Clasificarea tumorilor cerebrale din imagini RMN utilizând transfer learning, module de atenție și interpretabilitate prin Grad-CAM și segmentare slab supervizată prin GrabCut

Abstract

Clasificarea tumorilor cerebrale în imagini RMN reprezintă o provocare majoră în imagistica medicală, având impact direct asupra diagnosticului, tratamentului și evoluției pacienților. În această lucrare propunem un cadru complet de clasificare bazat pe transfer learning, îmbunătățit prin trei module originale de atenție și rafinare spațială: Mixed Attention Block, Residual Path Head și Spatial Boundary Refinement, integrate în modele CNN pre-antrenate (MobileNetV2, EfficientNet-B0 și VGG16). Complementar, introducem un Vision Transformer simplificat, utilizat pentru comparație. Metoda este completată de un pipeline de interpretabilitate bazat pe Grad-CAM și segmentare slab supervizată prin GrabCut.

Rezultatele experimentale arată o creștere semnificativă a performanței (Accuracy și Macro-F1), o robustețe superioară obținută prin ensemble-uri și o explicabilitate ridicată prin utilizarea segmentării ghidate. Lucrarea demonstrează că arhitecturile CNN îmbunătățite cu module ușoare de atenție, combinate cu mecanisme de interpretabilitate vizuală, pot constitui un instrument valoros în context clinic.

Clasificare tematică

AMS MSC: 68T07 - Artificial neural networks and deep learning
ACM Computing Reviews Categories: Computing methodologies → Artificial intelligence → Computer vision → Computer vision problems → Object identification / Object detection / Image segmentation

1 Introducere

Clasificarea tumorilor cerebrale din imagini RMN reprezintă un domeniu central în imagistica medicală asistată de inteligență artificială, având impact direct asupra diagnosticului, planificării tratamentului și monitorizării clinice. RMN-ul este modalitatea principală de imagistică pentru tumorile cerebrale, datorită contrastului excelent între țesuturile

moi, însă interpretarea acestora necesită expertiză avansată și este susceptibilă la variabilitate inter- și intra-observator. Din acest motiv, metodele automate bazate pe învățarea profundă câștigă importanță, oferind diagnostic mai rapid, reproductibil și potențial mai precis.

1.1 Background și formularea problemei

Problema abordată în această lucrare constă în clasificarea automată a imaginilor RMN în patru categorii clinice: glioma, meningioma, tumoare pituitară și absența tumorilor. Această sarcină este dificilă deoarece tumorile cerebrale prezintă o diversitate considerabilă în ceea ce privește forma, dimensiunea, textura și localizarea lor. Gliomele, de exemplu, pot avea margini difuze și contururi greu de identificat, în timp ce meningiomele sunt adesea bine delimitate. Tumorile pituitare au aspect distinct, însă pot fi confundate în anumite secvențe cu alte tipuri de leziuni.

Seturile de date medicale sunt în mod inerent mici și dezechilibrate, ceea ce face dificilă generalizarea modelelor. Mai mult, preprocesarea și augmentarea trebuie ajustate cu grijă pentru a evita introducerea de artefacte care ar putea induce erori de clasificare.

1.2 Importanța problemei

Diagnosticul corect al tipului tumoral este esențial, deoarece fiecare categorie necesită tratament distinct, iar întârzierea diagnosticării poate duce la agravarea stării clinice. Astfel, un sistem automat care poate sprijini radiologii în interpretarea imaginilor poate reduce semnificativ timpul de analiză, poate crește precizia și poate contribui la introducerea unor fluxuri de lucru standardizate în spitale și clinici.

1.3 Lucrări conexe (Related Work)

Transfer learning în imagistică medicală.

Transfer learning-ul reprezintă strategia dominantă în clasificarea imaginilor medicale, din cauza lipsei dataseturilor mari etichetate. Modelele pre-antrenate pe ImageNet (VGG16 [1], ResNet [2], MobileNetV2 [3], EfficientNet [4]) sunt adaptate ulterior pentru sarcini medicale precum analiza tumorilor cerebrale.

MobileNetV2, introdus de Sandler et al. [3], este eficient din punct de vedere computațional, având o arhitectură bazată pe inverted residual blocks și linear bottlenecks.

