Autoencoders to detect manifestation shift in medical images

Samuel Armbrust Freitas[1] **, Cristiano André da Costa**[1] **, Gabriel de Oliveira Ramos**[1]

[1]Graduate Program in Applied Computing
Universidade do Vale do Rio do Sinos
São Leopoldo, RS, Brazil

{samuelaf@edu.,cac@,gdoramos@}unisinos.br

Abstract. *Advanced intelligent models can potentially assist medical professionals with their routine tasks. Nonetheless, a recognized challenge is that these algorithms often excel in operating in-distribution settings but may underperform when confronted with Out-of-Distribution situations from diverse sources, mainly caused by manifestation shifts. This inconsistency underscores the disparity between lab-based findings and real-world clinical applications, necessitating reliable methods for measuring uncertainty. This study introduces a novel approach proposing self-awareness in detecting manifestation shifts within medical images, designing the challenge as a classification of lesion subtypes. The proposed method achieved competitive performance compared with recent literature utilizing reconstruction thresholds based on histogram analysis. Furthermore, a review of erroneous predictions uncovered various confounding factors, contributing to a better understanding of the models' limitations.*

Resumo. *Modelos inteligentes podem potencialmente ajudar os profissionais médicos em suas tarefas diárias. No entanto, um desafio encontrado é que estes algoritmos muitas vezes apresentam bons resultados operando dentro da sua distribuição, mas podem ter um desempenho inferior quando confrontados com situações fora de distribuição de diversas fontes, causadas principalmente por alterações de manifestação. Esta inconsistência esclarece a disparidade entre os resultados baseados em laboratório e as aplicações clínicas no mundo real, necessitando de métodos adequados para medir a incerteza. Este estudo apresenta uma nova abordagem que propõe a autoconsciência na detecção de mudanças de manifestação em imagens médicas, concebendo o desafio como uma classificação de subtipos de lesões. O método proposto alcançou desempenho competitivo em comparação com a literatura recente, utilizando limiares de reconstrução baseados na análise de histograma. Além disso, uma revisão de classificações erradas revelou vários fatores de confusão, contribuindo para uma melhor compreensão das limitações dos modelos.*

1. Introduction

In recent years, significant advances have been made in developing intelligent models to assist medical professionals during their daily routines [Tschuchnig and Gadermayr 2022]. A key challenge in integrating AI solutions into regular medical practice lies in their limited generalization capabilities [Esmaeili et al. 2023].

Deploying those solutions in clinical environments is a critical goal in medical image processing, requiring robustness to the full range of potential inputs encountered in real-world scenarios [Graham et al. 2023]. This inconsistency underscores the disparity between lab-based findings and real-world clinical applications, necessitating reliable methods for measuring uncertainty. Uncertainty methods can better rule out poor predictions when the data is near-out-of-distribution (OOD) [Graham et al. 2023]. However, gathering labeled data can be labor-intensive and expensive [Esmaeili et al. 2023, Tschuchnig and Gadermayr 2022]. As a result, semi-supervised and unsupervised techniques have frequently been applied, often leveraging the benefits of anomaly detection techniques [Tschuchnig and Gadermayr 2022]. While deep learning algorithms generally perform well in controlled distributions, their performance can significantly decline in OOD contexts [Graham et al. 2023].

Any image-processing method intended for clinical application must effectively handle various input scenarios. This leads to the importance of Domain Generalization (DG), which facilitates extending a learning model across various source domains to an unknown target domain [Chen et al. 2022]. The main goal of DG is to develop semantic representations that operate independently of domain-specific labels, such as disease-related patterns (e.g., lesion texture) rather than features tied to the imaging domain (e.g., scanner type) [Chen et al. 2022]. Traditional statistical learning methods often assume that source and target datasets are Independent and Identically Distributed (i.i.d), a perspective that neglects the complexities found in OOD data [Li et al. 2022]. To overcome this, approaches aim to establish domain-invariant representations by reducing the discrepancies among source domains.

