A Comparative Study of Deep Learning-Based Anomaly Detection for Intracranial Hemorrhage in Brain CT

Joon Hyuk Lee[1] LEEJOON2067@KHU.AC.KR
Hyeon Jun Park[2] GUSWNS0429@KHU.AC.KR
Tae-Seong Kim[3] TSKIM@KHU.AC.KR
Won Hee Lee[*1,2] WHLEE@KHU.AC.KR
[1] *Department of Software Convergence, Kyung Hee University, Yongin, Korea*
[2] *Department of Artificial Intelligence, Kyung Hee University, Yongin, Korea*
[3] *Department of Biomedical Engineering, Kyung Hee University, Yongin, Korea*

Editors: Under Review for MIDL 2026

Abstract

Deep learning-based anomaly detection (AD) offers a label-efficient approach to intracranial hemorrhage (ICH) detection in brain CT, yet systematic comparisons of AD methods on this clinically critical task remain scarce. We benchmark 11 reconstruction-based and self-supervised deep learning AD methods on the CQ500 dataset under consistent experimental conditions. Our results reveal a substantial performance gap between the two paradigms: reconstruction-based methods achieve up to AUC 90.8, while self-supervised approaches peak at AUC 64.7, suggesting that synthetic anomaly strategies designed for natural images do not transfer well to brain CT. Among reconstruction-based methods, uncertainty-aware modeling consistently provides the most discriminative anomaly scores. This study presents a systematic benchmark of deep learning AD methods for ICH detection, offering concrete guidance for method selection in automated brain CT screening.

Keywords: Anomaly Detection, Deep Learning, Intracranial Hemorrhage, Brain, CT

1. Introduction

Automated detection of intracranial hemorrhage (ICH) in brain CT is a clinically urgent task, as timely diagnosis is critical for patient outcomes (Choi et al., 2024). Deep learning has driven significant progress in medical image analysis (Litjens et al., 2017), and unsupervised anomaly detection (AD) has emerged as a particularly promising direction for pathology detection, as it does not require large annotated datasets of abnormal cases (Pang et al., 2021). Despite a growing body of work, systematic comparisons of deep learning-based AD methods on brain CT remain scarce, leaving practitioners with limited guidance on method selection.

In this work, we present a benchmark evaluation of deep learning-based anomaly detection methods for ICH detection in brain CT. We systematically compare representative approaches spanning reconstruction-based and self-supervised learning paradigms under consistent experimental conditions, using a publicly available brain CT dataset. This provides a structured analysis of slice-level detection performance across methods, offering practical insights for the development of automated ICH screening systems.

2. Methods

Dataset and Image Preprocessing We used the CQ500 dataset (Chilamkurthy et al., 2018), a publicly available brain CT dataset comprising 491 scans with ICH, fractures, midline shifts, and mass effect. Following the refined labeling protocol of CTFSH (Guerra et al., 2022), 364 patients were retained (117 normal, 247 pathological). Models were trained exclusively on normal data, with anomalies identified as deviations from the learned normal distribution. The dataset was split at the patient level to prevent data leakage, and each volume was processed at the 2D slice level. All DICOM files underwent HU conversion and windowing; slices were resized to 128×128 and normalized to $[-1, 1]$. We adopted a 3-channel input using complementary window settings – brain (40/80 HU), subdural (75/215 HU), and bone (600/2800 HU) – to capture tissue characteristics relevant to ICH detection (Wang et al., 2021).

Reconstruction AD We evaluated seven AutoEncoder-based models. The standard AE was trained with five loss variants: MSE, L1, gradient difference (Zimmerer et al., 2019), SSIM (Bergmann et al., 2019), and perceptual loss using a pretrained VGG-19 as a fixed feature extractor (Shvetsova et al., 2021). AE with Uncertainty (AE-U) (Mao et al., 2020) incorporates pixel-wise uncertainty estimation, scoring anomalies by uncertainty-weighted reconstruction error. Variational AE (VAE) (Kingma and Welling, 2022) learns a probabilistic latent space regularized by a standard normal prior, with anomaly scoring based on reconstruction error. Denoising AE (DAE) (Kascenas et al., 2022) reconstructs clean images from Gaussian-noised inputs using a UNet-based architecture (Ronneberger et al., 2015), improving robustness to local perturbations. Context-encoding AE (CeAE) (Zimmerer et al., 2018) enforces context-aware reconstruction through random patch masking during training. Constrained AE (Chen and Konukoglu, 2018) promotes latent consistency by penalizing discrepancies between original and re-encoded latent representations. GANomaly (Akcay et al., 2018) employs an adversarially trained encoder-decoder-encoder architecture, scoring anomalies by latent space discrepancy between input and reconstruction.

Self-supervised Learning AD All self-supervised methods share a ResNet-based encoder-decoder architecture (He et al., 2015) trained with pixel-level binary cross-entropy loss $\mathcal{L}_{\text{BCE}}(\hat{y},\ y_{\text{pseudo}})$, where pseudo-labels are derived from synthetic anomalies. Methods differ in their patch augmentation strategy. NSA (Schlüter et al., 2022) applies Poisson-blended patch transplantation with binary or intensity-weighted soft labels. FPI (Tan et al., 2022) linearly interpolates a foreign patch to yield continuous labels, with an optional Poisson-blended variant (FPI-Poisson). CutPaste (Li et al., 2021) swaps same-image patches with binary labels; its two-stage variant extends this by fitting a Gaussian density estimator on learned embeddings for anomaly scoring, replacing pixel-level prediction. AnatPaste (Sato et al., 2022) pastes a same-image patch onto an anatomically plausible region guided by a segmentation mask, with Gaussian-blended boundaries and binary labels.

Evaluation The dataset comprises 88,808 slices in total (27,562 normal, 61,246 ICH). We used 5-fold cross-validation with splits stratified at the patient level to prevent data leakage. Training folds contain only normal slices (18,187–19,460), while validation folds (normal: 2,692–3,578; ICH: 6,656–7,439) and test folds (normal: 4,797–6,683; ICH: 11,896–12,882) include both normal and pathological slices from held-out patients. AD performance was assessed at the slice level, and results are reported as mean area under the ROC curve (AUC) and average precision (AP) across five folds.

Table 1: Slice-level anomaly detection performance on brain CT. **Bold** denotes best per group

	Method	Loss Function	Backbone	Anomaly Score	Score	
					AUC	AP
Reconstruction AD	AE	$\mathcal{L}_0 = \ell_2(x, \hat{x})$	AE	$\|x - \hat{x}\|_2^2$	$86.6_{\pm 5.0}$	$89.3_{\pm 4.9}$
	AE-ℓ_l	$\ell_1(x, \hat{x})$	AE	$\|x - \hat{x}\|_1$	$78.3_{\pm 8.6}$	$80.5_{\pm 6.3}$
	AE-Grad	$\mathcal{L}_0$	AE	$\left\|\frac{\partial \|x - \hat{x}\|_2^2}{\partial x}\right\|$	$83.9_{\pm 3.3}$	$88.3_{\pm 3.8}$
	AE-SSIM	$1 - \ell_{\text{SSIM}}(x, \hat{x})$	AE	$(1 - \text{SSIM}(x, \hat{x}))$	$85.1_{\pm 6.9}$	$88.3_{\pm 5.4}$
	AE-PL	$\text{PeL}(x, \hat{x}).\text{mean}()$	AE+VGG_{19}	$\text{PeL}(x, \hat{x})$	$86.5_{\pm 4.3}$	$90.1_{\pm 4.4}$
	AE-U	$\ell_2(x, \hat{x}) \cdot \frac{10}{\sigma_i^2} + \log \sigma_i^2$	AE	$\|x - \hat{x}\|_2^2 \cdot \frac{1}{\sigma^2}$	$\mathbf{90.8}_{\pm 1.2}$	$\mathbf{93.7}_{\pm 2.1}$
	VAE	$\mathcal{L}_0 + \lambda \mathcal{L}_{\text{KL}}$	VAE	$\|x - \hat{x}\|_2^2$	$87.9_{\pm 3.6}$	$90.5_{\pm 4.4}$
	DAE	$\|x - f_\theta(\tilde{x})\|_2^2$	UNet	$\|x - f_\theta(x)\|_2^2$	$84.9_{\pm 6.8}$	$91.0_{\pm 2.9}$
	CeAE	$\|x - f_\theta(x_{\text{mask}})\|_2^2$	AE	$\|x - \hat{x}\|_2^2$	$84.7_{\pm 4.9}$	$88.2_{\pm 4.3}$
	Constrained AE	$\mathcal{L}_0(x, \hat{x}) + \ell_2(z, \hat{z})$	AE	$\|x - \hat{x}\|_2^2$	$84.7_{\pm 3.6}$	$89.2_{\pm 2.4}$
	GANomaly	$\lambda_{\text{Adv}}\mathcal{L}_{\text{Adv}} + \lambda_{\text{Rec}}\mathcal{L}_{\text{Rec}} + \lambda_{\text{Enc}}\mathcal{L}_{\text{Enc}}$	GAN	$\|z - \hat{z}\|_2^2$	$59.7_{\pm 8.4}$	$76.2_{\pm 6.4}$
Self-supervised AD	NSA	$\mathcal{L}_{\text{BCE}}(g_s(f(x^{\text{nsa}})),\ y_{\text{pseudo}})$	ResNet18	$\|g_s(f(x))\|$	$61.6_{\pm 6.2}$	$76.5_{\pm 4.2}$
	NSA-Intensity	$\mathcal{L}_{\text{BCE}}(g_s(f(x^{\text{nsa}})),\ \sigma(k \cdot \Delta I))$	ResNet18	$\|g_s(f(x))\|$	$62.9_{\pm 6.1}$	$76.8_{\pm 5.6}$
	FPI	$\mathcal{L}_{\text{BCE}}(g_s(f(x^{\text{fpi}})),\ \alpha)$	ResNet18	$\|g_s(f(x))\|$	$\mathbf{64.7}_{\pm 5.9}$	$\mathbf{79.1}_{\pm 3.9}$
	FPI-Poisson	$\mathcal{L}_{\text{BCE}}(g_s(f(x^{\text{fpi}})),\ \alpha)$	ResNet18	$\|g_s(f(x))\|$	$49.8_{\pm 5.5}$	$70.1_{\pm 3.7}$
	CutPaste	$\mathcal{L}_{\text{BCE}}(g_s(f(x^{\text{cp}})),\ y_{\text{pseudo}})$	ResNet18	$\|g_s(f(x))\|$	$63.5_{\pm 8.9}$	$77.4_{\pm 9.4}$
	CutPaste (2 stage)	$\mathcal{L}_{\text{CE}}(g_c(f(x)), 0) + \mathcal{L}_{\text{CE}}(g_c(f(x^{\text{cp}})), 1)$	ResNet18	$d_M(\phi(f(x)))$	$55.6_{\pm 5.9}$	$72.4_{\pm 6.8}$
	AnatPaste	$\mathcal{L}_{\text{CE}}(g_c(f(x)), 0) + \mathcal{L}_{\text{CE}}(g_c(f(x^{\text{ap}})), 1)$	ResNet18	$d_M(\phi(f(x)))$	$54.1_{\pm 3.8}$	$70.1_{\pm 2.0}$

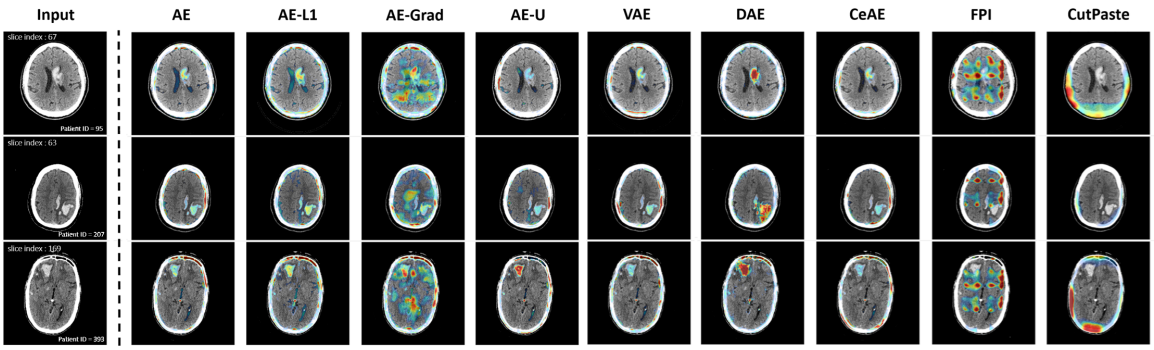

Figure 1: Qualitative comparison of anomaly heatmaps for three representative ICH cases. Warmer colors indicate higher anomaly scores: reconstruction error for AE-based models and predicted anomaly probability for SSL models. All maps are normalized independently per model.

3. Results and Discussion

Table 1 shows slice-level AD performance across all evaluated methods. Reconstruction-based methods consistently outperform self-supervised approaches, with AE-U achieving the best overall performance (AUC 90.8, AP 93.7). Most AE variants perform competitively in the AUC range of 83–88, indicating a moderate effect of loss function choice. GANomaly substantially underperforms (AUC 59.7, AP 76.2), likely due to adversarial training instability on the relatively small dataset. Among self-supervised methods, FPI achieves the best result (AUC 64.7, AP 79.1), yet the substantial gap with reconstruction-based methods suggests that synthetic anomaly strategies do not transfer well to brain CT, where ICH manifests as subtle intensity variations rather than texture anomalies. Qualitative results in Figure 1 further support this, with reconstruction-based methods producing better-localized heatmaps compared to the diffuse activations of self-supervised methods.

Acknowledgments This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korea government (MSIT), under grants RS-2024-00509257 (Global AI Frontier Lab) and IITP-2026-RS-2024-00438239 (ITRC); in part by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare (RS-2025-02293110); and in part by the National Research Foundation of Korea (NRF), funded by MSIT (RS-2026-25485605).

References

Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training, 2018. URL https://arxiv.org/abs/1805.06725.

Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, page 372–380. SCITEPRESS - Science and Technology Publications, 2019. doi: 10.5220/0007364503720380. URL http://dx.doi.org/10.5220/0007364503720380.

Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders, 2018. URL https://arxiv.org/abs/1806.04972.

Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G. Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. Development and validation of deep learning algorithms for detection of critical findings in head ct scans, 2018. URL https://arxiv.org/abs/1803.05854.

So Yeon Choi, Ji Hoon Kim, Hyun Soo Chung, Sona Lim, Eun Hwa Kim, and Arom Choi. Impact of a deep learning-based brain ct interpretation algorithm on clinical decision-making for intracranial hemorrhage in the emergency department. *Scientific Reports*, 14 (1):22292, 2024.

Michele Guerra, Abhijeet Parida, Daniel Rueckert, Mehmet Yigitsoy, and Shadi Albarqouni. CTFSH: Full head CT anomaly detection with unsupervised learning, 2022. URL https://openreview.net/forum?id=9qZUArz732o.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Antanas Kascenas, Nicolas Pugeault, and Alison Q. O'Neil. Denoising autoencoders for unsupervised anomaly detection in brain mri. In Ender Konukoglu, Bjoern Menze, Archana Venkataraman, Christian Baumgartner, Qi Dou, and Shadi Albarqouni, editors, *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pages 653–664. PMLR, 06–08 Jul 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization, 2021. URL https://arxiv.org/abs/2104.04015.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, December 2017. ISSN 1361-8415. doi: 10.1016/j.media.2017.07.005. URL http://dx.doi.org/10.1016/j.media.2017.07.005.

Yifan Mao, Fei-Fei Xue, Ruixuan Wang, Jianguo Zhang, Wei-Shi Zheng, and Hongmei Liu. *Abnormality Detection in Chest X-Ray Images Using Uncertainty Prediction Autoencoders*, pages 529–538. 09 2020. ISBN 978-3-030-59724-5. doi: 10.1007/978-3-030-59725-2_51.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, March 2021. ISSN 1557-7341. doi: 10.1145/3439950. URL http://dx.doi.org/10.1145/3439950.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Junya Sato, Yuki Suzuki, Tomohiro Wataya, Daiki Nishigaki, Kosuke Kita, Kazuki Yamagata, Noriyuki Tomiyama, and Shoji Kido. Anatomy-aware self-supervised learning for anomaly detection in chest radiographs, 2022. URL https://arxiv.org/abs/2205.04282.

Hannah M. Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization, 2022. URL https://arxiv.org/abs/2109.15222.

Nina Shvetsova, Bart Bakker, Irina Fedulova, Heinrich Schulz, and Dmitry V. Dylov. Anomaly detection in medical imaging with deep perceptual autoencoders. *IEEE Access*, 9:118571–118583, 2021. ISSN 2169-3536. doi: 10.1109/access.2021.3107163. URL http://dx.doi.org/10.1109/ACCESS.2021.3107163.

Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, and Bernhard Kainz. Detecting outliers with foreign patch interpolation. *Machine Learning for Biomedical Imaging*, 1 (April 2022):1–27, April 2022. ISSN 2766-905X. doi: 10.59275/j.melba.2022-e651. URL http://dx.doi.org/10.59275/j.melba.2022-e651.

Xiyue Wang, Tao Shen, Sen Yang, Jun Lan, Yanming Xu, Minghui Wang, Jing Zhang, and Xiao Han. A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head ct scans. *NeuroImage: Clinical*, 32:102785, 2021. ISSN 2213-1582. doi: 10.1016/j.nicl.2021.102785. URL https://www.sciencedirect.com/science/article/pii/S2213158221002291.

David Zimmerer, Simon A. A. Kohl, Jens Petersen, Fabian Isensee, and Klaus H. Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection, 2018. URL https://arxiv.org/abs/1812.05941.

David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. Unsupervised anomaly localization using variational auto-encoders, 2019. URL https://arxiv.org/abs/1907.02796.