POLITEHNICA BUCHAREST NATIONAL
UNIVERSITY FOR SCIENCE AND TECHONOLOGY

Faculty of Automatic Control and Computers
Computer Science and Engineering Department

Detectarea Automată a Tumorilor în Imagini CT Folosind Algoritmul Support Vector Machine (SVM)

Îmbunătățirea Acurateții Diagnosticului cu Ajutorul Învățării Automate (Machine Learning)

VĂDANA IOAN-GRIGORE

2023

CONTINUT

DESCRIEREA PROIECTULUI

Context și Justificare:

Proiectul se axează pe dezvoltarea unei soluții de învățare automată pentru clasificarea imaginilor Computer Tomograf (CT), cu scopul de a identifica prezența sau absența tumorilor cerebrale. Această inițiativă răspunde unei nevoi critice în domeniul diagnosticării medicale, unde precizia și rapiditatea sunt esențiale. Având în vedere volumul mare de imagini CT pe care radiologii trebuie să le analizeze, un astfel de instrument automatizat poate oferi un suport semnificativ, reducând riscul de eroare umană și accelerând procesul de diagnosticare.

Obiectivele Proiectului:

Principalul obiectiv este de a antrena un model de Machine Learning capabil să diferențieze eficient între imagini CT cu și fără tumori. Acest lucru nu doar că ar putea îmbunătăți acuratețea diagnosticării, dar ar putea și să ajute la prioritizarea cazurilor care necesită atenție medicală urgentă.

Metodologie:

Pentru atingerea acestui obiectiv, am ales să utilizez algoritmul SVM (Suport Vector Machine) datorită eficienței sale recunoscute în clasificarea datelor de înaltă dimensiune, cum sunt imaginile. Am implementat și un proces riguros de preprocesare și standardizare a datelor pentru a asigura calitatea inputului pentru modelul SVM.

```
# Train an SVM model
svm = SVC(kernel='linear')
svm.fit(X_train_scaled, y_train)
```

Importanța și Impactul Proiectului:

Implementarea cu succes a acestui model are potențialul de a transforma modul în care sunt evaluate și diagnosticate imaginile CT în context medical. Acest proiect demonstrează cum tehnologiile avansate pot fi utilizate pentru a sprijini și îmbunătăți procesele critice de îngrijire a sănătății.

ACHIZIȚIONAREA ȘI ORGANIZAREA DATELOR

Structurarea și Etichetarea Datelor:

Odată colectate, imaginile [1] (în număr de 3762) au fost structurate în directoare separate în funcție de prezența sau absența tumorilor - 'tumor' sau 'no tumor' - pentru a simplifica încărcarea și preprocesarea în etapele ulterioare ale proiectului. Această organizare a facilitat sarcina de încărcare și procesare a imaginilor în etapele ulterioare.

```
# Load images
tumor_images, tumor_labels = load_images_from_folder( folder: r'C:\Users\Grig\PycharmProjects\ML\Data\tumor',  label: 1)
no_tumor_images, no_tumor_labels = load_images_from_folder( folder: r'C:\Users\Grig\PycharmProjects\ML\Data\no_tumor',  label: 0)
```

Pregatirea seturilor de date:

Setul complet de date a fost împărțit în trei subseturi: antrenament, validare și testare. Această împărțire a fost esențială pentru evaluarea obiectivă a modelului, asigurând că acesta este antrenat și testat pe date diferite, ceea ce contribuie la evitarea suprapotrivirii și la îmbunătățirea capacității de generalizare.

```
X_train, X_temp, y_train, y_temp = train_test_split( *arrays: data, labels, test_size=0.3, random_state=42)
X_val, X_test, y_val, y_test = train_test_split( *arrays: X_temp, y_temp, test_size=0.5, random_state=42)
```

PREGĂTIREA DATELOR

Redimensionarea Imaginilor:

Toate imaginile au fost redimensionate la dimensiuni uniforme de 256x256 pixeli. Această standardizare asigură că fiecare imagine are același număr de caracteristici, facilitând procesarea ulterioară și analiza de către modelul SVM.

Normalizarea Valorilor Pixelilor:

Imaginile au fost apoi normalizate astfel încât valorile pixelilor să fie în intervalul 0-1. Acest pas este crucial pentru modelele de învățare automată, deoarece normalizează variația în intensitatea luminii și contrastul între diferite imagini.
Aplatizarea Imaginilor:

Pentru a le face compatibile cu modelul SVM, imaginile 2D au fost transformate în vectori 1D printr-un proces numit aplatizare. Aceasta este o etapă tehnică necesară, deoarece SVM-ul lucrează cu date în format vectorial

ALGORITM ȘI PARAMETRI

Am ales SVM ca algoritm de bază pentru acest proiect datorită eficienței sale recunoscute în clasificarea datelor de dimensiuni mari, precum imaginile. SVM este particular potrivit pentru acest tip de sarcină deoarece excellează în găsirea unui hiperplan optim care separă clasele de date, în cazul nostru, imaginile cu tumori de cele fără tumori.

