codes/) |
| ExpW | 91,793 IMG | ? | 6 + 1 neutră | [Click](http://mmlab.ie.cuhk.edu.hk/projects/socialrelation/index.html) |
| Google | ~156,000 IMG | ? | ? | [Click](https://research.google/tools/datasets/google-facial-expression/) |

Tabelul 6. Listă de seturi de date[6]

Dintre toate acestea, am ales într-un final să folosesc setul de date FER-2013. Acesta este alcătuit din poze cu fețe umane, poze care sunt grayscale și cu dimensiunea de 48x48. Fețele au fost prinse astfel încât să fie cât de cât centrate și să ocupe aproximativ aceelași spațiu din cadru.



Fig 4. Imaginile din setul de date

În total sunt 7 categorii de emoții, care sunt codificate astfel:

* 0 – Furios
* 1 – Dezgustat
* 2 – Înfricoșat
* 3 – Fericit
* 4 – Trist
* 5 – Surprins
* 6 – Neutru

Fișierul în care se află pozele conține trei coloane, prima care conține emoția codificată ca și mai sus, a doua coloană care conține valorile pixelilor pozelor, iar a treia coloană conține string care descrie dacă poza respectivă se folosește pentru antrenare sau testare (testare publică sau privată).

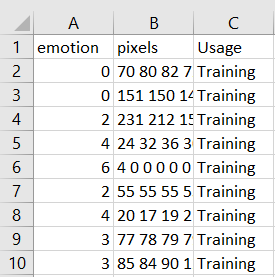


Fig 5. Fișierul care conține setul de date FER-2013

În total setul de date conține 35 887 de exemple. Dintre acestea 28 709 sunt în setul de antrenare, iar 7 178 se află în mulțimile de antrenare. Fiecare mulțime de antrenare conține 3 589 exemple.

Fig 6. Repartiția numărului de exemple pentru fiecare emoție

După cum se poate observa, repartiția datelor nu este egală și astfel algoritmii pot să dezvolte preferințe pentru anumite emoții (în cazul aceasta emoția dezgustat este defavorizată). În unele lucrări, categoria aceea a fost eliminată în întregime sau setul de date a fost unit cu altele pentru a avea un număr aproximativ egal de exemple per emoție. În cazul acesta am rămas cu setul implicit.

Același lucru se poate observa și doar pentru mulțimea de antrenament, unde numărul de exemple pentru emoția dezgustat este foarte mic. De asemenea, în timp ce 28 709 nu este un număr mic de exemple de integrat în algoritmi, ar fi ajutat folosirea unui set de date mai mare.

Fig 7. Repartiția numărului de exemple pentru fiecare emoție în mulțimea de antrenament

# **Implementarea programului software**

În total s-au testat 3 metode. Au fost implementate și testate folosirea unei mașini cu suport vectorial (atât cu extragere de caracteristici folosind HOG cât și folosind FAST) și folosirea unei rețele neuronale convoluționale.

Conform [8], o caracteristică este o bucată de informație care este importantă pentru a rezolva o anumită sarcină computațională într-un anumit program. Aceași definiției se poate da și la caracteristicile din contextul invățării automate sau recunoașterii de tiparuri. Astfel putem spune că un descriptor de caracteristici este un algoritm care pentru o imagine returnează locațiile zonelor importante din imagine.

Histograma Orientărilor Gradienților este un descriptor de caracteristici folosit pentru extragerea caracteristicilor din imagini, deobicei folosit pentru recunoașterea obiectelor.

Acesta funcționează astfel:

* În primul rând, opțional, se normalizează imaginea global – pentru reducerea influențării de către efecte de iluminare;
* Apoi se calculează gradientul imaginii în x și y – acestea capturează informații despre contur, siluete și texturi, oferind o rezistență mai mare la variații de iluminare;
* Calcularea histogramelor gradienților – producerea unei codificări sensibile la conținutul local din imagine. Aici se crează mai multe zone (numite celule) pentru care se crează o histogramă 1D a gradienților pentru toți pixelii din acea celulă;
* Normalizarea blocurilor – se i-au grupuri locale de celule și se normalizează contrastul lor;
* Crearea vectorului – ultimul pas colectează decriptorii din toate blocurile dintr-o grilă suprapusă de blocuri care acoperă fereastra de detectare, și le combină într-un vector de caracteristici.

