Generative Active Learning with Variational Autoencoder for Radiology Data Generation in Veterinary Medicine

In-Gyu Lee
Dept. Computer Science
Chungbuk National University
Cheongju, Republic of Korea
ingyu.lee@chungbuk.ac.kr

Jun-Young Oh
Dept. Computer Science
Chungbuk National University
Cheongju, Republic of Korea
jy.oh@chungbuk.ac.kr

Hee-Jung Yu
Dept. Veterinary Medical Imaging
Konkuk University
Seoul, Republic of Korea
hazel13@konkuk.ac.kr

Jae-Hwan Kim
Dept. Veterinary Medical Imaging
Konkuk University
Seoul, Republic of Korea
jaehwan@konkuk.ac.kr

Ki-Dong Eom
Dept. Veterinary Medical Imaging
Konkuk University
Seoul, Republic of Korea
eomkd@konkuk.ac.kr

Ji-Hoon Jeong*
Dept. Computer Science
Chungbuk National University
Cheongju, Republic of Korea
jh.jeong@chungbuk.ac.kr

***Abstract*—Recently, with increasing interest in pet healthcare, the demand for computer-aided diagnosis (CAD) systems in veterinary medicine has increased. The development of veterinary CAD has stagnated due to a lack of sufficient radiology data. To overcome the challenge, we propose a generative active learning framework based on a variational autoencoder. This approach aims to alleviate the scarcity of reliable data for CAD systems in veterinary medicine. This study utilizes datasets comprising cardiomegaly radiographic image data and chronic kidney disease ultrasound image data. After removing annotations and standardizing images, we employed a framework for data augmentation, which consists of a data generation phase and a query phase for filtering the generated data. The experimental results revealed that as the data generated through this framework was added to the training data of the generative model, the frechet inception distance decreased from 84.14 to 50.75 in the radiographic image and from 127.98 to 35.16 in an ultrasound image. Subsequently, when the generated data were incorporated into the training of the classification model, the true negative of the confusion matrix also improved from 0.16 to 0.66 on the radiograph and from 0.44 to 0.64 on the ultrasound image. The proposed framework has the potential to address the challenges of data scarcity in medical CAD, contributing to its advancement.**

***Index Terms*—Artificial intelligence, generative model, active learning, variational autoencoder, veterinary medicine**

This work was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00252624), was partly supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET) through Agriculture and Food Convergence Technologies Program for Research Manpower Development Program funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA)(RS-2024-00398561), also partly funded by Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2021-II212068, Artificial Intelligence Innovation Hub), and was partly supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0025399, Development of Intelligent Energy Management Platform).

*(*Corresponding author: Ji-Hoon Jeong)*

I. Introduction

Pets have become important members of our lives, forming strong bonds and emotional connections with their owners. The healthcare of pets has become a subject of increased interest. With the continuous evolution of artificial intelligence (AI), there is a noticeable trend towards incorporating AI into computer-aided diagnosis (CAD) systems for pet healthcare. The effectiveness of AI models is highly dependent on access to high-quality training data [1]. However, acquiring a substantial amount of medical data for CAD has challenges due to the sensitive personal information in such data. Consequently, there is a persistent effort to explore the application of generative models for the generation of medical data.

Among these efforts, numerous studies have leveraged generative adversarial networks (GAN) [2]. Yoon et al. [3] achieved a fretchet inception distance (FID) of 42.19 for Sessile serrated lesion images using style-based GAN. The FID is a metric of which calculates the disparity in features between generated and real images. A lower FID value signifies better performance. Salvia et al. [4] proposed the use of GAN to generate synthetic hyperspectral images of epidermal lesions, addressing the challenge of limited large datasets. These academic efforts demonstrate the potential of GAN in generating medical images to improve research and diagnostics. Zhu et al. [5] explored the concept of generative adversarial active learning, employing a generative model in active learning to improve the performance of classification models. Although this approach involved queries to augment the training dataset for labeling, they emphasized the proposed framework, rather than focusing on criteria for queries.

