Enhancing Cancer Cell Classification through Dataset Augmentation using Conditional Variational Autoencoder (CVAE)

Department of Mathematical Sciences
African Institute for Mathematical Sciences
Kigali, Rwanda
anita.esieshun@aims.ac.rw

Abstract

This paper uses Conditional Variational Autoencoder (CVAE) as a data augmentation technique for cancer cell classification using four different classification algorithm; Support Vector Machine (SVM), XGBoost, Decision Tree, and Logistic Regression. The main aim of the study is to enhance the accuracy of the classification models by generating synthetic data using CVAE. The results obtain shows that there is significant improvement when CVAE is used to generate additional data. The approach used indicates a promise in accurate cancer classification, even in the absence of adequate dataset. This will help to improve cancer diagnosis and treatment in the health industry

1 Introduction

The classification of cancer cells play a critical role in early diagnosis, prognosis and treatment decisions in oncology. The advent of machine learning and deep learning techniques have proven to accurately identify cancerous cells using histopathological images and the likes. However, the fundamental challenges in this domain have to do with limited availability of training data, and lack of the different forms in which cancerous cells can appear. These are major obstacles that hinders the performance of classification models and the ability to diagnose effectively. In order to address this issue, data augmentation techniques have gained considerable attention as they enable the generation of additional data instances, effectively expanding the size and diversity of the dataset without the need for expensive and time-consuming data collection procedure, or effort involved in annotation. In the past few years, Variational Autoencoders (VAEs) have emerged as a powerful tool used for generating synthetic data, and has proven to be both realistic and also capable of preserving the data distribution structure. The aim of the study is to explore the effectiveness of Conditional Variational Autoencoders (CVAEs) for augmenting cancer cell datasets, in order to address the concern of data size limitation, while ensuring high accuracy of cancer cell classification models. CVAEs, unlike VAEs leverage the given class labels during encoding and decoding, which in tern makes it possible to generate samples that are conditioned on particular classes. By including class information, CVAEs are equipped with the ability to generate cancer cell-like data that is not just visually similar to the real samples but also aligns with the intended classes, hence ensuring more meaningful augmentation for the task. In this paper, a comprehensive investigation of CVAE based data augmentation for cancer cell classification is presented. The architecture of CVAE model is

5th Deep Learning Indaba Conference (DLI 2023).

described, and a proposed methodology for training the CVAE on the dataset to generate the synthetic instances. The effectiveness of the augmented data is validated using state of the art classification models, and comparing their performance on the original dataset, and the augmented dataset The study proposed above promises to address the scarcity of labeled cancer cell data, and also enhance the robustness of cancer cell classification model through the generation of synthetic but informative data instances. The ultimate goal is to contribute to the advancement of medical image analysis and help in the early detection and treatment of cancer

2 Literature review

Literature Review Cancer cell classification is very crucial in early discovery and personalized treatment. However, a wide variety of well labelled datasets are necessary for accurate categorization. Classification models have difficulties when dealing with imbalanced datasets with limited positive samples. To generate synthetic data and improve model performance, data augmentation techniques have been explored. This review explore the use of Conditional Variational Autoencoders (CVAEs) for dataset augmentation to enhance cancer cell classification. Numerous studies have applied machine learning and deep learning for cancer cell classification. U-Net by Ronneberger et al. [2015] and transfer learning by Guo et al. [2019] show the potential of deep learning in classifying cancer cells. Nevertheless, the success story depends on large and balanced dataset, which are not the case most of the time in medical applications. The goal of Data augmentation approaches is to address the issue of imbalance dataset. Random oversampling, and SMOTE Chawla et al. [2002] generate synthetic samples for the minority class. Even though effective, the approach may lead to over-fitting or may fail to reflect the data distribution. CVAEs are extensions of VAEs by including data labels in the process of encoding and decoding. This ensures the generation of synthetic data that is conditioned on certain classes, resulting in meaningful augmentation. CVAEs have shown to be very successful in the analysis of medical images. CVAEs was used by Xiong et al. [2021] as a data augmentation technique in cardiac MRI, to enhance cardiac disease classification. This same augmentation technique is used by on skin lesion datasets for melanoma detection. CVAEs are capable of generating realistic labeled synthetic data in cancer cell classification. Even though CVAEs are capable of producing class conditioned synthetic data, hyper-parameter tuning and computation requirement can be hindrances. Also, the quality of the generated data is dependent on diversity and representation of the dataset. Deep learning has shown great capability in classification of cancer cells, even though labelled data is scares. CVAEs are versatile techniques for generating synthetic data.
This review motivates our research on enhancing cancer cell classification using CVAE data augmentation. By leveraging the strength of CVAE, we can improve the accuracy and robustness in cancer cell classification models, to ensure early detection of cancer

3 Methodology

3.1 Data description

The dataset used in this work is the breast cancer dataset obtained from kaggle. It consists of 569 data instances of biopsied breast cancer cells, with a total of 32 features representing the physical characteristics of cell nuclei. The dataset consists of two classes: Malignant (212 of the total sample) and Benign (357 of the total sample).

3.2 Data pre-processing

Pre-prossessing the dataset is the first stage before training the model. We check for missing values, and remove redundant features, which reduces the feature to 30. We then standardize the feature variables in order to improve convergence of the optimization algorithm.

3.3 Conditional Variational Autoencoder (CVAE), training, data augmentation process, and classification

In order to augment the breast cancer data, we apply CVAE on the training set (80% of the total dataset). Its architecture is adapted to generating new samples based on already existing data. The feature and their corresponding class label is encoded into latent space. There is also the decoder network, that reconstructs the features in the latent space to produce synthetic data, conditioned on a particular class.

