Clasificarea Imaginilor Radiologice pentru Diagnosticarea Pneumoniei Utilizând Tehnici de Machine Learning

Curcă Ștefan

Grupa 331 AA

Descrierea proiectului:

Proiectul propus se axează pe dezvoltarea și implementarea unui sistem de clasificare a imaginilor radiografice pentru diagnosticul eficient al pneumoniei, prin intermediul tehnologiilor de învățare automată. Scopul fundamental al acestui proiect constă în furnizarea unei soluții precise și eficiente pentru identificarea diferențelor semnificative între imagini radiografice asociate pacienților sănătoși și celor afectați de pneumonie.

Acest proiect reprezintă o inițiativă în aplicarea tehnologiilor de machine learning în domeniul medical, aducând beneficii semnificative în îmbunătățirea eficienței diagnosticului și tratamentului bolilor respiratorii.

Setul de date:

Modul de obținere și organizare a setului de date a implicat mai multe etape esențiale. Inițial, am colectat o varietate de imagini radiografice de la pacienți, incluzând atât radiografii ale persoanelor sănătoase, cât și ale celor diagnosticate cu pneumonie. Aceste imagini au fost supuse unui proces de selecție, eliminând imaginile necorespunzătoare.

Pentru a organiza eficient setul de date l-am divizat în subset de antrenare, validare și testare. Ulterior pentru fiecare subset am creeat subdirectoare distincte pentru categoriile de imagini, respectiv “Normal” și “Pneumonie”. Fiecare subdirector a fost asociat cu un set corespunzător de imagini relevante. Această abordare a permis o gestionare clară și ordonată a datelor, facilitând astfel procesul de încărcare și manipulare a acestora în cadrul algoritmilor de învățare automată.

De asemenea, am implementat un sistem de etichetare corespunzător pentru fiecare imagine, indicând dacă aceasta aparține categoriei "Normal" sau "Pneumonie". Etichetarea corectă a fost esențială pentru antrenarea modelului de clasificare, asigurându-ne că algoritmul învață să distingă între cele două tipuri de imagini cu precizie.

Fiecare fișier al setului de antrenare („Normal” și „Pneumonie”) conține 500 de imagini, în cazul setului de testare fiecare fișier conține 100 de imagini iar în cazul setului de validare fiecare conține 8 imagini.

Algoritmii utilizați:

1) k-Nearest Neighbors (k-NN) este un algoritm de învățare supervizată utilizat pentru clasificare și regresie. În cadrul clasificării, k-NN atribuie o etichetă unui punct de date pe baza etichetelor punctelor învecinate din setul de date de antrenare. Principiul fundamental al k-NN constă în căutarea celor mai apropiați k vecini pentru un punct de date de test și atribuirea clasei dominante din acești vecini punctului de test.
2) Naive Bayes este un algoritm probabilistic bazat pe teorema lui Bayes. Acesta este adesea utilizat în clasificarea datelor, inclusiv în domeniul medical. Algoritmul presupune

independența condiționată a atributelor (de aici și "Naive") și folosește probabilități pentru a face predicții.

Parametrii utilizați în model:

1) Dimensiunea imaginilor au un impact semnificativ în cadrul unui proiect de clasificare a imaginilor, iată câteva aspecte:
   - Redimensionarea imaginilor la dimensiuni mai mici poate fi utilă pentru a gestiona eficient resursele hardware disponibile.
   - Dimensiunea imaginilor influențează numărul total de caracteristici (pixeli) pe care modelul trebuie să le ia în considerare.
   - Imaginile redimensionate pot afecta performanța generală a modelului. Redimensionarea excesivă poate duce la pierderea de detalii semnificative.
   - În proiecte medicale, rezoluția și claritatea imaginilor pot fi cruciale pentru interpretarea rezultatelor.
2) Calea către seturile de date
   - dataset_folder reprezintă calea către setul de date de antrenare (“./dataset”)
   - test_folder reprezintă calea către setul de date de test (“./test”)
   - valid_folder reprezintă calea către setul de date de validare (“./valid”)
3) Constanta k (numărul de vecini pentru k-NN), în acest model este setată la 3.

