This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Generative Unsupervised Anomaly Detection with Coarse-Fine Ensemble for Workload Reduction in 3D Non-contrast Brain CT of Emergency Room

Jongjun Won[1] , Jihwan Kim[1] , Joonseo Oh[1] , Yereen Yoo[1] , Jieun Yum[1] , Joonsang Lee[1] , Joon Hyung Park[1] , Wooyoung Jo[1] , Yoojin Nam[1] , Hyunki Lee[1] , Gil-sun Hong[2(✉)] , Namkug Kim[1(✉)]

[1] Department of Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[2] Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, South Korea
{wjj910, hgs2013, namkugkim}@gmail.com

Abstract. Neurologic emergencies need to treat unspecified anomalies with various shapes, intensities, and locations in 3D non-contrast brain CT. However, in practice, patients with anomalies take a relatively small portion of total CT volumes. In this situation, excluding unremarkable scans could reduce radiologists' workload. We used a generative unsupervised anomaly detection (GUAD) with 3D Hierarchical Diffusion AutoEncoder (HDAE) model to develop this. In this study, we considered anomalies in two perspectives and made models. One is a Coarse-Morphological anomaly detection Model (CMM), and the other is a Fine-Grained anomaly detection Model (FGM). We ensembled these models' decisions for the exclusion of the unremarkable scans. Models were trained with normal scans of 28,510 from Asan Medical Center (AMC). For evaluation, we mainly used two consecutive test sets of 544 scans from AMC and 1,795 scans from Gangneung Asan Hospital (GNAH). Among clinically significant and unremarkable scans, our study showed [NPV (Negative Predictive Value)/workload reduction] of [98.1%/9.7%] and [96.7%/19.9%] for AMC and GNAH, respectively. Additionally, we used a public dataset (NPV of 98.5%) and five other external hospitals' hemorrhage sets (NPV of 96.0%) to evaluate robustness. Under the reasonable NPV, models showed the potential for workload reduction by omitting unremarkable scans. Compared to individual results of CMM or FGM, the ensembled decision usually shows NPV advantages. Also, with visual results, we observed our model could detect various types of anomalies.

Keywords: 3D Generative Model, Anomaly Detection, Workload Reduction, Coarse-Fine Ensemble.

1 Introduction

Generative models (GMs) achieved significant advancements in capturing the underlying data distribution, allowing them to generate realistic data. With various

applications of GMs in the medical domain [1, 2], sometimes GMs have been adopted in unsupervised anomaly detection tasks. Diffusion probabilistic models [3, 4] have been used in Generative Unsupervised Anomaly Detection (GUAD) [5-8]. In GUAD, GMs were trained only with normal data. They understood the normal data distribution and could capture the deviation of normal data. When we input normal data, GUAD usually reconstructs scans with little difference, but when we input abnormal data, reconstructed normal-like scans have a significant difference compared to abnormal input. After producing a difference map between input and reconstruction, post-processing yields anomaly scores in areas over the anomaly score threshold.

In the case of neurologic emergencies, various diseases should be treated in different sizes, shapes, locations, intensities, and prevalence rates. Non-contrast brain computed tomography (CT) is usually used as standard screening for fast scanning time. In practice, of the total medical scan readings, the volume of the abnormal group accounts for a smaller portion than that of the normal group. For this reason, excluding normal scans reduces the workload of radiologists. In chest radiographs, some previous studies show that commercial deep-learning software can reduce radiologists' workload by about 8 to 17% [9-12]. Using brain CT modality, we attempted to evaluate the workload reduction in neurological emergencies. We have assumed that GUAD may be an appropriate method for detecting non-specific abnormalities and, in the same sense, able to filter out unremarkable scans with no abnormalities. Also, GUAD has an efficiency that sets variations up as one class of "deviation from the normal group." Therefore, there is no need to build many supervised models for each anomaly. Our goal is to exclude some portion of the normal group by setting the anomaly decision cut-off as low as possible. Of course, achieving high Negative Predictive Values (NPV) is most important.