There are five distinct types of distribution shifts: population (covariate shift), annotation/labeling (concept shift), prevalence (target shift), manifestation (conditional shift), and acquisition (domain shift) [Castro et al. 2020]. Among these, the population shift is the most prevalent and can typically be addressed through data augmentation, while challenges like manifestation shift require more intricate solutions [Xu et al. 2022]. When the presentation of a disease is not uniformly predictable across a population, data augmentation becomes ineffective. Data augmentation cannot resolve a manifestation shift when the disease's manifestation is unpredictable for the entire population. This introduces the concepts of near-OOD and far-OOD [Yang et al. 2022], which refer to semantically similar and unrelated images. The state of the art for near-OOD has seen much less progress than far-OOD, as these images' shifts cannot be corrected except by using strong parametric assumptions on the nature of these differences [Castro et al. 2020]. Henceforth, we interchange the terms near-OOD and manifestation shift to refer to the same concept. Given these complexities and the challenges in forecasting target domains, self-awareness emerges as vital. Self-awareness pertains to the understanding of one's internal state and varying levels of expertise [Petrovska 2021]. A self-aware entity possesses knowledge about its environment and interactions within a broader system [Lewis et al. 2011]. It has been recognized that self-awareness is crucial for enabling the adaptability of intelligent models [Lewis et al. 2011, Petrovska 2021].

This study introduces an architecture leveraging normalization layers for identifying manifestation shifts in medical imaging through a detection system. As part of a framework to improve model performance through self-awareness, this preliminary re-

search presents findings on manifestation shift detection derived from a comprehensive evaluation of autoencoder models, framing the challenge around non-healthy MRI brain scans. The outcomes of this investigation seek to advance solutions for instances of manifestation shifts. The contributions of this work can be summarized as: (1) a challenge framework utilizing publicly available datasets to systematically address manifestation shift cases, enabling rigorous benchmarking and development of robust predictive models; (2) a simple autoencoder architecture to detect manifestation shift in medical images. This paper is structured in the following manner. Section 2 overviews the background and related work. Section 3 explains the proposed architecture in detail. The discussion of experimental results in Section 4. Finally, Section 5 presents the main findings and outlines directions for future work.

2. Background and Related Work

This section explores the theoretical foundation of our work. Section 2.1 introduces the crucial context of manifestation shifts. Section 2.2 reviews recent studies on out-of-distribution (OOD) data, highlighting manifestation shifts.

2.1. Manifestation shift

In recent years, there has been a growing focus on OOD detection, driven by the necessity for safe machine learning models, which often exhibit subpar performance on OOD samples [Graham et al. 2023]. The most prevalent case involves a population shift, where the marginal distribution of variables alters between training and test data, while the labeling remains consistent [Xu et al. 2022] (the distribution $P(x)$ changes, but the relationship between $x \rightarrow y$ remains intact).

Conversely, manifestation shift refers to a situation where the relationship $x \rightarrow y$ changes (with y representing the target prediction and x being the input image). This means that the conditional distribution $P(x|y)$ changes under certain constraints (conditional shift) [Zhang et al. 2013]. This shift in medicine illustrates how a given disease manifests differently in various patients and its unknown variables [Castro et al. 2020]. The capability of OOD detection to differentiate between normal and abnormal patterns is typically developed by learning from normal data [Cui et al. 2023]. However, abnormal images are often utilized during the validation phase [Cui et al. 2023, An and Cho 2015], which restricts its practical applicability in real-world situations.

2.2. Out-of-distribution detection

There has been a rising interest in OOD detection in recent years, driven by the need to enable safe machine learning models, which usually perform inferior on OOD samples [Graham et al. 2023]. However, none of the existing proposed methods recently evaluated achieved sufficient performance in different datasets to be applied in the daily practice [Cui et al. 2023]. In the recent literature, there are different strategies applied to improve performance in detection tasks, such as the use of only healthy exams to then detect OOD exams at test time (i.e. low likelihood) [Graham et al. 2023]; use of only healthy exams in both test and validation steps [Tschuchnig and Gadermayr 2022]; the self-supervised method that leverages self-distillation and negative samples for the task of abnormality detection without accessing labels [Rafiee et al. 2023].