In this study, we have used active learning in the context of generative models, incorporating a unique criterion for data

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00123

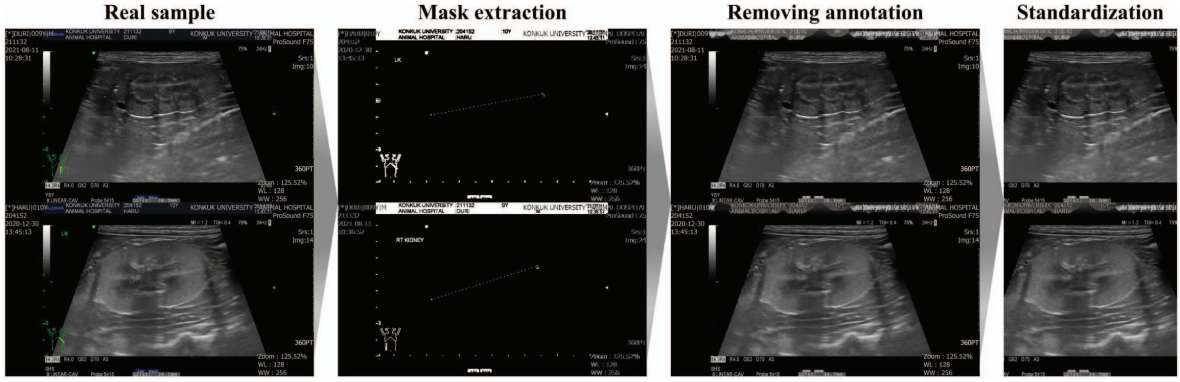

Fig. 1. The data preprocessing pipeline for training a generative model. If doctors draw annotations for diagnosis, the annotations are extracted to create masks. Then, image inpainting techniques are applied to remove the annotations. Subsequently, the resolution of the images is standardized.

filtration through a variational autoencoder (VAE) [6]. The approach of using the feature vectors generated by encoder as features has been utilized in other studies [7]. In this research, it has significantly enhanced the robustness of the generative model's performance. We focused on addressing the scarcity of medical data for CAD, especially in the field of veterinary medicine. The introduced approach leverages query processes facilitated by a VAE to improve the performance of generative models in generating medical image data. This method shows the solution to address the persistent challenge of limited medical image data in CAD applications.

II. Generative Active Learning with VAE

A. Dataset

This research used the pet's thorax medical image datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)' and ultrasound images of chronic kidney diseases (CKD) from Konkuk University Animal Hospital. Radiograph data information can be accessed through 'AI-Hub (www.aihub.or.kr)'. We selected 100 images from the cardiomegaly disease data and 100 images from stage 4 of CKD as the initial training data set, focusing on the visually prominent characteristics of the disease. Data on cardiomegaly disease can visually confirm that the heart is enlarged and CKD data can visually confirm that the surface of the kidney becomes rough [8], [9]. We used these selected data as initial training data for the generative model and VAE.

Before training the model, a preprocessing step was performed to standardize the data used for training. For ultrasound images, annotations made by veterinarians for diagnosis were present. As the model could potentially learn from these annotations as features of the data, we performed a task to remove the annotations. Initially, the data in RGB format was transformed into the HSV color space. Subsequently, since the color of the annotations was in kinds of green, we extracted the green color to create a binary mask. The mask obtained was then utilized in the image inpainting technique to restore the image. This method replaces pixels in the masked area using neighboring pixels.

Algorithm 1 Overall process of proposed framework

Step 1: Data Preprocessing
Input: $Raw\_dataset$
Output: $Preprocessed\_dataset$
if *Annotations exist in* $Raw\_dataset$ **then**
  Convert RGB images to HSV images
  Extract green tones to create masks
  $Inpainted\_dataset$ = Use masks for image inpainting
$Preprocessed\_dataset$ = Standardize resolution of $Inpainted\_dataset$

Step 2: Data Generation
Input: $Preprocessed\_dataset$
Output: $Augmented\_dataset$
while $Augmented\_dataset\ size < 500$ **do**
  for $epochs = 1$ *to* 20 **do**
    if $epochs$ *is even* **then**
      New_FID = Evaluate the generative model
      if $Saved\_FID > New\_FID$ **then**
        $Saved\_FID = New\_FID$
        $Saved\_weights$ = Save the generative model's weights
  $Generated\_dataset$ = Generate 1000 data using $Saved\_weights$
  Calculate the cosine similarity of $Generated\_dataset$
  $Filtered\_dataset$ = Select the top 10% $Generated\_dataset$
  $Augmented\_dataset+ = Filtered\_dataset$

Second, we standardized the resolution of both radiographic and ultrasound images. The raw data had diverse resolutions. Inconsistent image resolutions in the training dataset can lead to unstable learning due to variations in the size of the feature maps extracted by the neural network. To address this, we employ the center-cropping method, which uses center-based image cropping to ensure that essential organ information, such as the heart and kidney, is not lost. Radiographic images were resized to 256×256 pixels, while ultrasound images were resized to 512×512 pixels. The data preprocessing procedure is depicted in Fig. 1.

