Improved Classification Learning from Highly Imbalanced Multi-Label Datasets of Inflamed Joints in [^{99m}Tc]Maraciclatide Imaging of Arthritic Patients by Natural Image and Diffusion Model Augmentation

Robert Cobb[1(✉)], Gary J.R. Cook[1,2], and Andrew J. Reader[1]

[1] School of Biomedical Engineering and Imaging Sciences, King's College London
robert.cobb@kcl.ac.uk
[2] King's College London and Guy's and St Thomas' PET Centre, King's College London

Abstract. Gamma camera imaging of the novel radiopharmaceutical [^{99m}Tc]maraciclatide can be used to detect inflammation in patients with rheumatoid arthritis. Due to the novelty of this clinical imaging application, data are especially scarce with only one dataset composed of 48 patients available for development of classification models. In this work we classify inflammation in individual joints in the hands of patients using only this small dataset. Our methodology combines diffusion models to augment the available training data for this classification task from an otherwise small and imbalanced dataset. We also explore the use of augmenting with a publicly available natural image dataset in combination with a diffusion model. We use a DenseNet model to classify the inflammation of individual joints in the hand. Our results show that compared to non-augmented baseline classification accuracy, sensitivity, and specificity metrics of 0.79 ± 0.05, 0.50 ± 0.04, and 0.85 ± 0.05, respectively our method improves model performance for these metrics to 0.91 ± 0.02, 0.79 ± 0.11, 0.93 ± 0.02, respectively. When we use an ensemble model and combine natural image augmentation with [^{99m}Tc]maraciclatide augmentation we see performance increase to 0.92 ± 0.02, 0.80 ± 0.09, 0.95 ± 0.02 for accuracy, sensitivity, and specificity, respectively.

Keywords: Augmentation - Generative Modelling - Rheumatoid Arthritis

1 Introduction

Data scarcity and quality is a well known hurdle for AI adoption into the nuclear medicine imaging domain [2, 19, 15]. [^{99m}Tc]maraciclatide (Serac Healthcare, UK) which is imaged with a gamma camera is a novel radiopharmaceutical for detecting inflammation in the joints of patients with rheumatoid arthritis (RA) [3]. Due to the novelty of this application, data scarcity issues are even

more pronounced. There exists only one dataset of 48 patients. The dataset contains gamma camera scans of the hands, with one view of the palmar and dorsal aspects, and two of the obliques (Fig 1 a, b). The obliques are less useful for the classification of joints as it is hard to delineate overlapping joints from that view. The dorsal and palmar are taken in the same acquisition in a dual-headed scanner from above and below, so the areas of activity are the same in these two views, merely flipped and with a different noise observation. Within each hand we wish to categorise inflammation in 15 joint regions, the metacarpophalangeal and interphalangeal joints and the wrist, which is made up of many different joints but is treated as one joint for classification purposes. This leaves only 98 images for classification. Classifying the inflammation using a DenseNet [11] on the existing dataset creates a model that either has a high true negative rate (TNR) with low true positive rate (TPR) or a model with a modest performance for both TPR and TNR. In this work, by utilising diffusion models [10], 2D Perlin noise maps and a large natural image dataset, we aim to improve the performance of these baseline models.

