Deep neuroevolution for limited, heterogeneous data: proof-of-concept application to Neuroblastoma brain metastasis using a small “virtual pooled” image collection

Subhanik Purkayastha 1, Hrithwik Shalu 2, David Gutman 1, Shakeel Modak 3, Ellen Basu 3, Brian Kushner 3, Kim Kramer 3, Sofia Haque 1 and Joseph Stember 1

1Department of Radiology, Memorial Sloan Kettering Cancer Center, NY, NY, 10065
2Department of Aerospace Engineering, Indian Institute of Technology Madras, Chennai, India, 600036
3Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, NY, 10065

Abstract

Artificial intelligence (AI) in radiology has made great strides in recent years, but many hurdles remain. Overfitting and lack of generalizability represent important ongoing challenges hindering accurate and dependable clinical deployment. If AI algorithms can avoid overfitting and achieve true generalizability, they can go from the research realm to the forefront of clinical work. Recently, small data AI approaches such as deep neuroevolution (DNE) have avoided overfitting small training sets. We seek to address both overfitting and generalizability by applying DNE to a virtually pooled data set consisting of images from various institutions. Our use case is classifying neuroblastoma brain metastases on MRI. Neuroblastoma is well-suited for our goals because it is a rare cancer. Hence, studying this pediatric disease requires a small data approach. As a tertiary care center, the neuroblastoma images in our local Picture Archiving and Communication System (PACS) are largely from outside institutions. These multi-institutional images provide a heterogeneous data set that can simulate real world clinical deployment. As in prior DNE work, we used a small training set, consisting of 30 normal and 30 metastasis-containing post-contrast MRI brain scans, with 37% outside

*Corresponding Author : purkays@mskcc.org, Phone : 4849941222
images. The testing set was enriched with 83% outside images. DNE converged to a testing set accuracy of 97%. Hence, the algorithm was able to predict image class with near-perfect accuracy on a testing set that simulates real-world data. Hence, the work described here represents a considerable contribution toward clinically feasible AI.
1 Introduction

1.1 Biases from limited data still hinder clinical deployment in radiology AI

Successful clinical deployment remains a challenge for radiology AI due to limited ability to generalize. Generalizability refers to how well an AI algorithm can handle the domain shift problem [1]. The domain shift problem refers to the distribution shift between training set (source) and testing set (target) data.

Let us illustrate with the typical pipeline for radiology AI. We train and validate an algorithm from data at Institution A. The images from Institution A are the source data. We clinically deploy the trained algorithm at a different setting, Institution B. Institution B’s images are the target data. In general, we may expect different distributions between source and target image data. Most AI methods presume falsely that source and target data are independently and identically distributed (IID). The domain shift problem refers to the fact that target data is usually out-of-distribution (OOD) relative to source data. In other words, compared to Institution A, we can expect a myriad of subtle differences in image data at Institution B. The differences can result, for example, from varying scanner builds, settings, techniques, and post-processing as well as different patient populations and disease prevalences.

Because AI largely ignores the domain shift problem, algorithm performance often deteriorates upon real world clinical deployment to OOD data sets. This deployment failure results from data-induced bias, which reduces generalizability. The effect can be “spurious correlations” [2] and “shortcut learning” [3], failure modes that can accompany small or even modestly sized training sets. Since AI largely assumes no domain shift, features of IID data are falsely assumed to be present in the OOD data, which again, is typically not the case.

Generalizability thus is the ability to address and transcend the domain shift problem, thereby permitting methods trained on IID data sets to maintain performance when deployed on OOD data. A generalizable algorithm learns the essential intrinsic features of the underlying system. Such features represent both IID and OOD data. For radiologists to trust that a model will perform as expected in clinical practice, it should be adequately generalizable. In this paper, we will show what we believe is a new peak in generalizability, by the methods described in the following sections.

1.2 Source-target data mismatch underlie lack of generalizability

As previously stated, in the presence of small or limited data, generalizability is diminished due to data-induced bias. Data-induced bias can take two forms: overfitting and under-specification. Overfitting refers to when a model
is trained to fit perfectly to a specific dataset, then fails to generalize even narrowly when one resamples IID data as a validation set. This problem occurs commonly with small training sets for supervised deep learning (SDL), whose parameters are trained via stochastic gradient descent (SGD). Obtaining enough labeled data is a longstanding challenge in radiology [4]. To mitigate overfitting, researchers typically employ large training sets. However, producing large, expertly labeled training sets is tedious and time-consuming. Requiring copious data also excludes inherently small data sets, such as for rare disease and under-represented populations. Popular techniques to address the need for large data include transfer learning and data augmentation. Limitations to these techniques include source-target mismatch in transfer learning and limited generalization reached with data augmentation. Recently, small data techniques have been introduced to address overfitting of small training sets. In particular, deep reinforcement learning and deep neuroevolution have generalized well based on sparse data.

### 1.3 Deep reinforcement learning and deep neuroevolution can learn from small training sets

Another approach to avoid overfitting to limited training sets is to employ small-data AI techniques. In particular, deep reinforcement (DRL) [5, 6] and supervised deep learning trained via deep neuroevolution (DNE) [7, 8] have been applied to small training sets. Both techniques were able to glean essential information, achieving high testing set accuracies. The approaches were successful despite training from scratch, even without the benefit of transfer learning. DNE notably converged to 100% testing set accuracy. Given its early success for small training sets, DNE was the approach we pursued in this work.

Although recent work shows DNE’s resistance to overfitting small training sets, important questions remain. In particular, the issue of generalizability to OOD data was not addressed. Extending DNE’s potential for effective clinical deployment requires that we increase generalizability by including external data. In the following sections, we discuss how OOD images improve generalizability by alleviating under-specification. As previously mentioned in Section 1.2, under-specification is one of the two key factors that limit generalizability in AI.

Under-specification [9, 10] refers to the fact that multiple models can fit IID data. However, most of the possible models do not fit the OOD data. The mismatch between OOD and IID data, as nicely illustrated by Eche and Schwartz et al. [10], results in only the more generalized algorithms fitting OOD data. Two chief methods to address under-specification are: stress tests and external validation sets.

To make IID testing sets "new" in a manner that reduces under-specification, one can apply stress tests to simulate different ways in which domain shift can occur [9, 10]. For example, stress tests could be designed to simulate different
acquisition parameters, such as varying slice thickness or CT radiation dose. Alternatively, stress tests may introduce specific artifacts, such as motion or susceptibility [10]. Although they can address individual causes of data-induced bias, stress tests cannot account for all possible sources of bias.

External validation sets can fill in these gaps of unaccounted-for bias left over by stress testing. External validation sets generally come from institutions outside of where the original algorithm was trained. Unlike stress testing, external data can simultaneously address most domain shifts. The domain shifts, or sources of bias, might be visually apparent to the radiologists, whereas other could evade visual detection. However, one at least theoretical limitation applies to using a single external validation set. Although much more likely to represent OOD image data than images from the home institution, a single outside validation set could represent near-IID data [10].

1.4 Heterogeneous data sets can enhance generalizability

We propose the intuitive notion that the wider the variety of OOD image data used in training and testing, the less under-specified and hence more generalized the resulting algorithm. This idea is the rationale behind data pooling [11, 12]. We seek to simulate data pooling via images that have already been scanned into the local Picture Archiving and Communication System (PACS) for clinical purposes. Though we illustrate the approach at our center, we presume feasibility at most tertiary care facilities, which commonly store outside images from patients’ prior care. We describe our data enrichment with outside scans in forthcoming sections.

Although increasingly used in clinical oncological research [12, 13], data pooling is hindered by issues of data compatibility and privacy concerns [14, 15, 16]. These barriers have largely motivated the sub-field of federated learning [17], which is a separate field of study that is not our current focus. Because of the challenges associated with data pooling, we wish to simulate data pooling by employing outside images already available in our locals PACS system. Since we have numerous outside scans from various institutions, we have the equivalent of a highly pooled data set already within our grasp. We can take advantage of the ready availability of this heterogeneous image data set to essentially recreate the benefits of multi-institutional data pooling. In this work, we show that applying DNE to a heterogeneous training set permits state-of-the-art generalizability to OOD images.

DNE involves tuning the parameters of a convolutional neural network (CNN) classifier via random "mutations" followed by a selection criterion. As described in prior literature [7, 8], parameter tuning via random perturbations and selection criteria forms a radical departure from the near-ubiquitous approach to CNN training, which is dominated by SGD optimization. To summarize, we seek to address the longstanding challenge of generalizability with two novel in-
Ingredients: the recently introduced DNE approach to network training, and the "virtual pooled" data consisting of previously uploaded outside scans. We hypothesize that doing so will achieve the most generalizability currently demonstrated. As such, we posit that our approach provides a blueprint for truly reliable clinical deployment.

1.5 Neuroblastoma exemplifies rare diseases, requiring a small data AI approach

Neuroblastoma is an embryonal tumor of the sympathetic nervous system that typically originates in the adrenal medulla or paraspinal ganglia of the abdomen [18]. It is the most common pediatric extracranial solid tumor and accounts for approximately 15% of childhood cancer-related mortality [19]. Neuroblastoma usually occurs sporadically, but the 1-2% of familial cases that have been observed are associated with specific hereditary ALK and PHOX2B germline mutations [20, 21]. The incidence of neuroblastoma is 10.2 cases per million children under 15 years of age, and approximately 500 new cases are reported annually [22]. This cancer has been characterized by unexpected clinical behaviors, such as spontaneous regression, maturation, and metastases; more than 70% of patients with neuroblastoma have already developed malignant lesions outside the primary tumor at the time of diagnosis [23]. Metastasis to the brain is an especially consequential complication. The most common sites of neuroblastoma metastases include the bones and bone marrow of the cranium. Though less frequent, spread to the central nervous system (CNS), including brain parenchyma and leptomeninges, is more fatal [24, 25, 26]. Most CNS metastases are detected at recurrence rather than at diagnosis, with one study reporting an incidence rate of 6.3% [27, 28].

The occult nature and poor prognosis of CNS metastases necessitates robust surveillance. Over the past few years, artificial intelligence (AI) has been applied extensively to radiology to improve the classification of various lesion types. However, most AI models utilize Convolutional Neural Networks (CNN) that depend on large image sets to produce accuracies in the 80-90% range [29, 30, 31]. Given the inherently small incidence of neuroblastoma brain metastases, current state-of-the-art SDL models may not perform optimally due to minimal training data. Hence, neuroblastoma brain metastasis detection is a relevant use-case for small data techniques such as DNE.

2 Methods

2.1 Data collection

We obtained a waiver of informed consent for this retrospective study from our Institutional Review Board. We searched both the electronic medical record as well as that from our PACS image database to identify pediatric neuroblastoma
patients with MRI brain studies. These lists were narrowed down to those of interest via examining the report text, when available, and verified by two neuroradiologists viewing the full imaging exams and, in some cases, preceding or follow-up exams.

As mentioned earlier, a notable feature of our data set is the breadth of scan settings and techniques. Given our institution’s role as a tertiary care center, many patients have their initial diagnosis and treatment at outside institutions. To optimize their care, many patients scan prior outside imaging into our local PACS system. Resulting AI networks should be able generalize to most expected OOD data.

Our training set consisted of 60 post-contrast MRI images, half of which were judged by two neuroradiologists to be normal, or metastasis-free, while the other half contained enhancing parenchymal brain metastases. The testing set also consisted of 60 post-contrast MRI images, with a 1:1 normal:metastasis-containing split. The institutional split of our training and testing sets are shown in Table 1.

| Dataset          | Home institution image, % (n) | Outside institution image, % (n) |
|------------------|-------------------------------|---------------------------------|
| Training – Normal| 100 (30)                      | 0 (0)                           |
| Training – Metastasis | 27 (8)             | 73 (22)                         |
| Testing – Normal  | 3 (1)                         | 97 (29)                         |
| Testing – Metastasis | 30 (9)                    | 70 (21)                         |

Table 1: Breakdown of institutions represented in our training and testing sets.

Hence, the training set as a whole was enriched with 37% outside studies and the testing set as a whole with 83% outside exams. The 22 outside images of the training set were obtained at 20 unique institutions. The 50 outside images of the testing set were obtained at 42 unique institutions. In total, including our hospital, we sampled image data from 58 unique institutions. As such, our data set can represent the range of settings and techniques that AI may encounter in real-world deployment scenarios.

Figure 1 displays sample images that originated from various institutions, illustrating the broad range of image characteristics and reconstruction algorithms in our data set. For example, Figure 1B and 1F evince prominent vascular enhancement, which is less emphasized in the other displayed images. Figure 1E features higher tissue contrast, in particular gray-white differentiation, than other displayed images.
To enable transfer learning with a pre-trained VGG16 network, which is described later on, we captured the T1-weighted post-contrast image for each patient as a 2D image slice. Normal images were intended to reflect the distribution along craniocaudal dimension of the metastasis-containing images. The metastasis images were selected to coincide with the largest portion of the lesions.

2.2 We used a standard type of CNN architecture for both "from scratch" training approaches

The CNN architecture whose weights we trained both via deep neuroevolution (DNE) and stochastic gradient descent (SGD) is displayed in Figure 2. For both DNE and SGD, CNN weights were randomly initialized via the standard Glorot scheme. The CNN consisted of four convolutional layers, each with 32 output channels constituting feature maps. We used kernels composed of $3 \times 3$ weights. We employed a stride of 2 in the x and y directions, with padding of one applied to the input at each step. ReLu activation followed each 2D convolution. We then flattened the last convolutional layer’s output. The resulting flattened vector proceeded to fully connected layers of sizes 512, 256, and 128. We connected this last 128-node layer to the two output nodes representing our two class predictions, normal and metastasis-containing.
2.3 We trained a convolutional neural network with deep neuroevolution

DNE refers to employing evolutionary strategies in deep learning to train the network weights/parameters. The method thus represents an alternative to the prevailing approach of SGD to train the network weights. DNE has been hypothesized, and in two applications demonstrated [7, 8], to avoid overfitting even small training sets. DNE’s ability to resist overfitting has been attributed to its system-agnostic parameter tuning [7, 8].

Further examining why SGD overfits small data sets, while DNE can avoid doing so, we point out the following key difference in training: SGD updates CNN weights after evaluating an objective function. The objective function is typically the negative gradient of the output loss with respect to network weights. DNE reverses the order, updating the weights first and then evaluating an objective function. Because neither the training set nor the current weights influence DNE’s weight updates, DNE can better explore the complex loss surface. As a result, whereas SGD becomes trapped in a local minimum, DNE eventually escapes local minima en route to the global minimum. In evolutionary strategies, including DNE, we call the objective function the fitness. For classification tasks, the fitness is simply total training set accuracy.

As mentioned, DNE produces random perturbations to each network weight.
These perturbations can be thought of as "mutations" in the evolutionary analogy. The mutations of a forerunner "parent" CNN produce "children" CNNs. Then, following a scheme that is further described in the aforementioned prior DNE work [7, 8], the "fittest" children are selected for representation in the next generation. Because we seek highly accurate CNNs, we take as fitness criterion the summed training set accuracy. The parent CNN incorporates into their network the weight updates of the best 50% of children CNNs. The updated parent CNN then undergoes more mutations to produce the next generation of children CNNs, and the process continues for a large number of training generations.

2.4 For comparison with deep neuroevolution, we trained convolutional neural networks with simple stochastic gradient descent

Figure 3: VGG16 architecture used for transfer learning. The network has 263,169 trainable parameters and 14,714,688 frozen weights. Optimization mode used is Batch-based Adam (batch size: 32) with an initial learning rate of 1x10^{-4}. The number of epochs is set until divergence of the validation/testing set is prominent. The objective function used is standard Binary Cross-Entropy and learning rate was further regularized in plateau regions.
For the sake of comparison, we also trained CNNs via SGD. One CNN employed transfer learning from a pre-trained VGG16 network and another CNN was built from scratch. Training with transfer learning from a large, pre-trained VGG16 network is a well-known approach to address overfitting, and the architecture is illustrated in Figure 3. Details of the CNN trained from scratch can be found in Section 2.2.

3 Results

Figure 4 shows the accuracies that the three models achieved. We see that SGD trained without transfer learning badly overfits the small training set due to data-induced bias. The training and testing set accuracies diverge markedly during training. The CNN reaches high training set accuracy. On the other hand, within the first 10 epochs of training, the network attains a maximum testing set accuracy of 53% before quickly converging on 48%, which is essentially equivalent to a coin toss. As opposed to SGD without transfer learning, which overfits the training data, SGD with transfer learning underfits the training set. As is the case for SGD without transfer learning, SGD with transfer learning does not achieve a high testing set accuracy. However, in the case of transfer learning, the testing set under-performance results from transfer bias, which is essentially a mismatch between source and target tasks or datasets. We note, however, that SGD with transfer learning does not overfit due to frozen weights from the prior VGG16 optimization. As such, transfer learning SGD does slightly better than SGD from scratch, quickly reaching a maximum of 58% testing set accuracy, then converging on 57%. Standing in stark contrast to both SGD-based approaches, DNE neither overfits nor underfits the small training set and generalizes to the mostly OOD testing set, converging to a testing set accuracy of 97%.

To highlight the differences across methods in testing set performance, Figure 5 displays only testing set accuracies during training. Again, we see that DNE
generalizes well. DNE in fact converges to a testing set accuracy of 97%, wrongly predicting only two out of the sixty testing set images. Not shown in the figure is that DNE also converges to 96% accuracy for the 50-image subset of the testing set consisting of exclusively outside images. In contrast to DNE, both SGD approaches produce low testing set accuracy. Although SGD with transfer learning performs a bit better than SGD without transfer learning, neither reaches substantial testing set accuracy.

![Figure 5](image.png)

Figure 5: Demonstration that DNE generalizes well, whereas SGD does not generalize effectively. Testing set accuracy as a function of training time is displayed for DNE (blue), SGD using transfer learning (solid green), and SGD trained from scratch (dashed green). Note the two different training time scales: generations for DNE along the upper x-axis and epochs for the SGD calculations along the lower x-axis.

To compare DNE and SGD further, we start by computing confusion matrices, which allows calculation of several quality measures. Figure 6 juxtaposes the confusion matrices for DNE and SGD with transfer learning. We see that DNE produces a near-perfect confusion matrix, the ideal such matrix being an identity matrix with ones along the diagonal entries and zeros on the off-diagonals. The confusion for SGD shows relatively much high off-diagonal values, indicating an overall poor performance.
We used the confusion matrices to calculate multiple quality metrics, namely the sensitivity, specificity, positive predictive value, negative predictive value, and F1 score. The quality metrics are important when there is class imbalance, in which case high accuracy could conceal poor model performance. However, our study is class balanced, with equal numbers of normal and metastasis-containing images.

Table 2: Comparison of several model performance quantities between DNE and SGD with transfer learning.

| Quantity                  | DNE (%) | VGG (%) |
|---------------------------|---------|---------|
| Sensitivity               | 97      | 63      |
| Specificity               | 97      | 55      |
| Positive predictive value | 97      | 34      |
| Negative predictive value | 97      | 80      |
| $F_1$ score               | 97      | 43      |

Table 3: Similar comparison as for Table 2, but only for outside testing set images. The various computed quantities are noted to be essentially identical.

As such, accuracy is a sufficient measure to assess model performance. Nev-
ertheless, for the sake of completeness, we list the aforementioned additional quantities in Tables 2 and 3. Table 3 specifically concerns only the subset of 50 outside institution testing set images. As expected based on accuracy results and the overall appearance of the respective confusion matrices, DNE vastly outperforms SGD with transfer learning in all measures. The disparity is present in both the full testing set and the subset of only outside images.

Additionally, we found the difference in performance between DNE and SGD with transfer learning was statistically significant, with a McNemar’s test $p$-value of $3 \times 10^{-6}$ overall, and $1.9 \times 10^{-5}$ for outside testing set images only. Finally, DNE offers explainability in our application. Figure 7 displays the saliency maps, also called class activation maps, which reflect where the network focuses when determining class prediction. We see that SGD with transfer learning is unable to produce a focus of attention on any particular part of the image, indicating that the network has not extracted meaningful insights or patterns during training. On the other hand, DNE is able to concentrate on a particular part of the image. Moreover, the center of focus coincides with the lesion itself, thereby mirroring human interpretation and abstraction.

![Testing set image](image1)

| Testing set image | Deep neuroevolution | SGD transfer learning |
|-------------------|--------------------|----------------------|
| A                 | B                  | C                    |
| D                 | E                  | F                    |

Figure 7: DNE focuses appropriately, whereas SGD has no coherent center of attention. Class activation maps (CAMs) for two examples (A and D) of testing set images with neuroblastoma metastases. Figures B and E show the corresponding deep neuroevolution CAMs, while C and F display the corresponding CAMs obtained from VGG with transfer learning.
4 Discussion

DNE converges to near-perfect accuracy for a heterogeneous testing set simulating clinical deployment, despite a small training set. The results represent a new milestone for DNE; the approach generalizes as broadly as we were able to assess. Hence, DNE can extract the key intrinsic features of the images. Doing so brings us a step closer to how human radiologists acquire deep understanding and intuition for image interpretation. Such core understanding permits trained radiologists to read scans accurately even when the style and quality vary from their home institution.

However, despite the promise and potential to learn in generalizable manner with DNE, many limitations need to be addressed in future work. Among these is a need to process full 3D image stacks; in our current study, we analyzed 2D image slices. We also want our CNNs to process simultaneously other sequences, such as T2, T2 FLAIR, DWI and SWI. Doing so should improve performance by incorporating more disease state information.

Finally, we would like to go further with the notion of highly enriched heterogeneous data sets via virtual pooling to produce highly generalizable and ultimately enterprise-grade algorithms. Although the present study featured a majority of scans from outside our institution, a future goal is to build databases comprised exclusively of outside-institution images. This task is made more feasible with small data AI, because one would not need to assemble beyond a few dozen outside scans for a particular modality and disease state.

Enabling ourselves to analyze full-volume, multisequence image data from various institutions will continue to draw us closer to truly dependable AI. We anticipate that such algorithms will increasingly be able to assist radiologists and clinicians to improve patient care.

5 Conclusions

Our study is a proof-of-principle for DNE’s success predicting metastases in an inherently small-data context, a task that is currently unattainable for SGD-trained models. We show that DNE remains generalizable to a heterogeneous testing set, which helps us envision its use in the real world.

6 Conflicts of interest

Two of the authors have obtained provisional patents for a company, Authera Inc, related to the research presented in this work.
7 Funding

We gratefully acknowledge external support from the Radiological Society of North America (RSNA) and the American Society of Neuroradiology (ASNR).

References

[1] Kaiyang Zhou et al. “Domain generalization: A survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[2] Martin Arjovsky et al. “Invariant risk minimization”. In: *arXiv preprint arXiv:1907.02893* (2019).

[3] Robert Geirhos et al. “Shortcut learning in deep neural networks”. In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.

[4] Shelly Soffer et al. “Convolutional neural networks for radiologic images: a radiologist’s guide”. In: *Radiology* 290.3 (2019), pp. 590–606.

[5] Joseph Stember and Hrithwik Shalu. “Deep Reinforcement Learning Classification of Brain Tumors on MRI”. In: *Innovation in Medicine and Healthcare*. Springer, 2022, pp. 119–128.

[6] Joseph Nathaniel Stember and Hrithwik Shalu. “Deep reinforcement learning with automated label extraction from clinical reports accurately classifies 3D MRI brain volumes”. In: *Journal of Digital Imaging* (2022), pp. 1–10.

[7] Joseph N Stember and Hrithwik Shalu. “Deep Neuroevolution Squeezes More out of Small Neural Networks and Small Training Sets: Sample Application to MRI Brain Sequence Classification”. In: *arXiv preprint arXiv:2112.12990* (2021).

[8] Joseph Stember, Robert Young, and Hrithwik Shalu. “Direct evaluation of progression or regression of disease burden in brain metastatic disease with Deep Neuroevolution”. In: *arXiv preprint arXiv:2203.12853* (2022).

[9] Alexander D’Amour et al. “Underspecification presents challenges for credibility in modern machine learning”. In: *arXiv preprint arXiv:2011.03395* (2020).