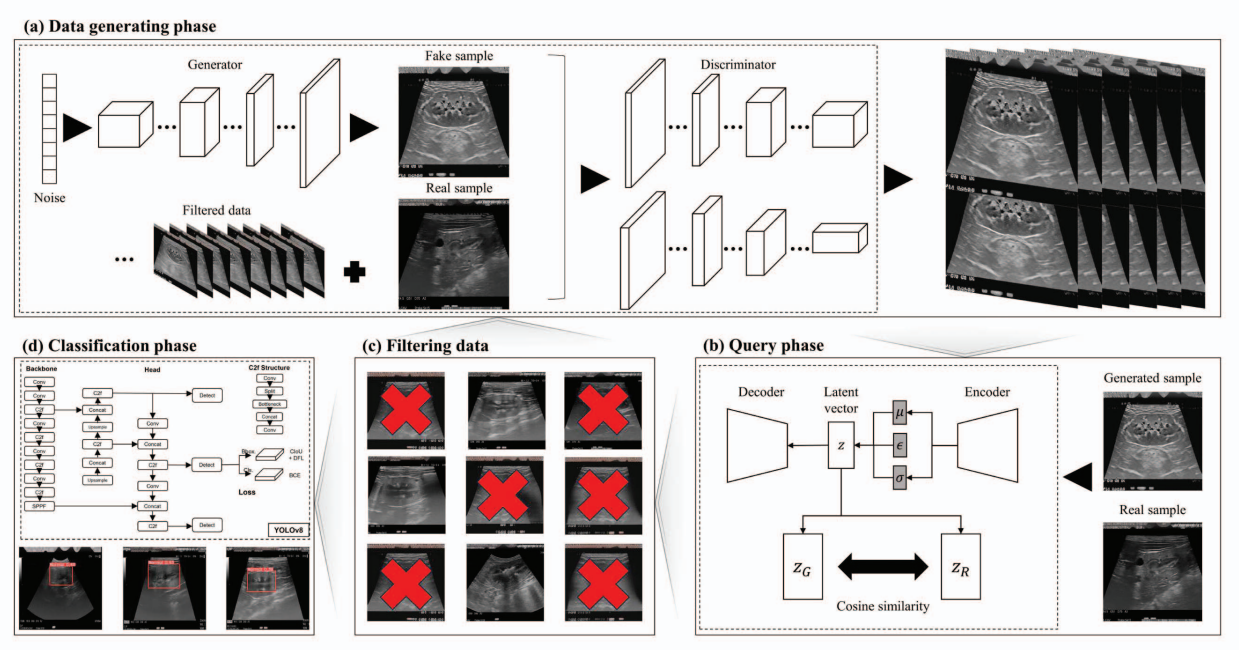

Fig. 2. Overall flow of the proposed framework. (a) The projectedGAN is trained with filtered image data and real image data to generate a new radiographic and ultrasound image. (b) VAE trained with 100 original data are used to calculate cosine similarity. (c) The top 10% cosine similarity of the data is added to the training dataset. (d) Finally, classification is performed using the object detection model after labeling to prove the usefulness of the data.

B. Proposed Framework

The framework is composed of the two phases. First, the data generating phase trains the generative model and generates data. Second, the query phase filters the generated data through the query strategy before incorporating them into the dataset for training the generative model. Algorithm 1 details the steps involved in this process with data preprocessing. The term cycle refers to the process of generating data in the data generating phase, filtering through the query phase, and adding the filtered data to the training dataset of the generative model. We repeated the cycle until the train dataset has 500 data. The overall flow is depicted in Fig. 2.

1) Data generating phase: The study utilized the projectedGAN model proposed by Sauel et al. [10] due to its state-of-the-art performance across various datasets during the experiment. ProjectedGAN comprises a generator and a discriminator. The generator is trained to learn the data features to generate images that can effectively deceive the discriminator. The discriminator learns the data in a way that distinguishes between ground truth data and generated data.