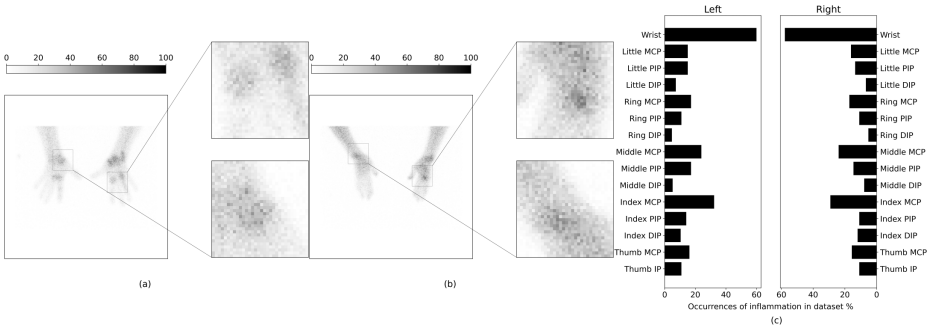

Fig. 1. Two example [^{99m}Tc]maraciclatide gamma camera images of the hands with zoomed sections of inflammation and occurrences of inflammation in the [^{99m}Tc]maraciclatide dataset. Panel (a) shows inflammation in the wrists and metacarpophalangeal joints, (b) shows an oblique view with inflammation in the metacarpophalangeal joints, and (c) shows the percentage of occurrences of inflamed joints in the dataset, broken down by 14 joints in the hands and one wrist joint region.

2 Methods

Dataset and Labels The [^{99m}Tc]maraciclatide images are single channel, 2D images of the size (256×256). The [^{99m}Tc]maraciclatide dataset was labelled by a clinician (G.J.R.C) with over 30 years of nuclear medicine experience, who provided binary labels for each joint/joint region and segmentation maps of the normal tissue as well as abnormalities as either low or high inflammation (Fig

2). More information on the labelling procedure can be found in previous work [7].

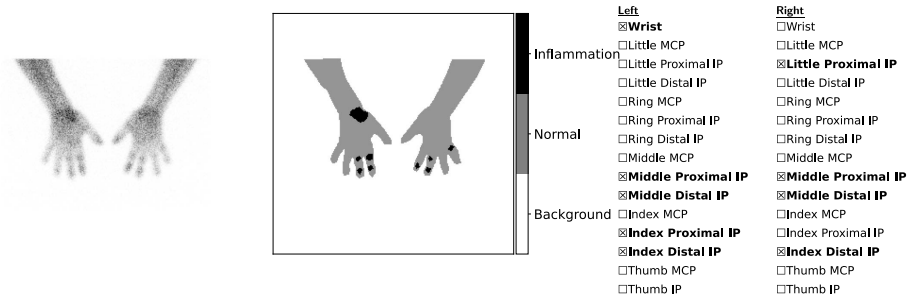

Fig. 2. Example [^{99m}Tc]maraciclatide image with clinician-defined segmentations of normal and inflamed tissue and clinician defined classification labels for 15 joints/joint regions for each hand.

Training and Validation Each model was trained in the same way but with different datasets. We used a DenseNet [11] implementation by MONAI [5] with a growth rate of 32 in a five-fold cross validation with 60% of our data for training, 20% for validation and 20% for test where each patient was present in only one of the three datasets per fold. The validation dataset is used for early stopping, stopping model training when there was no improvement in the G-Mean Score (GMS, Eq 1) score over the validation dataset in 500,000 training samples. GMS is chosen to encourage a classifier that balances TNR and TPR instead of using the more common accuracy metric as accuracy can be misleading for imbalanced datasets [13]. When reporting the results of our models we report a range of metrics we believe to be complementary such as TNR and TPR that show how the model performs with type I and II errors, respectively. The model is evaluated every 10 epochs. The training used the Adam optimiser with a fixed learning rate of 0.0001 and a mini batch size of 32.

$$\text{GMS} = \sqrt{TPR \times TNR} \tag{1}$$

The training used traditional data augmentation strategies of randomly rotating the images [-30, 30] degrees, horizontal and vertical flipping with 50% probability, and random cropping with 50% probability; the cropping procedure pads the image by (32, 32) and then randomly crops the images back to the original size (256×256). Unless stated otherwise, the models were trained 3 times each with different random initialisation and then the metrics were averaged. All metrics are calculated over the whole image (not per joint).

Baseline The baseline used the existing dataset of 98 images from 48 patients. One patient has 2 extra dorsal/palmar views.

