**ROMÂNIA**

**MINISTERUL APĂRĂRII NAŢIONALE**

**ACADEMIA TEHNICĂ MILITARĂ**

**„Ferdinand I”**

**FACULTATEA DE SISTEME INFORMATICE ȘI SECURITATE CIBERNETICĂ**

**Specializarea: Calculatoare și sisteme informatice pentru apărare**

**și securitate națională**



**Aplicație pentru extragerea si indexarea conținutului din documente scanate**

COORDONATOR ȘTIINȚIFIC: ABSOLVENT:

**Lector dr. ing. Stelian Spînu Sd.Plt. Tudose Alin-Romeo**

Conține \_\_\_\_\_\_\_\_\_\_\_ file Inventariat sub nr. \_\_\_\_\_\_

Poziția din indicator: \_\_\_\_

Termen de păstrare:\_\_\_\_\_

**BUCUREȘTI**

**2021**

PAGINĂ ALBĂ

PAGINĂ ALBĂ

PAGINĂ ALBĂ

PAGINĂ ALBĂ

# ABSTRACT

# Rezumat

**Cuprins**

[ABSTRACT 7](#_Toc73976965)

[Rezumat 8](#_Toc73976966)

[LISTĂ ABREVIERI 10](#_Toc73976967)

[LISTĂ FIGURI 11](#_Toc73976968)

[1. INTRODUCERE 12](#_Toc73976969)

[1.1. Importanța și actualitatea problematicii abordate 12](#_Toc73976970)

[1.2. Scopul și obiectivele proiectului 12](#_Toc73976971)

[1.3. Rezultate obținute 13](#_Toc73976972)

[1.4. Structura proiectului 13](#_Toc73976973)

[2. STADIUL CUNOAȘTERII ÎN DOMENIU 14](#_Toc73976974)

[2.1. Fundamente teoretice 14](#_Toc73976975)

[2.1.1. Inteligență artificială 14](#_Toc73976976)

[2.1.2. Machine Learning 14](#_Toc73976977)

[2.1.3. Deep Learning 15](#_Toc73976978)

[2.1.4. Problema proiectului 15](#_Toc73976979)

[2.2. Soluții și abordări similare 17](#_Toc73976980)

[2.2.1. Modelul propus de Alex Graves și Jurgen Schmidhuber, din cadrul Universității Tehnice din Munchen. [3] 17](#_Toc73976981)

[2.2.2. Modelul propus de Hanu Priya Indiran, din cadrul Universității de Tehnologie Kamaruguru, India. [2] 18](#_Toc73976982)

[2.2.3. Modelul propus de Arik Poznanski și Lior Wolf, din cadrul Universității Tel Aviv, „The Blavatnik School of Computer Science”. [4] 19](#_Toc73976983)

[3. Implementarea aplicației 20](#_Toc73976984)

[3.1. Descrierea proiectului 20](#_Toc73976985)

[6. Bibliografie 21](#_Toc73976986)

# LISTĂ ABREVIERI

* IA – Inteligență Artificială
* ML – Machine Learning
* DL – Deep Learning
* CV – Simularea vederii pentru un computer (Computer Vision)
* SP – Procesare de secvențe (Sequence Processing)
* CNN – Rețele Neuronale Convoluționale (Convolutional Neural Networks)
* RNN – Rețele Neuronale Recurente (Recurrent Neural Networks)
* MDRNN – Rețele Neuronale Recurente Multi-Dimensionale (Multi-Dimensional Recurrent Neural Networks)
* CTC – Clasificare Temporală Conexionistă (Connectionist Temporal Classification)
* LSTM – Long Short Term Memory, strat prezent în arhitecturile de RNN.
* HMM – Model Markov Ascuns (Hidden Markov Model)
* OCR – Recunoașterea Optică a Caracterelor (Optical Character Recognition)

# LISTĂ FIGURI

**No table of figures entries found.**

CAPITOLUL I

# INTRODUCERE

## Importanța și actualitatea problematicii abordate

În societatea modernă, automatizarea proceselor simple, de zi cu zi, a devenit, încetul cu încetul, un subiect de discuție foarte important. Începând de la micile sisteme de IoT, care preiau și excelează in desfășurarea activităților zilnice, precum încuietorile smart sau termostatele automate, care creează mediul optim, după preferințele fiecărei persoane, pana la întregi linii de asamblare industriale, automatizarea tuturor acestor procese a permis ridicarea standardului de viață pentru întreaga omenire, fie că vorbim despre confortul propriu, fie capacitatea de producție a fabricilor în diferite industrii.

