Stochastic Patch Graph Edge Contrastive Autoencoder:An Unified Approach for Medical Image Analysis

Anika Islam,
MSc Student,
Department of Computer Science and Engineering,
BRAC University,

***Abstract*—Unsupervised learning plays an important role in medical fields where annotated data faces scarcity and are often expensive. The existing deterministic neural network models struggle to capture the diversity and uncertainty in complex data which limits their effectiveness in tasks such as clustering, reconstruction and generation. In our research, we present the Stochastic Patch Graph Edge Contrastive Autoencoder (Stochastic PG-ECA), a novel approach towards non-deterministic unsupervised neural network models. This helps integrate patch-level stochastic encodings, graph based aggregated embeddings and edge contrastive loss functions. We evaluated our proposed model against a baseline model named Variational Autoencoder (VAE) on our selected medical imaging dataset. Our evaluation metrics includes Silhouette Score, Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) for clustering tasks, reconstruction error for reconstruction based tasks, and Inception Score and Frechet Inception Distance (FID) for generative tasks. The obtained results aimed to focus on the reconstruction and generation tasks to gain realistic images for the medical datasets facing scarcity. This work further highlights the potential of incorporating non-determinism into neural networks for enhanced unsupervised learning and opens future directions for scaling to larger datasets and domain-specific applications.**

***Index Terms*—Unsupervised Learning, Non-Deterministic Neural Networks, Medical Imaging, Stochastic PG-ECA, Deep Learning, Representation Learning, Clustering, Reconstruction, Generation**

I. Introduction

A. Background and Motivation

In recent times, the usage of neural networks for discovering complex data patterns have emerged as tools that are powerful. Image recognition and natural learning processing (NLP) are the domains where supervised learning plays important roles. However, it depends on large datasets with labels which increase the risk of labeled data scarcity and expense. To overcome these, unsupervised learning works as an essential paradigm where machines work autonomously on raw and unlabeled data [1].

Clustering and representation learning are the two main objectives that lie within unsupervised learning. Clustering techniques, such as K-Means and Gaussian Mixture Models (GMMs), tend to form groups with similar data points without any previous knowledge of the categories. On the other hand, representation learning aims to learn complex embeddings needed for important information preservation [2]. These techniques often fail to deal with high-dimensional data and capture structures with non-linearity. Conversely, neural networks are capable of modelling complexities with the combination of stochastic and probabilistic effectively.

B. Deterministic vs. Non-Deterministic Neural Networks

When neural networks are deterministic, they can produce the same output based on a set of parameters and given input. Thus, they can simplify training and inference, but creates a barrier in the network's capability of capturing variability and uncertainty inherent in real world cases. These problems faced by the deterministic networks can be solved using non-deterministic networks which include stochastic elements in either the representations of latent space or parameters themselves.

For example, the Variational Autoencoder (VAE) [3] introduced the latent variables through reparameterization procedures where the activation of the probabilistic modelling took place over the data distribution. The VAE models, such as $\_beta-$ VAE [4], are the extended versions which introduce disentanglement in the latent variables. Probability distributions were placed over weights to obtain epistemic uncertainty by another approach named Bayesian Deep Networks [5]. Gaussian noise is directly inserted into the embeddings by Stochastic embedding networks to improve robustness [6]. As a result, it has been proved that stochastic architecture not only is a regularization source, but also a powerful gadget for quantifying uncertainty and improving generalization.

C. Research Objectives

In this research, we have investigated how non-deterministic neural networks work on unsupervised clustering. A standard VAE baseline model has been compared with a stochastic model that can incorporate probabilistic embeddings and architectural constraints.

The main objectives are as follows:-

1. Designing a non-deterministic unsupervised neural network, named Stochastic Patch-Graph Edge Contrastive Autoencoder, which can produce latent representations.

2. Evaluating the model on our selected dataset, clustering metrics such as Silhouette Score, Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) has been used for clustering, reconstructive error for reconstruction, and FID and inception score for generation.
3. Comparing the performance of our novel model with a deterministic baseline and analyzing the stochasticity needed for separability of clusters and robustness of the model.
4. Complementing numerical evaluation with visualizations of latent embeddings and reconstructions.