Using Vision Transformers using teacher and student networks showed efficiency in different medical image modalities (such as pneumonia, polyp, and glaucoma detection from X-ray, colonoscopy, and ophthalmology) using only healthy exams during training, but they fall in the far-OOD scenario, which is easier than near-OOD due to the different between image modalities and contexts. Still, it is limited by small datasets and the creation of negative examples [Rafiee et al. 2023]. Nevertheless, abnormal images are often employed during the validation process [Cui et al. 2023]. The abnormality in medical image analysis is heterogeneous and complicated due to shape, the density of anomalies, and other sensory abnormalities [Cui et al. 2023]. For this reason, the successful detection happens for more apparent anomalies, such as tumors, hemorrhages, and ischemia [Kascenas et al. 2023], but lacks better performance for subtle anomalies [Kascenas et al. 2023].

As well as the use of standard data, it is expected to use reconstruction error as a metric to separate in-distribution (ID) and OOD [Guo et al. 2022, Kascenas et al. 2023], directly or via thresholding the residual map between input and reconstruction [Kascenas et al. 2023]. Reconstruction loss might be good for prominent anomalies but struggle with anomalies subtler in intensity contrast [Kascenas et al. 2023]. Given the limitations for specific problems, the use of decision-level ensemble takes advantage of different methods to capture multiple kinds of anomalies [Cui et al. 2023] as the majority of studies using autoencoders for anomaly detection are limited by the use of small dataset shifts, such as increased noise or lower quality scans [Graham et al. 2023].

3. Methodology

This section introduces our proposed deep-learning architecture for detecting manifestation shifts in medical images. This task is part of more extensive research to foster self-awareness in neural networks to improve their capability to adapt to novelty. Figure 1 presents the proposed architecture, containing an autoencoder for detecting manifestation shifts. Our proposed architecture starts as an autoencoder to reconstruct healthy brain Magnetic Resonance Imaging (MRI) exams. An image enters our model being evaluated. Our model outputs the reconstruction perceptual loss, and this result works as a classification, which defines if the sample is ID along with the class identified or an OOD sample. If the sample is OOD, it should go through an adaptation phase, composed by fine-tuning our base model to the manifestation shift context detected. In this research, we decided to model the manifestation shift detection using freely available datasets and compare it with recent research that classified brain tumor subtypes to compare the model's performance, even if this is not the fairest approach.

The proposed architecture is based on literature findings about the lack of real-world datasets for manifestation shift [Castro et al. 2020] together with the continuous inclusion of outliers [Yang et al. 2021] proposing a simple architecture [Zhou et al. 2022]. Figure 1 contains near-OOD detection composed of the autoencoder architecture to determine if the input is under distribution (ID) or manifestation shift (near-OOD). When the input is under the existing model's knowledge, we refer to the given class. Otherwise, when the input is outside the existing model's knowledge, it is classified as a manifestation shift. We obtain this output using a perceptual loss metric derived from a VGG19 architecture. Equation 1 is the euclidean distance between the feature representations of a reconstructed image G_{θ_G} (I^{LR}) and the reference image I^{HR} [Ledig et al. 2017], followed

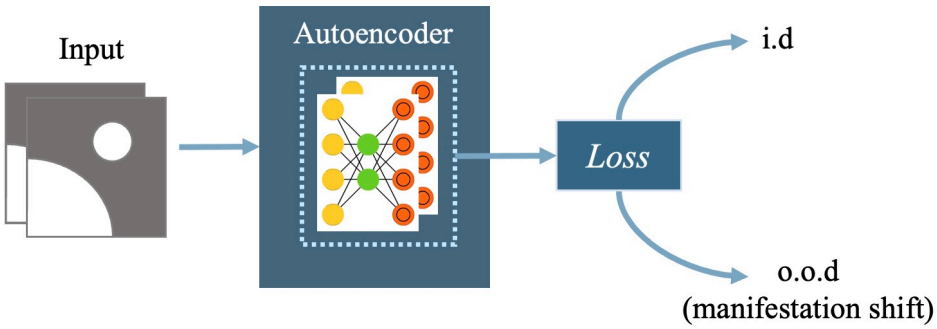

Figure 1. Proposed autoencoder architecture composed of the autoencoder architecture to determine if the input is under the model's knowledge

by Equation 2, where $W_{i,j}$ and $H_{i,j}$ describe the dimensions of the respective feature maps within the VGG network [Ledig et al. 2017].