[10] Thomas Eche et al. “Toward Generalizability in the Deployment of Artificial Intelligence in Radiology: Role of Computation Stress Testing to Overcome Underspecification”. In: *Radiology: Artificial Intelligence* 3.6 (2021).

[11] Arno Appenzeller et al. “Towards Distributed Healthcare Systems—Virtual Data Pooling Between Cancer Registries as Backbone of Care and Research”. In: 2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA). IEEE, 2021, pp. 1–8.

[12] John Kang et al. “National cancer institute workshop on artificial intelligence in radiation oncology: training the next generation”. In: *Practical radiation oncology* 11.1 (2021), pp. 74–83.
[13] Mark D Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: Scientific data 3.1 (2016), pp. 1–9.

[14] Arindam R Chatterjee et al. “Image sharing in radiology—A primer”. In: Academic radiology 24.3 (2017), pp. 286–294.

[15] Martin J Willemink et al. “Preparing medical imaging data for machine learning”. In: Radiology 295.1 (2020), pp. 4–15.

[16] Francesco Sardanelli et al. “To share or not to share? Expected pros and cons of data sharing in radiological research”. In: European Radiology 28.6 (2018), pp. 2328–2335.

[17] Curtis P Langlotz et al. “A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy Workshop”. In: Radiology 291.3 (2019), p. 781.

[18] John M Maris. “Recent advances in neuroblastoma”. In: New England Journal of Medicine 362.23 (2010), pp. 2202–2211.

[19] Atif Zafar et al. “Molecular targeting therapies for neuroblastoma: Progress and challenges”. In: Medicinal research reviews 41.2 (2021), pp. 961–1021.

[20] Nevim Aygun, “Biological and genetic features of neuroblastoma and their clinical importance”. In: Current pediatric reviews 14.2 (2018), pp. 73–90.

[21] Erin K Barr and Mark A Applebaum, “Genetic predisposition to neuroblastoma”. In: Children 5.9 (2018), p. 119.

[22] Nadja Colon and Dai Chung, “Neuroblastoma”. In: Advances in pediatrics 58 (Dec. 2011), pp. 297–311. DOI: 10.1016/j.yaped.2011.03.011.

[23] Tasnim Ara and Yves A DeClerck, “Mechanisms of invasion and metastasis in human neuroblastoma”. In: Cancer and Metastasis Reviews 25.4 (2006), pp. 645–657.

[24] Steven G DuBois et al. “Metastatic sites in stage IV and IVS neuroblastoma correlate with age, tumor biology, and survival.” In: Journal of pediatric hematology/oncology 21.3 (1999), pp. 181–189.

[25] Katherine K Matthay et al. “Central nervous system metastases in neuroblastoma: radiologic, clinical, and biologic features in 23 patients”. In: Cancer 98.1 (2003), pp. 155–165.

[26] Kim Kramer et al. “Compartmental intrathecal radioimmununotherapy: results for treatment for metastatic CNS neuroblastoma”. In: Journal of neuro-oncology 97.3 (2010), pp. 409–418.

[27] Kim Kramer et al. “Neuroblastoma metastatic to the central nervous system: The Memorial Sloan-Kettering Cancer Center experience and a literature review”. In: Cancer: Interdisciplinary International Journal of the American Cancer Society 91.8 (2001), pp. 1510–1519.

[28] Nicholas D’Ambrosio et al. “Imaging of metastatic CNS neuroblastoma”. In: American Journal of Roentgenology 194.5 (2010), pp. 1223–1229.
[29] Chenjie Ge et al. “Deep semi-supervised learning for brain tumor classification”. In: *BMC Medical Imaging* 20.1 (2020), pp. 1–11.

[30] Tariq Sadad et al. “Brain tumor detection and multi-classification using advanced deep learning techniques”. In: *Microscopy Research and Technique* 84.6 (2021), pp. 1296–1308.

[31] Ahmed M Gab Allah, Amany M Sarhan, and Nada M Elshennawy. “Classification of Brain MRI Tumor Images Based on Deep Learning PGGAN Augmentation”. In: *Diagnostics* 11.12 (2021), p. 2343.