EfficientNet, propus de Tan și Le [4], scalează optim arhitectura printr-un coeficient compus.

Modele CNN și Vision Transformers în imagistica cerebrală.

Vision Transformers (ViT) [5] modelează dependențe globale, însă necesită seturi mari pentru performanță ridicată.

Metode de interpretabilitate și segmentare slab supervizată.

Interpretabilitatea este esențială în medicina asistată de AI. Grad-CAM [6] este standardul pentru vizualizarea hărților de atenție. GrabCut [7] este folosit pentru segmentare slab supervizată.

Limitele literaturii existente și probleme nerezolvate

Din analiza lucrărilor conexe, reies câteva limitări:

- modelele CNN standard nu captează suficient informația spațială globală;
- modulele existente cresc numărul de parametri;
- contururile tumorale sunt rar exploatate;
- ViT-urile necesită seturi mari neaccesibile medical;
- segmentarea bazată pe CAM nu este integrată complet în pipeline.

Partea originală a abordării propuse

Lucrarea introduce o arhitectură CNN extinsă cu:

- Mixed Attention Block;
- Residual Path Head;
- Spatial Boundary Refinement.

Acestea cresc performanța fără cost computațional excesiv.
De asemenea introducem:

- un Vision Transformer minimal;
- un ensemble soft-voting;
- un ensemble CNN + ML (LR, SVM);
- pipeline Grad-CAM + GrabCut.

1.4 Research questions

- **RQ1:** Modulele Mixed Attention, ResPath și SBR cresc Macro-F1 în comparație cu backbone-urile standard?
- **RQ2:** Ensemble-urile îmbunătățesc robustețea și acuratețea sistemului?
- **RQ3:** Grad-CAM + GrabCut produce segmentări interpretabile clinic?
- **RQ4:** Fine-tuning-ul parțial este superior față de antrenarea exclusivă a capului de clasificare?
- **RQ5:** ViT-ul simplificat poate concura cu CNN-urile ușoare pe seturi moderate?

1.5 Cunoștințe noi introduse

Lucrarea contribuie prin:

- trei module originale de atenție și rafinare;
- un pipeline complet de segmentare slab supervizată;
- un ensemble hibrid CNN+ML;
- un ViT minimal pentru comparație;
- o analiză comparativă completă (CNN vs ViT, module vs fără module, ensemble vs modele singulare, ensemble cu clasificatori vs ensemble simplu);
- integrarea interpretabilității în procesul de evaluare.

1.6 Ce urmează în articol

Articolul continuă cu prezentarea abordării originale (secțiunea 2), validarea experimentală (secțiunea 3), rezultate și concluzii (secțiunea 4), urmate de un set de referințe (bibliografie).

2 Descrierea abordării originale

Abordarea propusă în această lucrare îmbină în mod coerent elemente avansate de transfer learning, module originale de atenție și rafinare spațială, arhitecturi moderne de rețele neuronale, tehnici de interpretabilitate și mecanisme de segmentare slab supervizată. Obiectivul principal al acestei secțiuni este prezentarea detaliată a întregului cadru metodologic utilizat, justificarea alegerilor arhitecturale și evidențierea contribuției originale în raport cu literatura de specialitate.

2.1 Viziunea generală a abordării

Modelul propus pleacă de la rețele CNN pre-antrenate pe ImageNet, alese pentru eficiența lor computațională și capacitatea de generalizare pe seturi mici. Acestea includ MobileNetV2, EfficientNet-B0 și VGG16, fiecare servind ca backbone pentru extragerea reprezentărilor spațiale ale imaginilor RMN.

Pe aceste arhitecturi se adaugă trei module originale concepute pentru a rafina activările intermediare, a crește sensibilitatea la contururi și a îmbunătăți focalizarea asupra regiunilor relevante pentru clasificare:

- **Mixed Attention Block (MAB)**, un modul de atenție spațială globală bazat pe proiecții Q/K/V 1×1;
- **Residual Path Head (ResPath)**, un traseu rezidual suplimentar către capul de clasificare;

- **Spatial Boundary Refinement (SBR)**, un modul care combină hărțile de margini cu activările CNN.