Una din problemele ce încă persistă în ziua de astăzi, este procesarea diferitelor tipuri de documente, în mod automat, și cât mai rapid. Rapiditatea, ușurința stocării, si eficiența nu sunt termeni asociați cu modalitatea tradițională de gestionare, unde una sau mai multe persoane erau însărcinate cu gestiunea arhivelor de documente. În plus, întreținerea unei persoane care să se ocupe de această gestiune poate fi costisitoare pe termen lung. Așadar, implementarea unei soluții eficiente, automate, care să reducă atât costurile, cât și timpul de răspuns la diferite tipuri de cereri este necesară.

## Scopul și obiectivele proiectului

Proiectul își propune realizarea unei aplicații software ce va realiza în mod automat extragerea informațiilor din documente formatate, completate de mână. Această aplicație, primind ca input o imagine scanată a unui document preformatat, va identifica secțiunile de interes, din imagine, și anume, informațiile completate de mână. Odată identificate, aceste secțiuni din imagine vor fi oferite unei mașini cu inteligență artificială, bazată pe tehnici de învățare profundă, ce va identifica textul din imagine la nivel de caracter. Caracterele prezise din imagine vor fi asamblate înapoi in cuvinte, si stocate într-o bază de date nerelațională, ElasticSearch. Pentru ușurința accesării și configurării formatelor documentelor, datele salvate vor putea fi accesate printr-o interfață web.

## Rezultate obținute

## Structura proiectului

CAPITOLUL II

# STADIUL CUNOAȘTERII ÎN DOMENIU

## Fundamente teoretice

În ultimii ani, inteligența artificială a fost un subiect de interes, atât pentru media, cât și pentru noii cercetători în domeniul Științei Calculatoarelor. Inteligența artificială, algoritmi de Machine Learning, chiar si algoritmi de Deep Learning, apar în tot mai multe articole, chiar și în afara publicațiilor științifice. Ne-a fost promis un viitor plin de autoturisme autonome și asistenți virtuali inteligenți, de care ne apropiem, din ce în ce mai rapid.

### Inteligență artificială

Inteligența artificială a luat naștere în anii 1950, când o grupare de cercetători în nou-creatul domeniul al științei calculatoarelor, și-au pus următoarea întrebare: *„Ar putea calculatoarele să fie învățate să gândească?”*. Această întrebare este în dezbatere până în ziua de azi. O definiție pentru domeniul IA, ar putea fi următoarea: *„Efortul de a automatiza procese desfășurate, în mod normal, de oameni.”.* Așadar, inteligența artificială este un domeniu cu un grad foarte mare de generalizare, ce cuprinde, dar nu se limitează la domeniile Machine Learning și Deep Learning.

### Machine Learning

În Anglia epocii victoriene, Lady Ada Lovelace, era o prietenă și colaboratoare a lui Charles Babbage, inventatorul mașinii analitice. Chiar dacă a fost mult înaintea timpului ei, mașina analitică nu a fost niciodată gândită ca un calculator de uz general, atunci când a fost proiectat, în anii 1830 – 1840, deoarece conceptul uzului general al unui calculator nu era încă inventat. A fost doar un mod de a folosi operații mecanice pentru a automatiza anumite calcule din domeniul analizei matematice, de unde și numele acesteia. Ada Lovelace a făcut atunci remarca, despre mașina analitică: *„Mașina Analitică nu are nicio pretenție sa aducă ceva nou. Ea poate să facă tot ce știm noi sa îi spunem să facă. Scopul ei este să ne asiste în găsirea rezultatelor problemelor cu care suntem familiari.”*

Domeniul Machine Learning are la bază această intrebare: *„Poate un calculator să treacă dincolo de granițele a ceea ce știm noi să îi spunem să facă?”*. Așadar, este introdusă o nouă paradigmă de programare. Dacă în programarea clasică, unui calculator îi erau prezentate reguli și date, acesta trebuia să furnizeze rezultate. În domeniul ML, un calculator primește date și rezultate, iar acesta trebuie să furnizeze regulile de obținere a rezultatelor din acele date. Un sistem ML nu este un sistem programat explicit, ci este un sistem antrenat.