The evaluation of the generative model was based on the FID [11], which measures the dissimilarity of the characteristics between the generated and ground truth images. The formula to calculate the FID is provided below. T represents ground truth images, and G represents generated images. Tr is defined as the sum of elements from the upper left to the lower right of the vector.

$$FID = \|\mu_T - \mu_G\|^2 - Tr((\Sigma_T + \Sigma_G - 2(\Sigma_T\Sigma_G)^{\frac{1}{2}}) \quad (1)$$

The GAN initiates training by using 100 selected ground truth data. Training progresses through a total of 20 epochs per cycle, with a performance evaluation conducted every 2 epochs. Consequently, each cycle yields a total of 10 FID assessments. During evaluation, if the current FID value is lower than the previously recorded FID value, we save the model's weights. We utilize these saved weights to generate 1,000 images per cycle.

2) Query phase: The evaluation of image similarity in this study used a VAE, comprising a decoder and a encoder. In VAE, the decoder aims to regenerate the input in a form that is most similar to when a latent vector is given. The encoder, on the other hand, seeks to find the mean and standard deviation of the input and generates a latent vector with noise epsilon in Gaussian distribution. The study focused on utilizing the latent vector generated by the encoder. The generated latent vector has a shape of 200×1. Training the VAE involved using 100 selected ground truth data for a total of 25 epochs.

To assess image similarity, cosine similarity was employed [12]. Unlike distance measures such as the Euclidean distance, which evaluate vectors based on their magnitudes, cosine similarity examines whether both vectors are aligned in the same direction. This characteristic makes cosine similarity particularly suitable for gauging significant similarities between

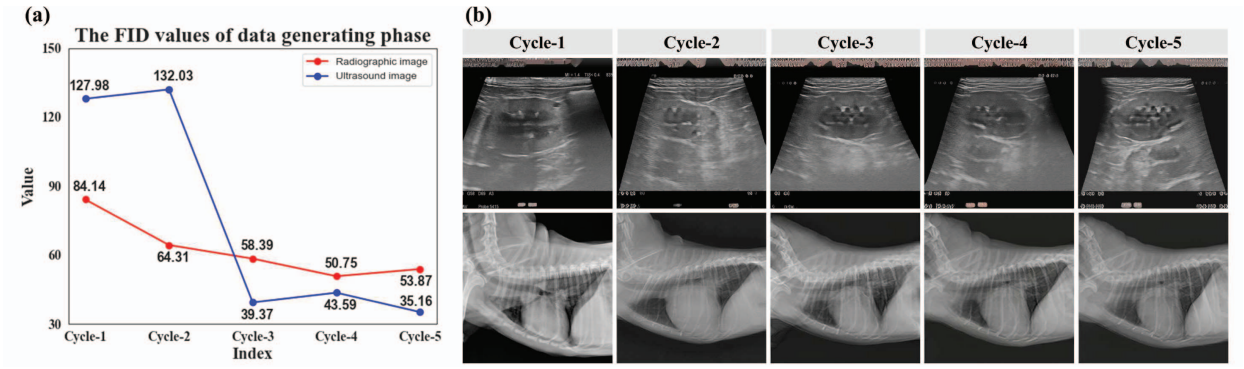

Fig. 3. Data generating phase's results of each cycle. (a) The graph of the FID value over the cycles. (b) The generation results of each cycle.

images. The formula for cosine similarity is provided below. In the given equations, T denotes the latent vector of the ground truth, whereas G represents the latent vector of the generated image.

$$Cosine\ similarity = 1 - \frac{T \times G}{\|T\| \, \|G\|} \tag{2}$$

The original images and the images generated through the data generating phase were passed to the VAE's encoder to obtain embeddings. The cosine similarity between the 100 original images and the generated image was calculated. The generated images with the top 10% of cosine similarity were selected and added to the training set.

C. Classification phase

To demonstrate the validity of our framework, we applied a object detection model for classification to images generated using our framework. The model we used for this purpose is YOLOv8, which is an enhancement of YOLOv5 based on additional layer modifications to improve the model's performance, achieving state-of-the-art results. The YOLO series is a well-known model extensively utilized in various CAD applications [13]. We considered the classification to be correct if the intersection over union (IoU) of the predicted bounding box by YOLO was above 0.7 and the class probability was higher than 0.5. IoU is a measure indicating the degree to which the predicted bounding box matches the ground truth bounding box.