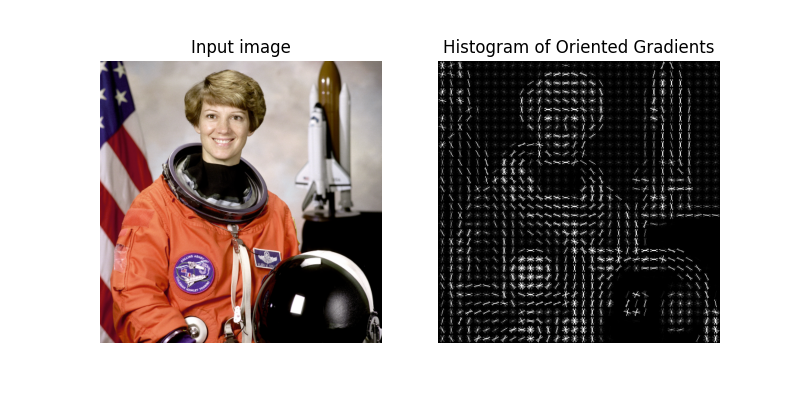


Fig 8. Imagine înainte și după ce a fost extrasă Histograma Orientărilor Gradienților[9]

Algoritmul FAST este folosit pentru detectarea punctelor de interes dintr-o imagine. Un punct de interes este un pixel care are o poziție bine definită și poate fi detectată ușor. Algoritmi asemănători existau de mult timp, însă acesta a fost create pentru aplicații în timp real pe dispozitive cu resurse computaționale reduse.

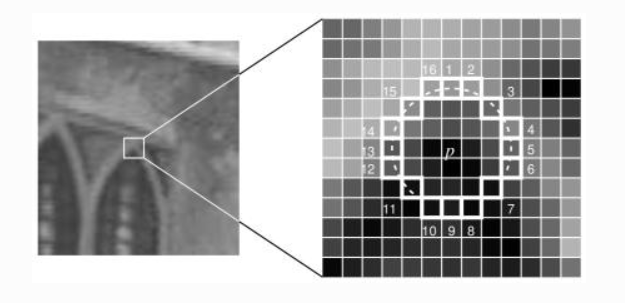


Fig 9. Imagine care arată punctul de interes și cei 16 pixeli pe un cerc[10]

O masină cu suport vectorial este o familie de algoritmi de învățare automată folosită pentru probleme matematice și inginerești. Clasificarea este execută prin construirea unui hiperplan de N-dimensiuni care separă în mod optim informațiile în categorii.

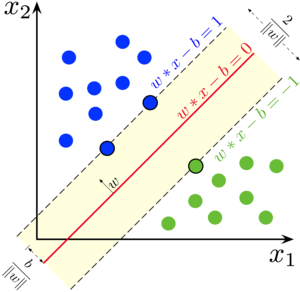


Fig 10. Separarea hiperplanelor[11]

Rețelele neuronale convoluționale sunt asemănătoare cu rețelele neuronale artificiale tradiționale în sensul că sunt formate din neuroni care se optimizează singuri prin învățare. Neuronii în continuare o să primească un input și o să efectueze o operație. De la primirea vectorului de imagine până la oferirea rezultatului, rețeaua în continuoare va folosi greutățile pentru a ține evidența scorului.

CNN se bazează că noi îi vom oferii imagini. Una din diferențe este că în acest caz, neuronii sunt organizați în 3D (inălțime, lățime și adâncime) .

În general arhitectura unei astfel de rețele este alcătuită din 3 tipuri de straturi:

* Straturi convoluționale
* Straturi de pooling
* Straturi inter-conectate în întregime

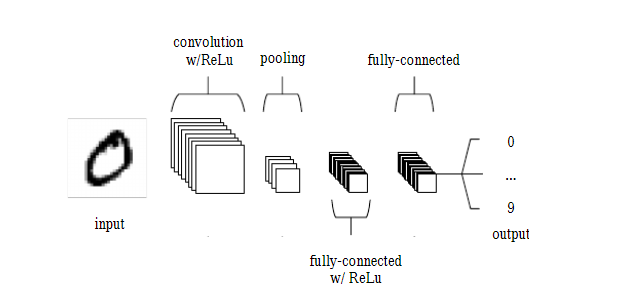


Fig 11. Arhitectură simplă a unei rețele CNN[12]

Xception este o rețea neuronală convoluțională insirată de Inception, unde modulele clasice din Inception au fost înlocuite cu convoluții separabile în întregime.

Arhitectura este următoarea, sunt 36 de straturi convoluționale care se ocupă cu extragerea de caracteristici, grupate în 14 module, toate care au conexiuni reziduale liniare între ele (mai puțin primul și ultimul modul).