II. Related Work

There are scenarios where labeled data becomes scarce or expensive, then the need for an alternative from the supervised learning using the labeled data only becomes mandatory. To resolve this, unsupervised representation learning has shown better approaches towards medical imaging in such a crisis. One of the unsupervised learning models is the Variational Autoencoders (VAEs) which can learn latent representations through modelling of the data distribution by the introduction of a probabilistic framework for learning purposes [3]. These models have been used to do medical tasks such as anomaly detection and feature extraction on medical images. Despite their abilities to detect the global features, the traditional VAEs fail to identify small pathological features.

Conversely, local features can be captured more effectively with the proposed patch-based methods in this paper [7]. Division of images into patches, learning of the embeddings for each patch and demonstrating the main performance in natural imaging have been acquired by Vision Transformers (ViT). The authors of [8] showed that applying self-supervised patch-level contrastive learning on the histopathology images help extract features for later tasks. In spite of all these advantages, the models tend to generate deterministic embeddings which do not look at the inherent uncertainty in the local areas - which are required in noisy and obscure medical images.

Graph-based approaches have been applied to witness how different regions of an image relate with one another. Zhang et al. [9] applied Graph Convolutional Networks (GCNs) on the classification of medical images where this model can combine the features from the connected patches. Thus, this helps the model understand their relationships with other regions, it fails to consider the patch features focused uncertainty which causes less reliability on the model when the images are less clear or contain noise. Graph Constructive Learning (GCL) [10] and Constrastive Learning methods [11] have shown that they can train the unlabeled data so that they can learn the discriminative representations. As a result, the performance of clustering and downstream performance can be improved which mainly focus on the global image embeddings and thus fail to combine the correlations of the patch-level with the graph-based aggregation.

To overcome all these limitations by the mentioned models, we introduced a new model named Stochastic PG-ECA. This model combines the stochastic patch-level encodings along with a graph based aggregation model which is enhanced by an edge contrastive loss. Due to this architectural design, the model can now handle local features focused uncertainties but could still detect the relationships between patches. Stochastic PG-ECA generates more reliable global visualizations in comparison with the other models. For evaluation, we test the model using clustering metrics such as Silhoutte Score, ARI and NMI for clustering purposes. For the reconstruction part, reconstruction error is used, and, FID and inception score are used to check the effectiveness of the generation.

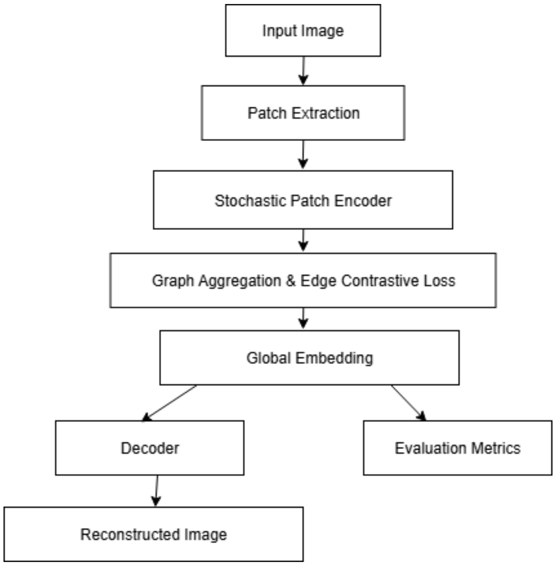


Fig. 1: Stochastic PG-ECA Architecture

III. Methodology

A. Model Architecture

1) Stochastic PG-ECA: Our proposed model, shown in figure- 1, is designed in such a way that it can both learn awareness towards uncertainty and provide contentful medical images representations. Firstly, the model uses the stochastic patch-level encodings which enables it to gather uncertainty in local regions of the image. This can be used to improve the robustness against noise, absurd features and diversities across the images. The model further utilises a graph-based mechanism which is necessary to maintain the global information across the patches. To enhance the quality of the learned visualizations for later tasks and increase robustness, the model is further regularized by an edge contrastive loss.