Diffusion Model Diffusion models have recently been used in a large variety of medical imaging tasks [14, 8]. We trained a diffusion model using the Palette framework [17], using an open source implementation [12]. The generative model was trained on the [^{99m}Tc]maraciclatide dataset with the clinician-defined labels. The model input was a three channel input of the normal tissue, low, and high inflammation. The model was trained with 500 timesteps and used downsampled images of size (128×128). The output images are upsampled with nearest neighbour interpolation to restore them to the original size of (256×256); then they are Poisson sampled. The generative model was trained on both the obliques, palmar, and dorsal views. Five diffusion models were trained, one for each of the five cross folds. All models (including the classification models) were randomly initialised using Kaiming uniform initialisation.

[^{99m}Tc]Maraciclatide Image Augmentation In addition to the clinically-defined labels, layperson-defined labels were also provided using the Label Studio platform [18] for the [^{99m}Tc]maraciclatide dataset, segmenting each joint and joint region that the final model seeks to classify. Once the diffusion model was trained it was then used to augment the train dataset by randomly inflaming different joints according to a random sampler. The sampler creates an inflammation vector of 30 binary labels by running a series of hierarchical binary decisions controlled by random probability. Firstly, it decided if there should be any inflammation in the hands, favouring inflammation with a 90%. Then, if the image was to contain inflammation it decided if each hand should be inflamed, favouring inflammation with 75% probability but ensuring at least one hand was inflamed. Then, if a hand was to contain inflammation, each joint in the hand was inflamed with a 50% probability. Once the inflammation vector is created, two segmentation maps were created from that vector by combining the layperson segmentations corresponding to each of the inflamed joints in the vector. Two segmentation maps were created, one for low and one for high inflammation, each inflamed joint was assigned to low or high with 50% random probability. Once the hand mask and two segmentation maps were created, they were then used as input into the diffusion model to generate a synthetic [^{99m}Tc]maraciclatide image, with the inflammation vector becoming the corresponding label for classification purposes. (Fig 3)

Natural Hand Image Dataset Augmentation The 11k [1] dataset is a publicly available dataset of hand pictures. There is a non-trivial domain shift between the [^{99m}Tc]maraciclatide hand images and the hands present in the 11k dataset. The 11k dataset often does not contain the wrist which is present in the [^{99m}Tc]maraciclatide images. The hands in the 11k images often have accessories and clothing and only contain one hand. In contrast, the diffusion model

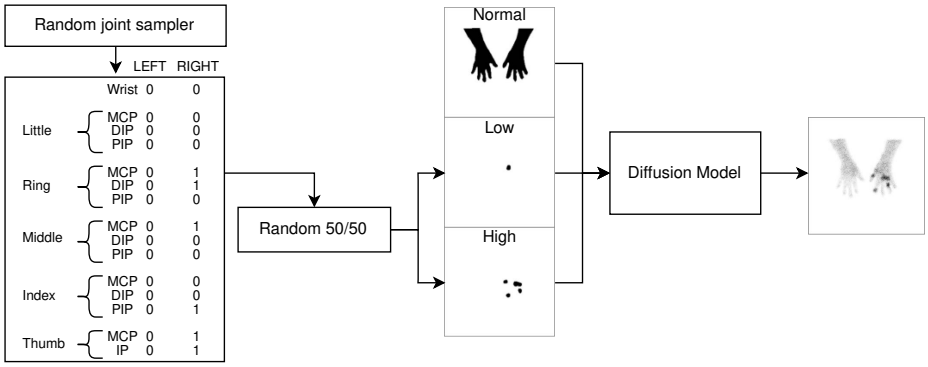

Fig. 3. A randomly generated inflammation vector was used to generate low and high inflammation maps, where each joint can be either low or high with 50% probability. The clinician-defined normal, with generated low and high inflammation were then used as inputs into the diffusion model and the corresponding [^{99m}Tc]maraciclatide output is generated.

takes masks of two hands along with joint segmentations to generate a synthetic [^{99m}Tc]maraciclatide image. To create the hands masks from the 11k images, a heuristic preprocessing algorithm was run on the images. The background of the 11k hand images was always white, and so the algorithm detects all pixels where the maximum difference between the RGB channels is 10% and flags them as background pixels. Any discontinuous regions of background pixels were set to non background and finally non-maximum suppression on the mask is done to ensure only one contiguous object is set to the foreground.