$$\alpha = \left(\phi_{i,j}\left(I^{HR}\right)_{x,y} - \phi_{i,j}\left(G_{\theta_G}\left(I^{LR}\right)\right)_{x,y}\right)^2 \tag{1}$$

$$l_{VGG/i.j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \alpha \tag{2}$$

Specifically, we employed the VGG19 network pre-trained on ImageNet [Ledig et al. 2017], analyzing the differences in the deep features of the images across the first three convolutional blocks of VGG19. Each trained autoencoder was assigned a threshold through histogram analysis in the latent feature space by comparing negative and positive samples. There are no manifestation shift datasets of medical images in the literature, and for this reason, we designed the challenge and evaluation based on publicly available datasets for disease subtypes, specifically using brain tumor subtype lesions. Based on a design decision, this study considered a baseline of 90% accuracy in classifying classes acceptable for comparison with literature findings.

3.1. Architecture

As part of a framework to improve model performance through self-awareness, this study proposes an architecture based on a convolutional autoencoder to detect near-OOD exam characteristics. Convolutional autoencoders are characterized by convolutional layers instead of fully connected ones, where the output is the reconstruction of the input image. Figure 2 presents the proposed autoencoder containing one encoder with hidden layers followed by a decoder, resulting in an output size equal to the input one. The proposed architecture has an input of $64 \times 64 \times 3$, containing three hidden layers, with $128, 64$, and 32 filters, respectively, resulting in a latent space of 2048. This latent space represents an image reduction of approximately 6 times.

The architecture also follows an experimental understanding from recent literature, which explains that Batch Normalization (BN) and Instance Normalization (IN) have shown improvements in slightly different directions for style transfer. The latter is implemented in higher layers to preserve the content, and any one of them in shallow layers to

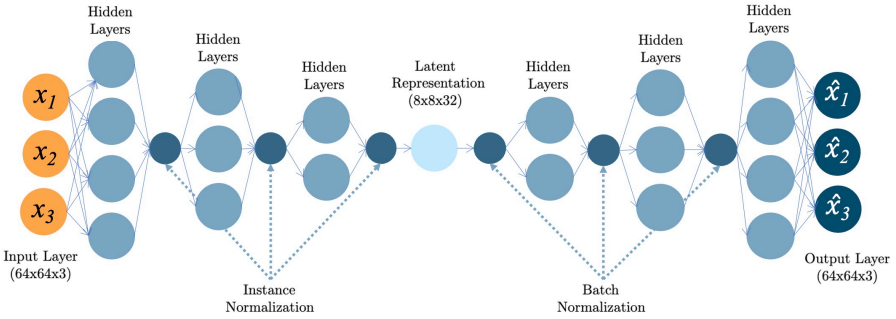

Figure 2. Autoencoder representation with 3-channel images as input and reducing input's dimension to 32-dimensional 8×8 image vectors in the output

preserve the style, improving performance and generalization at once [Pan et al. 2018]. So, we added IN between the shallow convolutional encoder layers down to the latent space to preserve image style. In contrast, we added BN between the deeper convolutional decoder layers to preserve content. We proposed using the Adam optimizer without weights initialization and perceptual loss for model training and evaluation.

4. Experimental Evaluation

We now present the empirical findings that support the theoretical outcomes of this research. Our objective is to demonstrate that our proposed architecture can accurately identify manifestation shift samples in medical images, irrespective of the dataset distribution. The study initiates with a model trained solely on healthy brain examinations, transitioning to unsupervised training using medical images. All experiments were conducted utilizing the computational resources available through Google Colaboratory, specifically employing NVIDIA's GPUs P100 and T4 with 52 GB of RAM.

Throughout the experiments, we assessed various architectures and methodologies from the literature. We ultimately achieved the best reconstruction results and latent space performance, characterized by more compressed versions of the original exams. This was accomplished by integrating BN and IN normalization layers within a straightforward autoencoder architecture. We separated the experimental evaluation into Section 4.1, where we detail the datasets to simulate manifestation shift, followed by Section 4.2, where we detail the overall performance of our proposed method compared with the literature.