Acest ansamblu formează o arhitectură plug-and-play, ce poate fi integrată ușor în diverse backbone-uri, fără o creștere semnificativă în numărul de parametri.

2.2 Backbone-urile folosite

MobileNetV2

MobileNetV2 este ales datorită raportului excelent între acuratețe și eficiență. Arhitectura sa folosește inverted residual blocks și linear bottlenecks, ceea ce permite obținerea unor reprezentări compacte, ideale pentru dispozitive medicale portabile sau sisteme cu resurse limitate.

EfficientNet-B0

EfficientNet-B0 scalează uniform adâncimea, lățimea și rezoluția printr-un coeficient compus și este cunoscut pentru performanțe superioare în raport cu numărul de parametri. Acest model extrage caracteristici robuste indiferent de dimensiunea datasetului.

VGG16

VGG16 este inclusă datorită simplității arhitecturale și a rolului ei istoric în transfer learning. În ciuda numărului mare de parametri, performează surprinzător de bine în contexte medicale.

Vision Transformer (ViT) simplificat

Pentru comparație, este utilizat un Vision Transformer minimal, construit pe principiile introduse de Dosovitskiy et al. ViT procesează imaginea împărțind-o în patch-uri, transformând problema într-o secvență pentru un encoder Transformer. Deși ViT necesită în general seturi mari, în lucrare este utilizat într-o formă simplificată pentru a evalua potențialul său pe date medicale moderate.

2.3 Modulele originale propuse

Contribuția centrală a lucrării constă în elaborarea a trei module originale, integrate post-backbone.

2.3.1 Mixed Attention Block (MAB)

MAB este un modul de atenție spațială globală ce folosește proiecții Q/K/V (query, key, value) realizate prin convoluții 1×1. Aceste componente sunt multiplicate pentru a captura relevanța globală a fiecărei zone a imaginii. Ulterior, activările rezultate sunt combinate cu harta originală printr-o cale reziduală, stabilizând antrenarea.

Acest modul îmbunătățește capacitatea rețelei de a identifica zone relevante ale tumorilor, permite captarea dependențelor spațiale, fără să adauge cost computațional semnificativ.

2.3.2 Residual Path Head (ResPath)

ResPath introduce o cale reziduală suplimentară între ultimele straturi convoluționale și capul final de clasificare. Această cale facilitează propagarea gradientului, reduce pierderea informațională și creează o legătură directă între extragerea caracteristicilor și decizia finală.

Rezultatul este o îmbunătățire observabilă a convergenței și o redresare a situațiilor în care backbone-ul este prea rigid în fine-tuning.

2.3.3 Spatial Boundary Refinement (SBR)

SBR este un modul inovator ce combină activările intermediare ale rețelei cu o hartă de margini generată prin filtre. Această fuziune permite evidențierea contururilor tumorale, în special în cazul gliomelor cu margini difuze.

Prin rafinarea activărilor în zonele de tranziție între tumoare și țesut sănătos, rețeaua devine mai sensibilă la structurile anatomice subtile.

2.4 Pipeline-ul complet propus

Întregul pipeline include:

- încărcarea datelor și preprocesarea (redimensionare, normalizare, augmentare);
- extragerea caracteristicilor prin backbone;
- rafinarea caracteristicilor prin modulele MAB, ResPath și SBR;
- clasificarea în patru clase;
- generarea hărților Grad-CAM pentru interpretabilitate;
- binarizarea hărților de atenție;
- segmentarea slab supervizată ghidată de GrabCut;
- evaluarea modelelor individuale și ensemble;
- compararea CNN-urilor modificate cu ViT-ul minimal;
- compararea ensemble-ului simplu cu un ensemble peste care se aplică clasificatori (LR, SVM).

Pipeline-ul este complet automatizat și poate fi aplicat oricărui set RMN similar.

