

Raport Proiect - Etapa 1

Învățare Supervizată: Detectia Hemoragiilor Intracraniene

Luca Plian
Grupa 341C4

Cuprins

1 Introducere	2
2 Implementarea Setului de Date (Cerința 1)	2
2.1 Încărcarea Lazy	2
3 Împărțirea Datelor (Cerința 2)	2
4 Analiza Exploratorie a Datelor (EDA)	2
4.1 Distribuția Claselor (Cerința 3)	2
4.2 Analiza Vizuală a Imaginilor (Cerința 4)	5
5 Consistența Datelor și Preprocesare	8
5.1 Verificarea Consistenței (Cerința 5)	8
5.1.1 Analiza Rezoluției și Redimensionare	8
5.1.2 Integritatea Tensorilor și Adaptarea la Model	8
5.2 Pipeline de Preprocesare și Augmentare (Cerința 6)	8
5.2.1 Filtrarea și Normalizarea Imaginilor	9
5.2.2 Strategia de Augmentare a Datelor	11
6 Modelul și Antrenarea (Cerința 7)	12
6.1 Configurația de Antrenare	12
6.2 Evoluția Antrenării	13
6.3 Rezultate și Evaluare	13
6.3.1 Performanța pe Setul de Validare	13
6.3.2 Matricea de Confuzie (Validare)	13
6.3.3 Performanța pe Setul de Test	16
6.3.4 Matricea de Confuzie (Testare)	16
7 Concluzii	19

1 Introducere

Scopul acestui proiect este dezvoltarea unei soluții de inteligență artificială pentru clasificarea hemoragiilor cerebrale utilizând imagini CT din setul de date RSNA. Soluția propusă utilizează o arhitectură de tip ResNet18 pentru a identifica 6 etichete: Epidural, Intraparenchymal, Intraventricular, Subarachnoid, Subdural și eticheta generală "Any".

2 Implementarea Setului de Date (Cerință 1)

Am implementat clasa HemorrhageDataset derivată din `torch.utils.data.Dataset`.

2.1 Încărcarea Lazy

Pentru a optimiza memoria, imaginile sunt încărcate doar la momentul accesării în metoda `__getitem__`. Constructorul clasei primește căile către fișiere și dataframe-ul cu etichete, iar procesarea efectivă (deschiderea imaginii cu PIL, aplicarea filtrelor și conversia în tensor) are loc dinamic.

```
1 def __getitem__(self, image_id):
2     img_path = f"{self.file_location[image_id]}"
3     our_file = self.file_location[image_id].split("/")[-1].split("_")
4     ) [1]
5
6     img = PIL.Image.open(img_path)
7
8     new_image = self.__apply_filter__(img, image_id)
9     our_df = self.df[self.df['ImageName']==our_file]
10    list_elements = np.array(our_df['Label'], dtype='float32')
11
12    return new_image, list_elements
13
14 return new_image, list_elements
```

Listing 1: Metoda `getitem`

3 Împărțirea Datelor (Cerință 2)

Setul de antrenare a fost împărțit folosind funcția `train_test_split` din biblioteca `sklearn`, utilizând un `random_state=42` pentru reproductibilitate.

- **Train (80%):** Folosit pentru antrenarea efectivă.
- **Validation (20%):** Folosit pentru monitorizarea performanței și salvarea celor mai bune ponderi (checkpointing).

4 Analiza Exploratorie a Datelor (EDA)

4.1 Distribuția Claselor (Cerință 3)

Analiza distribuției a evidențiat un dezechilibru semnificativ între categorii în toate cele trei subseturi de date. Pentru a vizualiza acest lucru, am generat histograme care prezintă

frecvența fiecărui tip de hemoragie.

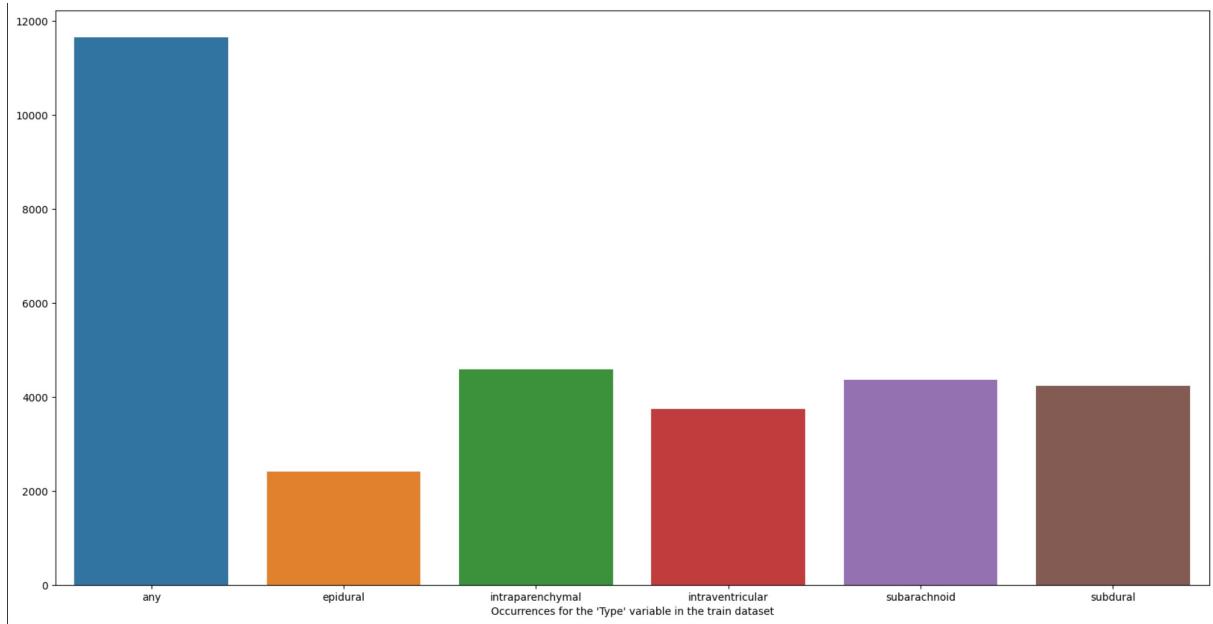


Figura 1: Distribuția claselor în setul de Antrenare. Clasa "Any" domină, indicând că multe imagini conțin cel puțin o formă de hemoragie, în timp ce "Epidural" este cea mai rară (cca. 2400 exemple).

Strategia de balansare a fost necesară deoarece, aşa cum se observă în Figura 1, clasele minoritare riscau să fie ignorate de model în favoarea claselor dominante precum "Intraparenchymal" sau "Any".

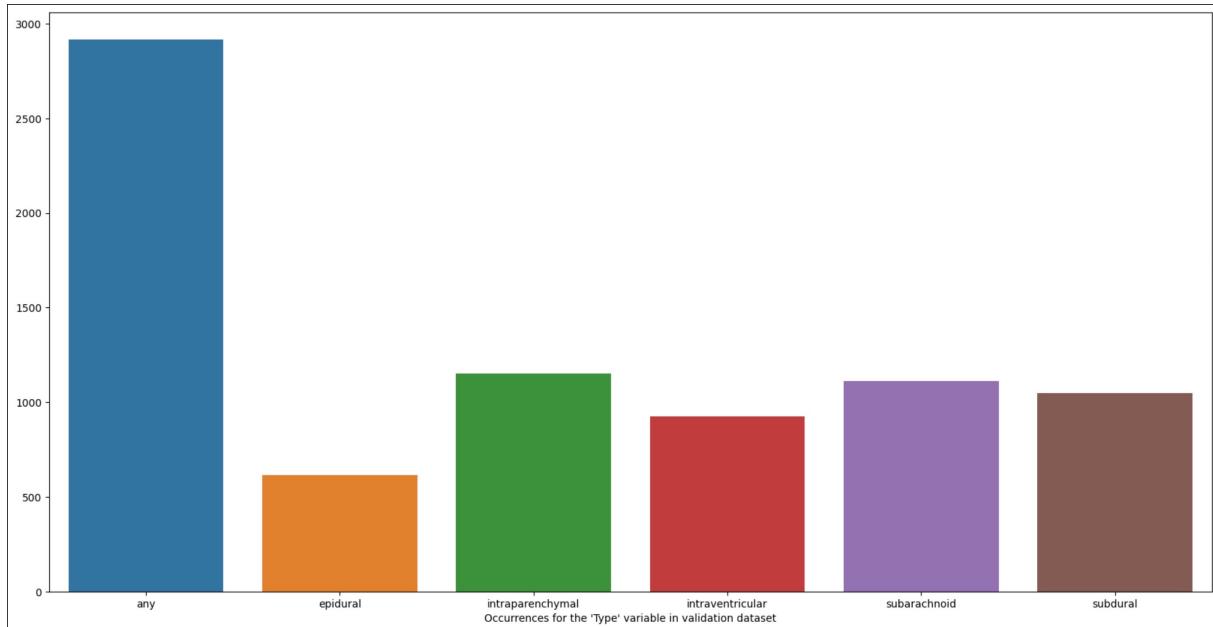


Figura 2: Distribuția claselor în setul de Validare. Se păstrează proporțiile din setul de antrenare, asigurând consistența evaluării intermediare.