### Deep Learning

Deep Learning-ul este un subdomeniu al domeniului ML, o nouă abordare asupra învățării reprezentărilor din date disponibile, ce pune accent pe straturi succesive de învățare. Cuvântul „Deep” din Deep Learning nu presupune vreun fel de înțelegere în profunzime a problemelor, ci doar face referire la modelele cu straturi succesive de învățare. Adâncimea modelului reprezintă numărul de straturi succesive de învățare pe care le are un model. Modelele actuale de DL adesea cuprind zeci, chiar sute de astfel de straturi succesive, și fiecare este antrenat automat prin prezentarea datelor si a rezultatelor așteptate pentru acestea. [1]

### Problema proiectului

Traducerea automată a imaginilor ce conțin scris de mână în seturi de date ce pot fi recunoscute de un computer este un proces destul de dificil. În cadrul acestei teme apar probleme complicate, precum problema segmentării sau problema dicționarului. Pentru a găsi o soluție acestei probleme, este necesar să folosim una din tehnologiile recente descoperite în domeniul DL, și anume Computer Vision (Simularea vederii pentru un computer).

Pentru rezolvarea problemelor de Computer Vision, se folosesc, în majoritatea cazurilor, rețele neuronale convoluționale. Acestea introduc operația de convoluție, în straturile lor. Diferența fundamentală între un strat „Dense” și un strat convoluțional este următoarea: Un strat dens învață trăsături reprezentative în întreg spațiul său de intrare, in timp ce un strat convoluțional învață aceste trăsături local, în cazul imaginilor, in mici ferestre bidimensionale, de obicei de dimensiuni 3 x 3 sau 5 x 5.

Această caracteristică conferă unui CNN două proprietăți interesante:

1. Trăsăturile învățate de acestea sunt invariante la translatări. După ce a fost învățată o trăsătura în colțul din dreapta, jos al unei imagini, acesta poate fi recunoscut în orice altă parte a imaginii, de exemplu în stânga, sus.
2. Pot învăța ierarhii spațiale ale trăsăturilor. Un prim strat convoluțional va învăța trăsături mici, locale, precum colțuri, sau muchii, un al doilea strat va învăța trăsături mai generale, alcătuite din trăsăturile învățate de primul strat, și așa mai departe. Acest lucru permite CNN-urilor să învețe trăsături din ce în ce mai complexe și concepte vizuale abstracte.

## Soluții și abordări similare

### 2.2.1. Modelul propus de Alex Graves și Jurgen Schmidhuber, din cadrul Universității Tehnice din Munchen. [3]

Cei doi cercetători propun un model aparte. În loc să trateze problemele CV si SP separat, precum modelele HMM, folosite pentru transcrieri, aceștia, folosindu-se de două descoperiri recente în domeniul rețelelor neuronale, RNN multidimensionale, și CTC, introduc un model antrenat offline pentru recunoaștere a scrisului de mână, care acceptă la intrare informație sub forma de pixeli de imagine.

Pentru MDRNN se folosesc straturi LSTM. Un strat LSTM este format din celule de memorie conectate in mod recurent, ale căror activări sunt controlate de 3 porți: poarta de intrare, poarta de uitare, și poarta de ieșire. Aceste porți permit celulelor de memorie să rețină, să șteargă și să acceseze informații reținute anterior, obținând recunoașteri dependente de un context destul de larg.

Forma standard a straturilor LSTM, este unidimensionala, deoarece fiecare celulă de memorie conține o singură conexiune recurentă, controlată de o singură poartă de uitare. Totuși, aceste straturi se pot extinde la n dimensiuni, folosind n conexiuni recurente, una pentru fiecare stare anterioară de-a lungul fiecărei dimensiuni, cu n porți de uitare.