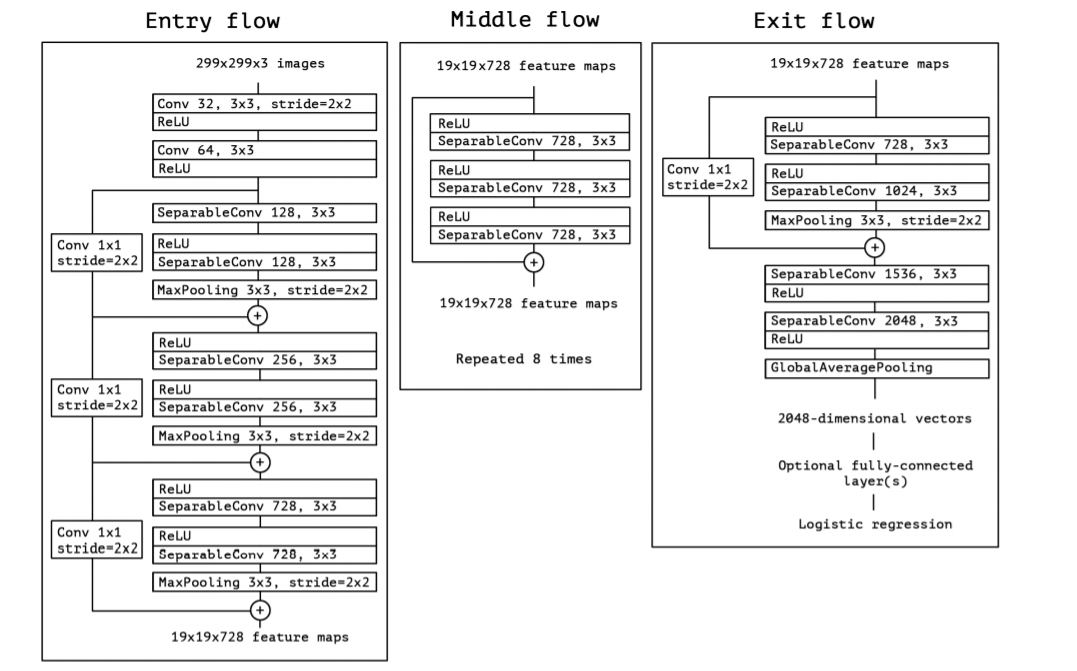


Fig 12. Arhitectura Xception[13]

Pentru clasificarea folosind SVM (atât cu HOG cât și FAST) imaginile inserate au fost de dimeniune 48x48, grayscale (1 channel). După ce caracteristicile au fost extrase și salvate local, acestea au fost introduse pentru învățare. În ambele cazuri condițiile de învățare au fost identice, diferind doar metoda de extragere folosită.

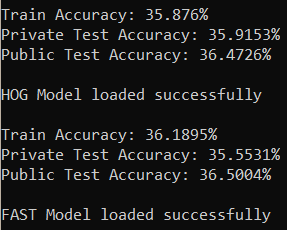
****

Fig 13. Acuratețea folosind SVM+HOG(sus) și SMV+FAST(jos)

În ambele cazuri au fost create și matrici de confuzie.

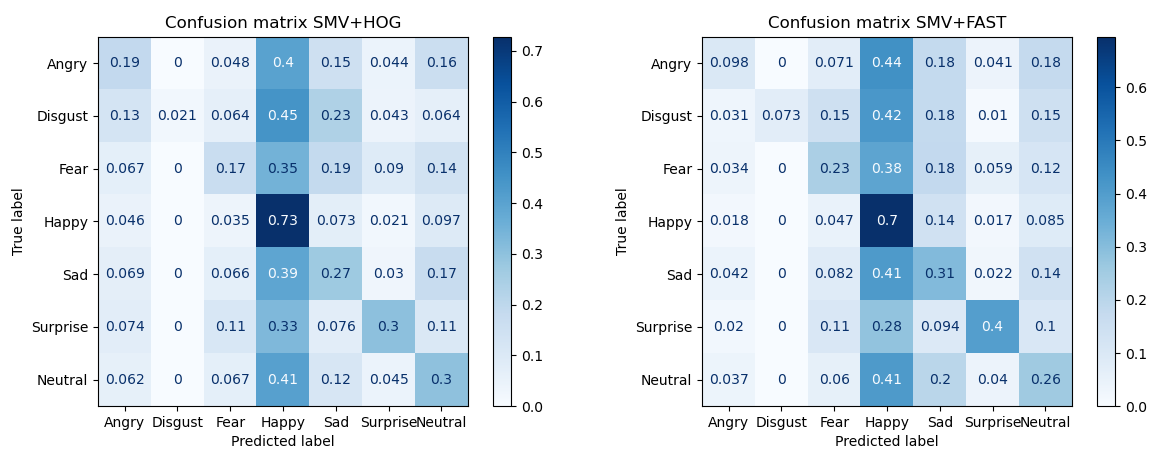


Fig 13. Matricile de confuzie

După cum se poate observa, rezultatele nu sunt uimitoare, însă avem o preferință de caracterizare pentru emoția de fericire. Iar diferențele de acuratețe dintre cele 2 sunt minore.