Parametrii SVM

- Kernel SVM: Pentru acest modelul, am ales un kernel liniar, adecvat pentru datele mele care nu necesită transformări complexe ale spațiului de caracteristici. Kernelul liniar este eficient și reduce riscul de overfitting pentru seturi de date mari.

- Parametrii de Regularizare: In modelul acesta este folosită valoarea implicită din **scikit-learn**, care este 1.0. Acest parametru controlează trade-off-ul între clasificarea corectă a punctelor de antrenament și maximizarea marginii de decizie.

Standardizarea Datelor:

Pentru a optimiza performanța SVM, am standardizat setul de date folosind **StandardScaler**, care ajustează caracteristicile la o distribuție cu medie zero și deviație standard unitară.

```
# Standardize the data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_val_scaled = scaler.transform(X_val)
X_test_scaled = scaler.transform(X_test)
```

ANTRENAREA MODELULUI

Procesul de Antrenare:

Antrenarea modelului SVM este un pas crucial în acest proiect. Acest proces implică furnizarea setului de date de antrenament la model, permițându-i să învețe cum să diferențieze între imagini cu și fără tumori.

```
# Train an SVM model
svm = SVC(kernel='linear')
svm.fit(X_train_scaled, y_train)
```

Importanța Datelor de Antrenament:

Calitatea și cantitatea datelor de antrenament sunt esențiale. Prin utilizarea unui set de date divers și reprezentativ, ne asigurăm că modelul poate generaliza bine pe date noi, o caracteristică vitală pentru aplicabilitatea clinică.

Evaluarea Performanței în Timpul Antrenării:

Pe parcursul antrenării, monitorizăm performanța modelului pe setul de validare. Aceasta ne ajută să detectăm și să evităm overfitting-ul (suprapotrivirea) - o problemă comună în învățarea automată, unde modelul se potrivește prea bine cu datele de antrenament și nu performează bine pe date noi.

EVALUAREA MODELULUI

Utilizarea Setului de Validare:

Odată ce modelul este antrenat, următorul pas esențial este evaluarea performanței acestuia. Folosim setul de validare, care nu a fost implicat în antrenarea modelului, pentru a testa cât de bine poate modelul să clasifice date noi și necunoscute. Acest lucru ne oferă o estimare realistă a modului în care modelul ar putea performa în aplicații practice.

```
# Validate the model
y_val_pred = svm.predict(X_val_scaled)
print("Validation Results:")
print(classification_report(y_val, y_val_pred))
```

Analiza Performanței:

Evaluăm modelul folosind mai multe metrici, cum ar fi acuratețea, precizia, recall-ul și scorul F1. Aceste metrici ne oferă o imagine completă a performanței modelului, permițându-ne să identificăm punctele forte și aspectele care necesită îmbunătățiri.

```
print("Test Results:")
print("Test Accuracy:", accuracy_score(y_test, y_test_pred))
print(classification_report(y_test, y_test_pred))
```

Matricea de Confuzie:

Utilizăm, de asemenea, o matrice de confuzie pentru a vizualiza performanța modelului în clasificarea fiecărei clase. Aceasta ne ajută să înțelegem tipurile de erori comise de model.