CTC este un strat de ieșire creat special pentru probleme de procesare de secvențe folosind RNN-uri. Spre deosebire de alte straturi de ieșire, stratul CTC nu necesită date de antrenare pre-segmentate, și nici post-procesarea ieșirilor, pentru a obține datele recunoscute. Acesta antrenează direct rețeaua să estimeze probabilitățile condiționale ale etichetelor posibile prezente in secvențele de intrare.

Acest model nu are nevoie de niciun fel de preprocesare specifică fiecărui alfabet, deci poate fi folosit neschimbat pentru orice limbă. Dovezile generalității si a eficienței acestuia sunt asigurate de datele puse la dispoziție de o competiție recentă de recunoaștere a caracterelor arabice, unde a obținut o acuratețe de 91.4%, (câștigătorul concursului a obținut 87.2%), chiar dacă niciunul din autori nu cunosc alfabetul arabic.

### 2.2.2. Modelul propus de Hanu Priya Indiran, din cadrul Universității de Tehnologie Kamaruguru, India. [2]

Proiectul lui Indiran dorește să clasifice un cuvânt individual, astfel încât textul scris de mână să poată fi translatat într-o forma digitală, și demonstrează folosirea rețelelor neuronale pentru a proiecta un sistem de recunoaștere a caracterelor din alfabetul englez. Acest sistem preia la intrare imaginile binarizate cu literele ce trebuie să fie recunoscute. Imaginile sunt introduse într-un sistem de extragere de trăsături, al cărui ieșire este luat ca intrare în rețeaua neuronală. Indiran abordează doua metode pentru rezolvare: clasificarea cuvintelor în mod direct, dar și segmentarea acestora și recunoașterea la nivel de caracter.

Pentru prima metodă, diferite arhitecturi de CNN sunt folosite pentru a antrena un model cu acuratețe ridicată, pentru costuri computaționale relativ mici. Dezavantajul acestei metode este însă faptul ca este dependentă de un dicționar de cuvinte inițial, pentru a putea clasifica intrările.

Pentru cea de-a doua metodă, un RNN cu straturi LSTM, sunt folosite, împreună cu convoluția, pentru a crea delimitări pentru fiecare caracter, în imaginile de intrare. După ce au fost extrase imaginile pentru fiecare caracter din imagine, acestea sunt date către un CNN pentru clasificare. În urma clasificării, se reconstruiește fiecare cuvânt, în funcție de rezultatele obținute. Rezultatele obținute vorbesc de la sine, obținându-se o eroare de 1% în recunoașterea cuvintelor. Dacă se folosesc straturi de dimensiune foarte mare pentru CNN, eroarea de recunoaștere poate scădea pana la 0.2%.

### 2.2.3. Modelul propus de Arik Poznanski și Lior Wolf, din cadrul Universității Tel Aviv, „The Blavatnik School of Computer Science”. [4]

Proiectul celor doi profesori adopta o metoda relativ aparte. Aceștia folosesc un CNN pentru a estima, plecând de la imaginea unui cuvânt scris de mână, frecvențele grupărilor de n caractere ce apar în cuvânt. Frecvențele monogramelor, digramelor și trigramelor, sunt folosite pentru a asocia profilul estimat al cuvântului de recunoscut, cu profilul real al cuvintelor dintr-un dicționar de dimensiuni mari. Din nou, metoda cu dicționarul nu este o metodă facilă, pentru tema proiectului, din cauza necesității recunoașterii datelor care nu se afla neapărat într-un dicționar, precum nume, prenume, adrese e-mail, etc..

Modelul este bazat pe o arhitectura de tip VGG, formată din straturi convoluționale mici, de dimensiune (3 x 3). Rețeaua are un total de 12 straturi, 9 straturi convoluționale, si 3 straturi dense. Performanțele acestui model sunt destul de ridicate, obținându-se o acuratețe de aproximativ 95%, pe mai multe seturi de date, printre care se numără și seturile IAM, SVT și RIMES.

CAPITOLUL III

# Implementarea aplicației

## Descrierea proiectului

Având o aplicație ce îmbină tehnologii din mai multe câmpuri de cunoaștere din domeniul Computer Science, am hotărât că o abordare modulară a implementării este cea mai potrivită metodă.