In this study, we suggest a perspective that anomalies need to be treated as two sides. One is a type of coarse morphological anomaly, and the other is a type of fine-grained anomaly. We build two models of coarse morphological anomaly detection model (CMM) (**Fig.1.a**) and fine-grained anomaly detection model (FGM) (**Fig.1.b**). Now, we describe our suggestion with visual results. On one side, we describe the need for CMM. As the **Fig.2. a, b**, when a model only focuses on reconstructing abnormal "intensity" areas to normal, its reconstruction cannot be considered normal images. Indeed, reconstruction of CMM with modification of the brain's morphology seems like normal images. CMM is necessary to catch morphological anomalies (**Fig.2. c**. Hydrocephalus, **Fig.2. d**. mass effect of hemorrhage). On the other side, we describe the need for FGM. Although the CMM model has a reasonable aspect for anomaly detection, it could change normal brain anatomy (e.g., ventricle or subarachnoid space) as a variation of the normal brain and sometimes yield many false positives (**Fig.2. e**). Detecting fine-grained anomalies like small infarctions among false positive areas may be challenging. So, when we train FGM, we give hard conditions such as brain anatomy segmentation mask, which consists of subarachnoid space, ventricle, and brain boundary [13]. Hard conditioning makes FGM yield relatively few false positives. As a result, as shown in **Fig.2. f**, we could set a low anomaly score threshold in the processing stage and find fine-grained anomalies. In sum, we use complementary models to treat anomalies. CMM is relatively easy to modify input scans' morphology to detect coarse morphological anomalies. FGM detects slight intensity gap anomalies or small-size anomalies.

As part of the anomaly decision process (**Fig.1.c**), we exclude scans predicted as unremarkable from both models simultaneously and consider that they do not need to be referred to radiologists. For the models' architecture (**Fig.1.a, b**), we used a 3D expanded hierarchical diffusion autoencoder [14], which consists of a semantic encoder to extract CT's feature and a DDPM [15] U-Net to output the same image as the input CT. HDAE is suitable for our task because it treats images from coarse to fine levels and has high reconstruction performance.

We evaluated our model's ability to distinguish remarkable cases with critical findings in brain CTs from unremarkable cases, which included normal and benign cases. Our model was tested using one internal dataset (AMC: Asan Medical Center) and three external datasets: external dataset 1 (GNAH: Gangneung Asan Hospital), external dataset 2 (an open dataset from the 2019 RSNA Brain Hemorrhage Challenge [16]), and external dataset 3 (a multi-institutional dataset of brain hemorrhages). Notably, the internal dataset and external dataset 1, both collected consecutively from emergency departments, represent real-world clinical data. To evaluate the clinical utility of our model, we focused on negative predictive value (NPV) as the primary evaluation metric and analyzed workload reduction (WLR) across various anomaly decision thresholds. Additionally, we assessed the benefits of using a decision ensemble of FGM and CMM to improve NPVs. Our code is available at https://github.com/Krying/WLR_ANO_3D. In summary, our study makes two key contributions:

- We demonstrate the potential of GUAD for workload reduction by excluding cases without critical findings using real-world datasets
- We propose an anomaly detection approach that considers coarse morphological and fine-grained anomalies through qualitative assessment.

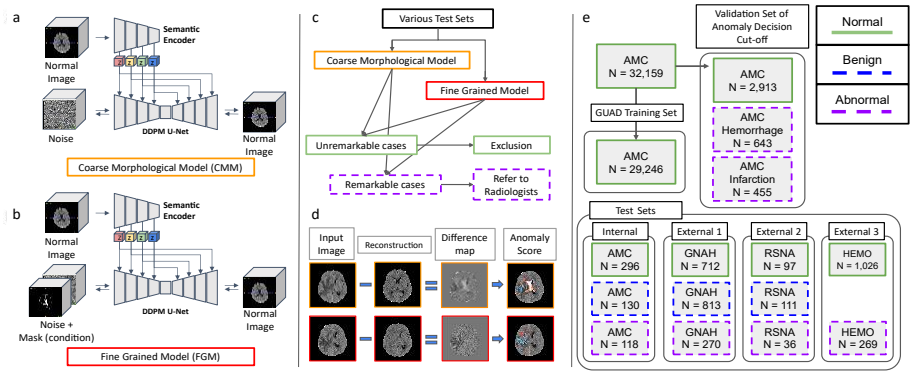

Fig. 1. The overall workflow of our study. (a) and (b) are the model architectures (HDAE) of CMM and FGM. (c) is the process of making anomaly decisions using FGM and CMM. (d) is the process of deriving anomaly scores with the difference map between input and reconstruction images. (e) is the data flow of the train, valid (for anomaly decision cut-off), and various test sets.