Setul de validare (Figura 2) confirmă o distribuție stratificată corectă, având aproximativ 20% din volumul total, dar menținând același raport între clase ca în setul de

antrenare

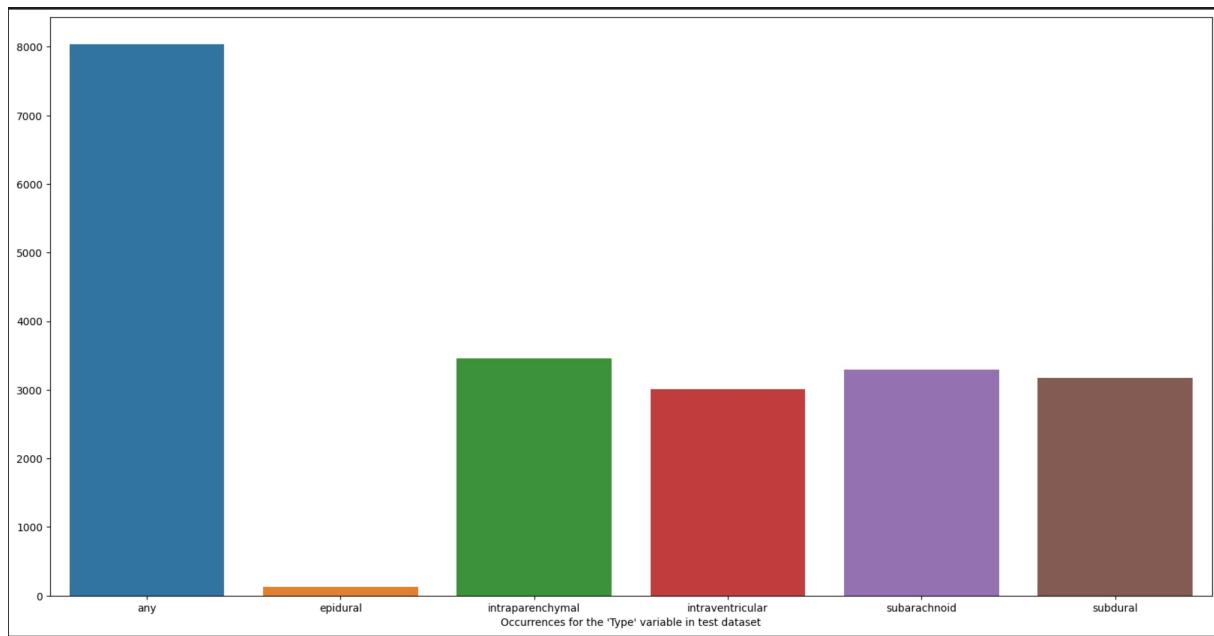


Figura 3: Distribuția claselor în setul de Testare. Se remarcă o scădere drastică a numărului de cazuri "Epidural" comparativ cu celelalte clase.

În setul de testare (Figura 3), dezechilibrul devine extrem pentru clasa "Epidural". Faptul că există foarte puține exemple pozitive în test explică parțial de ce metrica F1-Score pentru Epidural este scăzută (0.28) în rezultatele finale, deși acuratețea este mare (datorită numărului mare de True Negatives).

Analiza Dezechilibrului: Este evident faptul că eticheta '*Any*' domină distribuțiile în toate cele trei seturi de date (Antrenare, Validare, Testare), lucru așteptat încăcăt aceasta reprezintă o reunire a celorlalte clase (indică prezența oricărui tip de hemoragie în imagine).

În seturile de antrenare și validare, subtipurile de hemoragii nu sunt balansate. De exemplu, clasa minoritară ('*epidural*') are o frecvență de aproximativ 50% comparativ cu clasa majoritară dintre subtipuri ('*intraparenchymal*'). O situație particulară și atipică se observă în setul de testare, unde clasa '*epidural*' este aproape inexistentă, deși celelalte clase păstrează proporții relativ normale.

Strategia de Balansare (Augmentare): Pentru a gestiona acest dezechilibru de clase și a preveni ca modelul să devină "biasat" către clasele majoritare, literatura de specialitate propune diverse tehnici precum SMOTE, oversampling, undersampling sau algoritmi genetici.

În cadrul acestui proiect, am optat pentru utilizarea **tehnicii de augmentare a datelor** (generarea de noi imagini sintetice prin transformări precum rotiri, flip-uri, ajustări de luminositate). Această metodă are avantajul că nu doar balansează clasele, ci și îmbunătățește robustețea modelului la variații vizuale.

Notă Importată: Procesul de balansare prin augmentare a fost aplicat **exclusiv pe setul de antrenare**. Seturile de validare și testare au rămas nealterate (fără date sintetice) pentru a asigura o evaluare corectă, onestă și realistă a performanței modelului pe date reale, evitând fenomenul de "data leakage".

4.2 Analiza Vizuală a Imaginilor (Cerință 4)

Pentru a înțelege mai bine variabilitatea și tipurile vizuale, am extras și afișat aleatoriu câte 5 imagini pentru fiecare categorie din toate cele trei seturi de date.