Astfel, se pot distinge 4 mari componente, și anume:

* Modulul de preprocesare *(python)*;
* Clasificatorul *(python)*;
* Baza de date *(ElasticSearch)*;
* Interfața web *(PHP + CSS + JavaScript)*.

Folosind rețeaua neuronală, putem realiza recunoașterea textului prin două metode: recunoașterea la nivel de caracter, și recunoașterea la nivel de cuvânt.

Abordările bazate pe recunoașterea la nivel de caracter, împart, mai întâi, cuvântul în caractere sau subcaractere. Caracterele scrise de mână fără constrângeri sunt adesea conectate, sau chiar se suprapun cu caracterele învecinate, ceea ce face dificil de spus unde se termină un caracter, și începe următorul. În asta constă problema segmentării, acest tip de abordări fiind susceptibil la erori de segmentare.

Abordările bazate pe recunoașterea la nivel de cuvânt nu utilizează segmentare, ci recunosc cuvintele ca entități. Astfel, intervine problema dicționarului, deoarece aceste sisteme au nevoie de cel putin un model, sau șablon pentru fiecare cuvânt recunoscut. Aceasta abordare nu poate fi implementată în proiect, deoarece, de obicei, în formulare sunt prezente nume de persoane, adrese, numere de telefon, sau alte informații pentru care nu se poate genera un șablon eficient.

Pentru ușurarea implementării, și evitarea întâmpinării problemelor precum problema dicționarului, sau problema segmentării caracterelor scrise de mână, am impus, fără a restrânge generalitatea aplicației, ca formularele ce vor fi recunoscute să fie completate doar cu majuscule, cifre, sau simboluri, introduse, fiecare caracter într-o căsuță separată. Astfel, impunem ca recunoașterea textului scris de mână să fie făcută la nivel de caracter, iar problema segmentării caracterelor dispare.

De asemenea, pentru a putea îmbunătăți calitatea recunoașterii, fara a introduce inconveniențe prea mari, am hotărât ca, pentru recunoașterea unui formular, să fie introdusă, mai întâi, o poză cu formularul necompletat, pentru a putea extrage informații suplimentare despre imaginile ce urmează a fi recunoscute, fără ca acestea să fie influențate de scrisul de mână.

În secțiunile următoare, vom parcurge, în detaliu, fiecare dintre aceste componente, prezentând părțile componente, si funcționalitățile implementate.

## Modulul de preprocesare

Preprocesarea este utilizată, în principal, pentru reducerea zgomotului prezent în datele de intrare ale clasificatoarelor, îmbunătățind astfel performanțele sistemelor.

În aplicația prezentată, modulul de preprocesare este folosit în două scenarii, procesarea formularelor completate, și procesarea formularelor goale, ce urmează a fi recunoscute.

### Preprocesarea formularelor goale

Pentru recunoașterea oricărui document completat, este necesar ca, mai întâi, să fie introdusă o imagine cu formulatul necompletat, pentru a obține informații despre document ce vor facilita extragerea și recunoașterea ulterioară a caracterelor. Această acțiune trebuie efectuată o singură data, pentru fiecare tip de formular, recunoașterile ulterioare folosind informațiile extrase de prima dată.

Modulul de preprocesare, în această situație, are rolul de a construi un document *.json*  ce va conține informații despre formular, precum mărimea originală a formularului, titlul, o scurtă descriere, textul tipărit pe formular, și numele și tipul fiecărui câmp de completat, pentru selectarea modelului de rețea neuronală folosit la clasificare.

Toate caracteristicile prezentate mai sus, în afară de textul tipărit, sunt introduse de către utilizator în interfața web a aplicației, și trimise către modulul de preprocesare, pentru a fi îmbinate la final, deoarece se pot introduce rapid si fără inconveniențe. Introducerea textului tipărit, este o cu totul altă situație, deoarece aceste formulare pot conține volume mari de text, iar introducerea acestuia poate fi deranjantă. De aceea am hotărât ca extragerea textului să se facă în mod automat folosind tehnologii de OCR.