Biblioteci Python utilizate:

1) Numpy
2) Scikit-learn
3) Scikit-image
4) Matplotlib
5) Pandas
6) Os

Descriere procese:

Procesul de antrenare și validare al modelului implică mai multe etape fundamentale:

1) <u>Încărcarea datelor</u>

   Seturile de date de antrenare și validare, compuse din imagini radiografice și etichete corespunzătoare, sunt încărcate folosind funcția „load_datase”. Imaginile sunt redimensionate la o rezoluție specifică pentru a uniformiza dimensiunile.

2) Preprocesarea datelor

Imaginile sunt redimensionate și aplatizate pentru a forma vectori de caracteristici. În acest proces, datele sunt aduse la o formă compatibilă cu algoritmii de învățare automată: k-NN și Naive Bayes.

3) Inițializarea și antrenarea modelului k-NN

Un clasificator k-NN este inițializat cu un număr de trei vecini. Modelul este antrenat pe setul de date de antrenare utilizând funcția fit.

4) Predicția pe setul de date de testare și validare

Modelul antrenat este utilizat pentru a face predicții pe setul de date de testare și validare. Rezultatele sunt stocate și utilizate ulterior pentru evaluare.

5) Inițializarea și antrenarea modelului Naive Bayes

Un clasificator Naive Bayes este inițializat și apoi antrenat pe setul de date de antrenare.

6) Predicția și evaluarea pentru modelul Naive Bayes

Similar cu k-NN, modelul Naive Bayes face predicții pe setul de date de testare și validare, iar rezultatele sunt evaluate pentru a determina performanța acestuia.

7) Salvarea rezultatelor

Rezultatele predicțiilor și informațiile despre performanță sunt stocate în fișierele CSV și imaginile generate pentru a permite o analiză ulterioară și o comunicare eficientă a rezultatelor obținute.

8) Validare încrucișată

Se efectuează o validare încrucișată pentru a evalua robustețea modelului la variații ale setului de date și pentru a obține o estimare mai precisă a performanței acestuia.

Procesul de testare implică următoarele etape:

1) Încărcarea datelor de testare

Imaginile radiografice din setul de date de testare sunt încărcate utilizând funcția „load_dataset”. Asemenea setului de antrenare, acestea sunt redimensionate și pregătite pentru a fi utilizate în procesul de testare.

2) Predicția pe datele de testare

Modelul antrenat este folosit pentru a face predicții pe setul de date de testare. Pentru clasificarea k-NN si Naive Bayes, se utilizează funcția predict.

3) Evaluarea performanței

Performanța modelului este evaluată pe baza predicțiilor făcute pe setul de date de testare. Acuratețea și alte metrici de evaluare sunt calculate pentru a cuantifica eficiența modelului în clasificarea imaginilor.

4) Compararea cu etichete reale

Rezultatele obținute prin predicții sunt comparate cu etichetele reale din setul de date de testare. Această comparație oferă o măsură clară a preciziei modelului în recunoașterea și clasificarea corectă a imaginilor.

Rezultate obținute:

Pentru imaginile redimensionate la 64x64 :

```
69.5% of test samples were correctly classified -- k-NN
68.75% of validation samples were correctly classified -- k-NN
49.5% of test samples were correctly -- NB
68.75% of validation samples were correctly classified -- NB
Average accuracy for k-NN: 91.80%
Average accuracy for NB: 87.60%
```

Pentru imaginile redimensionate la 100x100:

```
70.0% of test samples were correctly classified -- k-NN
62.5% of validation samples were correctly classified -- k-NN
50.5% of test samples were correctly -- NB
68.75% of validation samples were correctly classified -- NB
Average accuracy for k-NN: 91.50%
Average accuracy for NB: 87.40%
```

Pentru imaginile redimensionate la 200x200 :

```
70.5% of test samples were correctly classified -- k-NN
56.25% of validation samples were correctly classified -- k-NN
49.0% of test samples were correctly -- NB
62.5% of validation samples were correctly classified -- NB
Average accuracy for k-NN: 89.50%
Average accuracy for NB: 87.60%
```

Interpretarea rezultatelor:

În mod general se observă faptul că algoritmul k-NN pare să aibă o performanță mai bună decât algoritmul Naive Bayes pe aceste seturi de date și cu parametrii respectivi aleși.