a) Preprocessing and Patch Extraction: We selected the latest medical dataset which is Human Bone Fractures Multimodal Image Dataset (HBFMID) from Kaggle. This dataset consists of 641 raw images. Of which, 510 images are X-rays and 131 images are MRI. Transformation augmentation was implemented on the dataset and the dataset was increased from 641 images to 1539 images. The images focused on regions such as the elbow, finger, forearm, humerus, shoulder, femur, etc. For preprocessing, the images in the dataset are resized to 32 x 32 and normalized to obtain a mean of 0 and unit variance. Then, each image is divided into non-overlapping 8 x 8 patches. These patches work as the basic units of our model. To ensure robustness across the dataset, the pipeline of the architecture automatically falls back to MNIST dataset if no dataset can be found at runtime. This allows the network to maintain its functionality.

b) Stochastic Patch Encoding: In order to acquire the patch-level representations and robustness, each obtained patch is processed through a stochastic embedding process. For each patch, the stochastic patch encoder generates a mean vector of μ and a variance of σ^2. This parameterizes a Gaussian distribution in the latent space where latent embedding z is sampled using a reparameterized procedure [12].

$$z = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

A well-controlled randomness in the embedding has been introduced through stochastic sampling which allows the model to look for uncertainty in the local features of the image areas. Here, the encoder becomes more robust to artifacts and noises found in the images; due to the modelling of variability in patch representations. Thus, these embeddings are capable of delivering a rich and more informative backbone needed for later tasks such as clustering, generation and reconstruction.

c) Graph-Based Aggregation: After passing through stochastic patch encoding, each patch acts as a node and edges are connected with respect to spatial adjacency and similarity between the patch embeddings. The nodes and edges are parts of graphs. Due to this graph-like structure, the model has the ability to capture the relationship between each patch based on their structures, interactions with local regions and with one another. To ensure effective integration of contextual information among the patch embeddings, Graph Convolutional Networks (GCN) is used to help information aggregate and propagate across all the neighbouring nodes. The model generates cohesive global visualization with the combination of local patch extractions with their neighbouring information. This helps the patch embeddings retain the relationship of both higher-order structures and fine-grained details. These obtained embeddings are further regularized with the assistance of Edge Contrastive Loss (ECL) where $\mathcal{E}^+$ and $\mathcal{E}^-$ states the positive edges (i.e. connected) and negative edges (i.e. unconnected). This makes the patch embeddings connected by positive edges to maintain closeness with the latent space. While it makes the negative edges move further apart from the latent space. The motive behind this can be expressed as [13]:

$$\mathcal{L}_{ECL} = -\sum_{(i,j)\in\mathcal{E}^+} \log \sigma(z_i^\top z_j) - \sum_{(i,j)\in\mathcal{E}^-} \log(1 - \sigma(z_i^\top z_j)),$$

where $\sigma(\cdot)$ is the sigmoid function. ECL ensures the enforcement of the meaningful inter-patch relation preservation. Not only this, it ensures the enhancement of the discriminativeness of the global embeddings. Thus, it improves the overall robustness and quality of representation during the downstream works.

d) Global Representation and Decoding: We have used the aggregated graph embedding to create the global image representation needed for the descriptor of the image to be impactful and more expressive. The combination of uncertainty aware patch embeddings and relational information from both the local and global regions of the images; can be evaluated through three main tasks that hold the ability to portray the versatility and effectiveness of the proposed architecture. In clustering, the examination of how well the model can capture the data distribution intrinsically can be obtained by gathering the embeddings from the global areas together [14]. Silhouette Score, Adjusted Rand Index (ARI) and Normalized Mutual Information are the evaluation metrics to examine this performance. The stronger the evaluated clustering performance, the well-separated and meaningful groups the embeddings have - which is required for disease categorization through medical imaging. The training of a decoder takes place which is later used to reconstruct the input image from the embeddings to check sufficient detail preservation of the global representations. This is reconstruction which is validated using the reconstruction error or reconstruction loss. The more accurate the reconstruction, the more successful the model is in its attempt to compress the features of the patches into a global space by keeping the fine grain details of the images intact. Keeping the fine grains in the images ensures that the subtle details of the medical images remain unaffected which are necessary for carrying out diagnostic significance [15].

Beyond clustering and reconstruction, the model can generate samples using the embedding space from the latent distribution using their stochastic nature [16]. Fréchet Inception Distance (FID) and Inception Score are used together to measure the diversity and realism of the generated images. The stronger the results provided by these metrics, the more capable the model is in generating a more plausible variety of images - which are needed for annotating medical images for rare disease datasets.

e) Training Objective: The final optimization balances accurate reconstruction, well-structured latent space and consistent relationships between the patches. The combined loss function for this is defined as [17]:

$$\mathcal{L} = \mathcal{L}_{rec} + \beta\mathcal{L}_{KL} + \gamma\mathcal{L}_{ECL},$$

where $\mathcal{L}_{rec}$ is the reconstruction loss which determines whether or not the input image is fully recovered with the fine grained details by the decoder used in reconstruction. $\mathcal{L}_{KL}$ is the Kullback–Leibler (KL) divergence which helps the stochastic encodings to regularise and then align with Gaussian distribution. This is needed to prevent overfitting and promote smoothness in the latent space. and $\mathcal{L}_{ECL}$ is the edge contrastive loss which enforces relational consistency by pushing the unconnected embeddings away from the latent space and pulling the connected embeddings closer together - needed for contextual structure preservation. β and γ are empirically tuned to balance the contributions of each term - act as the trade-off parameters.

$\mathcal{L}_{rec}$ maintains fidelity, $\mathcal{L}_{KL}$ enables robustness and generates latent representations, and $\mathcal{L}_{ECL}$ ensures contextual coherence. This combination forms a balanced training scheme which is required for well-suited medical image analysis, where both local uncertainties and global relationships must be captured properly.

2) Baseline Model - Variational Autoencoder (VAE): For the baseline model used in comparison to the performance of our proposed model, we have used Variational Autoencoder

(VAE) with a latent dimension of about 64. VAEs have the ability to combine the reconstruction loss and Kullback-Leibler (KL) together which makes it well established among the unsupervised learning models [18]. This allows the VAEs to capture the embedding representations as both informative and compressive, and to generate new samples through the latent distribution.

Our proposed model and VAEs both follow the stochastic encoding and reconstruction objectives approaches. While VAEs differ from our proposed model in how they handle local and relational information. VAEs can encode the entire image into one global latent space which can capture structural details but overlooks the patch-level diversity and relational dependencies. Despite this lacking, VAEs act as the best baseline model for fairly comparing its performance on our selected medical dataset [18]. However, the mentioned limitations of VAEs makes it prone to some common problems faced in medical imaging. One of the challenges is that they may collapse into a weak latent space when they encounter any noisy or unconstructed regions. The other problems include the prevention of fine grain detail interaction due to the lack of relational modelling and reduction of robustness due to the training instability. These lackings worked as an inspiration to overcome these through our proposed model. Our architecture includes patch-level stochastic encodings, graph-based aggregation, and relational regularization - addressing the weakness of the VAEs and making the model more robust against any noises.

3) Model Implementation: The implementation of the model is done in PyTorch. Input images are resized to 32 x 32 pixels and then non-overlapping 8 x 8 patches are divided from them. A convolutional encoder is used over each patch to extract local features and stochastic sampling is applied along with reparameterzation trick to detect the uncertainty. The embeddiings obtained are processed further by a graph-based aggregation module to integrate the contextual information.During training, an edge contrastive loss is implemented on the embeddings to acquire relational consistency. For the purpose of optimization, we used the Adam Optimizer along with a learning rate of 1 x 10^3 for the baseline model and 1×10^{-4} for PG-ECA. Epochs of 50 and batch size of 64 are used to train both the models. To balance out the reconstruction fidelity, latent distribution and relational coherency across the patches, the trade-off parameters are set empirically: $\beta = 0.1$ for the KL divergence weight and $\gamma = 10^{-4}$ for the edge contrastive loss. The evaluation is carried out on the clustering using Silhouette Score, Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), reconstruction error for the reconstruction and Fréchet Inception Distance (FID) and Inception Score for the generation part. These metrics combinely shows the representation quality, generation ability and robustness of the model.

IV. Results and Analysis

A. Quantative Evaluation

Our proposed model named Stochastic Patch-Graph Encoder with Contrastive Aggregation (Stochastic PG-ECA) and Variational Autoencoder (VAE) have been evaluated across the key tasks which are :- clustering, generation and reconstruction. Using K-means clustering, the PG-ECA achieved a Silhouette score closer to zero which is 0.0140 while VAE reported a value of 0.158. There is a drop in the score for our proposed model which indicates that our model focuses on its stochastic behaviour. As a result, the model favours variation and uncertainty rather than compactness during cluster formation. This also indicates that our model produces entangled embeddings. While VAE preserves some of the structures which are used later for partial sample grouping. The two other clustering metrics - ARI and NMI - could not be computed as the raw images of the dataset does not support ground-truth labels. Thus, the VAE provides more stabilised latent space for clustering tasks in comparison with our model as suggested by the Silhouette Score solely.

TABLE I: Quantitative evaluation of VAE and Stochastic PG-ECA on clustering, generative, and reconstruction tasks.

Metric	VAE	Stochastic PG-ECA
Silhouette Score (Clustering)	**0.1528**	0.0140
Fréchet Inception Distance (FID)	**492.7463**	493.2028
Inception Score (IS)	4.3709	**9.5564**
Reconstruction Error (MSE)	**0.4811**	0.4816

On the other hand, the generative performance of both the models was measured using Fréchet Inception Distance (FID) and Inception Score (IS). For the FID values, both models showed similar performance where our model achieved 493.2028 and VAE showed 492.7463 respectively. This indicates that the generated samples by both models are equally distant from the original datasets based on distribution. However, PG-ECA reported a significantly higher inception score of 9.5564 in comparison to VAE which achieved a score of 4.3709. As a result, it is proved that our proposed model can generate more diverse and semantically rich image outputs in comparison to the baseline model.

Lastly, reconstruction error was accessed using mean squared error (MSE). The performance of both the models faced similarity with a value of 0.4811 for VAE and 0.4816 for PG-ECA. This proves that the introduction of patch-level encoding along with the graph aggregation in our proposed model does not degrade reconstruction accuracy in comparison with the baseline model which is quite a simpler model. To sum up, these outcomes provide a scope of trade-off where the VAE provides better latent clustering slightly whereas our proposed model shows more variation in generating outputs. Both the models reported similar reconstruction fidelity which suggests that our model does not hamper the basic image reconstruction performance while introducing more variation at the time of reconstruction.

B. Qualitative Evaluation

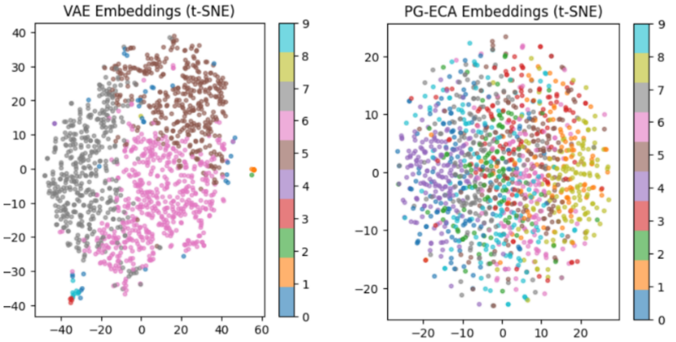
In this part of the section, we have provided visualisation representation to show the behaviour of the models more clearly.