To generate the hand masks and joint labels for the 11k images, 50 images from the 11k dataset were labelled. Then a U-Net was trained with a Dice loss to segment the joints. Then these segmentations were randomly sampled just as was done for the [^{99m}Tc]maraciclatide dataset in order to create images with varying levels of inflammation (Fig 4). Using the same probabilities as with the [^{99m}Tc]maraciclatide augmentation, samples were drawn from the 11k dataset and were probabilistically inflamed and passed to the diffusion model. The wrists for the 11k dataset were never marked as inflamed, as most of the 11k images did not contain the wrists so the labels were not deemed to be reliable. 500 images from the 11k dataset are used that have been filtered to exclude those with accessories and only to include those where the hands were open like those in the [^{99m}Tc]maraciclatide dataset. The images were then reshaped, duplicated and rotated to match the [^{99m}Tc]maraciclatide dataset. Figure 4 shows the 11k augmentation process and figure 5 shows some example images.

Perlin noise The [^{99m}Tc]maraciclatide and 11k augmented dataset used deterministic atomic labels for each joint/joint region. The augmentation previously described combines these atomic inflammation masses in different combinations.

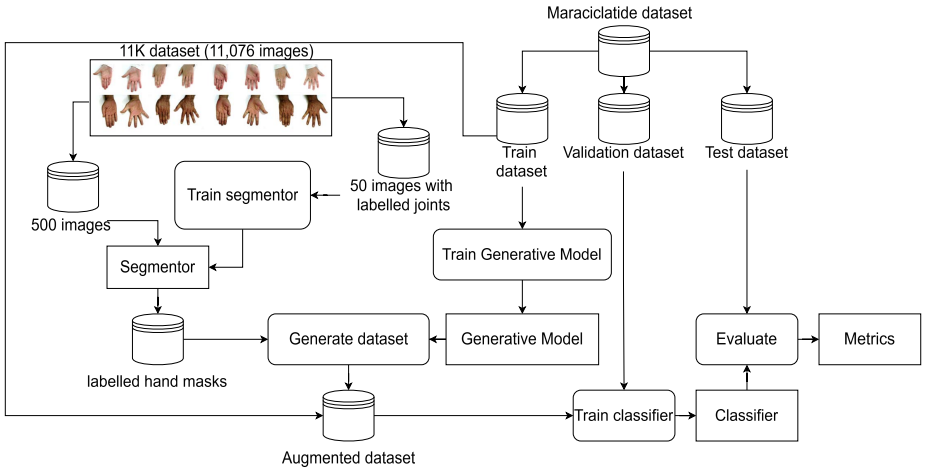

Fig. 4. Overview of the augmentation method using the 11k publicly available dataset for a single cross fold. The [^{99m}Tc]maraciclatide dataset is split into a 60% train, 20% validation, 20% test; the validation and test datasets were used in the model training and testing without modification. The diffusion model was trained on the [^{99m}Tc]maraciclatide images in the train dataset with clinician defined labels. The images from the 11k dataset were labelled with a segmentor network that labels all the joint regions in the hands; the hand masks were generated through a preprocessing algorithm. These labelled 11k images with joints are then randomly sampled to generate [^{99m}Tc]maraciclatide synthetic version. The augmented train dataset is comprised of the [^{99m}Tc]maraciclatide train dataset as well as the 11k augmented images. The trained classifier was then evaluated on the test dataset.

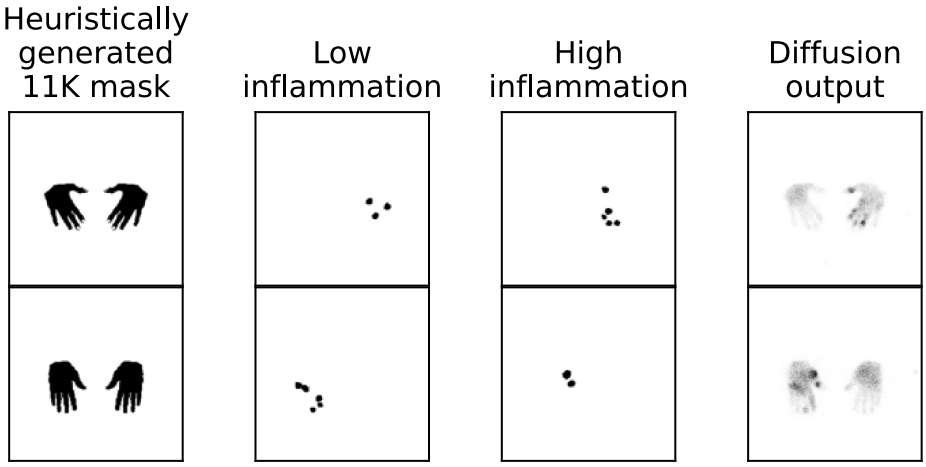

Fig. 5. Two example 11K image masks, with the low and high inflammation and [^{99m}Tc]maraciclatide synthetic image.

Thus if a specific joint is inflamed in different variations of the same image the inflammation pattern passed into the diffusion model is the same. This is especially problematic for the wrists, where the layperson-defined labels encompass a large area where various joints in the wrist exist, whereas in a real [^{99m}Tc]maraciclatide image only one or a few joints in the wrist might be inflamed. To increase the variation, Perlin noise [16] was used to vary the inflammation pattern.

Perlin noise augmentation has been used in previous works [4, 9] where the noise was added to train images much like Gaussian noise is, here we use the noise map to vary the segmentation maps, similar to elastic deformation augmentations [6]. The noise is not added to the images but it is used to define a 2D map with smooth contours, these maps are then multiplied by the joint segmentation maps to vary the segmentations whilst maintaining smooth edges of the inflammation mass.

For each individual joint/joint region that was to be inflamed in the given sample, a random Perlin map in the range 0-1 was generated and multiplied by the inflamed segmentation. A random threshold is then generated to turn the region into a binary label. This new map is then checked to ensure it has a minimum of 25 pixels and at most 90% of the original inflammation map. If the modified segmentation map meets the criteria it is used, else a new random Perlin map is generated. This is tried 5,000 times, and if no valid map is found, then the default joint segmentation is used. The Perlin variation was applied to both the 11k and [^{99m}Tc]maraciclatide augmentation. The hyperparameters for the Perlin noise augmentation were chosen as a trade-off between computational complexity and variability of the segmentaitons.

3 Results

Table 1. Results of augmentation datasets for a 5-fold cross validation for 8,192 augmented images. Each fold was trained 3 times and the results averaged, then these results across all folds were averaged and the standard deviation calculated.

Method	Accuracy	TPR	TNR	F1	GMS
Baseline	0.79±0.05	0.50±0.04	0.85±0.05	0.45±0.07	0.65±0.02
11k	0.89±0.04	0.65±0.08	0.93±0.03	0.66±0.10	0.78±0.05
11k-Perlin	0.88±0.03	0.64±0.08	0.93±0.03	0.64±0.10	0.77±0.05
[^{99m}Tc]maraciclatide	0.91±0.03	0.72±0.11	0.94±0.03	0.72±0.11	0.82±0.06
[^{99m}Tc]maraciclatide-Perlin	0.91±0.02	0.79±0.11	0.93±0.02	0.74±0.12	0.85±0.06

11k At 8,192 images, the 11k augmentation shows improved performance over the baseline on all metrics charted. The use of Perlin variation slightly decreases the performance of the model. (Table 1). Analysis showed that this is partly but not completely explained by the lack of wrist augmentation in the 11K augmentation dataset.