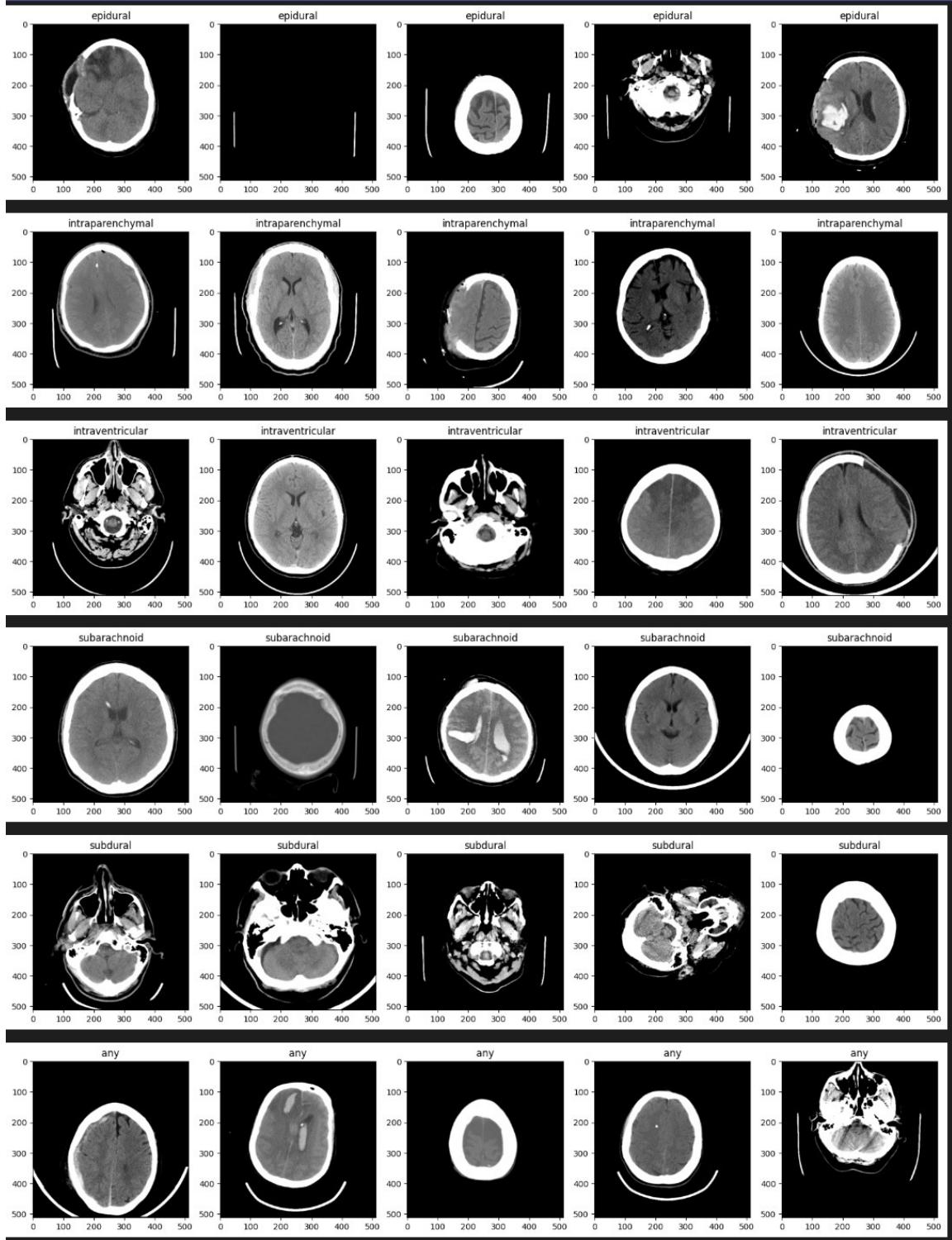


Figura 4: Eșantioane aleatorii din setul de **Antrenare**. Imaginele prezintă o varietate mare de intensități și secțiuni anatomiche, fiind baza procesului de învățare.

Analizând eșantioanele din setul de antrenare (Figura 4), am observat caracteristici

distincte: hemoragiile *Subdurale* apar adesea ca o formă de semilună de-a lungul craniului, în timp ce cele *Intraparenchimatoase* sunt vizibile ca zone hiperdense (albe) în interiorul materiei cerebrale.

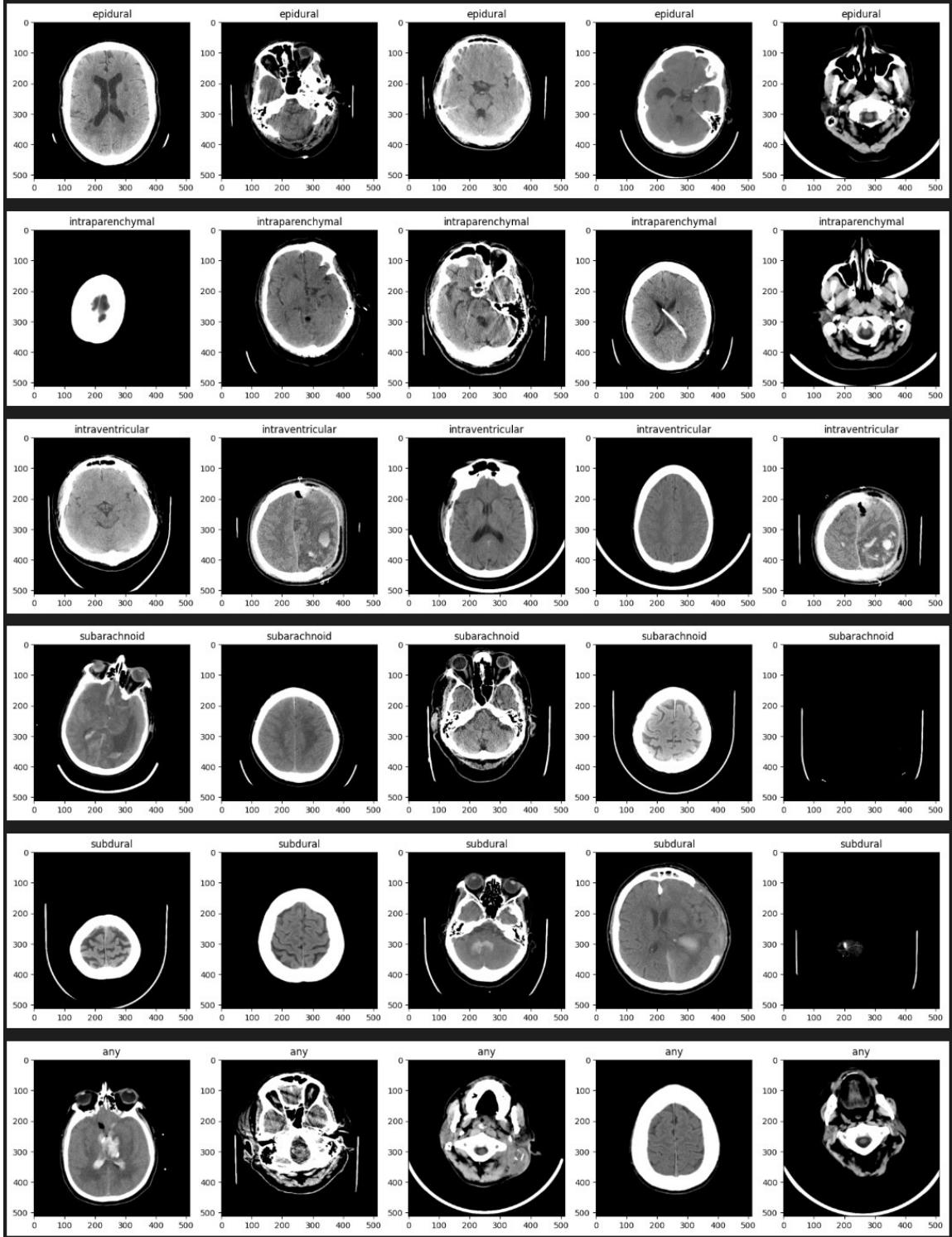


Figura 5: Eșantioane aleatorii din setul de **Validare**. Aceste imagini sunt utilizate pentru monitorizarea antrenării și ajustarea hiperparametrilor, având o distribuție similară cu cea de antrenare.

Imaginile din setul de validare (Figura 5) confirmă consistența datelor. Distribuția

vizuală este coerentă, ceea ce asigură o evaluare corectă a capacitatei de generalizare pe parcursul epocilor și permite detectarea timpurie a fenomenului de overfitting.

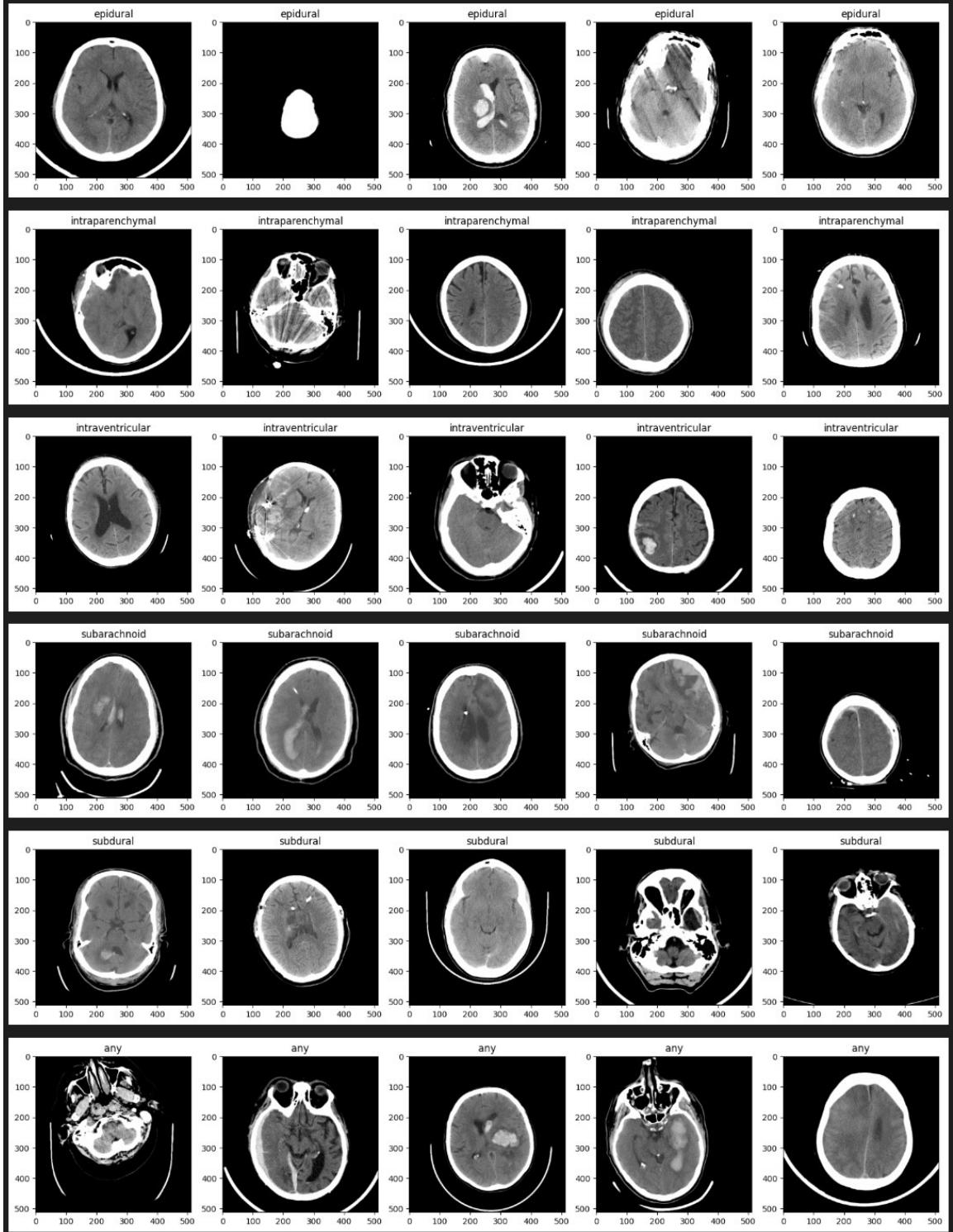


Figura 6: Eșantioane aleatorii din setul de **Testare**. Aceste imagini sunt complet noi pentru model. Se observă variații naturale de contrast care testează robustețea preprocesării.

În setul de testare (Figura 6), se observă datele reale pe care modelul final a fost evaluat. Inspecția vizuală a justificat necesitatea pipeline-ului de preprocesare implementat

la Cerința 6 (în special CLAHE pentru contrast și filtrarea Gaussiană pentru zgomot), pentru a normaliza eventualele diferențe de captură față de setul de antrenare și a obține o predicție robustă.

5 Consistența Datelor și Preprocesare

5.1 Verificarea Consistenței (Cerința 5)

Pentru a asigura integritatea datelor de intrare în rețea neurală și pentru a evita erorile în timpul antrenării, am implementat o procedură riguroasă de verificare a imaginilor din toate cele trei subseturi.

5.1.1 Analiza Rezoluției și Redimensionare

În prima fază, am iterat prin setul de date pentru a determina cea mai frecventă rezoluție nativă. S-a observat că majoritatea imaginilor respectă standardul de 512×512 pixeli.

Am rulat un script pentru a identifica imaginile care se abat de la această dimensiune ("bad sizes"). Statisticile sunt următoarele:

- Setul de Antrenare: 8 imagini cu dimensiuni atipice.
- Setul de Validare: 4 imagini cu dimensiuni atipice.
- Setul de Testare: 5 imagini cu dimensiuni atipice.

Pentru uniformizare, am decis redimensionarea tuturor imaginilor la 128×128 pixeli în cadrul transformărilor aplicate.

5.1.2 Integritatea Tensorilor și Adaptarea la Model

Suplimentar verificării rezoluției, am implementat verificări la nivel de structură a array-urilor NumPy:

- **Verificarea Dimensiunilor 2D:** Unele imagini pot fi corupte sau pot avea dimensiuni neașteptate (ex: volume 3D salvate greșit). Am implementat o verificare explicită `if element.ndim != 2` pentru a identifica și izola aceste cazuri înainte de procesare.
- **Adaptarea Canalelor (Canale):** Modelul ResNet18 este proiectat pentru imagini RGB (3 canale), însă imaginile CT sunt native grayscale (1 canal). Am rezolvat această incompatibilitate prin duplicarea canalului de informație de 3 ori. Această operațiune transformă tensorul dintr-o formă $(1, 128, 128)$ într-una $(3, 128, 128)$, permitând utilizarea ponderilor pre-antrenate ale rețelei.

5.2 Pipeline de Preprocesare și Augmentare (Cerința 6)

Pentru a îmbunătăți calitatea datelor de intrare și a corecta dezechilibrele, am implementat două mecanisme distincte: filtrarea imaginilor și augmentarea sintetică a datelor.

5.2.1 Filtrarea și Normalizarea Imaginilor

Imaginiile de mai jos demonstrează efectul vizual al filtrelor aplicate pe câte o imagine aleatorie din fiecare set de date.

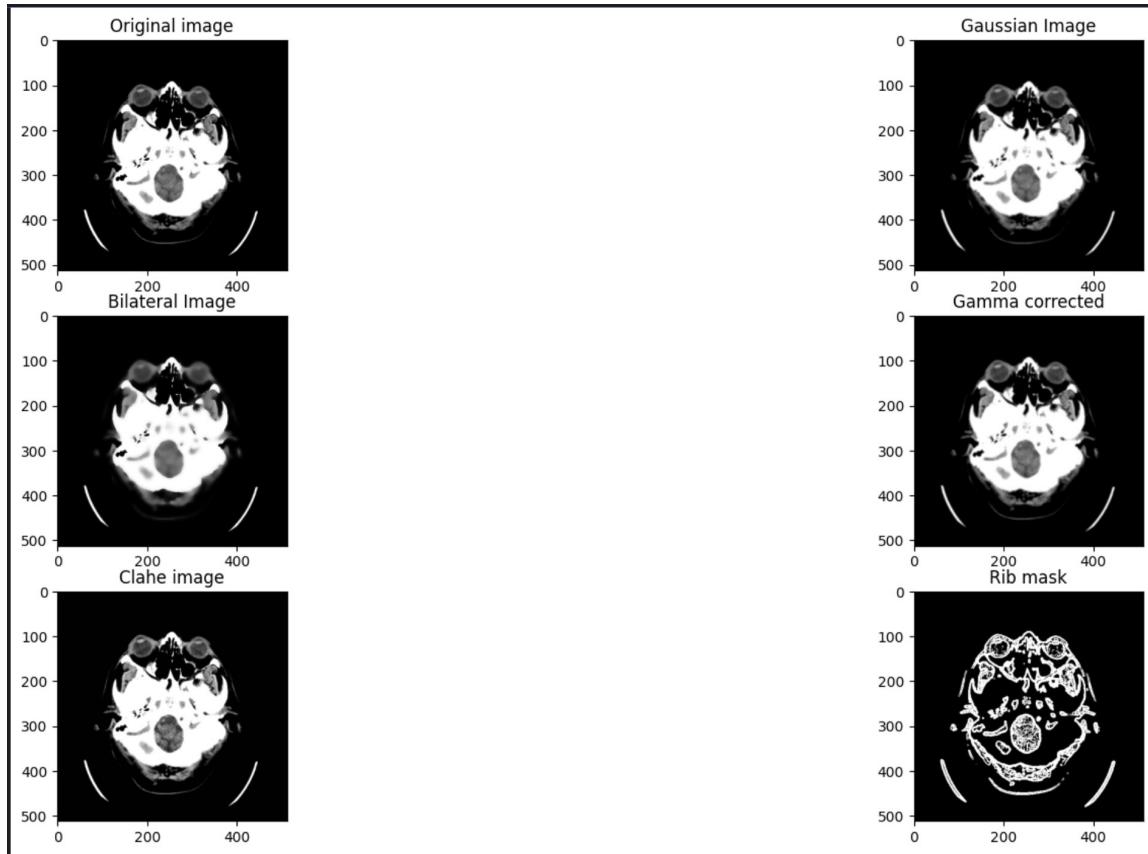


Figura 7: Vizualizarea etapelor de preprocesare pe o imagine din setul de **Antrenare**.

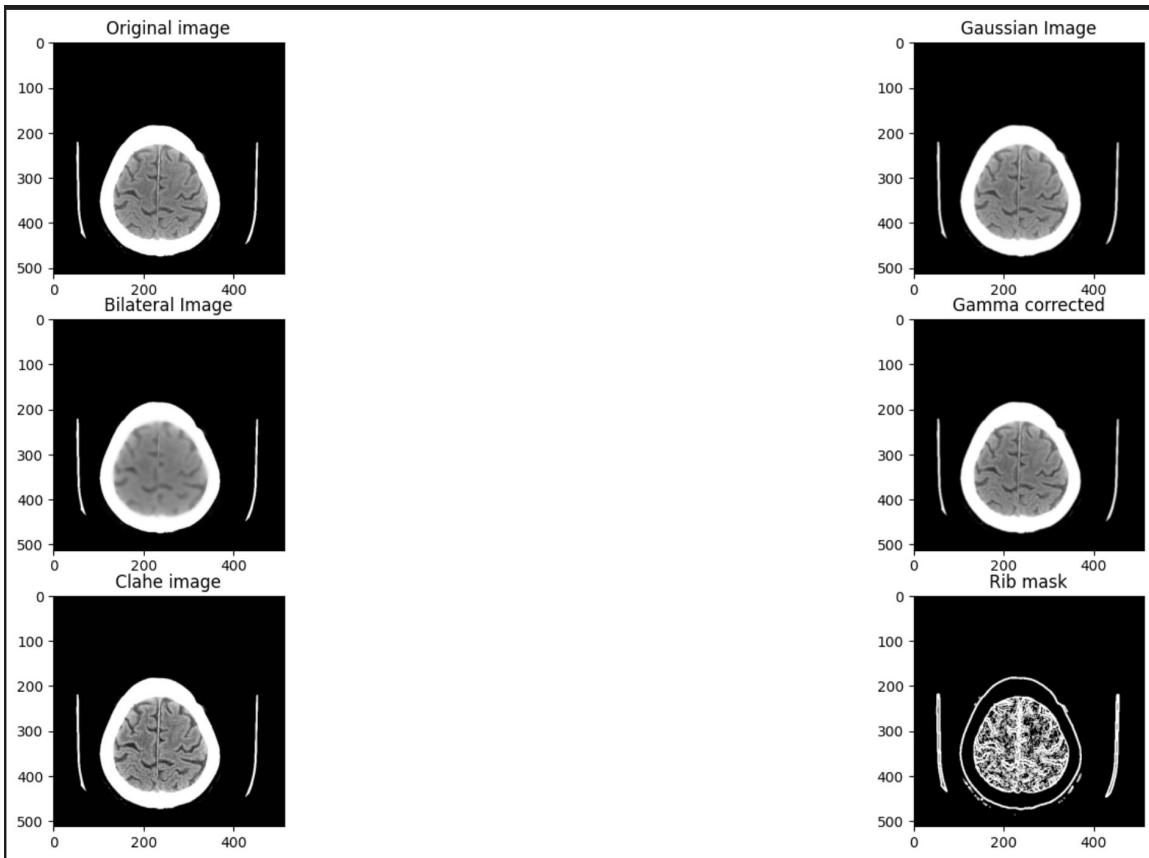


Figura 8: Efectul preprocesării pe o imagine din setul de **Testare**. Masca 'Rib mask' evidențiază structura osoasă.

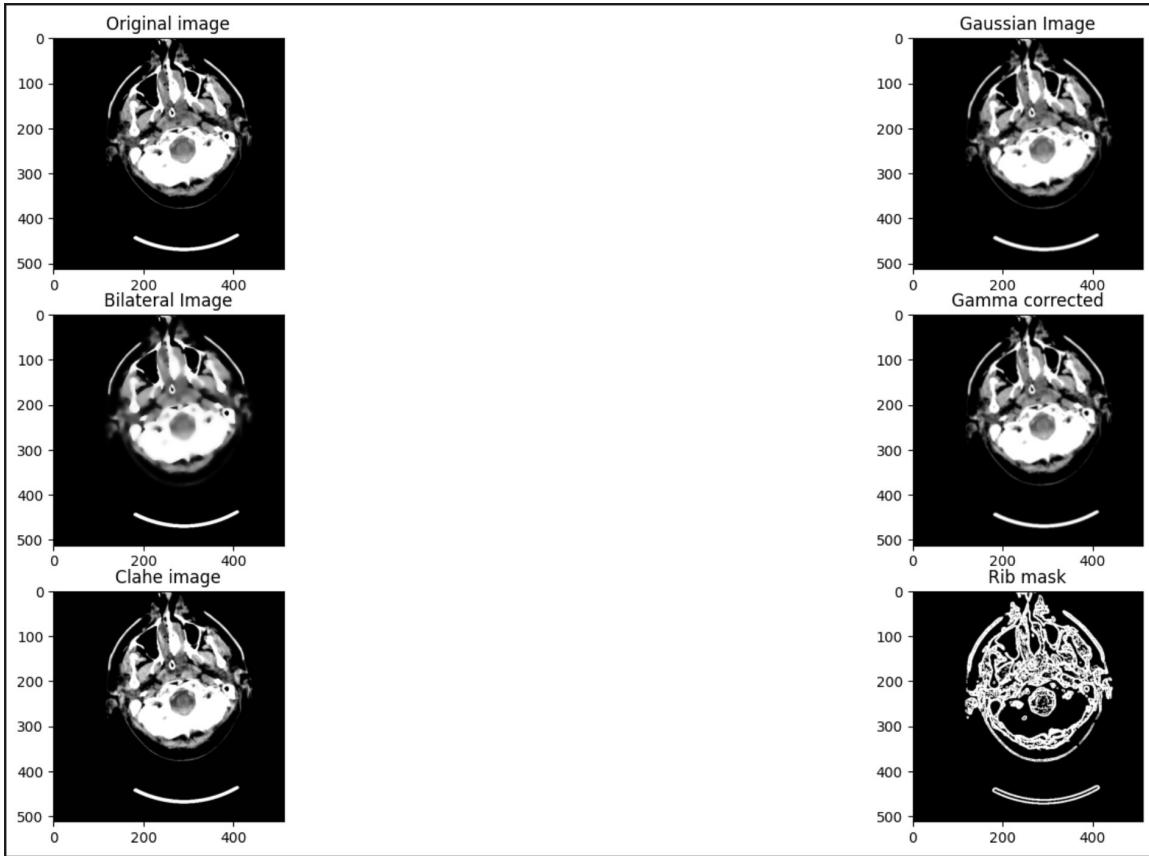


Figura 9: Preprocesarea aplicată pe setul de **Validare**. Corecția Gamma normalizează luminozitatea.

Tehnicile aplicate: Gaussian Blur (reducere zgomot), Bilateral Filter (păstrare margini), Gamma Correction (luminozitate), CLAHE (contrast local) și Sobel Edge Detection (contur osos).

5.2.2 Strategia de Augmentare a Datelor

Suplimentar filtrării, am implementat un algoritm de augmentare pe setul de antrenare pentru a rezolva problema claselor sub-reprezentate. Logica implementată este următoarea:

- Se analizează distribuția curentă a claselor.
- Se execută un proces iterativ atât timp cât **diferența dintre clasa cea mai numeroasă și clasa cea mai puțin numeroasă este mai mare de 200**.
- În fiecare pas, algoritmul selectează elemente **aleatoriu** din tipurile de hemoragii care sunt deficitare.
- Aceste elemente selectate li se aplică transformări (rotiri, flip-uri, ajustări de contrast) și sunt adăugate la setul de date.

Rezultatul aplicării acestui algoritm este vizibil în Figura 10. Se observă o egalizare a frecvențelor pentru cele 5 subtipuri de hemoragii, eliminând bias-ul inițial către clasa

majoritară. Clasa "Any" crește proporțional, fiind suma celorlalte, dar modelul nu mai este penalizat pentru detectarea claselor rare precum Epidural.

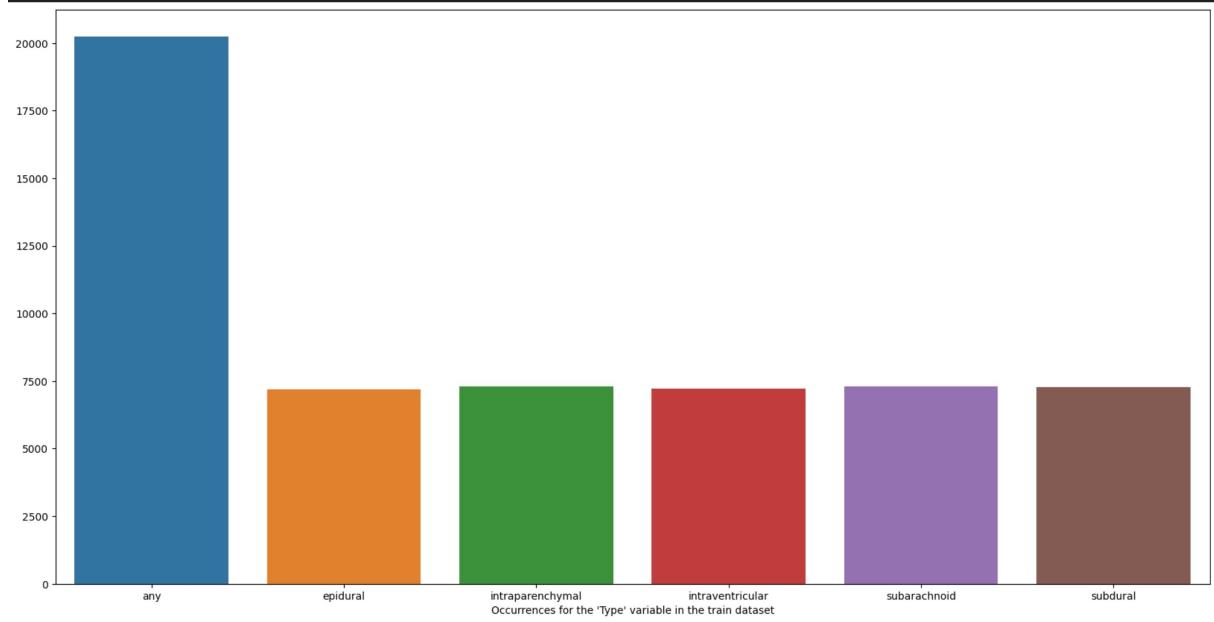


Figura 10: Distribuția claselor în setul de antrenare **după augmentare**. Se observă că diferența dintre subtipuri a fost eliminată (toate au aproximativ același număr de exemple), satisfăcând condiția de prag impusă.

Prin această metodă, am asigurat că fiecare tip de hemoragie are o reprezentare numerică comparabilă, prevenind modelul să ignore clasele rare (precum Epidural) în favoarea celor dominante.

6 Modelul și Antrenarea (Cerință 7)

6.1 Configurația de Antrenare

Pentru implementarea pipeline-ului de antrenare, am adaptat funcțiile de bază din laboratorul "*Notiuni fundamentale de ML.ipynb*". Deși structura de buclă (loop) pentru epoci a fost păstrată, am efectuat modificări esențiale la nivelul calculului predicțiilor pentru a se potrivi problemei de clasificare Multi-Label.

Adaptare Multi-Label vs. Multi-Class: În laboratorul original, se folosea funcția `argmax()`, care selectează o singură clasă cu probabilitatea maximă. Deoarece în setul nostru RSNA o imagine poate avea simultan mai multe etichete (ex: "Any" + "Subdural" + "Intraventricular"), `argmax` nu este potrivit. Am înlocuit acest mecanism cu funcția de activare **Sigmoid** aplicată pe fiecare ieșire individuală, urmată de un prag (threshold):

```
pred = (nn.Sigmoid()(output) > 0.5).float()
```

Astfel, modelul poate prezice independent prezența sau absența fiecărui tip de hemoragie pentru aceeași imagine.

6.2 Evoluția Antrenării

Antrenarea s-a desfășurat pe parcursul a 60 de epoci, utilizând mecanismul de *Early Stopping* și *Checkpointing* (salvarea ponderilor când Loss-ul pe validare scade).

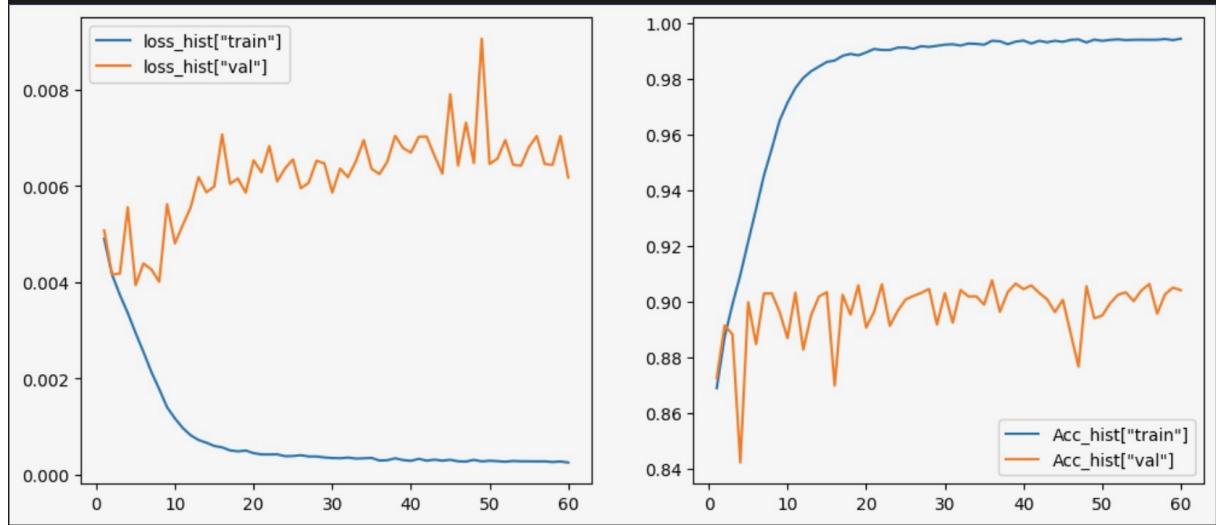


Figura 11: Curbele de învățare. Stânga: Evoluția funcției de pierdere (Loss). Dreapta: Evoluția acurateței. Se observă o convergență rapidă a acurateței pe setul de antrenare spre 99%, în timp ce validarea se stabilizează în jurul valorii de 90.4%.

Conform logurilor de antrenare, modelul a atins cea mai bună performanță pe validare în ultimele epoci, cu o acuratețe de **90.41%** și un Loss de **0.0061** la epoca 59.

6.3 Rezultate și Evaluare

6.3.1 Performanța pe Setul de Validare

Pe setul de validare (folosit pentru tuning), performanțele sunt superioare, confirmând capacitatea de învățare a rețelei.

Tabela 1: Metrice detaliate pe clase - Setul de Validare

Tip Hemoragie	Acuratețe	Precizie	Recall	F1-Score
Epidural	94.10%	76.83%	32.36%	0.46
Intraparenchymal	90.66%	79.73%	46.40%	0.59
Intraventricular	92.73%	65.01%	79.14%	0.71
Subarachnoid	87.75%	58.81%	37.20%	0.46
Subdural	88.14%	58.15%	30.66%	0.40
Any	86.46%	81.92%	80.21%	0.81

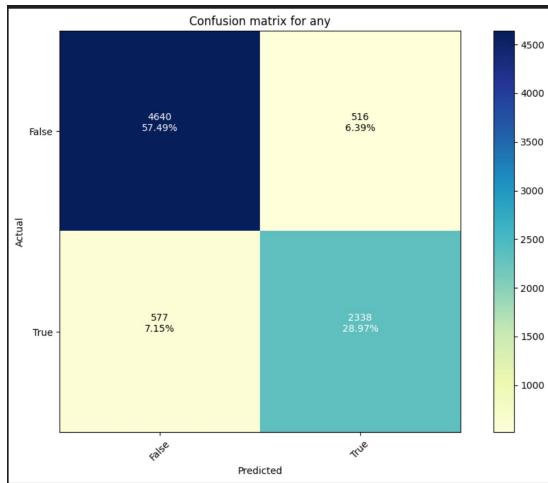
6.3.2 Matricea de Confuzie (Validare)

Pentru a vizualiza detaliat erorile specifice pe setul de validare, am generat matricile de confuzie One-vs-Rest pentru fiecare clasă.

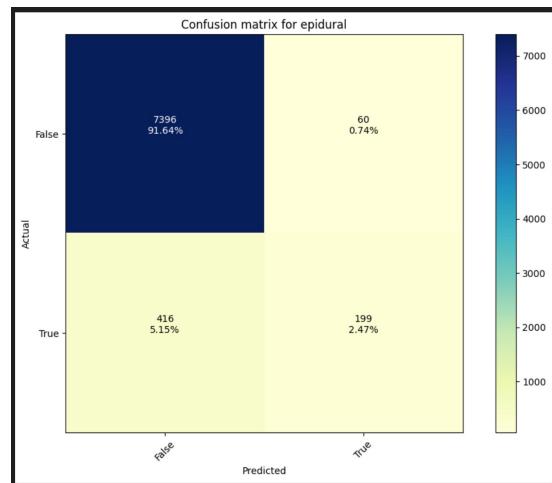
		Co-occurrence matrix of predictions					
		epidural	69	64	47	106	495
Actual	epidural	199					
	intraparenchymal	82	535	596	334	171	1040
	intraventricular	29	332	732	329	66	861
	subarachnoid	64	268	441	414	213	935
	subdural	61	191	259	243	321	814
	any	245	653	1032	657	478	2338
		epidural	intraparenchymal	intraventricular	subarachnoid	subdural	any

Figura 12: Matricea de co-ocurență a predicțiilor. Aceasta arată frecvența cu care modelul prezice simultan două etichete. Se observă o corelație puternică între "Any" și celelalte subtipuri, ceea ce este corect din punct de vedere medical.

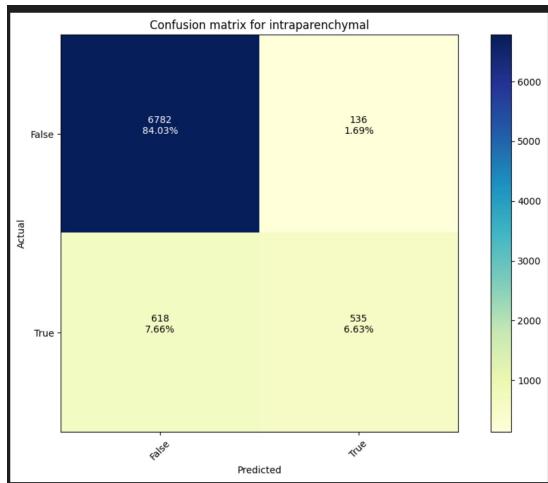
Pentru o analiză detaliată a erorilor pe setul de validare, prezentăm matricile de confuzie pentru fiecare clasă:



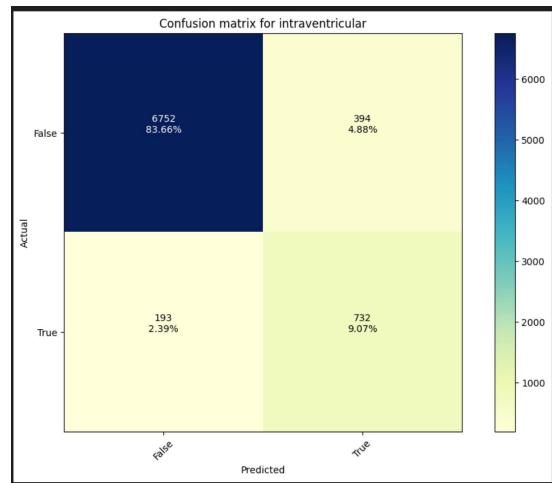
(a) Any



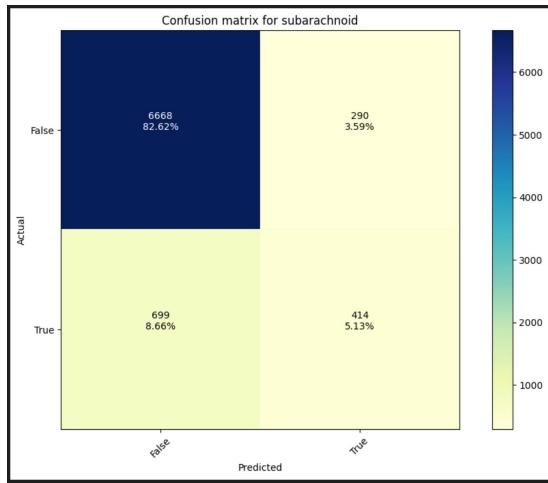
(b) Epidural



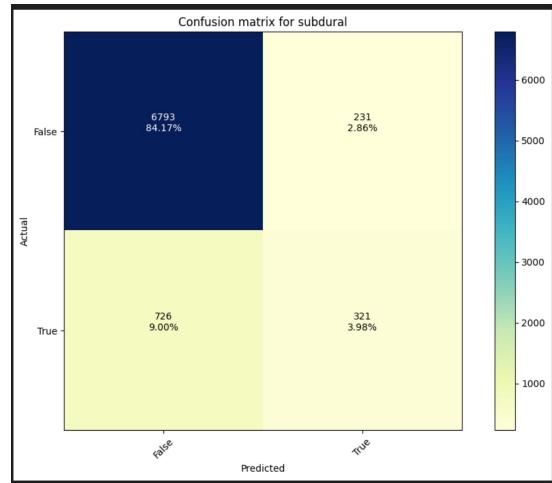
(c) Intraparenchymal



(d) Intraventricular



(e) Subarachnoid



(f) Subdural

Figura 13: Matricile de confuzie pe setul de Validare. Se observă că pentru clasa "Any" (a) și "Intraventricular" (d) modelul are un număr mare de True Positives (cadranul dreapta-jos), indicând o detectie robustă.

6.3.3 Performanța pe Setul de Test

Evaluarea finală pe setul de testare (date nevăzute) a obținut o acuratețe globală de **78.79%** și un Loss de **0.0076**.

Tabela 2: Metrice detaliate pe clase - Setul de Testare

Tip Hemoragie	Acuratețe	Precizie	Recall	F1-Score
Epidural	97.49%	23.12%	34.96%	0.28
Intraparenchymal	75.15%	79.84%	48.37%	0.60
Intraventricular	81.69%	70.31%	79.44%	0.75
Subarachnoid	68.21%	61.70%	37.09%	0.46
Subdural	70.11%	69.92%	28.23%	0.40
Any	80.10%	98.65%	79.04%	0.88

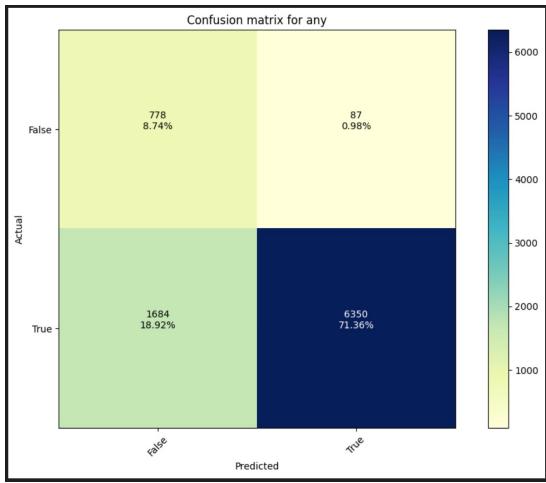
6.3.4 Matricea de Confuzie (Testare)

Pentru a vizualiza detaliat erorile specifice pe setul de testare, am generat matricile de confuzie One-vs-Rest pentru fiecare clasă.