2.5 Modelarea matematică și formularea problemei

Setul de date este formalizat astfel:

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \qquad x_i \in \mathbb{R}^{224\times224\times3}, \quad y_i \in \{1, 2, 3, 4\}.$$

Modelul transformă o imagine într-o distribuție de probabilitate peste clase:

$$f_\theta : \mathbb{R}^{224\times224\times3} \to \Delta^3, \qquad p_\theta(y \mid x) = \mathrm{softmax}(f_\theta(x)).$$

Funcția de pierdere folosită este cross-entropy:

$$\mathcal{L}(\theta) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{4} \mathbf{1}[y_i = c] \log p_\theta(y = c \mid x_i).$$

2.6 Algoritmii și metodele utilizate

Pe lângă modulele originale și backbone-uri, sunt utilizate:

- augmentări moderate pentru robustețe;
- warm-up și fine-tuning parțial;
- callback-uri precum *ReduceLROnPlateau*, *EarlyStopping* și *ModelCheckpoint*;
- ensemble soft-voting între modelele cu performanță ridicată;
- ensemble hibrid CNN+ML (extractor de feature-uri + Logistic Regression / SVM);
- Grad-CAM pentru interpretabilitate;
- GrabCut pentru segmentare slab supervizată.

2.7 Diferențiere față de literatura existentă

Abordarea propusă se diferențiază prin:

- integrarea simultană a trei module de atenție optimizate pentru arhitecturi mobile;
- combinația dintre Grad-CAM și GrabCut într-un pipeline complet pentru segmentare slab supervizată;
- analiza comparativă CNN vs ViT minimal;
- construcția unui ensemble hibrid CNN+ML;
- faptul că modulele sunt plug-and-play și scalabile.

3 Validare experimentală

Validarea experimentală urmărește evaluarea performanței sistemului propus, a modulelor originale introduse, a ensemble-urilor și a pipeline-ului de interpretabilitate. Această secțiune prezintă experimentele desfășurate pe un set de date complet și pe un subset de date redus, ilustrarea etapelor metodologice pe un exemplu simplu, evaluarea calitativă a Grad-CAM și GrabCut, precum și rezultatele măsurate prin Accuracy și Macro-F1.

3.1 Setul de date utilizat

Setul de date utilizat în această lucrare conține imagini RMN cerebrale aparținând celor patru clase principale: *glioma*, *meningioma*, *pituitary* și *no_tumor*. Aceste imagini provin din colecții publice de imagistică medicală, prelucrate astfel încât să fie compatibile cu rețelele folosite.

Fiecare imagine este redimensionată la 224×224 pixeli, păstrând raportul de aspect pentru a evita deformările anatomice.

Imaginile sunt normalizate pe canalele RGB și sunt augmentate prin:

- rotații moderate de până la ± 0.05 radiani,
- zoom de 0.1,
- translații de 0.05,
- ajustări de contrast de ± 0.1.

Augmentarea ajută la extinderea artificială a datasetului și la creșterea robustetii modelului.

Împărțirea datasetului se realizează stratificat, în proporție:

- 80% pentru antrenare,
- 20% pentru validare și testare.

3.2 Experiment simplu pe subset artificial pentru ilustrarea pipeline-ului

Pentru a demonstra clar funcționarea metodologiei propuse, este utilizat un subset restrâns de date, similar celui folosit în notebook-ul de dezvoltare. Acest subset conține câteva zeci de imagini din fiecare clasă, suficient pentru a ilustra comportamentul modelelor într-un cadru controlat.

Pe acest subset mic se aplică toate etapele pipeline-ului:

- încărcarea și etichetarea imaginilor;
- preprocesare (redimensionare, normalizare);
- augmentări minimale pentru a preveni supraînvățarea;

- antrenarea modelului în faza de warm-up cu backbone înghețat;
- fine-tuning-ul ulterior al ultimelor straturi;
- generarea hărților Grad-CAM;
- binarizarea hărților de atenție;
- segmentarea ghidată prin GrabCut;
- analiza vizuală a predicțiilor.

Acest experiment simplificat demonstrează ușor de urmărit efectul fiecărei componente din pipeline și justifică includerea modulelor originale prin vizualizări clare ale zonelor activate.

3.3 Experimente complete pe setul de date real

Experimentele principale se realizează pe întreg setul de date și includ:

Arhitecturi evaluate

- MobileNetV2 standard și extins cu module (MAB, ResPath, SBR)
- EfficientNet-B0 standard și extins cu module (MAB, ResPath, SBR)
- VGG16 standard și extins cu module (MAB, ResPath, SBR)
- Vision Transformer (ViT) minimal

Strategia de antrenare

- Faza 1: warm-up cu backbone înghețat
- Faza 2: fine-tuning selectiv al ultimelor straturi
- Optimizare cu EarlyStopping, ReduceLROnPlateau, ModelCheckpoint

Metricile utilizate

- Accuracy
- Macro-F1

Acuratețea reflectă procentul imaginilor corect clasificate, iar Macro-F1 asigură evaluarea echilibrată pentru fiecare clasă, fiind esențială într-un dataset dezechilibrat.

Macro-F1 este definit ca:

$$\text{Macro-F1} = \frac{1}{4} \sum_{c=1}^{4} \frac{2 \cdot \text{Prec}_c \cdot \text{Rec}_c}{\text{Prec}_c + \text{Rec}_c},$$

unde:

$$\text{Prec}_c = \text{precizia pentru clasa } c, \qquad \text{Rec}_c = \text{rechemarea (recall) pentru clasa } c.$$

3.4 Formularea matematică a experimentului

Datasetul este formalizat ca:

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \qquad x_i \in \mathbb{R}^{224\times224\times3}, \qquad y_i \in \{1, 2, 3, 4\}.$$

Modelul este o funcție parametrică:

$$f_\theta : \mathbb{R}^{224\times224\times3} \to \Delta^3,$$

care produce:

$$p_\theta(y \mid x) = \text{softmax}(f_\theta(x)).$$

Funcția de pierdere utilizată este cross-entropy:

$$\mathcal{L}(\theta) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{4} \mathbf{1}[y_i = c] \log p_\theta(y = c \mid x_i).$$

3.5 Rezultatele experimentelor

Rezultatele obținute confirmă:

- modulele Mixed Attention, ResPath și SBR cresc semnificativ Macro-F1 pentru toate arhitecturile;
- ensemble-urile soft-voting cresc robustețea predicțiilor;
- ensemble-ul hibrid CNN+ML îmbunătățește clasificarea față de modelele singulare, dar este inferior ca performanță ensemble-ului fără ML;
- ViT-ul simplificat obține rezultate moderate, inferioare CNN-urilor îmbunătățite.

Clasificările se îmbunătățesc sistematic atunci când modulele sunt activate, iar convergența în timpul fine-tuning-ului este mai rapidă și mai stabilă.

3.6 Validare vizuală: Grad-CAM și GrabCut

Hărțile Grad-CAM arată zonele din imagine care influențează decizia modelului.

Pentru tumorile *glioma*, aceste hărți evidențiază regiuni difuze, corespunzând contururilor neregulate ale tumorii. Pentru meningiome, activările se concentrează pe mase rotunjite, bine delimitate.

Binarizarea hărților Grad-CAM permite definirea unei zone inițiale pentru segmentare. Aplicarea ulterioară a algoritmului GrabCut rafinează această zonă și produce o segmentare slab supervizată care urmărește conturul tumorii într-un mod vizual coerent.

Această combinație Grad-CAM + GrabCut reprezintă o contribuție importantă deoarece permite obținerea unei aproximări a segmentării fără a fi nevoie de etichete pixel-level.

3.7 Comparații și interpretări vizuale

Comparând backbone-urile standard cu variantele îmbunătățite:

- performanța crește în mod semnificativ în prezența modulelor propuse;
- ensemble-urile depășesc orice model individual;
- CNN-urile ușoare îmbunătățite depășesc ViT-ul minimal.

Interpretarea vizuală confirmă că:

- există coerență și focalizare în hărțile Grad-CAM;
- segmentările GrabCut sunt stabile și identificate corect.

4 Rezultate și concluzii

Această secțiune sintetizează rezultatele obținute în urma experimentelor, interpretează performanțele modelelor, compară abordarea propusă cu metodele existente și răspunde la întrebările de cercetare formulate în Introducere. Rezultatele sunt analizate atât din perspectiva metricilor de clasificare (Accuracy, Macro-F1), cât și din perspectiva interpretabilității prin Grad-CAM și a segmentării GrabCut, pentru a confirma atât performanța, cât și explicabilitatea modelului.