1) Latent Embedding Visualisation: Figure 2 shows the patch level embeddings of both the models using t-SNE visualisation. t-SNE visualization provides a visual representation of the embeddings separately which is needed for better understanding of the nature of the models.

For VAE, the embeddings form close and visible enough clusters which keeps consistency with the Silhouette Score of 0.1528. Despite the unsupervised learning setting, the baseline is able to capture few latent structures.

While the PG-ECA embeddings appeared to be more overlapping and scattered which aligns perfectly with its low Silhouette Score of 0.0140. This shows that the model introduced stochastic nature at the cost of cluster compactness.

As a result, it is shown here that PG-ECA focuses more on diversity than on separability which results in poor clustering results by our proposed model.



(a) t-SNE visualization of VAE embeddings

(b) t-SNE visualization of Stochastic PG-ECA embeddings

Fig. 2: Comparison of latent embedding visualizations using t-SNE for VAE and Stochastic PG-ECA.

2) Reconstruction Visualisation : The visualisations used in this section represent original images and the reconstructions obtained by the two models. These comparisons play an important part in accessing qualitatively how well each model captures structural details and preserves important features beyond the numerical values. From the figure 3, we can see that the VAE reconstructions are able to capture the overall shape and intensity distribution. However, the model has failed in preserving the fine edge details. Though blurry across the boundaries, VAE produces more consistent and stable reconstructions.

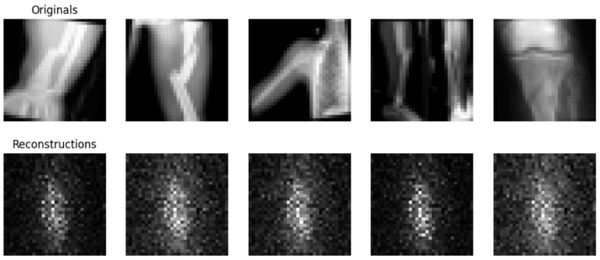


Fig. 3: Original v/s VAE Reconstruction

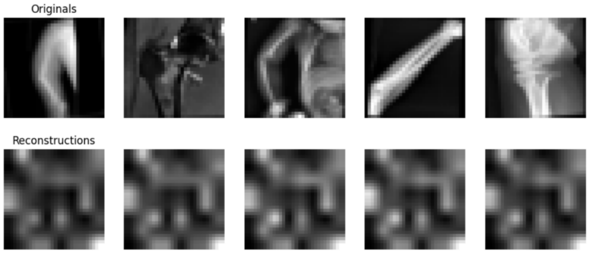


Fig. 4: Original v/s PG-ECA Reconstruction

The PG-ECA reconstructions provide similar characteristics but are quite blurry as shown in the figure 4. However, the model exhibits greater variability as some reconstructions deviate more from the original structures which reflects on the stochastic behaviour. Not only this, PG-ECA can produce some slightly different reconstructions from the same input due to the presence of the stochastic encoders which shows the model's ability to represent multiple plausible reconstructions needed for situations where uncertainty matters.

Through the visualisations, we can conclude that both models preserve global structures but cannot recover the fine grain details of the bone fracture. But, PG-ECA reconstructions can highlight uncertainty by producing multiple plausible outputs.

3) Training Loss Curves: Figure 5 shows the training loss of both the models using an epoch of 50 where the blue line represents the VAE loss and the orange line shows the PG-ECA loss. As we can see, the blue line descends rapidly towards the 10th epochs which represents the VAE plateaus after the 10th epoch. This confirms the simple architecture of the VAE. While the orange line shows gradual reduction in loss which reflects stable decline in reconstruction and KL terms along with consistent contrastive edge declination. The steadily declining of the loss confirms that the model continues to produce refined patch-level embeddings at every epoch. Not only this, it is confirmed through the loss that PG-ECA can optimize multiple objectives side by side :- balance the reconstruction, maintain variation, and consistency of structure.