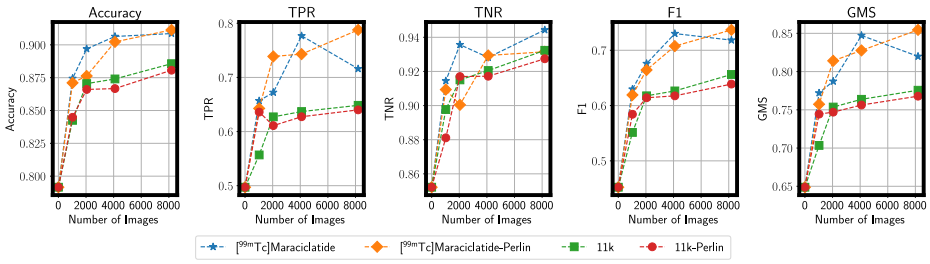

Fig. 6. Performance metrics over the test dataset for a five-fold cross validation for different train augmentation strategies and different train augmentation sizes. 0 augmented images is the baseline and the results are averaged over three runs, At 8,192 augmented images the models are also averaged over three runs. Intermediate data points are run only once.

[^{99m}Tc]Maraciclatide The [^{99m}Tc]maraciclatide augmentation shows improved performance over the baseline and the 11k augmentation strategy. In our results, the Perlin variation increases performance over the non Perlin version in all metrics except TNR which decreases slightly. (Fig, 6, Table 1).

Table 2. Results of augmentation datasets for ensemble models with 8,192 images in the augmented train dataset on a 5-fold cross validation. MP x% 11kP y% represent an augmented train dataset of 8,192 images where x% of the images comes from the [^{99m}Tc]maraciclatide Perlin augmentation strategy and y% of the images comes from the 11k Perlin augmentation strategy. Each fold was trained 3 times and the results averaged, then these results across all folds were averaged and the standard deviation calculated.

Method	Accuracy	TPR	TNR	F1	GMS
[^{99m}Tc]Maraciclatide	0.92±0.02	0.74±0.11	0.96±0.02	0.76±0.09	0.84±0.06
11kP	0.89±0.04	0.66±0.08	0.93±0.03	0.67±0.09	0.78±0.05
MP 20% 11kP 80%	0.91±0.04	0.77±0.11	0.94±0.04	0.74±0.11	0.85±0.06
MP 40% 11kP 60%	0.92±0.03	0.80±0.12	0.94±0.05	0.75±0.12	0.86±0.06
MP 60% 11kP 40%	0.92±0.03	0.78±0.12	0.94±0.04	0.76±0.09	0.86±0.06
MP 80% 11kP 20%	0.92±0.02	0.80±0.09	0.95±0.02	0.77±0.11	0.87±0.05
[^{99m}Tc]Maraciclatide-Perlin	0.92±0.02	0.79±0.15	0.94±0.02	0.76±0.14	0.86±0.08

Ensembles For all configurations where the train augmented dataset had 8,192 images, three models were trained. We also use these three trained models to create an ensemble model. The results for these ensembles are shown in table 2 and show an increase in performance compared to the non ensembled models. The best ensemble model by TPR, F1, and GMS is an ensemble model trained on 80% [^{99m}Tc]maraciclatide Perlin and 20% 11k Perlin.

4 Conclusion

We present a novel method that allows us to classify inflammation in the individual joints of the hands and wrists of patients with RA using [^{99m}Tc]maraciclatide images. We used diffusion models with extra labels over the dataset to create patterns of inflammation that were not present in the dataset. We also explored augmenting with natural images from a large open source dataset of hand images, and finally we used 2D Perlin maps to vary segmentation maps used to create synthetic training samples. Our results show that augmenting with natural hand images improves performance over the test dataset compared to our baseline. Our results also show that using extra labels over the [^{99m}Tc]maraciclatide dataset improved results more than the natural hand augmentation. Lastly our results show that using 2D Perlin maps to vary the segmentation maps used to generate the synthetic [^{99m}Tc]maraciclatide images with the extra labels performed the best, improving upon the baseline in the GMS metric by 89%.