During training, the input features and their corresponding class labels are both incorporated to form conditional data pair. There are two loss functions that need to be minimized: thus the Reconstruction loss, which is responsible for reconstructing the latent space, and the Kullback-Leibler (KL) divergence loss, also responsible for regularizing the latent space distribution. The CVAE model is trained in order to minimize the combined losses. After the CVAE is trained, we sample points randomly from the latent space and then condition them on the respective labels to generate new samples associated with the class labels. Four different classification algorithms are trained for cancer classification using the augmented dataset.

3.4 Evaluation metrics

The performance of the classifiers is evaluated using four evaluation metrics: Accuracy, Precision, Recall, and F1 score. The algorithms are implemented in python using libraries such as Tensorflow, Scikit-learn, and Pandas. The CVAE is trained for 100 epochs with batch size of 32.

4 Results and discussion

This section presents the results of our experiment on the cancer classification using Conditional Variational Autoencoder (CVAE) for data augmentation. Also, four different classifcation algorithms: Support Vector Machine (SVM), Decision Tree, Logistic Classifier, and XGBoost, are used to train the augmented data. The main objective of this work is to evaluate how synthetic data generated using CVAE is effective in improving the accuracy of classification models.

4.1 Original dataset distribution

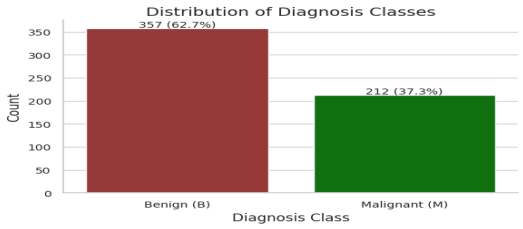

Figure 1: Original Dataset Distribution

Figure 1 provides an insight on the data distribution of the two class labels of the dataset. It shows that, the dataset consists of 62.7% Benign cells, and 37.3% Malignant cells. That is to say that, about 34% of the individuals involved in the study have tumors that is cancerous and gets worst over time.

4.2 Performance comparison of classifiers

We compare the performance of the different classification algorithms on the augmented cancer dataset using accuracy, precision, recall, and F1-score. Nevertheless, due to the imbalance nature of our dataset, we will put more emphasis on the F1 score (which is a better metrics for the case where the class labels in the dataset is not balanced)

Table 1: Performance of the Classifiers with and without data Augmentation

Classifier	Accuracy	Precision	Recall	F1-score
LogReg(with CVAE)	0.9825	0.9767	0.9767	0.9767
LogReg(without CVAE)	0.9737	0.9762	0.9535	0.9647
SVM (with CVAE)	0.9737	0.9545	0.9767	0.9655
SVM (without CVAE)	0.9561	0.9318	0.9535	0.9425
D Tree (with CVAE)	0.9737	0.9545	0.9767	0.9655
D Tree (without CVAE)	0.9474	0.9302	0.9302	0.9302
XGBoost (with CVAE)	0.9737	0.9762	0.9535	0.9647
XGBoost (without CVAE)	0.9561	0.9524	0.9302	0.9412

Tabel 1 shows the performance of each of the four classifiers on the original dataset and the augmented dataset. As shown in Table 1, the logistic classifier obtained the highest performance of 0.9825,0.9767,0.9767, and 0.9767 in Accuracy, Precision, Recall, and F1-score respectively. It is very important to note that, the classification algorithms performed significantly better when the augmented dataset is used compared to the original dataset. This means that, synthetic data obtained from CVAE appears to contribute to the improvement in the performance across all classifiers. This is to further say that, the classification models are more robust and much more capable of handling the class imbalance in the original dataset. The augmented data samples provide additional training instances for the minority class; malignant class, hence reducing the possibility of over-fitting, therefore improving the accuracy

4.3 Future works

With reference to our findings, there are several areas that needs further improvement and investigation. First of all, advance hyper-parameter tuning and a more sophisticated architectures can result in an even more better data augmentation results. Also, we may explore other techniques, such as Generative Adversarial Networks (GANs) for better diversity in the nature of synthetic data generated

5 Conclusion

In conclusion, the study provides an overview on the benefits of using Conditional Variational Autoencoder as a data augmentation technique for the cancer cell classification task. From the results above, it is evident that, the augmented data significantly improve the performance of the classification algorithms, particularly the Logistic Regression Classifier, which attained the highest accuracy score. Due to the imbalance nature of the dataset, we rely on the F1 score as an appropriate metric for measuring the performance of our model. The results highlights the capability of data augmentation in improving the accuracy of cancer cell classification tasks.
Even though, the study provides promising results, further research,and optimization procedure are a necessity to unlocking the whole capabilities of data augmentation in cancer research.

Acknowledgments and Disclosure of Funding

I would like express my sincere gratitude to all who contributed to this work. Their input contributed to a significant improvement in the structure and content of this work

First of all, I would like to express my heartfelt gratitude to the Lord Almighty, for granting the strength and wisdom I need to produce this work.

Secondly, I would like to express my heartfelt appreciation to my supervisor, Prof Guy Degla, who contributed to providing me with the foundational background from which I developed this paper.

I am also thankful to my mother, Ms Georgina Addowah, for providing me with the resources I need to make this work a success. She is my biggest motivation.

Finally I would like to thank all my friends and loved ones for their continuous effort and criticism in the progress of this work

References

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4805–4814, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67:101832, 2021.