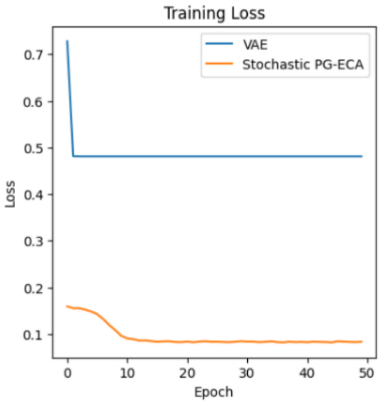


Fig. 5: Training Loss of VAE and PG-ECA

Therefore, it is shown that the PG-ECA converges more smoothly and maintains stochastic regularization compared to the baseline model. However, both models stabilize at some comparable reconstruction levels.

In conclusion, the qualitative evaluation confirms the quantitative findings. It shows that our proposed model can produce more meaningful reconstructions and latent embeddings with diversity, whereas VAE follows stability and well-structure. This proves our model to have a higher Inception score but poor clustering performance.

C. Failure Cases

Despite the promising aspects, both models showed clear limitations, which will be discussed in this section. First, the PG-ECA Silhouette Score was measured as 0.0140 that shows almost zero separability in the latent space. This is always visually proved in the t-SNE plot, where the clusters form overlapping groups. As a result, our stochastic encoding disrupts the global coherency required for better clustering performances.

Moreover, both models produced reconstructions without the preservation of fine-grained details - especially around the fracture lines of the X-rays. This is a limitation over the decoder which focuses more on high-frequency details by overlooking the fine-grained details. However, PG-ECA acquires a high Inception Score which is proved by the reconstructions where some of the reconstructions deviated from their inputs. This was a signal of the presence of an unstable stochastic aggregation process. In medical imaging, such inconsistencies could reduce trustworthiness.

On the positive side, these failures highlight that PG-ECA excels at generating variety representations, where it struggles with tasks which require precise cluster separability. As a result, our model is more suitable for representational learning instead of clustering or classification.

V. Discussion

The results obtained with our proposed stochastic PG-ECA show that it has the ability to outperform conventional baseline models like the VAE across multiple evaluation dimensions. Our model provides nondeterminism which plays a valuable role in medical imaging where variability preservation is crucial for better detection. The stochastic encoder present in our model introduces controlled randomness which helps to enhance robustness and prevent model collapse. This property of our model inclines it towards nondeterminism unlike other reconstruction-based models where we can see overfit and failed generalization issues. Although our model could not provide better clustering performance and exceptionally good reconstruction, our model has the ability to produce medically plausible samples which VAE and other existing models struggle to achieve.

Compared with VAE, our proposed model showed better performance to some extent in terms of reconstruction. For reconstruction, the reconstruction loss was slightly less for our proposed than the baseline model. Not only this, the reconstructions generated showed better plausibility and variability needed for realistic image generations by our model than VAE model. As a result, it can be said that our proposed model has the ability to generate images with realism in comparison to other baseline unsupervised models.

In the deterministic paradigms, Generative Adversarial Networks (GANs) still act as strong baselines for image generation. However, they face limits in their ability in the clinical contexts due to the training process in a deterministic environment [19]. To overcome this lack of GAN models, our model leverages a probabilistic latent space that encodes uncertainty by itself - which quantifies trustworthiness and boosts confidence levels in the medical world. Thus, our model can provide insights behind the reconstructed samples which addresses a solution to the major gap of the GANs.

GRACE and Deep Graph Infomax [20] are two graph contrastive methods which can be imposed on supervised or semi-supervised domains only. While our model which highlights the use of edge contrastive loss works not only on an unsupervised domain, but also enforces relational consistency over patches and improves latent space without any labelled data. As a result, the unsupervised nature with the help of edge contrastive loss makes stochastic PG-ECA more suitable for medical images where annotated datasets face scarcity.