We generated data for cardiomegaly and CKD diseases during the data generating phase. The number of data increased from 100 in the Cycle-1 to 500 in the Cycle-5. We labeled these data as abnormal. Eventually, since the data increased to 500 images, we prepared 500 normal data to train YOLOv8. For example, in the classification phase, Cycle-3 refers to training with 300 abnormal data and 500 labeled normal data. For testing, we extracted 50 data samples per class from ground truth data that were not duplicated with the training data.

For evaluation metrics, we utilized the confusion matrix, along with accuracy, precision, recall, and F1-score [14]. The confusion matrix is a table used in machine learning to assess the performance of a classification model, summarizing the relationship between the model's predictions and ground truth values. From this matrix, accuracy, precision, recall, and F1-score can be calculated using the following formulas. In these formulas, true positive (TP) represents the number of correctly predicted positive observations, while true negative (TN) denotes the number of correctly predicted negative observations. False positive (FP) indicates instances predicted as positive, but actually negative. False negative (FN) indicates the number of instances that are actually positive but are incorrect.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1\ score = \frac{2 \cdot (Precision \cdot Recall)}{Precision + Recall} \tag{6}$$

III. Experiments

A. Data generating phase

The FID values obtained from the experiment on generating radiographic image data and ultrasound image data are illustrated in Fig 3. (a). The FID values are the best results obtained from a cycle's iteration. In Cycle-4, the generated radiographic image data showed the best performance with a FID value of 50.75. For the generation of ultrasound image data, Cycle-5 produced the best results with an FID value of 35.16. Additionally, a tendency was observed in which performance tended to be less favorable with a smaller amount of data. On the contrary, as the amount of data increased, there was an improvement in performance, although the final cycle did not consistently yield the best results for radiographic images.

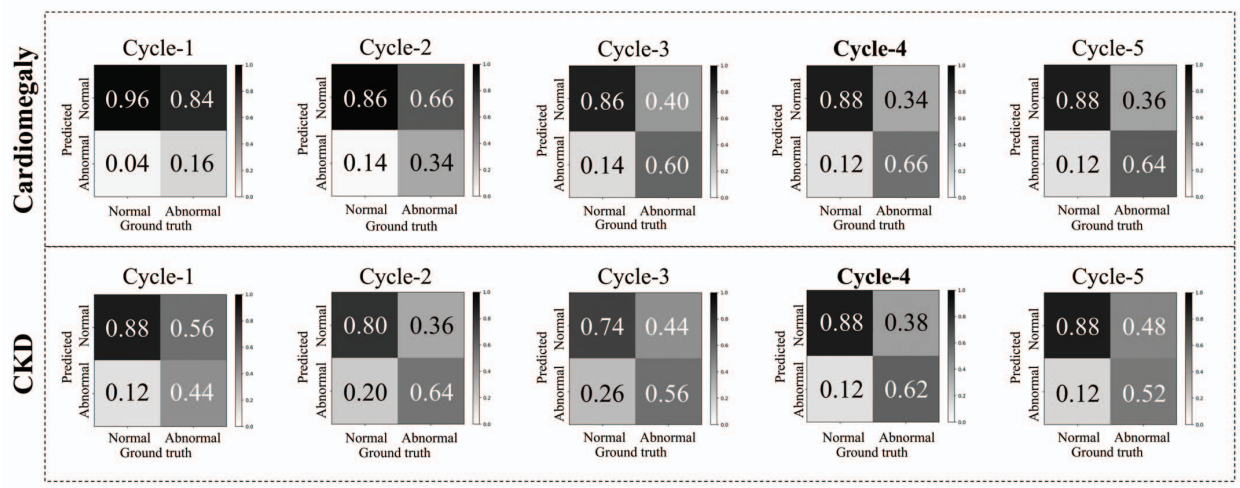

Fig. 4. Confusion matrix results of classification phase. The above results represent training and testing results using cardiomegaly data at the top and CKD data at the bottom. The cycle with the highest accuracy among the five cycles is bolded in the figure.