4.1. Use case

We demonstrate the application of our proposed method leveraging three datasets of brain tumor diseases. This study leveraged the following datasets: The Brain tumor segmentation dataset [Haq et al. 2023] contains 3064 images from 233 patients with three kinds of brain tumor MRI: meningioma (708 slices), glioma (1426 slices), and pituitary tumor (930 slices) [Haq et al. 2023]. The Br35H dataset [Gómez-Guzmán et al. 2023] contains 3000 brain MRI images of tumor and no-tumor classes (1500 for each class) [Alanazi et al. 2022]. The Brain Tumor Classification dataset [Kang et al. 2021] contains four MRI classes: glioma (826 slices), meningioma (822 slices), pituitary (827 slices), and absence of tumor (395 slices) [Kang et al. 2021]. Finally, we created a unique dataset containing 1530 images of meningioma, 2252 images of glioma, 1757 images of pituitary, and 1895 images of healthy brains.

A preprocessing phase was essential for effectively working with the datasets due to the differences in image dimensions. We began by resizing the images from their original size of $256 \times 256 \times 1$ pixels [Amran et al. 2022] to a standardized size of $64 \times 64 \times 1$ pixels. The preprocessing process concluded with contrast stretching, which enhances the grayscale properties of the images. This technique involves mapping the minimum intensity in the image to the lowest value in a specified range (e.g., $85 \rightarrow 0$) and the maximum intensity in the image to the highest value in the range (e.g., $189 \rightarrow 255$). Following the resizing, we conducted image normalization to scale all pixel values to a range between 0 and 1 [Huang and Belongie 2017]. We separated the dataset into training, testing, and validation with three classes: glioma, meningioma, and pituitary. We excluded the original non-lesion brain from validation due to the design decision to use it as the base for newly created models. The validation dataset contains 125 samples of each class, and we divided the remaining samples into 60% for training and 40% for testing.

4.2. Model evaluation

The proposed autoencoder for near-OOD detection was trained using only healthy brain exams. As already stated, we evaluated the model's capacity to identify lesions using a compressed version of input images in an autoencoder architecture. Based on extensive tests, we proposed using the following parameters for the new autoencoders created during the adaptation to new manifestation shift: $5e-3$ as the learning rate and 50 epochs of training leveraging transfer learning from our healthy brain's base model. Given the defined brain tumor subtype scope, we reached three autoencoders, one for each lesion subtype, and compared their performance.

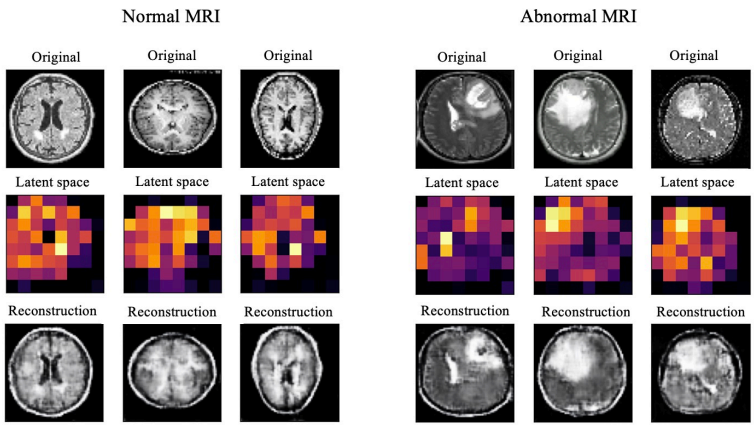

Figure 3. The latent space with a compressed version of brain exams.

Our experiments indicate that a perceptual reconstruction loss exceeding the value of 0.08 effectively differentiates whether a sample is within the ensemble's knowledge or classified as an OOD sample. The latent space with a compressed version with brain exams correlates visually with the input images, which validates the reduction hypothesis. The latent space of $8 \times 8 \times 32$ retained the most essential features from healthy images. Figure 3 shows examples of reconstructions, latent space, and the original exams for both normal and abnormal MRI, where our proposed architecture was able to reduce the brain image dimensionality by 6 times showing the visual correlation with the input images, so retaining important features from images.

Our experiments showed that the feature space needs to be adjusted to maintain enough of the image's input shape and style to reconstruct a specific class. However, it shouldn't retain so much detail that it can reconstruct any image input from different datasets and classes. In Table 1, we present high-performance methods with almost complete confusion matrix-based metrics, reaching accuracy over 95% using ResNet-50 architecture [Haq et al. 2023], and accuracy of 95% specifically in brain tumor datasets [Alanazi et al. 2022].

Table 1. Comparison of our proposed architecture with studies using brain tumor MRI datasets

Reference	Classification			
	Ac.	**Se.**	**Sp.**	**F1-score**
Ours	0.91	**0.95**	0.87	0.90
[Alanazi et al. 2022]	0.95	0.94	0.95	-
[Haq et al. 2023] w/o aug	**0.95**	**0.95**	**0.96**	**0.94**

Our proposed architecture reached a competitive general performance with an accuracy of 0.9022, a sensitivity of 0.9471, a specificity of 0.8668, and an F1-score of 0.8952, which is over the baseline performance but still lacks generalization. From this perspective, our proposed architecture competently classifies brain tumor lesions with a simpler architecture and fewer parameters compared with the recent findings in the literature. Given the high sensitivity, it might be a triage method for manifestation shift detection.

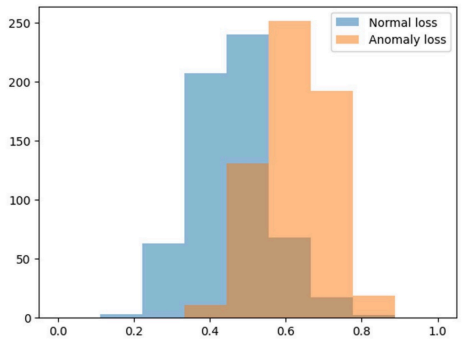

Figure 4. Resulting threshold histogram comparing lesion subtypes meningioma and Glima.

After using the identified threshold as criteria to determine OOD samples, we reached histogram distributions of threshold presented in Figure 4, specifically when comparing glioma and meningioma, which shows that it is possible to determine a perceptual loss threshold. However, we still have some border cases due to image similarity when comparing lesion subtypes. It is essential to highlight that some lesion characteristics are also appropriately reconstructed, such as massive lesions in MRI exams. Small lesions were less reconstructed, which can be due to the convolutional characteristics of our proposed architecture. We identified that using this architecture, it is possible to locate a consistent threshold for the proposed OOD problem, which will be specific for each kind

of task and requires an analysis of the histogram to determine the appropriate minimum perceptual reconstruction loss to determine an image was correctly classified.

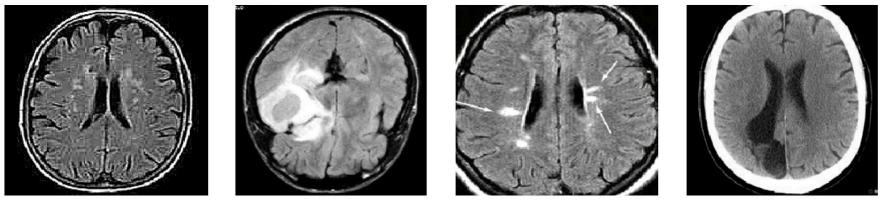

Figure 5. Prediction error analysis caused by brain bleeding at (a) and (c); white arrows at (c); ischemia area at (d); and lesions rounded by bleeding at (b)

Aside from the numerical evaluation, a visual analysis of classification error was performed to identify confounding factors and provide awareness on where models might fail. We noticed four main confounding factors for healthy exams: the presence of different lesions, which are not brain tumors, such as brain bleeding in Figure 5 (a), (c) and ischemia area (previous stroke) in Figure 5 (d). Also, our proposed autoencoder fails in lesions rounded by bleeding in brain tumor cases in Figure 5 (b). Beyond that, we also refer to Figure 5 (c), where we noted white arrows caused by inaccurate sample inclusion in the dataset. These confounding factors might be attenuated when better acquisition and storage protocols are available to avoid adding non-curated samples to freely available datasets.

Table 2. Comparison of our proposed architecture for each subtype brain tumor lesion

Reference	Classification			
	Ac.	**Se.**	**Sp.**	**F1-score**
Ours (healthy vs non-healthy)	0.91	0.95	0.87	0.90
Ours (meningioma)	0.72	0.90	0.55	0.76
Ours (glioma)	**0.88**	0.92	**0.83**	**0.90**
Ours (pituitary)	0.84	**0.96**	0.73	0.86

Any different sample from the class-specific autoencoder is classified as OOD. Given that, comparing each lesion subtype against the others, we could identify the lesion subtypes with better performance detecting near-OOD samples. With this evaluation, we identified that the reconstruction of meningioma is the most challenging one compared with glioma and pituitary. This finding is interesting, and the literature indicates that identifying meningioma can be difficult because it often grows very slowly and may not cause noticeable symptoms until it reaches a significant size [Haq et al. 2023]. Given the visual analysis of freely available datasets, gliomas and pituitary lesions usually appear in specific skull areas, which may generate a bias. In contrast, meningioma can appear in the whole skull area and generate false negatives on the other brain tumor subtypes.

Beyond that, given the best classification results using ResNet-50 reported in the recent literature, an analysis regarding the computational cost and the number of parameters is relevant. ResNet-50 usually has around 25.6 million parameters, while our proposed architecture has only 201 thousand parameters and presented competitive results.

Another critical factor is the use of data augmentation in the recent architectures, which improves the model's performance but ignores the nature of the shifts and might not be suitable for real-life scenarios. This study's limitations include the lack of manifestation shift datasets and proper baselines for comparison and reproducibility. However, selecting a wide study subtype disease such as a brain tumor proves the evaluation robustness in approximating the real-world cases to theoretical evaluation. Finally, the problem design and the dataset choices intend to facilitate solving manifestation shift cases and pave the way for manifestation shift detection and the appropriate adaptation for new shifts when the target domain is unknown.

5. Concluding Remarks

This brief study has shown promising results using an autoencoder to detect manifestation shifts for medical imaging. Although representing anomalies as OOD is dangerous [Tschuchnig and Gadermayr 2022], this study employed brain MRI exams as a study case to detect manifestation shifts under challenging situations. On the other hand, our proposal relies on using the autoencoder result as an awareness metric to enable decisions to be referred to humans when they are presented with difficult or OOD data samples [Graham et al. 2023]. One crucial direction identified is the use of an ensemble of neural networks for combined degrees of agreement as a measure of their certainty [Graham et al. 2023], powered by the use of skip connections and dropout layers to improve reconstruction and allow uncertainty to be measured via dropout stochasticity [Kascenas et al. 2023].

The vast existing brain MRI community paved the way for making different datasets available and boosting recent years' research [Tschuchnig and Gadermayr 2022] but still lacks manifestation shift-specific datasets, which makes fair comparison difficult. This highlights the importance of accessible and comparable datasets as a high priority together with acquistion and storage protocols to avoid confounding factors generated by inaccurate samples. Another future direction for this research is automating threshold identification to improve the proposed architecture's applicability in the community. Finally, the comprehensive annotation of anomalies also becomes an important avenue for future work, as annotated rare pathologies are not available and are potentially more important than common pathologies since this is where training traditional supervised approaches might be infeasible [Kascenas et al. 2023].

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. This research was partially supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (grants 313845/2023-9, 307345/2023-8, and 404572/2021-9).

References

Alanazi, M. F., Ali, M. U., Hussain, S. J., Zafar, A., Mohatram, M., Irfan, M., AlRuwaili, R., Alruwaili, M., Ali, N. H., and Albarrak, A. M. (2022). Brain tumor/mass classification framework using magnetic-resonance-imaging-based isolated and developed transfer deep-learning model. *Sensors*, 22(1):372.

Amran, G. A., Alsharam, M. S., Blajam, A. O. A., Hasan, A. A., Alfaifi, M. Y., Amran, M. H., Gumaei, A., and Eldin, S. M. (2022). Brain tumor classification and detection using hybrid deep tumor network. *Electronics*, 11(21):3457.

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18.

Castro, D. C., Walker, I., and Glocker, B. (2020). Causality matters in medical imaging. *Nature Communications*, 11(1):3673.

Chen, C., Tang, L., Liu, F., Zhao, G., Huang, Y., and Yu, Y. (2022). Mix and reason: Reasoning over semantic topology with data mixing for domain generalization. In *Advances in Neural Information Processing Systems*.