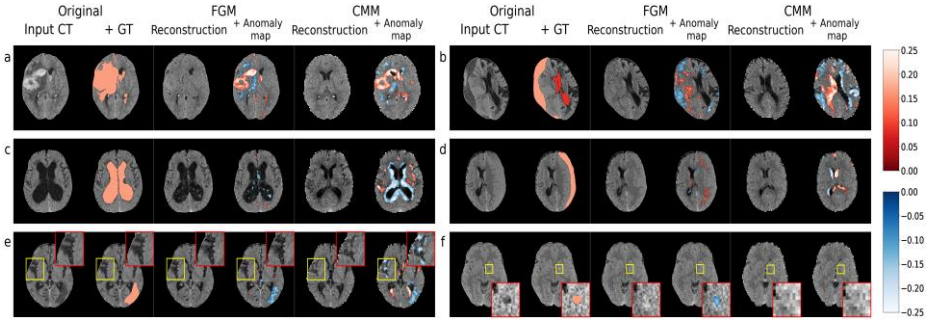

Fig. 2. Various cases of anomalies. (a) intracranial hemorrhage, (b) hemorrhage with hydrocephalus, (c) hydrocephalus, (d) subdural hemorrhage, (e) infarction, (f) hypoxic brain injury. Each case shows the original with lesion mask, reconstruction of FGM and CMM with anomaly map. The red and blue map indicates the hyper and the hypo intensity anomalies, respectively.

2 Material and Methods

2.1 Study Population

In **Fig1.e**, we summarized the data used for training, validation, and test. All training and validation data were collected retrospectively (from Jan. 2000 to Aug. 2018) from AMC. Among 32,159 normal scans, 29,246 were used for GUAD model training (age: 44.2±18.6 (Mean±SD) years; female: 53.5%). The remaining 2,913 normal cases, 643 hemorrhage cases, and 455 infarction cases form a validation set. This validation set was used to determine the cut-off of the anomaly decision. Test datasets consist of one internal test set and 3 external test sets. Test sets of internal (AMC) and external 1 (GNAH) were collected consecutively Feb. 2019 (one month) and from Jan. to May 2019 (3 months), respectively. For external dataset 2, we randomly selected normal cases of 97 from any-hemorrhage group among RSNA dataset [16]. Resembling the class ratio of external dataset 1, we randomly selected benign of 111 and abnormal of 36 cases. External dataset 3 comprised five other external hospitals that had hemorrhage and normal cases. Although this dataset is unsuitable for our theme, which aims to treat various diseases, we used it to confirm our models' robustness for other CT scanners (GE, SIMENS, PHILIPS, and TOSHIBA). The internal set included participants aged 58.7 ± 17.9 years, of whom 51.5% were female. External set 1 had ages of 61.5 ± 17.5 years with 47.7% female participants, and external set 3 had ages of 53.8 ± 18.1 years with 49.7% female participants. Abnormal scans consist of cases with hemorrhage, infarction, mass, hydrocephalus, and others.

2.2 Pre-processing and Post-processing for Anomaly Scoring

For preprocessing, the Brain Extraction Tool [17] was applied, and we performed depth-wise padding or crop to set all the scans' depth as 32 and then resized to 256 × 256 × 32 (for FGM) and 96 × 96 × 32 (for CMM) respectively. Intensity normalization

was performed Hounsfield unit [-10, 90] to [0, 1]. For postprocessing of anomaly scoring, in the difference map, logits under HU 7.5 (FGM) and 10 (CMM) were removed to suppress minor variations, and median smoothing (radius=1) was applied twice. Small objects with fewer than 75 voxels were removed, and all the absolute values of voxels were summed up.