Pe setul de testare, în cazul în care imaginile au fost redimensionate la 64x64, algoritmul k-NN a clasificat corect 69.5% dintre imaginile, pe setul de validare, procentul de clasificare corectă a fost de 68.75%, astfel acest algoritm oferă rezultate constante între setul de testare și cel de validare. În același caz, pe setul de testare, algoritmul Naive Bayes a avut o performanță de 49.5%, iar pe setul de validare, procentul de clasificare corectă a fost de 68.75%. Există o diferență semnificativă între performanța pe setul de testare și cel de validare, indicând o posibilă neadecvare a modelului sau necesitatea de optimizare.

Un factor important ce influențează performața modelului este asemănarea semnificativă între imaginile din clase diferite. Imaginele din clasele “normal” și “pneumonie” sunt foarte asemănătoare și au trăsături comune, ceea ce face dificil pentru modelul de machine learning să facă diferența între ele. Câteva posibilăți pentru a remedia această situație ar fi : utilizarea tehnicilor de augmentare de date pentru a crea variații ale imaginilor existente prin rotire, scalare, inversare, altă posibilate o reprezintă optimizarea hiperparametrilor.

Se observă că performanța modelului scade odată cu creștere rezoluției imaginii. Pentru k-NN, care se bazează pe distanțele dintre puncte în spațiul caracteristicilor, o creștere a dimensiunii poate duce la o dispersie crescută a punctelor, ceea ce face mai dificilă găsirea vecinilor apropriați. Naive Bayes presupune independența condiționată a trăsăturilor date, iar o creștere a dimensiunii poate conduce la asumarea unei independențe condiționate mai slabe, ceea ce poate afecta performanța. Imaginile cu rezoluții mai mari pot conține mai mult zgomot sau detalii irelevante. Acest zgomot poate afecta negativ performanța algoritmilor, în special a celor sensibili la variații sau la date irelevante. Cand spațiile caracteristice au o dimensionalitate mare, distanțele dintre puncte devin mai puțin semnificative, iar conceptul de vecini apropriați devine relativ. Acest fenomen este cunoscut sub numele de "Curse of Dimensionality."

Cea mai mare valoare pentru acuratețea medie este 91.80% și este întâlnită în cazul imaginilor redimensionate la rezoluția 64x64, care sunt clasificate cu algoritmul k-NN. Această

valoare indică faptul că modelul este eficient în clasificarea imaginilor radiografice pentru diagnosticul pneumoniei, având o capacitate semnificativă de a face distincții precise între pacienții cu pneumonie și cei sănătoși.

Pentru a îmbunătăți performanța se recomandă extinderea volumului de date de antrenament și optimizarea parametrilor modelului în conformitate cu particularitățile setului de date specific.

Avantajele metodologiei alese:

Automatizarea procesului de diagnostic: Utilizarea algoritmilor de machine learning oferă posibilitatea automatizării procesului de diagnostic, reducând astfel dependența de evaluarea manuală și accelerând timpul de răspuns în identificarea pneumoniei în imagini radiografice.

Scalabilitate și eficiență: Metodologia este scalabilă, putând gestiona o cantitate mare de date radiografii într-un mod eficient. Aceasta poate deveni o soluție viabilă în contextul unui volum crescut de imagini medicale ce necesită analiză rapidă și precisă.

Suport pentru luarea deciziilor clinice: Furnizarea unor rezultate rapide și precise poate oferi sprijin semnificativ medicilor în procesul de diagnostic și decizie. Metodologia poate servi ca un instrument de asistență în luarea deciziilor clinice.

Implicații și soluții propuse:

Reducerea erorilor umane: Prin eliminarea sau reducerea influenței subiective a evaluărilor umane, metodologia poate ajuta la diminuarea erorilor de diagnostic asociate cu interpretările subiective ale imaginilor radiologice.

Monitorizarea evoluției afecțiunilor: Sistemul poate fi extins pentru a oferi și funcționalități de monitorizare a evoluției afecțiunilor respiratorii în timp. Acest aspect poate fi deosebit de util în gestionarea pacienților și în ajustarea tratamentelor în funcție de evoluția clinică.

Posibilitatea integrării în sistemele de sănătate: Această metodologie poate fi integrată în sistemele de sănătate existente, contribuind la îmbunătățirea proceselor de diagnostic și gestionare a pacienților în cadrul unităților medicale.

Câteva rezultate ale testării: