Demo: Harnessing Generative AI for Comprehensive Evaluation of Medical Imaging AI

Yisak Kim[1,2,3,*], **Seunghyun Jang**[1,2,3,*], **Soyeon Kim**[1,3], **Kyungmin Jeon**[3], and **Chang-Min-Park**[1,2,3]

[1]Interdisciplinary Program in Bioengineering, Seoul National University Graduate School
[2]Integrated Major in Innovative Medical Science, Seoul National University Graduate School
[3]Department of Radiology, Seoul National University Hospital
{yisakk, jjaa0113, ksy_9396, kmjeon, morphius}@snu.ac.kr
[*]These authors contributed equally to this work.

Abstract

Evaluating AI models in the field of medical imaging, particularly for tasks such as nodule detection, is a challenging endeavor due to the scarcity of large, diverse, and well-annotated datasets. These constraints hinder the ability to accurately assess model performance and limit generalization to real-world scenarios. To address these challenges, we introduce SynNodBench, a generative AI-based demo that generates synthetic lung nodules of varying sizes, shapes, and locations on chest X-rays. Our method employs a diffusion-based inpainting model trained on the NODE21 dataset, allowing for the creation of realistic and customizable synthetic nodules.

We conducted multiple experiments to illustrate the utility of the demo in understanding nodule detection model behavior. First, by generating a large-scale synthetic test set, we were able to identify a positive correlation between nodule size and model confidence, a relationship that was not observed with smaller real-world datasets. We also demonstrated how the number of nodules in an image influences detection sensitivity, finding that the presence of additional nodules can increase sensitivity in detecting otherwise missed lesions. In another experiment, we examined whether a nodule detection model would correctly ignore nodules in anatomically impossible regions, such as air-leak areas, and confirmed the model's robustness in these cases. Our findings show that using synthetic data provides a scalable and effective solution for evaluating AI models in healthcare.

1 Introduction

Evaluating AI models for biomedical image analysis requires large and diverse datasets[1], but obtaining such data is challenging due to ethical, legal, geographical, and financial constraints. This scarcity can lead to biased testing datasets, limiting the accuracy of model performance evaluation and restricting their ability to generalize to real-world scenarios[2, 3]. Furthermore, ensuring the transparency and robustness of medical AI models demands comprehensive evaluations[4], which typically require densely annotated datasets that are both costly and time-consuming to produce.

For instance, a recent study[5] on chest X-ray nodule detection, using a test set of 144 nodules, found no significant correlation between nodule size and detection rate, despite the expectation that larger nodules should be easier to detect. This unexpected result highlights the difficulty of analyzing correlations between detection rates and factors like nodule size in small datasets, where other variables such as nodule location, opacity, and shape can have a significant impact.

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

Several studies utilize synthetic data to address the issue of data scarcity in the medical field[6]. Most of this research has focused on employing synthetic data for model training, and it has been reported that this approach can significantly enhance performance[7, 8, 9]. We aim to demonstrate that synthetic data can also be valuable for evaluation. Some recent research has explored the use of generative AI to create synthetic test sets[10, 11, 12]. For example, studies have used synthetic tabular data to reduce evaluation errors in minority groups[11], or to enhance the interpretability of models that assess skin cancer[12]. In our research, we trained a diffusion-based model[13] to inpaint nodules in chest X-rays, generating synthetic nodule datasets to evaluate nodule detection models. We created a demo called *SynNodBench*, which allows users to generate nodules with desired size, shape, and location in any chest X-ray. The tool also enables conditioning on radiomic features, providing a comprehensive framework for evaluating nodule detection models. Additionally, we have made YOLOv8 and YOLOv9[14] models, trained on the NODE21 dataset[15], available as examples for evaluation.

In our demonstration of SynNodBench, we analyzed the correlation between nodule size and model confidence. Unlike previous research, which found no correlation in 144 nodules, our large-scale synthetic test set revealed a clear positive correlation between nodule size and model confidence. Our synthetic approach also highlighted the impact of other factors, such as decreased detection accuracy near the hilar region and diaphragm.

The SynNodBench demo offers a robust solution to data scarcity and annotation challenges in chest X-ray lung cancer AI research, enabling more thorough and reliable model assessments.

2 Methods

2.1 Dataset

We utilized the NODE21 dataset[15] to train the nodule inpainting and detection models. NODE21 dataset is a comprehensive collection of frontal chest x-rays specifically curated for pulmonary nodule detection. It comprises 4,882 images, of which 1,134 contain 1,476 annotated nodules, each delineated by bounding boxes. The remaining 3,748 images serve as negative samples, containing no nodules. This dataset is constructed from publicly available repositories that permit remixing and redistribution, including the Japanese Society of Radiological Technology (JSRT) dataset[16], PadChest[17], ChestX-ray14[18], and the Open-I dataset[19]. The diverse origins of these images contribute to a robust and representative sample, essential for developing and evaluating effective nodule detection and generation tasks.

For the nodule inpainting model, we augmented the dataset by extracting 256x256 pixel patches from the original images. This resulted in a larger dataset comprising 14,764 patches for training (13,492 normal, 1,272 with nodules) and 1,652 patches for testing (1,500 normal, 152 with nodules). For the nodule detection model, we employed a dataset split of 4,393 images for training (3,373 normal, 1,020 with nodules) and 489 images for testing (375 normal, 114 with nodules).

2.2 Lung Nodule Inpainting Model

We employed a diffusion-based architecture for nodule inpainting, based mainly on RadiomicsFill-Mammo[13], an inpainting framework designed for generating mass lesions in masked regions of mammograms. To enhance tumor generation, the standard text encoder is replaced with a tabular encoder (MET)[20] that leverages radiomics features[21] such as shape, histogram, and textures. The denoising U-Net is fine-tuned for mass inpainting, improving the realism and clinical relevance of the generated images.

Our research adapts this approach for chest X-ray images and nodule inpainting as shown in figure 1. We utilized the NODE21 dataset to segment nodules and extract radiomics features. The structure of the tabular encoder is identical to the original MET[20]. We trained it using the extracted radiomics features while progressively increasing the masking rate up to 0.9 to enhance robustness. Unlike breast masses, where clinical conditions are categorized, the presence or absence of nodules is indicated.

After training the tabular encoder, we froze its parameters and used its features to condition the denoising U-Net, which was then fine-tuned for nodule inpainting. We maintained the original

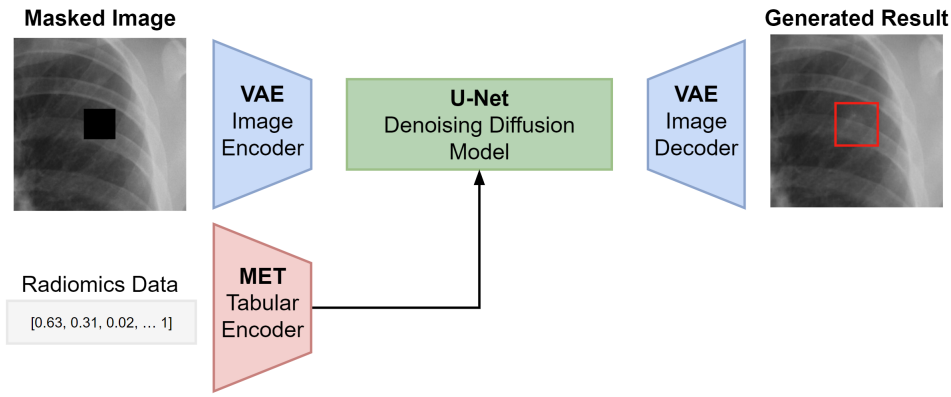


Figure 1: Overview of the nodule inpainting model, a stable diffusion-based architecture. This model performs iterative denoising for nodule inpainting within masked regions on a noisy latent vector, utilizing information from unmasked regions and specific radiomics conditions.

training configuration from [13], fine-tuning the "stabilityai/stable-diffusion-2-inpainting" pretrained model [22] using nodule data from CXR images under the same settings—a batch size of 64 over 1,000 epochs on four NVIDIA RTX 6000 GPUs.

For comparison, in addition to the method using the diffusion model, we also experimented with more straightforward approaches, such as applying Gaussian blur to mimic nodules or creating synthetic nodules using segmented nodule patches from the CT scan provided in the NODE21 dataset, similar to approaches explored in the previous study[23]. These methods have the advantage of generating synthetic nodules more easily without the diffusion process, but unlike the diffusion method, they are limited in their ability to adjust nodule characteristics under various conditions. Details of these methods can be found in the appendix A.2.

2.3 Lung Nodule Detection Model

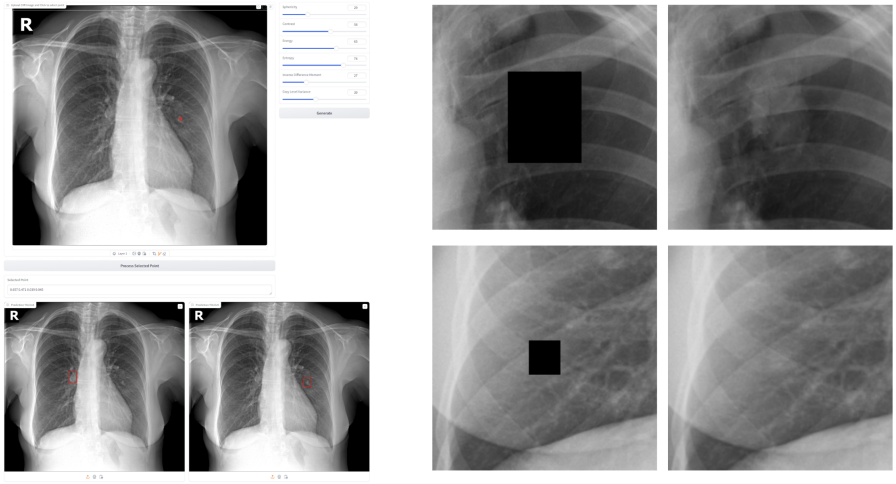
We utilize YOLOv8 and YOLOv9[14] as nodule detection models, which are enhanced versions of the original YOLO[24]. Both models were utilized with their baseline configurations without any modifications. We chose the YOLOv8-x version, which was trained using a single RTX6000 GPU with a batch size of 8 for 100 epochs. Similarly, we selected the YOLOv9-C version, which was trained using two RTX6000 GPUs with a batch size of 16 for 500 epochs. Input images were resized to a resolution of 1024x1024 pixels for both models, and mosaic data augmentation was applied to improve the generalization of the training process.

3 Results

3.1 SynNodBench Demo

We have developed SynNodBench, a demo that allows users to generate synthetic nodules in chest X-rays. In this demo, users can upload a chest X-ray image as the background, select the location and size of the nodule using a bounding box, and adjust radiomics features to generate the nodule. The adjustable radiomics features include Sphericity, Contrast, Energy, Entropy, Inverse Difference Moment, and Gray Level Variance, with the full list provided in the appendix A.1.

The generated chest X-ray image with the synthetic nodule can be exported at a resolution of 1024x1024, and users can also view the inference results from YOLOv8 and YOLOv9 models, trained on the NODE21 dataset. On the NODE21 test set, mAP scores of 0.641 and 0.673 were achieved, while on the Synthetic test set, scores of 0.566 and 0.578 were obtained, respectively. Figure 2 shows an example of a synthetic nodule created using the nodule inpainting model, along with the interface of SynNodBench.



(a) (b)

Figure 2: **(a)** SynNodBench, users can upload a desired chest X-ray, specify the region for nodule generation using a bounding box, adjust radiomics features, and generate a nodule. The results of YOLOv8 and YOLOv9 can also be reviewed. **(b)** Example of synthetic nodule using the proposed demo.

3.2 Exploring the Relationship between Nodule Size and Model Confidence

As an example of analysis using our proposed method, we examined the relationship between model confidence and nodule size, which previous study[5] failed to identify. We segmented the region of interest (ROI) where nodules can be found in chest X-rays and generated 5,710 samples, each containing one nodule ranging in size from 4mm to 40mm, for evaluation. The FID score of the generated samples was 3.22. The evaluation was conducted using one of the commercial models(Lunit-CXR) employed in prior research[5], and we calculated the Spearman's correlation between nodule size and model confidence. For comparison, we also assessed 863 nodules from the NODE21 dataset[15] that were 4mm to 40mm.

The results showed a Spearman's correlation of 0.5779 for the 5,710 synthetic nodules, indicating a moderate positive correlation. In contrast, the test using 863 real nodules yielded a Spearman correlation of 0.1228, showing no significant correlation. These findings suggest that using synthetic data to generate nodules on a large scale under controlled conditions opens up the possibility of analyzing relationships between specific variables, such as nodule size and model confidence, which may not be easily detected with traditional methods. Figure 3 illustrates how model confidence changes as nodule size increases.

3.3 Additional Insights into Model Behavior

Two experiments were conducted using the proposed demo to gain further insights into how the nodule detection model operates.

Effect of Multiple Nodules on Detection Sensitivity Figure 4 (a) illustrates a case where the behavior of the detection model was examined based on the number of nodules. First, a nodule was generated near the heart on a normal chest X-ray, and an additional nodule was created in the lung area for the image on the left. As a result, the detection model was able to identify the nodule in the left image, while it failed to detect it in the right image. This indicates that when a clearly visible nodule is present, the AI model becomes more sensitive in detecting regions that could be suspected as nodules.

Detection Behavior of the Model in Air-leak Regions Figure 4 (b) demonstrates an experiment designed to determine whether the given nodule detection model would detect nodules in anatomically impossible locations. For example, lung nodules occur in the lung parenchyma, and in cases of pneumothorax, nodules cannot be found in air-leak regions. Thus, nodules were generated across the

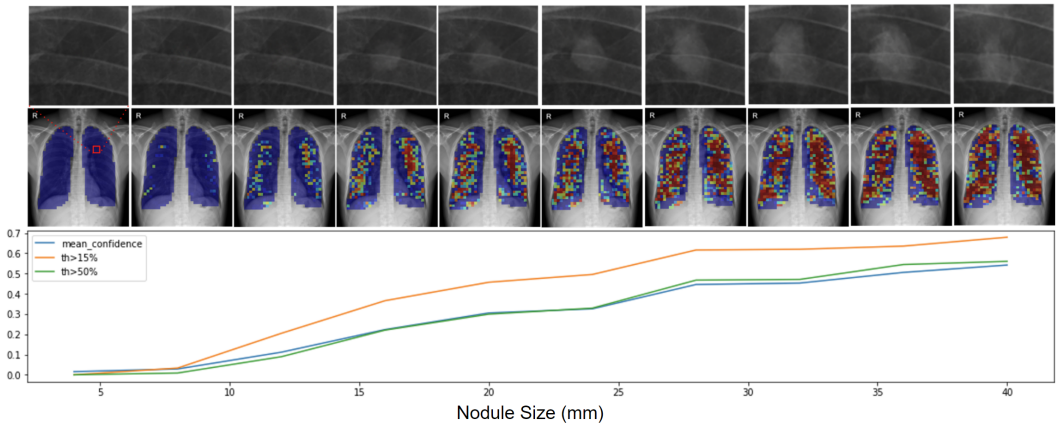


Figure 3: Sample image and graph illustrating how model confidence changes as nodule size increases. The first row shows examples of nodules generated at the same location with different sizes. The second row visualizes the model confidence across various regions when nodules are generated throughout the entire chest X-ray. The graph at the bottom shows the average model confidence for each nodule size (blue line), while the orange and green lines represent the detection rates when model confidence exceeds 15% and 50%, respectively.

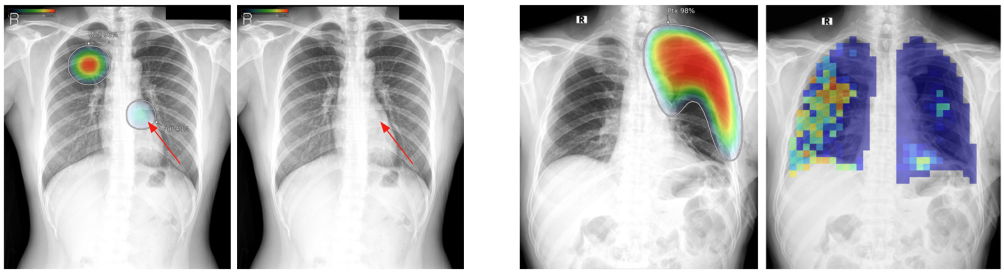


Figure 4: **(a)** Changes in nodule detection sensitivity based on the number of nodules. In the left image, where two nodules are present, the model sensitively detects the nodule near the heart, while in the right image, where only one nodule is present near the heart, the model fails to detect it. **(b)** When a nodule was generated in the air-leak region, the detection model did not classify it as a nodule. The left image shows the pneumothorax region and the right image visualizes the model confidence when nodules were generated across various regions of the pleural cavity.

entire pleural cavity on a chest X-ray with pneumothorax, and we tested whether the model would detect nodules in the air-leak region. Interestingly, the model did not classify the nodule in the air-leak area as a nodule. This shows that the model produces robust detection results, even within the pleural cavity, by correctly identifying that nodules cannot exist in air-leak regions.

4 Discussion

We have demonstrated how generative AI can be effectively used for model evaluation in healthcare through our proposed method. The generative AI-based approach shows potential for efficiently creating large-scale test datasets for medical imaging AI evaluations. Moreover, it allows data generation with specific variables controlled, enabling analysis of the relationship between lesion characteristics and model performance. This approach could also be extended to other areas, such as detecting lesions beyond nodules in chest X-rays, mass detection in mammography, or lung nodule detection in chest CT scans, making further studies highly valuable.

However, the inherent limitation of using synthetic data is that it is not equivalent to real data. While we used the FID score to evaluate the realism of the synthetic data, more robust assessments may be necessary, such as a visual Turing test by clinicians.

References

[1] Choong Ho Lee and Hyung-Jin Yoon. Medical big data: promise and challenges. *Kidney research and clinical practice*, 36(1):3, 2017.

[2] Johannes Rueckel, Christian Huemmer, Andreas Fieselmann, Florin-Cristian Ghesu, Awais Mansoor, Balthasar Schachtner, Philipp Wesp, Lena Trappmann, Basel Munawwar, Jens Ricke, et al. Pneumothorax detection in chest radiographs: optimizing artificial intelligence system for accuracy and confounding bias reduction using in-image annotations in algorithm training. *European radiology*, pages 1–13, 2021.

[3] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[4] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

[5] Kicky G van Leeuwen, Steven Schalekamp, Matthieu JCM Rutten, Merel Huisman, Cornelia M Schaefer-Prokop, Maarten de Rooij, Bram van Ginneken, Bas Maresch, Bram HJ Geurts, Cornelius F van Dijke, et al. Comparison of commercial ai software performance for radiograph lung nodule detection and bone age prediction. *Radiology*, 310(1):e230981, 2024.

[6] Lennart R Koetzier, Jie Wu, Domenico Mastrodicasa, Aline Lutz, Matthew Chung, W Adam Koszek, Jayanth Pratap, Akshay S Chaudhari, Pranav Rajpurkar, Matthew P Lungren, et al. Generating synthetic data for medical imaging. *Radiology*, 312(3):e232471, 2024.

[7] Shaoyan Pan, Tonghe Wang, Richard LJ Qiu, Marian Axente, Chih-Wei Chang, Junbo Peng, Ashish B Patel, Joseph Shelton, Sagar A Patel, Justin Roper, et al. 2d medical image synthesis using transformer-based denoising diffusion probabilistic model. *Physics in Medicine & Biology*, 68(10):105004, 2023.

[8] Mauro Giuffrè and Dennis L Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ digital medicine*, 6(1):186, 2023.

[9] Yisak Kim, Kyungmin Jeon, Soyeon Kim, and Chang Min Park. Lesion in-and-out painting for medical image augmentation. In *Deep Generative Models for Health Workshop NeurIPS 2023*.

[10] Fernando Pérez-García, Sam Bond-Taylor, Pedro P Sanchez, Boris van Breugel, Daniel C Castro, Harshita Sharma, Valentina Salvatelli, Maria TA Wetscherek, Hannah Richardson, Matthew P Lungren, et al. Radedit: stress-testing biomedical vision models via diffusion image editing. *arXiv preprint arXiv:2312.12865*, 2023.

[11] Boris van Breugel, Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Can you rely on your model evaluation? improving model evaluation with synthetic test data. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] Chanwoo Kim, Soham U Gadgil, Alex J DeGrave, Jesutofunmi A Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. Transparent medical image ai via an image–text foundation model grounded in medical literature. *Nature Medicine*, pages 1–12, 2024.

[13] Inye Na, Jonghun Kim, Eun Sook Ko, and Hyunjin Park. Radiomicsfill-mammo: Synthetic mammogram mass manipulation with radiomics features. *arXiv preprint arXiv:2407.05683*, 2024.

[14] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*, 2024.

[15] Ecem Sogancioglu, Bram Van Ginneken, Finn Behrendt, Marcel Bengs, Alexander Schlaefer, Miron Radu, Di Xu, Ke Sheng, Fabien Scalzo, Eric Marcus, et al. Nodule detection and generation on chest x-rays: Node21 challenge. *IEEE Transactions on Medical Imaging*, 2024.

[16] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American journal of roentgenology*, 174(1):71–74, 2000.

[17] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

[18] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[19] Dina Demner-Fushman, Sameer Antani, Matthew Simpson, and George R Thoma. Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6(2):168–177, 2012.

[20] Kushal Majmundar, Sachin Goyal, Praneeth Netrapalli, and Prateek Jain. Met: Masked encoding for tabular data. *arXiv preprint arXiv:2206.08564*, 2022.

[21] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[23] Manuel Schultheiss, Philipp Schmette, Jannis Bodden, Juliane Aichele, Christina Müller-Leisse, Felix G Gassert, Florian T Gassert, Joshua F Gawlitza, Felix C Hofmann, Daniel Sasse, et al. Lung nodule detection in chest x-rays using synthetic ground-truth data comparing cnn-based diagnosis to human performance. *Scientific Reports*, 11(1):15857, 2021.

[24] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

A Appendix

A.1 Radiomics Features Used in the Nodule Inpainting Model

The proposed nodule inpainting model can accept a total of 67 radiomics features as input. The full list of radiomics features is provided below.

- **Shape Features (2D):** Elongation, Major Axis Length, Maximum Diameter, Mesh Surface, Minor Axis Length, Perimeter, Perimeter Surface Ratio, Pixel Surface, Sphericity.
- **First Order Features:** 10 Percentile, 90 Percentile, Energy, Entropy, Interquartile Range, Kurtosis, Maximum, Mean Absolute Deviation, Mean, Median, Minimum, Range, Robust Mean Absolute Deviation, Root Mean Squared, Skewness, Total Energy, Uniformity, Variance.
- **Gray Level Co-occurrence Matrix (GLCM) Features:** Autocorrelation, Cluster Prominence, Cluster Shade, Cluster Tendency, Contrast, Correlation, Difference Average, Difference Entropy, Difference Variance, Id, Idm, Idmn, Idn, Imc1, Imc2, Inverse Variance, Joint Average, Joint Energy, Joint Entropy, MCC, Maximum Probability, Sum Average, Sum Entropy, Sum Squares.
- **Gray Level Size Zone Matrix (GLSZM) Features:** Gray Level Non-Uniformity, Gray Level Non-Uniformity Normalized, Gray Level Variance, High Gray Level Zone Emphasis, Large Area Emphasis, Large Area High Gray Level Emphasis, Large Area Low Gray Level Emphasis, Low Gray Level Zone Emphasis, Size Zone Non-Uniformity, Size Zone Non-Uniformity Normalized, Small Area Emphasis, Small Area High Gray Level Emphasis, Small Area Low Gray Level Emphasis, Zone Entropy, Zone Percentage, Zone Variance.

A.2 Comparison with methods utilizing Gaussian blur and segmented nodule patches from CT scans

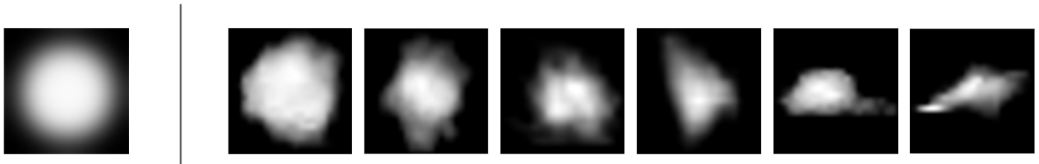


Figure 5: Examples of synthetic nodules generated using **(a)** Gaussian blur and **(b)** projected nodule patches.

In addition to the proposed method using the diffusion model, we evaluated the model with simpler approaches for generating synthetic nodules, such as Gaussian blur and segmented nodule patches from CT scans. Gaussian blur was used to create synthetic nodules with precise circular shapes, which were then blended onto chest X-rays. Segmented nodule patches were obtained from 3D nodules provided by CT scans from the NODE21 dataset and projected onto X-rays before blending. Previous studies inserted 3D nodules into chest CT scans and used Digitally Reconstructed Radiographs (DRR) for evaluation. However, because DRR images have lower resolution and differ in shape from real chest X-rays, we opted to use blending techniques instead. Examples of synthetic nodules generated using Gaussian blur and projected nodule patches can be seen in Figure 5.