2.3 Experiment Details

For here x is the input image, x_t is a noisy image, Z_{enc} are feature vectors, and p is the reverse process [15], HDAE [14] consists of a semantic encoder $Z_{enc} = Enc(x)$ and a conditional Denoising Diffusion Implicit Model (DDIM) [18] $p(x_{t-1} \mid x_{t,} Z_{enc})$ Instead of DDIM, we used the Denoising Diffusion Probabilistic Model (DDPM) [15]. The semantic encoder extracts semantic vectors from the input images, which are then fed into the corresponding DDPM U-Net layers as condition hierarchically. For the semantic conditioning method, we followed Diffusion AutoEncoder [19] and HDAE as adaptive group normalization.

For training FGM, we gave the condition of the segmentation mask (C_{mask}) to DDPM $p(x_{t-1} \mid x_{t,} Z_{enc}, C_{mask})$. We used a linear beta noise scheduler $T = 1000$ and only used 550 timesteps. Because our task is just the reconstruction of input image as a normal-like image, we did not need to focus on noise part $T > 550$. Where ϵ is Gaussian noise, and t is the time of a Gaussian diffusion process, we used the loss function defined in equations (2), and (3) for noise prediction, like DDPM. More mathematical formulas can be found in [15, 19]. For inference, we started inference from T=500 (FGM) and T=400 (CMM) to preserve patients' identical traits as possible. FGM was trained using normal scan size of 256 × 256 × 32, 10 epochs, learning rate (lr) of 3e-5, and batch size of 3. CMM was trained using normal scan size of 96 × 96 × 32, 22 epochs, lr of 4e-5, and batch size of 8. Experimentally, resolution of 96 × 96 × 32 is enough to find coarse morphological anomalies. Both models used a cosine decay lr scheduler [20] and were trained using an A100 GPU. We used PyTorch framework v2.5.1, CUDA 12.1, and MONAI [21] library v1.4.0.

$$Z_{enc} = Enc(x) \tag{1}$$

$$L_{CMM} = \sum_{t=1}^{T} \mathbb{E}_{x_0,\epsilon_t} \left[\left\| \epsilon_\theta \left(x_t, t, z_{enc\ of\ CMM} \right) - \epsilon_t \right\|_2^2 \right] \tag{2}$$

$$L_{FGM} = \sum_{t=1}^{T} \mathbb{E}_{x_0,\epsilon_t} \left[\left\| \epsilon_\theta \left(x_t, t, z_{enc\ of\ FGM}, C_{mask} \right) - \epsilon_t \right\|_2^2 \right] \tag{3}$$

2.4 Evaluation: NPV and Workload reduction

We used two evaluation metrics NPV and WLR. Metrics' formulation is described in equations (4) and (5). We considered unremarkable cases as negative class.

$$NPV = \frac{(True\ Negative)}{(True\ Negative + False\ Negative)} \tag{4}$$

$$WLR = \frac{(True\ Unremarkable\ scnas)}{(Unremarkable\ scans + Remarkable\ scans)} \tag{5}$$