Acknowledgments. The [^{99m}Tc]maraciclatide dataset was collected in accordance with the Declaration of Helsinki and approved by the NHS HRA on 7 April 2016 (IRAS ID: 188145 REC: 16/LO/0309). Institutional approval was obtained for the analysis of the anonymized data. This research was funded in whole, in part, by the Wellcome Trust [WT203148/Z/16/Z]. This research was also funded by Serac Healthcare.

Disclosure of Interests. G.J.R.C. has received funding from Serac Healthcare for research projects and has acted as a consultant in the past for Serac Healthcare.

References

1. Afifi, M.: 11k hands: Gender recognition and biometric identification using a large dataset of hand images. Multimedia Tools and Applications **78**, 20835–20854 (2019)
2. Aktolun, C.: Artificial intelligence and radiomics in nuclear medicine: potentials and challenges. European journal of nuclear medicine and molecular imaging **46**, 2731–2736 (2019)
3. Attipoe, L., Chaabo, K., Wajed, J., Hassan, F.U., Shivapatham, D., Morrison, M., Ballinger, J., Cook, G., Cope, A.P., Garrood, T.: Imaging neoangiogenesis in rheumatoid arthritis (inira): whole-body synovial uptake of a 99mtc-labelled rgd peptide is highly correlated with power doppler ultrasound. Annals of the Rheumatic Diseases **79**(9), 1254–1255 (2020)
4. Bae, H.J., Kim, C.W., Kim, N., Park, B., Kim, N., Seo, J.B., Lee, S.M.: A perlin noise-based augmentation strategy for deep learning with small data samples of hrct images. Scientific reports **8**(1), 17687 (2018)
5. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
6. Castro, E., Cardoso, J.S., Pereira, J.C.: Elastic deformations for data augmentation in breast cancer mass detection. In: 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). pp. 230–234. IEEE (2018)

7. Cobb, R., Cook, G.J., Reader, A.J.: Deep learned segmentations of inflammation for novel ^{99m}tc-maraciclatide imaging of rheumatoid arthritis. Diagnostics **13**(21), 3298 (2023)
8. Dorjsembe, Z., Odonchimed, S., Xiao, F.: Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In: Medical Imaging with Deep Learning (2022)
9. Haekal, M., Septiawan, R., Haryanto, F., Arif, I.: A comparison on the use of perlin-noise and gaussian noise based augmentation on x-ray classification of lung cancer patient. In: Journal of Physics: Conference Series. vol. 1951, p. 012064. IOP Publishing (2021)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
12. Jiang, L., Belousov, Y.: Palette: Image-to-image diffusion models. https://github.com/Janspiry/Palette-Image-to-Image-Diffusion-Models (2022)
13. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data **6**(1), 1–54 (2019)
14. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Medical Image Analysis p. 102846 (2023)
15. Pan, S., Wang, T., Qiu, R.L., Axente, M., Chang, C.W., Peng, J., Patel, A.B., Shelton, J., Patel, S.A., Roper, J., et al.: 2d medical image synthesis using transformer-based denoising diffusion probabilistic model. Physics in Medicine & Biology **68**(10), 105004 (2023)
16. Perlin, K.: Improving noise. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques. pp. 681–682 (2002)
17. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
18. Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: Data labeling software (2020-2022), https://github.com/heartexlabs/label-studio, open source software available from https://github.com/heartexlabs/label-studio
19. Visvikis, D., Lambin, P., Beuschau Mauridsen, K., Hustinx, R., Lassmann, M., Rischpler, C., Shi, K., Pruim, J.: Application of artificial intelligence in nuclear medicine and molecular imaging: a review of current status and future perspectives for clinical translation. European journal of nuclear medicine and molecular imaging **49**(13), 4452–4463 (2022)