Pentru acest lucru, am folosit *Pytesseract*, un utilitar de OCR, deținut si dezvoltat de Google, fiind un framework peste motorul Google Tesseract-OCR [5]. Astfel, cu ajutorul acestui utilitar, realizăm recunoașterea și extragerea textului tipărit.

În urma extragerii informațiilor, se creează obiectul *.json*, care este mai apoi introdus în baza de date ElasticSearch pentru a fi folosit ulterior la recunoașterea formularelor completate.

Poza cu formular necompletat si rezultatul recunoașterii aici!!!

### Preprocesarea formularelor completate

Dacă preprocesarea formularelor goale a fost o sarcină relativ simplă, putând fi folosite utilitare pentru părțile complicate, și intervenția utilizatorului pentru părțile simple, preprocesarea formularelor completate este cu totul altfel.

Modulul de preprocesare primește imaginea cu documentul dorit și obiectul *.json* al tipului de formular adăugat anterior, și are sarcina de a extrage imagini ce conțin fiecare caracter ce urmează a fi recunoscut. Pentru a realiza această sarcină, au fost parcurși pașii următori:

#### Extragerea pixelilor ce conțin informația scrisă de mână

Pentru a diferenția cu ușurință textul scris de mână de textul de tipar prezent în formular, pentru a face o extragere a caracterelor cât mai bună, trebuie să fie impus următorul lucru: Formularul trebuie să fie completat cu o altă culoare decât culoarea textului tipărit, negru. Aici apare întrebarea „Ce culoare ar fi potrivită pentru completare?”. Răspunsul evident ar putea să pară că este albastru, însă, după cum se poate vedea în imaginea de mai jos, pentru imagini scanate, unde scanarea imaginilor nu este ideală, negrul poate fi confundat foarte ușor cu albastru închis, și să fie interpretat ca fiind scris de mână.

Aici poza cu formular completat cu albastru si rezultatul extragerii!!!

Prin urmare, am ales să folosim pentru completare culoarea roșu, fiind ușor diferențiabilă de negru în imagini, și, în același timp, fiind o culoare utilizată în scris. Putem observa diferențele majore în separarea culorilor în imaginea următoare.

Aici poza cu roșu!!!

Separarea culorilor se face folosind biblioteca *opencv-python*, cu ajutorul căreia putem extrage dintr-o imagine doar pixelii ce au valorile cuprinse într-un interval setat de noi. Rezultatele din imaginea precedentă au fost obținute pentru valorile RGB (0,0,100) pentru minim, și (100,100,255) pentru maxim.

După extragerea pixelilor roșii, au loc următoarele operații:

* Aplicarea unui filtru median, de dimensiune 3, pentru eliminarea eventualului zgomot prezent în imagine, datorită scanării cu defecte a imaginii, sau prezența altor artefacte;
* Conversia imaginii extrase în tonuri de gri, pentru interpretarea ulterioară a imaginii;
* Efectuarea binarizării tip OTSU.

#### Extragerea regiunilor de interes

Pe baza imaginii obținute anterior, următorul pas este selectarea din imagine a regiunilor ce conțin informația completată. Aici se vede utilitatea convertirii imaginii în tonuri de gri, si binarizarea. În urma binarizării, pixelii din imagine vor avea doar două valori: 255 (alb) – pentru zonele ce conțin scris de mână, și 0 (negru) – pentru zonele de fundal. Astfel, extragerea zonelor de interes devine o problemă banală.

Capitolul VI

# 6. Bibliografie

[1] F. Chollet, *Deep Learning with Python,* Manning Publications, Shelter Island, NY, Nov. 2017.

[2] H. P. Indiran, *Handwritten Character Recognition using Convolutional Neural Networks in Python with Keras*, IEEE, Asian Journal of Convergence In Technology, Tamil Nadu, India, Vol. 5 Is. III, 2020.

[3] A. Graves, J. Schmidhuber, *Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks*, NIPS, München, Germania, 2008.

[4] A. Poznanski, L. Wolf, *CNN-N-Gram for HandwritingWord Recognition*, IEEE, Tel Aviv, Israel, 2016.

[5] Python Software Foundation , Google Tesseract-OCR, Accesibil: <https://pypi.org/project/pytesseract/>, Accesat la 07.06.2021 18:20.