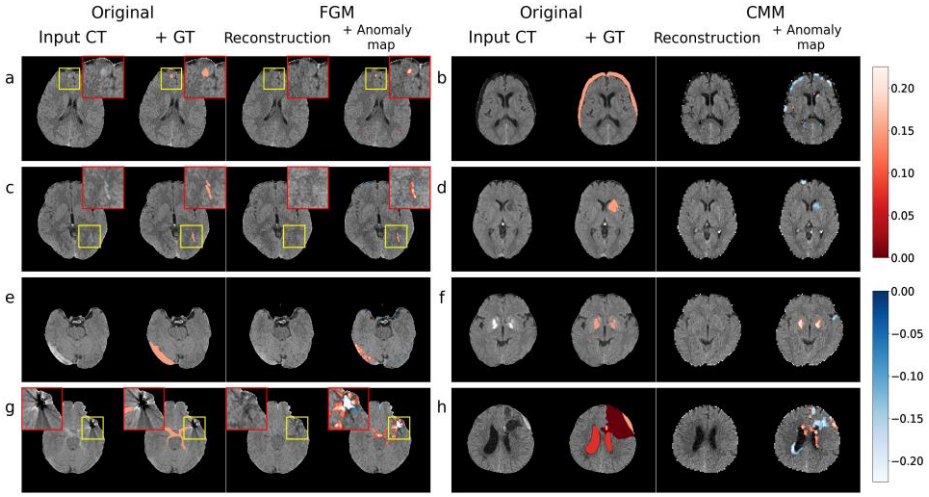

Fig. 3. Various cases of anomalies. (a) cavernoma, (b) subdural hygroma, (c) arteriovenous malformation, (d) infarction, (e) cerebral venous thrombus, (f) metabolic disease, (g) subarachnoid hemorrhage with metal artifact, (h) hemorrhage with hydrocephalus and intraparenchymal mass. Results of (a), (c), (e), and (g) are from FGM, and results of (b), (d), (f), and (h) are from CMM. The red and blue map indicates the hyper and the hypo intensity anomalies, respectively.

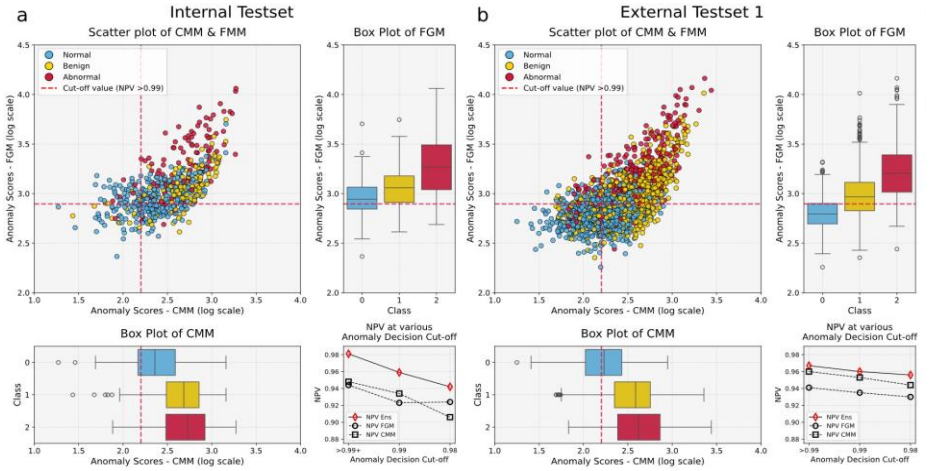

Fig. 4. Scatter plots and box plots of Anomaly Scores and plots of NPVs at various anomaly thresholds. (a) and (b) are the plots of internal and external 1 respectively. Scatter plots indicate the log scaled anomaly scores (x-axis means scores of CMM, and y-axis means scores of FGM). In box plots of anomaly score, class 0, 1, and 2 means 'normal', 'benign', and 'abnormal' respectively. Plots of "NPV at various anomaly decision cut-off values" (lower right of each figure) display the NPVs of FGM, CMM, and ensembled decision. The red markers indicate the NPVs of ensembled decision from CMM and FGM. The black markers indicate each model's NPVs.

Table 1. NPVs and WLRs for test sets at various anomaly decision cut-off (NPVs of 0.98, 0.99, and more than 0.99) from validation set.

Cut-off	Metrics	Internal test	External test 1	External test 2	External test 3
>99%	NPV	98.1% (53/54)	96.7% (357/369)	98.5% (64/65)	96.0% (360/375)
	WLR	9.7% (53/544)	19.9% (357/1795)	26.2% (64/244)	27.8% (360/1295)
99%	NPV	95.9% (71/74)	96.0% (433/451)	97.1% (68/70)	96.4% (429/445)
	WLR	13.1% (71/544)	24.1% (433/1795)	27.9% (68/244)	33.1% (429/1295)
98%	NPV	94.2% (97/103)	95.6% (503/526)	96.5% (83/86)	96.2% (477/496)
	WLR	17.8% (97/544)	28.0% (503/1795)	34.0% (83/244)	36.8% (477/1295)

Table 2. Comparison of NPVs of FGM, CMM and Ensembled results for test sets at various anomaly decision cut-off (NPVs of 98%, 99%, and more than 99%) from validation set.