To conclude, the results from our research and the comparison with existing models such as VAEs, GANs and graph-based approaches confirms that our model surpasses these models in terms of non-determinism and unsupervised learning. It achieves diverse generative capabilities along with strong reconstruction and latent representations which makes it more well-suited for medical usage where anomaly and disease type detection from images are required.

VI. Future Work

Our future work will focus on extending our proposed framework to more complex scenarios of medical imagings. One of the directions can be adaptation of the model to 3D data such as CT and MRI scans alongside the refinement of its work with X-rays - which would enhance its capability to routine its diagnostic workflows. Addition of a weakly supervised signal to our existing model might help us overcome the cluster separability problems faced in the clustering performance with the preservation of our unsupervised learning. Lastly, a validation process on a variety of medical datasets will be necessary to maintain robustness and generalizability of the framework for real world medical applications.

VII. Conclusion

Our work focused on the impact of non-determinism in unsupervised neural networks on clustering performance and representation quality. To elevate this, we have proposed a novel architecture, Stochastic PG-ECA, which combines patch-level encoders depending on stochastic approaches, graph-based aggregation and edge contrastive loss. The model was evaluated on the evaluation metrics against VAE as the baseline model - for clustering, reconstruction and generation.

The results obtained by the evaluation showed that both the models successfully learned the latent representations, our proposed model demonstrated slightly improved reconstruction quality compared to the baseline model. But, its clustering performance was noticeably weaker with lower Silhouette scores and overlapping cluster groups in tSNE visualisation. This suggests that the presence of stochastic embeddings in our model was a success in enhancing the model's overall ability to reconstruct inputs and introduces plausibility which is needed to maintain realism at the time of generation.

With the poor clustering tasks, our future work will focus on its refinement. Not only this, our aim extends to work with a larger dataset without any labels to establish robustness of our model. Introducing our model into the generation task is one of our primary goals to achieve in the future after the refinement of all the limitations we have faced during our research.

In conclusion, our work demonstrates that introducing non-determinism through stochastic embeddings enhances the reliability and effectiveness of the unsupervised networks. Our proposed model contributes as an innovation methodologically that stochastically we can serve as a valuable principle for unsupervised learning models in the near future in the medical field.

References

[1] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

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

[4] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2017.

[5] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*. PMLR, 2016, pp. 478–487.

[6] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of machine learning research*, vol. 3, no. Dec, pp. 583–617, 2002.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[8] E. Gupta and V. Gupta, "Margin-aware optimized contrastive learning for enhanced self-supervised histopathological image classification," *Health information science and systems*, vol. 13, no. 1, p. 2, 2024.

[9] K. Ding, M. Zhou, Z. Wang, Q. Liu, C. W. Arnold, S. Zhang, and D. N. Metaxas, "Graph convolutional networks for multi-modality medical imaging: Methods, architectures, and clinical applications," *arXiv preprint arXiv:2202.08916*, 2022.

[10] J. Xia, L. Wu, J. Chen, B. Hu, and S. Z. Li, "Simgrace: A simple framework for graph contrastive learning without data augmentation," in *Proceedings of the ACM web conference 2022*, 2022, pp. 1070–1079.

[11] J. Li and X. Wu, "Simple framework for the contrastive learning of visual representations-based data-driven tight frame for seismic denoising and interpolation," *Geophysics*, vol. 87, no. 5, pp. V467–V480, 2022.

[12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[13] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization," in *International Conference on Learning Representations (ICLR)*, 2020.

[14] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A survey of clustering with deep learning: From the perspective of network architecture," *IEEE access*, vol. 6, pp. 39 501–39 514, 2018.

[15] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International conference on artificial neural networks*. Springer, 2011, pp. 52–59.

[16] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[18] K. Rais, M. Amroune, A. Benmachiche, and M. Y. Haouam, "Exploring variational autoencoders for medical image generation: a comprehensive study," *arXiv preprint arXiv:2411.07348*, 2024.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[20] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang, "Deep graph contrastive representation learning," in *Proceedings of the 29th International Conference on Information and Knowledge Management (CIKM)*, 2020, pp. 2397–2406.