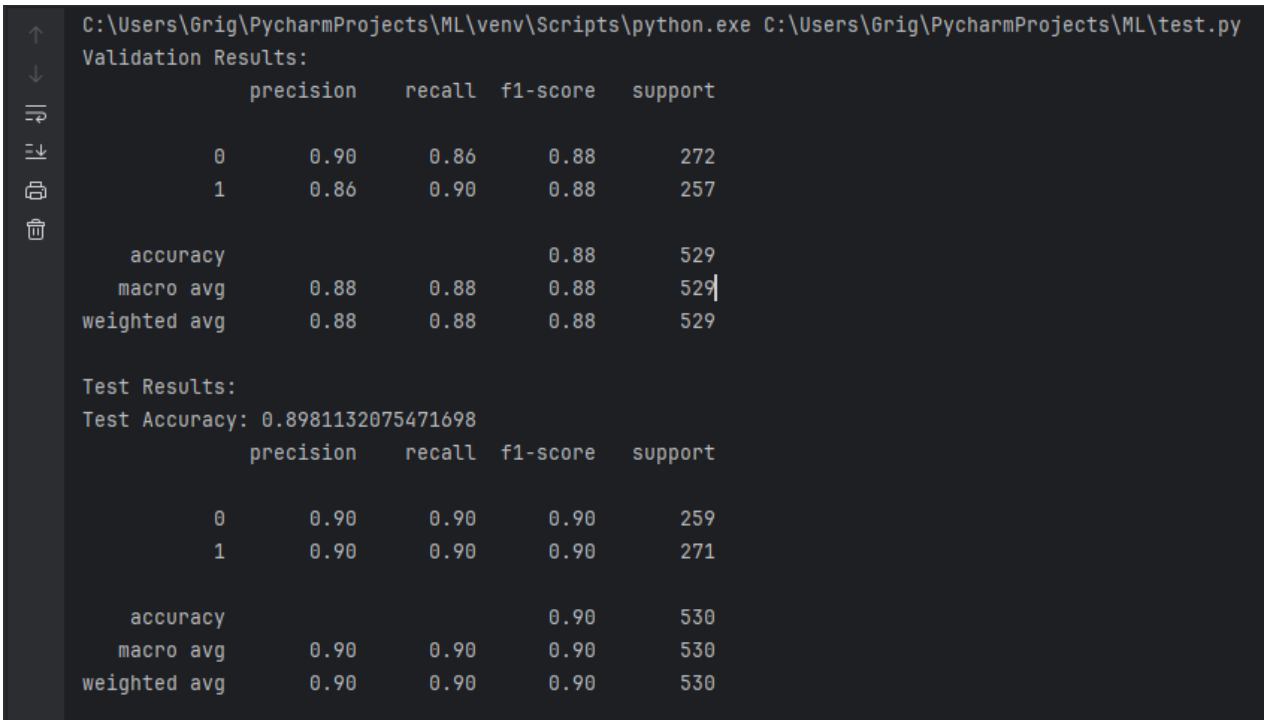
```
from sklearn.metrics import confusion_matrix
import seaborn as sns

cm = confusion_matrix(y_test, y_test_pred)
plt.figure(figsize=(10,7))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('Truth')
plt.title('Confusion Matrix')
plt.show()
```

Rezultatele evaluării ne ajută să determinăm dacă modelul este gata să fie testat pe setul de date de test sau dacă sunt necesare ajustări suplimentare.

REZULTATE

Analiza Rezultatelor de Validare si Testare:



```
C:\Users\Grig\PycharmProjects\ML\venv\Scripts\python.exe C:\Users\Grig\PycharmProjects\ML\test.py
Validation Results:
              precision    recall  f1-score   support

           0       0.90      0.86      0.88       272
           1       0.86      0.90      0.88       257

    accuracy                           0.88       529
   macro avg       0.88      0.88      0.88       529
weighted avg       0.88      0.88      0.88       529

Test Results:
Test Accuracy: 0.8981132075471698
              precision    recall  f1-score   support

           0       0.90      0.90      0.90       259
           1       0.90      0.90      0.90       271

    accuracy                           0.90       530
   macro avg       0.90      0.90      0.90       530
weighted avg       0.90      0.90      0.90       530
```

Rezultatele de Validare

Rezultatele de validare ne oferă informații despre cum și-a făcut modelul predicțiile pe datele pe care nu le-a văzut în timpul antrenamentului, dar care nu sunt complet noi (deci, datele de validare).

- **Precizie**: Pentru clasa '0' (nu tumoră), modelul a avut o precizie de 0.90, ceea ce înseamnă că 90% din cazurile pe care modelul le-a prezis ca fiind fără tumoră au fost corecte. Pentru clasa '1' (tumoră), precizia a fost de 0.86.
- **Recall**: Pentru clasa '0', recall-ul a fost de 0.86, indicând că modelul a identificat corect 86% din toate cazurile reale fără tumoră. Pentru clasa '1', recall-ul a fost de 0.90.
- **F1-score**: Scorul F1 este media armonică între precizie și recall, și este de 0.88 pentru ambele clase, indicând un echilibru bun între precizie și recall.
- **Support**: Numărul 'support' reprezintă numărul de exemple reale din fiecare clasă care au fost folosite pentru calcularea acestor metrici. Așadar, pentru clasa '0' au fost 272 de exemple, iar pentru clasa '1' au fost 257 de exemple.

Rezultatele validării indică o performanță echilibrată a modelului, cu o acuratețe generală de 88%. Precizia și recall-ul sunt aproape simetrice pentru ambele clase, sugerând că modelul nu este părtinitor față de niciuna dintre categorii. Un scor F1 de 0.88 pentru ambele clase indică o bună armonie între precizie și recall, semn că modelul gestionează bine atât cazurile pozitive, cât și cele negative.

Rezultatele de Testare

Acestea sunt rezultatele pe setul de testare, care sunt cu adevărat importante deoarece ne arată cum ar putea să performeze modelul în practică.

- **Acuratețe**: Acuratețea totală a modelului pe setul de testare a fost de aproximativ 0.89 sau 89%, ceea ce este destul de bine. Acest lucru înseamnă că, în general, modelul a făcut predicții corecte pentru 89% din cazuri.
- **Precizie, Recall, F1-score**: Aceste valori sunt foarte similare cu cele ale setului de validare, ceea ce sugerează că modelul are o performanță consistentă și nu suferă de overfitting.
- **Support**: Numărul total de exemple în setul de testare a fost de 530, împărțit aproape egal între cele două clase (259 pentru clasa '0' și 271 pentru clasa '1').

Rezultatele testului arată o ușoară îmbunătățire, cu o acuratețe totală de aproape 90%. Acest lucru confirmă robustețea modelului, arătând că poate generaliza bine pe date noi, o caracteristică esențială pentru aplicarea în scenarii reale.

Interpretare Generală

Modelul SVM pare să aibă o performanță solidă, cu o balanță bună între precizie și recall. Precizia ridicată indică faptul că, când modelul prezice prezența unei tumori, este de încredere. Un recall ridicat pentru clasa '1' este crucial în context medical, deoarece vrei să minimizezi numărul de tumori care nu sunt detectate.

Scorul F1 echilibrat pentru ambele clase sugerează că modelul nu favorizează niciuna dintre clase și că gestionează bine echilibrul între precizia și recall-ul pentru ambele.

Acuratețea globală de 89% este un rezultat bun, dar este esențial să recunoaștem că în aplicațiile medicale, chiar și un procent mic de erori poate avea consecințe serioase, deci trebuie să ne străduim să îmbunătățim aceste metrici cât mai mult posibil.

Interpretarea Metricilor Statistice:

- **Precision (Precizie):** Este măsura cât de precis modelul este atunci când face o predicție pozitivă. Pentru clasa "0"(no_tumor), o precizie de 0.90 înseamnă că 90% din imaginile pe care modelul le-a clasificat ca neavând tumori au fost clasificate corect. Pentru clasa "1"(tumor), o precizie de 0.86 înseamnă că 86% din imaginile clasificate ca având tumori au fost corect clasificate. Precizia este importantă în situații unde costul unei predicții false pozitive este mare.

Clasa Tumor: O precizie mare înseamnă că, atunci când modelul prezice prezența unei tumori, este de obicei corect. Dacă această precizie este ridicată, înseamnă că modelul este eficient în a minimiza alarmele false.

Clasa No Tumor: O precizie mare aici înseamnă că, când modelul prezice lipsa unei tumori, este adesea corect. Este important, deoarece vrem să evităm stresul și intervențiile inutile pentru pacienții care nu au tumori.

- **Recall (Sensibilitate):** Arată cât de bun este modelul în a detecta toate cazurile pozitive reale. Un recall de 0.90 pentru clasa "1" indică faptul că modelul a identificat corect 90% din toate cazurile reale de tumori din setul de date evaluat. Recall-ul este crucial în contexte unde este important să nu scăpăm niciun caz pozitiv, de exemplu, în diagnosticarea medicală.

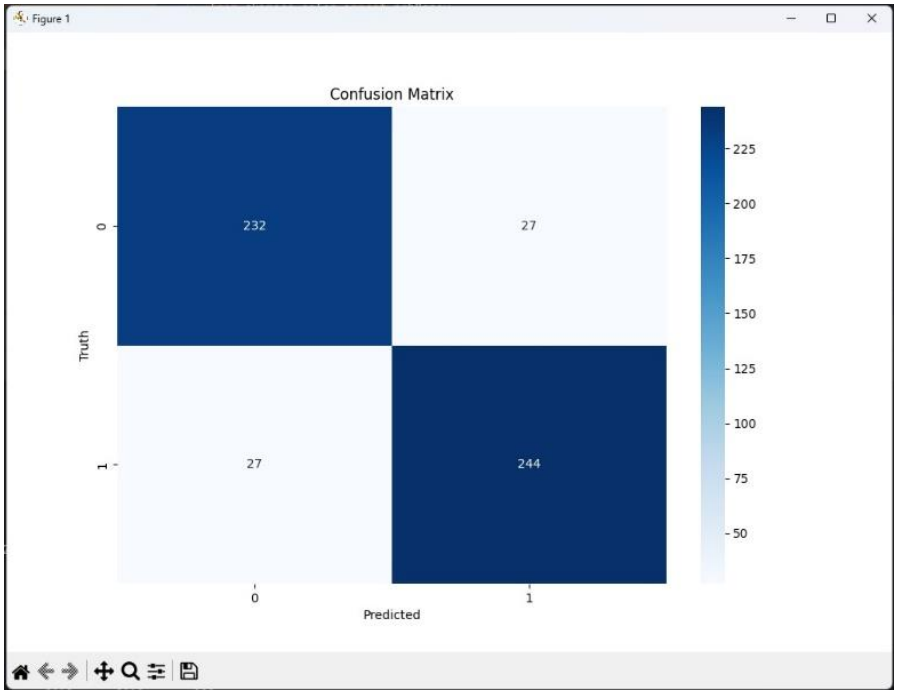
Clasa Tumor: Un recall mare sugerează că modelul identifică corect majoritatea imaginilor reale cu tumori. Acest lucru este vital pentru a nu rata diagnosticarea unei tumori reale, ceea ce ar putea avea consecințe grave pentru pacient.

Clasa No Tumor: Un recall mare indică faptul că modelul a identificat corect majoritatea imaginilor care nu au tumori. Acest lucru ajută la evitarea tratamentelor inutile sau a investigațiilor suplimentare pentru pacienții sănătoși.

- **F1-score:** Oferă o balanță între precizie și recall. Este un scor util când avem nevoie de un singur număr pentru a compara performanța modelului, mai ales atunci când distribuția claselor este neuniformă. F1-score este util atunci când costurile false pozitive și false negative sunt aproximativ echivalente.
- **Support:** Pentru fiecare clasă, "support" arată câte exemple reale (cazuri) sunt prezente în setul de date. Aici, "272" pentru clasa "0" și "257" pentru clasa "1" în setul de validare, și "259" pentru clasa "0" și "271" pentru clasa "1" în setul de testare. Aceste numere sunt folosite pentru a calcula precizia și recall-ul pentru fiecare clasă.
- **Accuracy (Acuratețe):** Reprezintă procentajul total de predicții corecte (atât pozitive, cât și negative) din totalul predicțiilor făcute. "Accuracy" oferă o viziune generală a performanței modelului, dar nu trebuie să fie singurul factor luat în considerare, mai ales când setul de date are un dezechilibru al claselor.
- **Macro Avg:** Calculul "macro average" ia media neponderată a preciziei, recall-ului și F1-score-ului pentru fiecare clasă. Aceasta nu ține cont de "support", așa că fiecare clasă contribuie egal la media finală, indiferent de cât de multe exemple are fiecare clasă în setul de date.
- **Weighted Avg:** "Weighted average" ia în considerare "support" pentru fiecare clasă, astfel încât clasele cu mai multe exemple au o greutate mai mare în calculul mediei. Acesta este adesea un reprezentant mai precis al performanței modelului pe un set de date neechilibrat.

Interpretarea imaginilor pentru performanței acestuia:

- **Matricea de Confuzie**: Arată că modelul a confundat o mică proporție de imagini pentru fiecare clasă, cu un număr egal de fals pozitive și fals negative, indicând echilibru. Acaeasta compară predicțiile modelului cu valorile adevărate (adevărul de teren) și este structurată astfel:



- **Axa verticală (Truth)**: Reprezintă clasele reale ale datelor, unde '0' corespunde imaginilor fără tumori și '1' imaginilor cu tumori.
- **Axa orizontală (Predicted)**: Reprezintă clasele prezise de model.

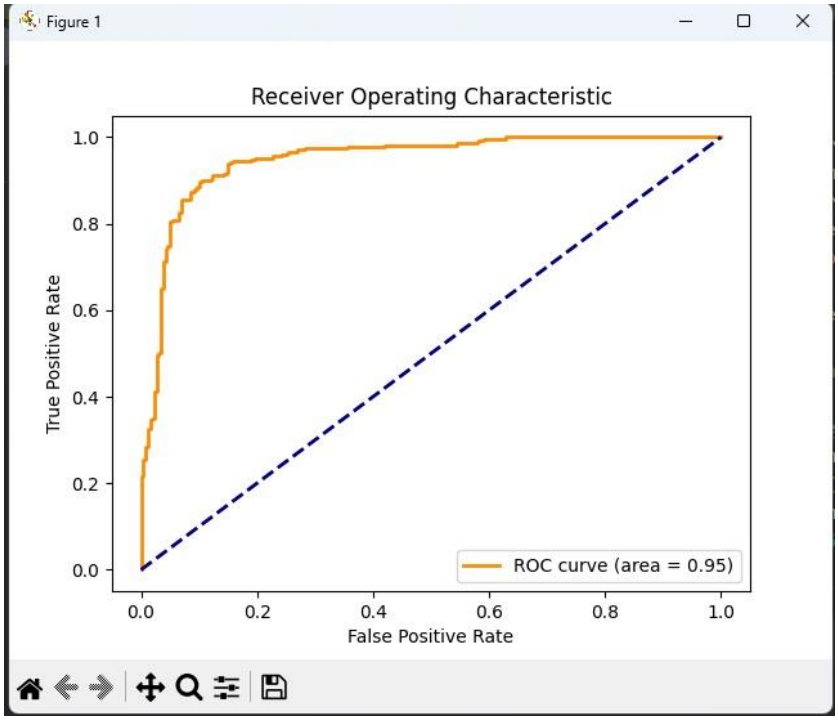
În matricea prezentată, avem patru secțiuni:

1. **True Negatives (TN)**: Poziția din stânga sus (232) reprezintă numărul de adevărate negative, adică numărul de imagini fără tumori corect identificate de model.
2. **False Positives (FP)**: Poziția din dreapta sus (27) indică numărul de false pozitive, cazuri în care modelul a prezis incorect că există o tumoră.
3. **False Negatives (FN)**: Poziția din stânga jos (27) reprezintă numărul de false negative, cazuri în care modelul a prezis incorect că nu există o tumoră.
4. **True Positives (TP)**: Poziția din dreapta jos (244) arată numărul de adevărate pozitive, adică numărul de imagini cu tumori corect identificate de model.

Interpretarea matricei de confuzie:

- **Sensibilitatea (Recall) pentru clasa '0'**: TN / (TN + FP) = 232 / (232 + 27) reprezintă probabilitatea ca modelul să clasifice o imagine fără tumoră corect.
- **Specificitatea pentru clasa '1'**: TP / (TP + FN) = 244 / (244 + 27) reprezintă probabilitatea ca modelul să clasifice o imagine cu tumoră corect.
- **Rata erorii de tip I (False Positive Rate)**: FP / (FP + TN) = 27 / (232 + 27) reprezintă probabilitatea ca modelul să clasifice greșit o imagine fără tumoră ca având o tumoră.
- **Rata erorii de tip II (False Negative Rate)**: FN / (FN + TP) = 27 / (244 + 27) reprezintă probabilitatea ca modelul să clasifice greșit o imagine cu tumoră ca fiind fără.

Curba ROC: Cu o valoare AUC (Aria de Sub Curba ROC) de 0.95, curba ROC demonstrează o performanță excelentă a modelului, cu o capacitate foarte bună de a distinge între clasele pozitive și negative.



Imaginea prezentată arată curba caracteristicii de operare a receptorului (ROC) pentru modelul de clasificare SVM.

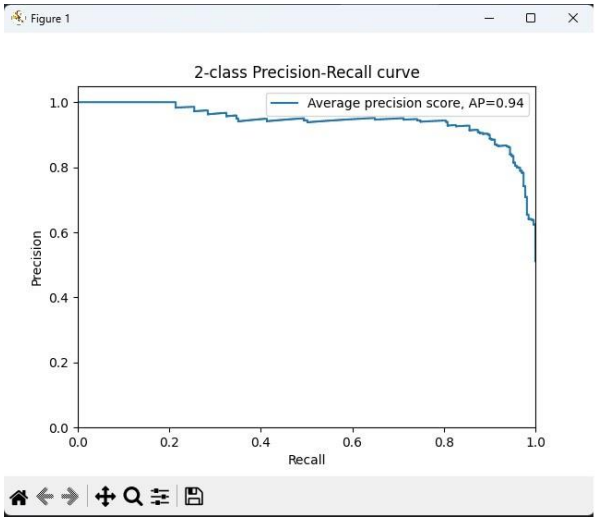
Curba ROC este un grafic care ilustrează capacitatea de diagnostic a unui sistem de clasificare binar pe măsură ce pragul de discriminare este variat. Acesta trasează Rata Adevărat Pozitivă (True Positive Rate - TPR) în funcție de Rata Fals Pozitivă (False Positive Rate - FPR) la diferite praguri de decizie.

Elementele cheie din grafic:

- **Axa X - Rata Fals Pozitivă (FPR)**: Reprezintă proporția de rezultate negative (fără tumori) care sunt incorect clasificate ca fiind pozitive (cu tumori). Cu cât această rată este mai mică, cu atât modelul este mai bun în a evita alarme false.
- **Axa Y - Rata Adevărat Pozitivă (TPR)**: Reprezintă proporția de rezultate pozitive (cu tumori) care sunt corect identificate de model. Ideal, vrem ca această rată să fie cât mai aproape de 1, indicând că modelul detectează corect toate cazurile pozitive.
- **Curba ROC**: Traseul portocaliu reprezintă performanța modelului la diferite praguri. Un model perfect ar avea o linie care merge direct în sus pe axa Y și apoi direct la dreapta pe axa X.
- **Linia Punctată Albastră**: Reprezintă o clasificare aleatorie. Orice performanță a modelului care se află deasupra acestei linii este considerată mai bună decât o alegere la întâmplare.
- **AUC (Area Under the Curve)**: Este măsura care rezumă întreaga curba ROC. Aria este de 0.95 în acest caz, ceea ce indică o performanță foarte bună a modelului. AUC variază între 0 și 1, unde 1 indică o performanță perfectă, iar 0.5 nu ar fi nicio diferență de la alegerea aleatorie.

Deci, această curba ROC și AUC de 0.95 sugerează că modelul are o capacitate excelentă de a distinge între clasele "cu tumoră" și "fără tumoră". Cu cât curba este mai apropiată de colțul din stânga sus al graficului, cu atât este mai mare capacitatea modelului de a clasifica corect pozitivele și negativele.

Curba Precision-Recall: Un scor mediu de precizie de 0.94 pe curba Precision-Recall subliniază că modelul are o rată mare de așteptare a cazurilor pozitive, un factor important în contextul medical.



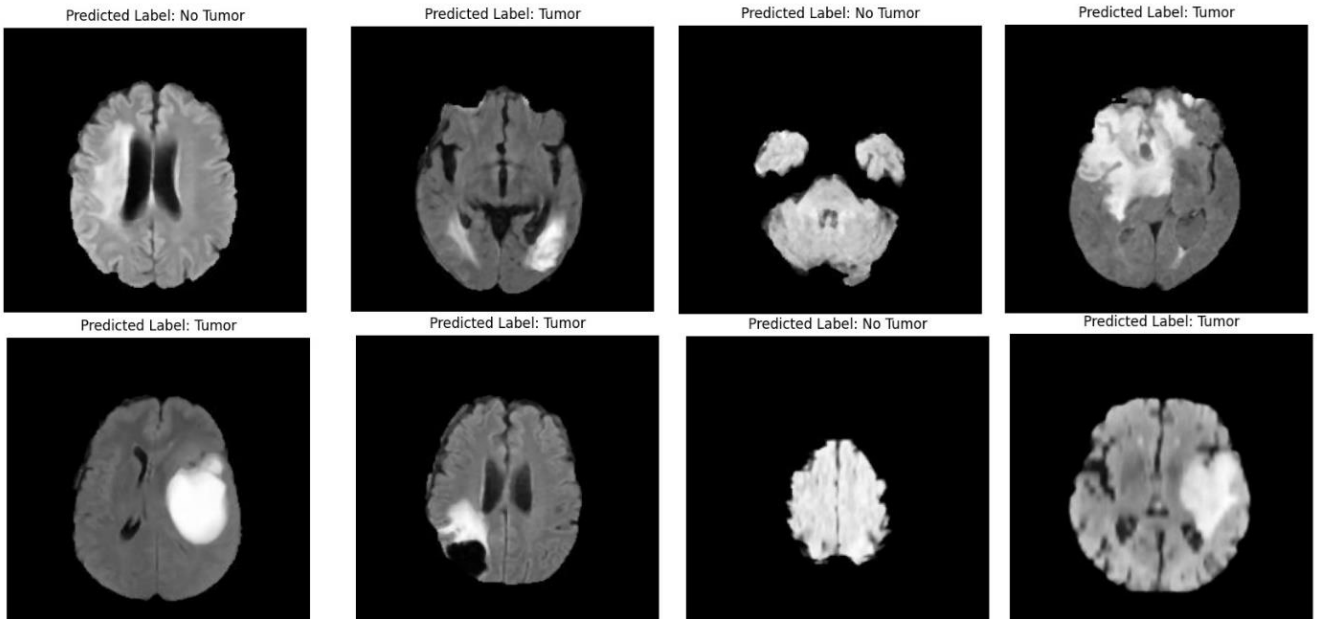
Această imagine prezintă curba Precision-Recall, care este o altă metrică folosită pentru a evalua performanța unui model de clasificare, în special în contextul în care clasele sunt dezechilibrate (adică numărul de exemple în fiecare clasă nu este aproximativ egal).