Test Sets	Internal test			External test 1			External test 2			External test 3		
Cut-off	>99%	99%	98%	>99%	99%	98%	>99%	99%	98%	>99%	99%	98%
FGM	0.949	0.934	0.906	0.960	0.953	0.944	0.953	0.956	0.927	0.949	0.946	0.939
CMM	0.944	0.923	0.924	0.941	0.935	0.930	**0.989**	**0.971**	**0.971**	0.945	0.944	0.946
Ensemble	**0.981**	**0.959**	**0.942**	**0.967**	**0.960**	**0.956**	0.985	**0.971**	0.965	**0.960**	**0.964**	**0.962**

3 Results and Discussion

3.1 Workload Reduction with Reasonable NPV

We evaluated the NPV and WLR of our models at different anomaly decision cut-offs (98%, 99%, and >99%) across internal and external test datasets (**Table 1**). For the highest threshold (>99%), the NPV across the four test sets ranged from 96.0% to 98.5%, while the corresponding WLR varied between 9.7% and 27.8%. As expected, lowering the anomaly decision cut-off to 99% led to a slight decrease in NPV (ranging from 95.9% to 97.1%), but with an increase in WLR (13.1%–33.1%). Further reducing the cut-off to 98% resulted in a more pronounced trade-off, with NPVs between 94.2% and 96.5%, accompanied by WLRs of 17.8%–36.8%. These results indicate that higher cut-offs ensure higher NPVs but limit the extent of workload reduction. Conversely, lower thresholds yield greater workload reduction but at the cost of reduced NPV.

3.2 The Effect of Model Ensemble

To assess the impact of our ensemble approach, we compared NPVs of the individual models (FGM and CMM) against the ensembled model at different cut-offs (**Table 2**). For critical findings, excluding external test 2, the ensembled model outperformed the individual models across all thresholds. For the real-world dataset (internal, external dataset 1), at NPV of >99%, the ensemble achieved an NPV of 98.1% for internal test, compared to 94.9% for FGM and 94.4% for CMM. 96.7% and, for external test 1, the

ensemble achieved an NPV of 96.7%, compared to 96.0% for FGM and 94.1% for CMM. The coarse-fine ensemble improved NPV performance over FGM and CMM alone. These results suggest that combining FGM and CMM enhances model reliability, in identifying critical cases with high NPV. This approach would be particularly valuable in settings where reducing false negatives is paramount, such as emergency departments and acute care settings.

3.3 Balancing Sensitivity and Workload Reduction

An important consideration in deploying AI models for clinical decision support is the balance between sensitivity (ensuring high NPV) and WRL. Our study shows that setting a higher anomaly decision cut-off (e.g. > 99%) maintains NPVs close to 98% while still achieving a reasonable WLR (~10–35%). However, a more relaxed threshold (e.g., 98%) increases WLR (~20–44%) at the cost of a slight NPV decline. Institutions must determine the optimal threshold based on their specific patient population and clinical workflow constraints. These findings highlight the importance of carefully selecting decision thresholds to minimize missed critical cases while maintaining workload efficiency.

3.4 Various Anomaly Cases and Misses Cases

Fig. 3 shows the cases of anomaly detection exist in real-world conditions. In real-world conditions, at NPV of >99%, our model classified 54 cases as unremarkable in the internal test dataset, among which one case was later identified as critical. In the external dataset, 12 out of 369 cases classified as unremarkable were later found to be critical. This underscores the need for human oversight and potential hybrid models where AI serves as an assistive tool rather than a sole decision-maker.

3.5 Limitation

The primary limitation of our study is the lack of comparative analysis. To the best of our knowledge, no prior studies have specifically addressed workload reduction (WLR) on brain CT. Our study may serve as one of the initial steps in this research direction. Given the novelty of the topic, our primary goal was to demonstrate the clinical relevance and potential impact of applying deep learning in this context. Additionally, the model's inference time per scan was approximately 88 seconds for CMM and 120 seconds for FGM on an RTX 3090 GPU (VRAM < 6 GB). While these times are reasonable, there is potential for further optimization.