TABLE I
PERFORMANCE EVALUATION OF CLASSIFICATION PHASE IN RADIOGRAPHIC IMAGE

Cycle	Accuracy	F1-score	Precision	Recall
Cycle-1	0.56	0.53	**0.96**	0.69
Cycle-2	0.60	0.57	0.86	0.68
Cycle-3	0.73	0.68	0.86	0.76
Cycle-4	**0.77**	**0.72**	0.88	**0.79**
Cycle-5	0.76	0.71	0.88	**0.79**

TABLE II
PERFORMANCE EVALUATION OF CLASSIFICATION PHASE IN ULTRASOUND IMAGE

Cycle	Accuracy	F1-score	Precision	Recall
Cycle-1	0.66	0.61	**0.88**	0.72
Cycle-2	0.72	0.69	0.80	0.74
Cycle-3	0.65	0.63	0.74	0.68
Cycle-4	**0.75**	0.70	**0.88**	**0.78**
Cycle-5	0.70	**0.71**	**0.88**	0.75

The generated radiographic image data and ultrasound image data are illustrated in Fig 3. (b). The generated images were generated using the model with the lowest FID within each cycle and represent images with the highest cosine similarity to the original data. Radiographic images exhibit an average cosine similarity of 80.2762, while ultrasound images demonstrate an average cosine similarity of 80.8931. During individual comparisons, the highest cosine similarity for radiographic image data was observed at 90.9575 in Cycle-3, whereas for ultrasound image data, it was recorded at 90.2917 in Cycle-5.

B. Classification phase

The results of the classification using YOLOv8 are presented in Fig. 4. Additionally, the cycle-wise accuracy, F1-score, precision, and recall for each confusion matrix are presented in Table I and Table II. Fig. 4 includes the confusion matrix, where the upper part represents the results tested on radiographic images of dogs with cardiomegaly, and the lower part represents the results tested on ultrasound images of dogs with CKD. The confusion matrix is commonly employed as an evaluation metric in various research studies involving classification tasks [15]. Each classification task was conducted to demonstrate the validity of the data and the classification model training utilized the dataset used for the generative model.

In the cardiomegaly dataset, Cycle-4 exhibited the highest accuracy and F1-score, with values of 0.77 and 0.72, respectively. Furthermore, Cycle-4 and Cycle-5 demonstrated the highest recall, reaching 0.79. The precision was maximized when trained with the original data. The Cycle-1 recorded the lowest accuracy and F1-score, with values of 0.56 and 0.53, respectively. The cycles with the lowest precision were Cycle-2 and Cycle-3, both with values of 0.86. Lastly, Cycle-2 exhibited the lowest recall, with a value of 0.68.

In the CKD dataset, Cycle-4 displayed the highest accuracy and recall, recording values of 0.75 and 0.78, respectively. Additionally, Cycle-1, Cycle-4, and Cycle-5 showed the highest precision of 0.88. The highest F1-score was achieved in Cycle-5, reaching 0.71. The cycles with the lowest accuracy, precision, and recall were Cycle-3, with values of 0.65, 0.74, and 0.68, respectively. In addition, the lowest F1-score was observed in Cycle-1, with a value of 0.61.

IV. Discussion

In this paper, we propose a generative active learning framework that automates the query process using VAE. This framework generates data during the data generating phase and incrementally augments the training dataset of the generative model by filtering data through the query phase. Unlike previous research, we adopt the VAE to enhance the robustness of the query process. The query process has the filtering step by calculating the cosine similarity between the generated images and the real images using 10% of the generated data. Experimental results demonstrate that iterative repetition of this process leads to an improved performance of the generative model.

Observing the change in FID during the data generating phase, we observed a consistent tendency for FID reduction as the cycles progressed. In the case of radiographic data generation, the FID value decreased from 84.14 to 50.75. For ultrasound image data generation, the FID value decreased from 127.98 to 35.16. These results demonstrate that our proposed framework effectively enhances the robustness of the generative model's performance.

To validate the data's effectiveness, we performed a classification phase. The data used for generation encompassed all disease classes, including cardiomegaly and CKD. The experimental results, as observed in the confusion matrix, demonstrate that as the cycle increases, the increase in disease data leads to an increase in TN. For cardiomegaly, TN increased from 0.16 to 0.66, and for CKD, it increased from 0.44 to 0.64. These findings indicate the utility of the data generated using our framework in improving the performance of the classification model.

However, there were also experimental results that raised suspicions of overfitting. When generating radiographic image data, the FID increased as we progressed from Cycle-4 to Cycle-5, and similarly, when generating ultrasound data, there was an increase in FID from Cycle-3 to Cycle-4. Additionally, during the classification phase, the TN values were highest for radiographic images at Cycle-4, and decreased in Cycle-5 compared to Cycle-4. Moreover, when classifying ultrasound images, the TN values were highest at Cycle-2 and decreased thereafter.