We first experimented with synthetic nodules created using Gaussian blur, increasing the size from 4mm to 40mm to observe changes in model confidence, as described in Section 3.2. As shown in Figure 6, similar to the diffusion model, model confidence tended to increase with nodule size. However, since the Gaussian blur nodules have consistent and well-defined shapes, they appear more easily detected on the graph.

We also attempted to observe model confidence variations using projected nodule patches by altering their sizes. However, we observed a substantial drop in detection performance when the nodule size deviated significantly from its original dimensions. Therefore, we selected six nodule patches with similar sizes between 21mm and 25mm, standardized them to 20mm, and analyzed how their shapes

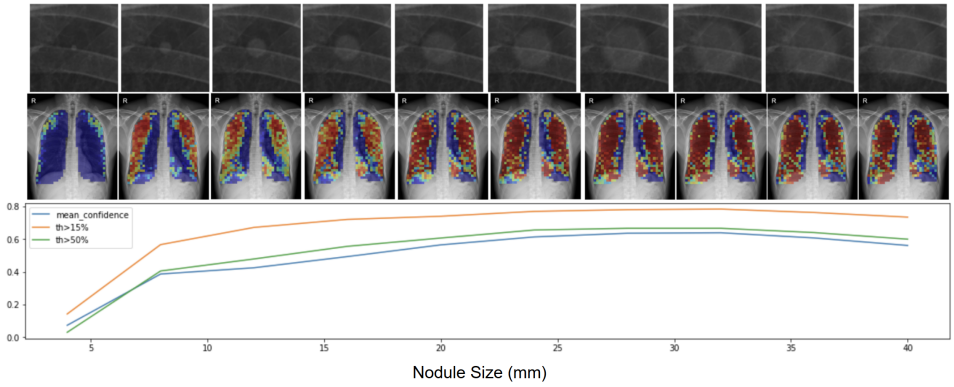


Figure 6: Sample image and graph illustrating how model confidence changes as nodule size increases. The first row shows examples of Gaussian blur nodules generated at the same location with different sizes. The second row visualizes the model confidence across various regions when nodules are generated throughout the entire chest X-ray. The graph at the bottom shows the average model confidence for each nodule size (blue line), while the orange and green lines represent the detection rates when model confidence exceeds 15% and 50%, respectively.

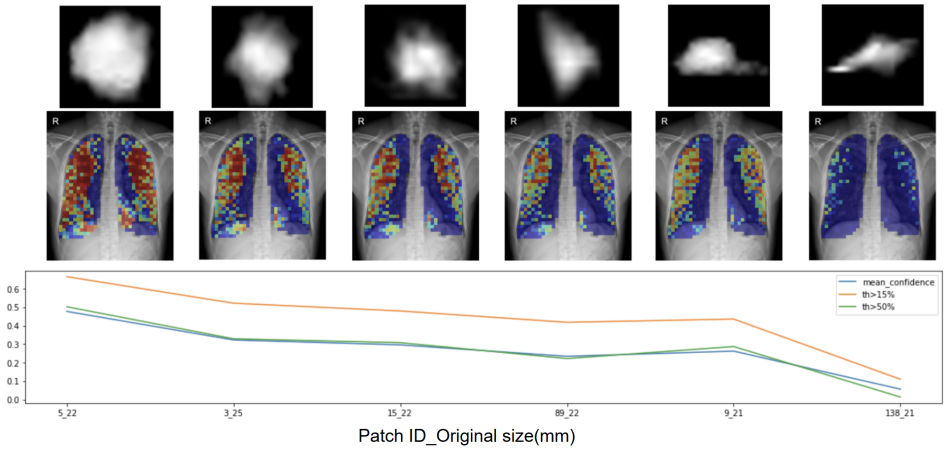


Figure 7: Sample image and graph illustrating how model confidence changes as nodule shape changes. The first row shows examples of projected nodule patches from CT scans. The second row visualizes the model confidence across various regions when nodules are generated throughout the entire chest X-ray.

affected model confidence. As shown in Figure 7, nodules with more circular and larger shapes exhibited higher model confidence, while elongated and smaller nodules tended to result in lower model confidence.