Pentru clasificarea folosind Xception, imaginile inserate au fost de dimensiune 71x71, convertite în RGB (3 canale). În cazul aceasta, imaginile au fost augmentate înainte să fie inserate în rețea (în fiecare epocă, aleatoriu se executau anumite operații cum ar fi rotire, schimbare înalțime/lățime, zoom, flip orizontal, etc...) și apoi au fost încărcate într-o rețea preantrenată per general și au foar doar ajustați parametri pentru cazul nostru.

Imaginile au fost inserate în rețea în batch-uri de 128, pentru 10 epoci. Pentru numărul de pași de validare și de antrenare, s-a luat numărul total de date și a fost împărțit la dimensiunea batch-ului.

Pentru a reduce problema de overfitting, sa folosit o rată de învățare foarte mică și au fost aplicate straturi de dropout.

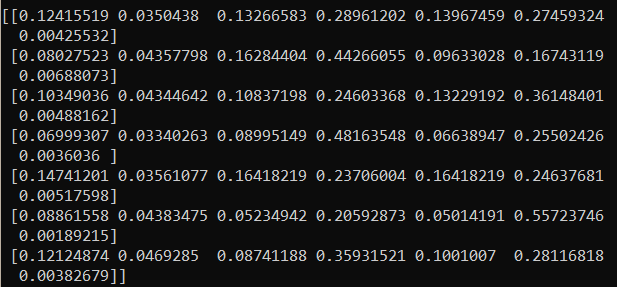


Fig 14. Matricea de confuzie pentru rețeaua Xception

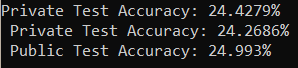


Fig 15. Acuratețea folosind Xception

După cum se poate observa, în cazul acesta am obținut rezultate mai slabe cu rețeaua Xception. Depinde foarte multe setul de date folosit și configurarea rețelei. În unele configurări am reușit să obțin o acuratețe de 31%, și cu optimizarea parametrilor această rețea ar putea să întreacă lejer rezultatele obținute de SVM. Cu toate acestea se poate observa că și în acest caz, unele emoții au fost favorizate.

# **Concluzii**

În concluzie, chiar dacă rețeaua proprie Xception nu a dat cele mai favorabile rezultate, din alte lucrări ștințiifice studiate și din unele observații făcute în timpul creării acestuia, este clar că Xception este o alegere foarte bună pentru prezicerea emoțiilor folosind expresiile faciale.

Principalul motiv pentru care cred că rețeaua se oprește la un moment dat din învățat este overiftting, unde rețeaua se obișnuiește prea mult cu datele. Adaugând mai multe date în setul de date ar fi o metodă de a evita această problemă. De asemenea metode de augmentare ale pozei mai agresive ar putea ajuta la creșterea acurateții. Într-un final modificarea parametrilor rețelei ar oferii probabil cea mai bunp acuratețe în cazul în care setul de date ar rămâne acesta.

Per total, Xception oferă un potențial mult mai mare decât mașina cu suport vectorial.

# **Bibliografie**

[1]. EKMAN, Paul; CORDARO, Daniel. What is meant by calling emotions basic. Emotion review, 2011, 3.4: 364-370.

[2]. RÁZURI, Javier G., et al. Automatic emotion recognition through facial expression analysis in merged images based on an artificial neural network. In: 2013 12th Mexican International Conference on Artificial Intelligence. IEEE, 2013. p. 85-96.

[3]. CONNIE, Tee, et al. Facial expression recognition using a hybrid CNN–SIFT aggregator. In: International Workshop on Multi-disciplinary Trends in Artificial Intelligence. Springer, Cham, 2017. p. 139-149.

[4]. SARAVANAN, Akash; PERICHETLA, Gurudutt; GAYATHRI, Dr KS. Facial Emotion Recognition using Convolutional Neural Networks. arXiv preprint arXiv:1910.05602, 2019.

[5]. GRAVES, Alex, et al. Facial expression recognition with recurrent neural networks. In: Proceedings of the International Workshop on Cognition for Technical Systems. 2008.

[6]. LI, Shan; DENG, Weihong. Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing, 2020.

[7]. CARRIER, Pierre-Luc, et al. FER-2013 face database. Universit de Montral, 2013.

[8]. <https://en.wikipedia.org/wiki/Feature_(computer_vision)>

[9]. <https://scikit-image.org/docs/dev/auto_examples/features_detection/plot_hog.html>

[10]. VISWANATHAN, Deepak Geetha. Features from accelerated segment test (fast). Homepages. Inf. Ed. Ac. Uk, 2009.

[11]. <https://en.wikipedia.org/wiki/Support-vector_machine>

[12]. O'SHEA, Keiron; NASH, Ryan. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458, 2015.

[13]. CHOLLET, François. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 1251-1258.