4 Conclusion

Our model has the potential to significantly alleviate the workload in emergency radiology by prioritizing cases requiring urgent review. The ability to maintain high NPVs across diverse datasets suggests that the model could be deployed across different institutions with minimal performance degradation. Additionally, we suggest a new

ensemble approach to enhance the model's performance. Future work should focus on real-time deployment simulations and evaluating the impact on radiologist efficiency. In conclusion, our study provides strong evidence that deep learning-based anomaly detection can serve as a reliable tool for optimizing radiology workflows, improving efficiency, and maintaining diagnostic accuracy in emergency and acute care settings.

Acknowledgments. This research was supported by grants from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI22C0471, RS-2022-KH130266); the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT) (RS-2024-00355370, RS-2025-00553009); and the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Korea (2024IP0058).

Disclosure of Interests. The authors have no competing interests.

References

1. Kazerouni, A., et al., *Diffusion models in medical imaging: A comprehensive survey.* Medical image analysis, 2023. **88**: p. 102846.
2. Yang, L., et al., *Diffusion models: A comprehensive survey of methods and applications.* ACM Computing Surveys, 2023. **56**(4): p. 1-39.
3. Sohl-Dickstein, J., et al. *Deep unsupervised learning using nonequilibrium thermodynamics*. in *International conference on machine learning*. 2015. pmlr.
4. Song, Y., et al., *Score-based generative modeling through stochastic differential equations.* arXiv preprint arXiv:2011.13456, 2020.
5. Wyatt, J., et al. *Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
6. Bercea, C.I., et al., *Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models.* arXiv preprint arXiv:2305.19643, 2023.
7. Fontanella, A., et al., *Diffusion models for counterfactual generation and anomaly detection in brain images.* IEEE Transactions on Medical Imaging, 2024.
8. Behrendt, F., et al., *Guided reconstruction with conditioned diffusion models for unsupervised anomaly detection in brain mris.* Computers in Biology and Medicine, 2025. **186**: p. 109660.
9. Schalekamp, S., et al., *Performance of AI to exclude normal chest radiographs to reduce radiologists' workload.* European Radiology, 2024: p. 1-9.
10. Plesner, L.L., et al., *Using AI to identify unremarkable chest radiographs for automatic reporting.* Radiology, 2024. **312**(2): p. e240272.
11. Dyer, T., et al., *Diagnosis of normal chest radiographs using an autonomous deep-learning algorithm.* Clinical radiology, 2021. **76**(6): p. 473. e9-473. e15.
12. Plesner, L.L., et al., *Autonomous chest radiograph reporting using AI: estimation of clinical impact.* Radiology, 2023. **307**(3): p. e222268.

13. Cai, J.C., et al., *Fully automated segmentation of head CT neuroanatomy using deep learning.* Radiology: Artificial Intelligence, 2020. **2**(5): p. e190183.
14. Lu, Z., et al. *Hierarchical diffusion autoencoders and disentangled image manipulation*. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024.
15. Ho, J., A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models.* Advances in neural information processing systems, 2020. **33**: p. 6840-6851.
16. Flanders, A.E., et al., *Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge.* Radiology: Artificial Intelligence, 2020. **2**(3): p. e190211.
17. Akkus, Z., et al., *Robust brain extraction tool for CT head images.* Neurocomputing, 2020. **392**: p. 189-195.
18. Song, J., C. Meng, and S. Ermon, *Denoising diffusion implicit models.* arXiv preprint arXiv:2010.02502, 2020.
19. Preechakul, K., et al. *Diffusion autoencoders: Toward a meaningful and decodable representation*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
20. Loshchilov, I. and F. Hutter, *Sgdr: Stochastic gradient descent with warm restarts.* arXiv preprint arXiv:1608.03983, 2016.
21. Pinaya, W.H., et al., *Generative AI for medical imaging: extending the MONAI framework. arXiv.* arXiv preprint arXiv:2307.15208, 2023.