Furthermore, this study has some limitations. First, our methodology involved the utilization of an existing GAN variant instead of the proposed model. In particular, contemporary image generative models lean toward diffusion models [16] rather than GAN. Second, while there is a plethora of medical image data available, we restricted our application of the framework to radiographic and ultrasound image data. Given the limited dataset that encompasses only these two modalities, further exploration is essential across diverse datasets to establish the generalizability of our findings. These limitations highlight avenues for future research and improvements in our approach.

V. Conclusion and Future works

This study proposed the VAE-based generative active learning framework. The potential of this framework to address the issue of medical data scarcity in CAD was demonstrated through experimental results, including the FID of the generative model and the confusion matrix, accuracy, F1-score, precision, and recall of the classification model. Future research will extend to proposing generative models such as diffusion models and using various types of data, such as computerized tomography (CT), magnetic resonance imaging (MRI), etc. It will have a positive impact on the performance improvement of the CAD system in the future and provide an opportunity to promote the development of the medical AI field.

References

[1] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan, "Do we need more training data?," *International Journal of Computer Vision*, vol. 119, no. 1, pp. 76–92, 2016.

[2] J.-Y. Oh, I.-G. Lee, H.-H. Chang, E. Lee, and J.-H. Jeong, "Application of a dual-stage deep learning framework to detect left atrial enlargement for pet heart failure," in *Proceedings of International Conference on Systems, Man, and Cybernetics*, 2023, pp. 2972–2978.

[3] D. Yoon, H.-J. Kong, B. S. Kim, W. S. Cho, J. C. Lee, M. Cho, M. H. Lim, S. Y. Yang, S. H. Lim, and J. Lee, "Colonoscopic image synthesis with generative adversarial network for enhanced detection of sessile serrated lesions using convolutional neural network," *Scientific Reports*, vol. 12, no. 1, p. 261, 2022.

[4] M. La Salvia, E. Torti, R. Leon, H. Fabelo, S. Ortega, B. Martinez-Vega, G. M. Callico, and F. Leporati, "Deep convolutinal generative adversarial networks to enhance artificial intelligence in healthcare: A skin cancer application," *Sensors*, vol. 22, no. 16, p. 6145, 2022.

[5] J.-J. Zhu and J. Bento, "Generative adversarial active learning," *arXiv preprint arXiv:1702.07956*, 2017.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[7] J.-H. Jeong, J.-H. Cho, B.-H. Lee, and S.-W. Lee, "Real-time deep neurolinguistic learning enhances noninvasive neural language decoding for brain–machine interaction," *IEEE Transactions on Cybernetics*, vol. 53, no. 12, pp. 7469–7482, 2023.

[8] C. Lam, B. J. Gavaghan, and F. E. Meyers, "Radiographic quantification of left atrial size in dogs with myxomatous mitral valve disease," *Journal of Veterinary Internal Medicine*, vol. 35, no. 2, pp. 747–754, 2021.

[9] N. Bragato, N. C. Borges, and M. C. S. Fioravanti, "B-mode and doppler ultrasound of chronic kidney disease in dogs and cats," *Veterinary Research Communications*, vol. 41, pp. 307–315, 2017.

[10] A. Sauer, K. Chitta, J. Müller, and A. Geiger, "Projected GANs converge faster," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 480–17 492, 2021.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] F. Rahutomo, T. Kitasuka, and M. Aritsugi, "Semantic cosine similarity," in *Proceedings of International Student Conference on Advanced Science and Technology*, vol. 4, no. 1, 2012, p. 1.

[13] J. Estrada, Y. Zhigang, S. Datta, N. Duraisamy, J. De Guia, O. Cheng Hun, G. Opina, and A. Tripathi, "AV in action: A development of robust and efficient planning and perception system for autonomous food delivery vehicle," in *Proceedings of IEEE International Conference on Artificial Intelligence*, 2023, pp. 19–20.

[14] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2020.

[15] J.-H. Jeong, B.-W. Yu, D.-H. Lee, and S.-W. Lee, "Classification of drowsiness levels based on a deep spatio-temporal convolutional bidirectional LSTM network using electroencephalography signals," *Brain Sciences*, vol. 9, no. 12, p. 348, 2019.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.