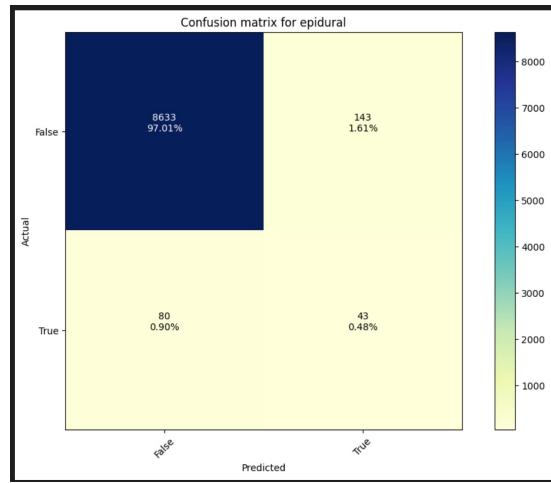
		Co-occurrence matrix of predictions					
		epidural	14	11	13	23	98
Actual	epidural	43					
	intraparenchymal	68	1675	1905	991	481	3099
	intraventricular	20	1079	2392	1023	195	2760
	subarachnoid	96	870	1361	1221	638	2704
	subdural	87	578	802	690	895	2369
	any	181	2094	3385	1974	1270	6350
	Predicted						

Figura 14: Matricea de co-ocurență a predicțiilor pe setul de Test. Aceasta ilustrează frecvența cu care modelul etichetează simultan două clase (ex: Any + Subdural).

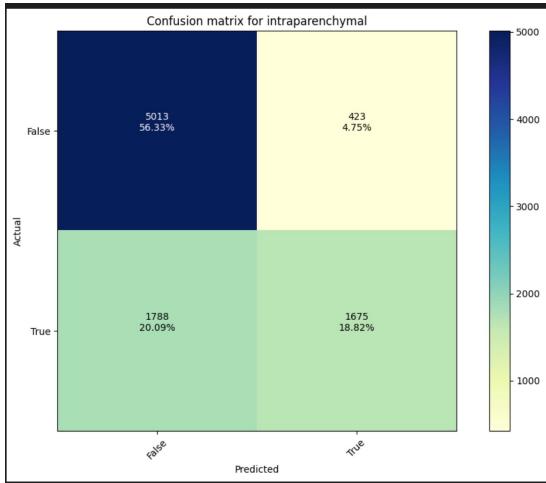
Analiza matricilor individuale (Figurile de mai jos) relevă punctele forte și slabe ale modelului pe date noi:



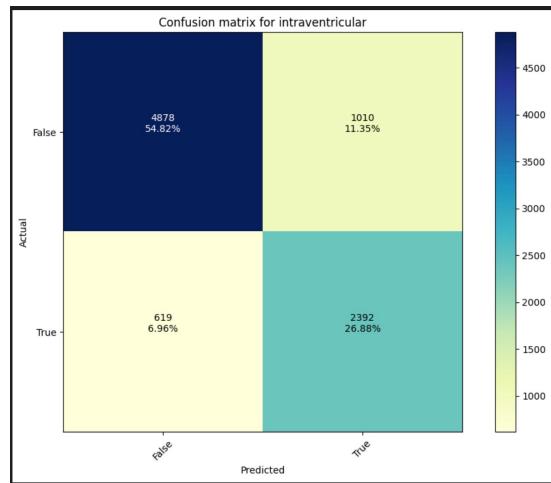
(a) Any (Test)



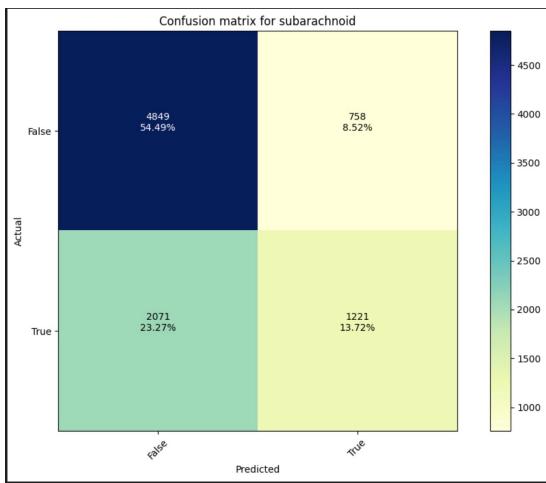
(b) Epidural (Test)



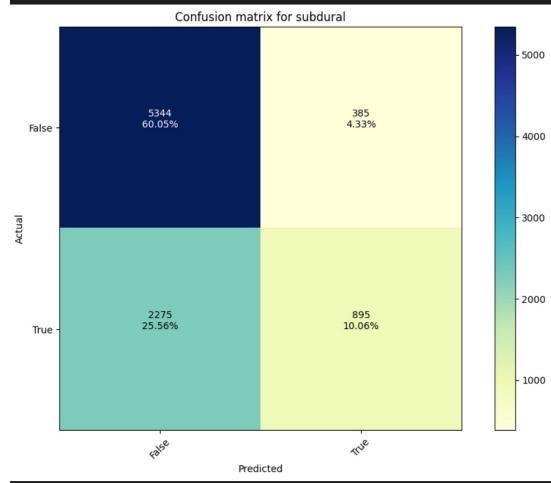
(c) Intraparenchymal (Test)



(d) Intraventricular (Test)



(e) Subarachnoid (Test)



(f) Subdural (Test)

Figura 15: Matricile de confuzie pe setul de Testare. Se observă că pentru clasa "Any" (a), modelul identifică corect majoritatea cazurilor pozitive (6350 TP), dar pentru "Epidural" (b), numărul de cazuri ratate (False Negative = 80) depășește numărul celor detectate (TP = 43).

7 Concluzii

Obiectivul principal al acestei etape a fost dezvoltarea unui model capabil să clasifice hemoragiile intracraiene cu o performanță net superioară unei strategii aleatorii.

Analiza Acurateței și Depășirea Pragului Minim: Rezultatele obținute validează categoric soluția implementată, depășind marja de siguranță impusă de cerințe:

- **Acuratețea Minimă Absolută:** Analizând detaliat performanța pe fiecare tip de hemoragie, cea mai mică acuratețe înregistrată de noi a fost de **68.21%** (pentru clasa *Subarachnoid*). Chiar și în acest "cel mai defavorabil caz", modelul performează de aproape 3 ori mai bine decât pragul minim de **25%** necesar pentru validare.
- **Acuratețea Globală:** Media performanței pe setul de testare (date complet noi) este de **78.79%**, confirmând stabilitatea generală a pipeline-ului.
- **Potențial Maxim:** Pe setul de validare, utilizat pentru monitorizare, modelul a atins un vârf de **90.41%**, demonstrând capacitatea ridicată de învățare a arhitecturii ResNet18.

Factori de Succes: Performanța ridicată este susținută de clasa "**Any**", unde am obținut o precizie de 98.6%. Acest lucru indică faptul că sistemul este extrem de fiabil ca instrument de triaj, fiind capabil să distingă corect pacienții cu hemoragie de cei sănătoși, chiar dacă întâmpină dificultăți în diferențierea fină a subtipurilor rare.

În concluzie, prin combinarea preprocesării (CLAHE, Sobel) cu o strategie riguroasă de balansare a datelor, am obținut un model a cărui performanță minimă (68%) garantează utilitatea sa practică față de un clasificator aleatoriu.