4.1 Interpretarea rezultatelor experimentale

Evaluările realizate arată că integrarea modulelor originale Mixed Attention Block, Residual Path Head și Spatial Boundary Refinement produce o creștere consistentă a performanței față de backbone-urile standard. Creșterea Macro-F1 este observată pentru arhitecturile CNN testate, atât pentru MobileNetV2, cât și pentru EfficientNet-B0.

Fine-tuning-ul selectiv al ultimelor straturi stabilizează antrenarea și reduce considerabil riscul de supraînvățare.

Ensemble-ul soft-voting produce performanțe superioare oricărei arhitecturi individuale.

Ensemble-ul hibrid CNN+ML oferă îmbunătățiri față de modelele individuale, dar nu depășește ensemble-ul simplu, bazat doar pe CNN-uri.

ViT-ul minimal oferă rezultate moderate, în general inferioare CNN-urilor îmbunătățite, în acord cu literatura actuală.

4.2 Comparații cu abordări existente

Literatura recentă arată rezultate între 93–96% acuratețe pe seturi similare de antrenare. Prin includerea modulelor originale, performanțele obținute în această lucrare se aliniază sau depășesc multe rezultate raportate anterior, în special prin îmbunătățirea Macro-F1.

Modulele MAB, ResPath și SBR sunt eficiente computațional, adăugând foarte puțini parametri în comparație cu SE-Blocks sau CBAM.

Pipeline-ul Grad-CAM + GrabCut extinde utilitatea interpretabilității către o formă apropiată de segmentare clinică.

4.3 Validarea research questions

- **RQ1:** Da — modulele cresc Macro-F1.
- **RQ2:** Da — ensemble-urile îmbunătățesc performanța.
- **RQ3:** Da — Grad-CAM + GrabCut generează segmentări stabile și interpretabile.
- **RQ4:** Da — fine-tuning-ul parțial produce rezultate superioare.
- **RQ5:** Parțial — ViT minimal este inferior CNN-urilor îmbunătățite, dar ar putea fi îmbunătățit prin pre-antrenare medicală.

4.4 Argumente pentru validitatea concluziilor

- Metodologia este evaluată pe un set RMN real, etichetat în patru clase clinice;
- Rezultatele sunt consistente între antrenare, validare și testare;
- Vizualizările Grad-CAM validate prin GrabCut confirmă coerența modelului;
- Comparațiile între backbone-uri întăresc generalitatea metodei;
- Ensemble-urile cresc stabilitatea și reduc variabilitatea rezultatului.

4.5 Direcții viitoare de cercetare

- Utilizarea unui set RMN clinic mult mai mare, incluzând multiple secvențe (T1, T2, FLAIR);
- Pre-antrenarea Vision Transformers pe imagistică medicală;
- Optimizarea modulelor prin Neural Architecture Search;
- Integrarea modelelor multimodale;
- Extinderea pipeline-ului de segmentare către soluții complet automate (U-Net, nnU-Net);
- Evaluarea timpului de inferență pentru integrare clinică în timp real.

4.6 Concluzia generală

Lucrarea prezintă un sistem modern, eficient și explicabil pentru clasificarea tumorilor cerebrale din imagini RMN. Contribuțiile originale (modulele de atenție și rafinare, ensemble-urile și pipeline-ul Grad-CAM + GrabCut) demonstrează îmbunătățiri semnificative față de arhitecturile standard.

Rezultatele confirmă că metodele propuse pot constitui o bază solidă pentru dezvoltarea unor instrumente clinice automate de sprijin în diagnostic, care combină performanța cu interpretabilitatea vizuală.

Bibliografie

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, 2018.

[4] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and X. Zhai, "An image is worth 16×16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

[7] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[8] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[9] L. Prechelt, "Early stopping — but when?," in *Neural Networks: Tricks of the Trade*, pp. 55–69, Springer, 1998.

[10] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, weakly-supervised and unsupervised medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.

[11] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[12] Q. Zhang and Y. Yang, "Understanding the role of attention in convolutional neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 32–40, 2021.

[13] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016.

[14] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[15] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.