Curba Precision-Recall pune accent pe performanța modelului cu privire la clasa pozitivă (clasa '1' - prezicând prezența tumorii) și este definită astfel:

- **Axa X - Recall (Sensibilitate)**: Este aceeași cu Rata Adevărat Pozitivă de la curba ROC și măsoară proporția tuturor cazurilor pozitive reale care sunt corect identificate de model. Ideal, dorim ca această valoare să fie cât mai aproape de 1.
- **Axa Y - Precision (Precizie)**: Măsoară procentul de predicții pozitive care sunt corecte. Cu alte cuvinte, din toate cazurile pe care modelul le-a prezis ca fiind pozitive, câte sunt cu adevărat pozitive. O valoare mai mare indică un număr mai mic de alarme false.
- **Curba**: Reprezintă relația dintre precizie și recall pentru diferite praguri. O curba care se menține în apropierea valorii 1 pentru precizie indică un model care menține o rată înaltă de predicții corecte chiar și pe măsură ce încearcă să identifice toate cazurile pozitive.
- **Scorul Mediu de Precizie (Average Precision Score - AP)**: Este o singură măsură care rezumă curba Precision-Recall. Scorul AP este calculat ca media ponderată a preciziei obținute la fiecare prag, cu ponderarea dată de schimbarea recall-ului de la un prag la altul. Un scor AP mai mare indică o performanță mai bună a modelului, cu un AP=1 reprezentând un model perfect. În această imagine, AP este 0.94, ceea ce sugerează că modelul are o performanță foarte bună în clasificarea pozitivelor.

Concluzia este că, pe baza curbei Precision-Recall și a scorului AP de 0.94, modelul pare să aibă o precizie excelentă în predicțiile de clasă pozitivă, ceea ce este adesea mai important în contexte precum diagnosticarea medicală, unde este critic să se identifice toate cazurile reale pozitive.

Analiza Vizuală a Imaginilor (exemplu pentru 8 imagini):



- Primele două imagini marcate ca "No Tumor" și "Tumor" par să fie clasificate corect, bazându-ne pe etichetele prezise. Prima imagine arată un creier fără anomalii evidente, în timp ce a doua arată o anomalie întunecată care poate reprezenta o tumoră.
- Următoarele două imagini marcate ambele ca "No Tumor" arată o separare clară între țesuturile sănătoase și zonele afectate, ceea ce sugerează că modelul a învățat să identifice anumite trăsături specifice tumorilor.
- Imaginile marcate ca "Tumor" prezintă diverse forme și dimensiuni ale anomaliilor, ceea ce indică faptul că modelul nu se bazează pe un singur tipar pentru a clasifica o imagine ca prezentând o tumoră.

BIBLIOTECILE PYTHON

Pentru construirea, antrenarea și evaluarea modelului SVM, am utilizat biblioteca **scikit-learn**. Aceasta a oferit instrumentele necesare pentru a implementa SVM și a calcula metricile de performanță, precum și pentru a împărți datele și a efectua standardizarea.

Biblioteca **scikit-image** a fost esențială pentru preprocesarea imaginilor, oferind funcții pentru citirea imaginilor, conversia lor în tonuri de gri și redimensionarea lor într-un format standardizat.

NumPy este o bibliotecă fundamentală pentru calcul științific în Python, permițându-ne să lucrăm eficient cu matricile de imagini și să efectuăm operații matematice necesare în preprocesare și evaluare.

Pentru vizualizarea rezultatelor, am utilizat **Matplotlib** și **Seaborn**, care sunt biblioteci de construire a graficelor în Python. Acestea au fost folosite pentru a genera matricea de confuzie, curba ROC și curba Precision-Recall.

CONCLUZII

În acest proiect, am reușit să dezvoltăm un model SVM care clasifică cu succes imagini CT în funcție de prezența sau absența tumorilor cerebrale. Modelul a arătat o capacitate solidă de discriminare între imagini pozitive și negative, cu metrici de performanță care indică o acuratețe, precizie și recall echilibrate. Vizualizările generate, inclusiv matricea de confuzie și curbele ROC și Precision-Recall, au furnizat o confirmare vizuală a competențelor modelului.

Un aspect cheie al SVM este abilitatea sa de a găsi un hiperplan care separă eficient clasele. În cazul setului de date:

- Imaginile CT sunt reprezentate într-un spațiu multidimensional, fiecare pixel contribuind la o dimensiune.
- SVM caută să găsească un hiperplan în acest spațiu multidimensional care separă cel mai bine imaginile cu tumori de cele fără tumori.
- Conceptul de "margine" este esențial aici: SVM nu doar că separă clasele, dar încearcă să maximizeze distanța dintre cele mai apropiate puncte ale fiecărei clase și hiperplan.

Performanța Algoritmului SVM: Algoritmul Support Vector Machine (SVM) cu un kernel liniar a demonstrat o capacitate robustă de a clasifica imaginile CT în contextul prezentat. Acest lucru este indicat de scorurile înalte ale metricilor de acuratețe, precizie și recall. SVM-ul este recunoscut pentru eficiența sa în spațiile cu dimensiuni mari, cum ar fi seturile de date de imagini, ceea ce explică de ce a fost o alegere adecvată pentru acest proiect.

Bibliografie:

[1]:

https://www.kaggle.com/datasets/jakeshbohaju/brain-tumor