Cui, C., Wang, Y., Bao, S., Tang, Y., Deng, R., Remedios, L. W., Asad, Z., Roland, J. T., Lau, K. S., Liu, Q., et al. (2023). Feasibility of universal anomaly detection without knowing the abnormality in medical images. In *Workshop on Medical Image Learning with Limited and Noisy Data*, pages 82–92. Springer.

Esmaeili, M., Toosi, A., Roshanpoor, A., Changizi, V., Ghazisaeedi, M., Rahmim, A., and Sabokrou, M. (2023). Generative adversarial networks for anomaly detection in biomedical imaging: A study on seven medical image datasets. *IEEE Access*, 11:17906–17921.

Gómez-Guzmán, M. A., Jiménez-Beristaín, L., García-Guerrero, E. E., López-Bonilla, O. R., Tamayo-Perez, U. J., Esqueda-Elizondo, J. J., Palomino-Vizcaino, K., and Inzunza-González, E. (2023). Classifying brain tumors on magnetic resonance imaging by using convolutional neural networks. *Electronics*, 12(4):955.

Graham, M. S., Tudosiu, P. D., Wright, P., Pinaya, W. H. L., Teikari, P., Patel, A., U-King-Im, J. M., Mah, Y. H., Teo, J. T., Jäger, H. R., Werring, D., Rees, G., Nachev, P., Ourselin, S., and Cardoso, M. J. (2023). Latent transformer models for out-of-distribution detection. *Medical Image Analysis*, 90.

Guo, X., Gichoya, J. W., Purkayastha, S., and Banerjee, I. (2022). Margin-aware intraclass novelty identification for medical images. *Journal of Medical Imaging*, 9.

Haq, A. U., Li, J. P., Kumar, R., Ali, Z., Khan, I., Uddin, M. I., and Agbley, B. L. Y. (2023). Mcnn: a multi-level cnn model for the classification of brain tumors in iot-healthcare system. *Journal of Ambient Intelligence and Humanized Computing*, 14(5):4695–4706.

Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510.

Kang, J., Ullah, Z., and Gwak, J. (2021). Mri-based brain tumor classification using ensemble of deep features and machine learning classifiers. *Sensors*, 21(6):2222.

Kascenas, A., Sanchez, P., Schrempf, P., Wang, C., Clackett, W., Mikhael, S. S., Voisey, J. P., Goatman, K., Weir, A., Pugeault, N., et al. (2023). The role of noise in denoising models for anomaly detection in medical images. *Medical Image Analysis*, 90:102963.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-

resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690.

Lewis, P. R., Chandra, A., Parsons, S., Robinson, E., Glette, K., Bahsoon, R., Torresen, J., and Yao, X. (2011). A survey of self-awareness and its application in computing systems. In *2011 Fifth IEEE conference on self-adaptive and self-organizing systems workshops*, pages 102–107. IEEE.

Li, C., Lin, X., Mao, Y., Lin, W., Qi, Q., Ding, X., Huang, Y., Liang, D., and Yu, Y. (2022). Domain generalization on medical imaging classification using episodic training with task augmentation. *Computers in biology and medicine*, 141:105144.

Pan, X., Luo, P., Shi, J., and Tang, X. (2018). Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479.

Petrovska, A. (2021). Self-awareness as a prerequisite for self-adaptivity in computing systems. In *2021 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, pages 146–149. IEEE.

Rafiee, N., gholamipoorfard, R., and Kollmann, M. (2023). Abnormality detection for medical images using self-supervision and negative samples. *bioRxiv*, pages 2023–05.

Tschuchnig, M. E. and Gadermayr, M. (2022). *Anomaly Detection in Medical Imaging - A Mini Review*, pages 33–38. Springer Fachmedien Wiesbaden.

Xu, R., Zhang, X., Shen, Z., Zhang, T., and Cui, P. (2022). A theoretical analysis on independence-driven importance weighting for covariate-shift generalization. In *International Conference on Machine Learning*, pages 24803–24829. PMLR.

Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al. (2022). Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611.

Yang, J., Zhou, K., Li, Y., and Liu, Z. (2021). Generalized out-of-distribution detection: A survey.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013). Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. PMLR.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.