A GAN Solution for Data imbalance in Atherosclerosis Screening

Amel LAIDI[1,*], Mohammed AMMAR[2], Mostafa EL HABIB DAHO[3]

[1]LIMOSE Laboratory M'Hamed Bougara University Boumerdes, Algeria
[2]LIST Laboratory, University M'Hamed Bougara, Boumerdes, Algeria
[3]Biomedical Engineering Laboratory, University of Tlemcen, Tlemcen, Algeria

Abstract

In our paper "GAN Data Augmentation for Improved Automated Atherosclerosis Screening from CCTA" [1], we introduced an end-to-end model for atherosclerosis detection straight from CCTA images. the model used transfer learning to overcome the issue of small training data, it achieved 95.2% accuracy, 60.8% sensitivity, and 90.48% PPV.
The model's performance was then improved by generating training data using a Generative Adversarial Network, to overcome the issue of class imbalance in the test set. The new model recorded a small drop in accuracy to 93.2%, but an improvement in overall performance with 89.0% sensitivity, and 97.13% PPV .

Keywords: Atherosclerosis, CCTA, Transfer learning, GAN, Data augmentation

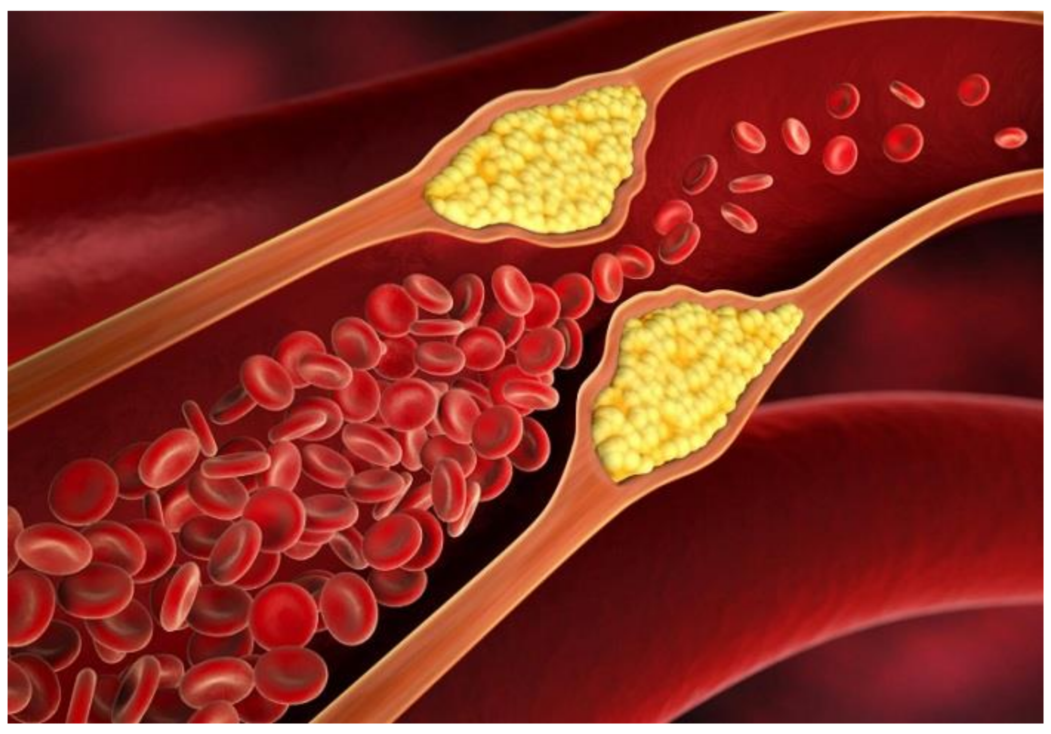


Figure 1. Fat buildup in coronary arteries

Introduction

Coronary artery disease generally, and atherosclerosis particularly, are the leading cause of world mortality [2].
Coronary artery disease is generally defined as the gradual narrowing of the lumen of the coronary arteries due to atherosclerosis; The buildup of plaque in the inner lining of the coronary arteries causes them to narrow in size and lose rigidity.
CCTA is the preferred modality for evaluation of patients with coronary artery diseases. It is a non-invasive imaging technique that permits thorough description of coronary artery plaque and grading of coronary artery stenosis [3].

Class imbalance is a rather common issue in medical datasets; solutions include: simple image transformations, increasing the dataset size via data augmentation and image synthesis [4], random oversampling, random undersampling [5], and informed undersampling .

Generative Adversarial Networks

Data Augmentation Using Generative Adversarial Networks (GAN) was first introduced in a paper on 2019 [6], The idea behind GAN is to generate images similar to the original data, enough to fool a discriminator.
training a GAN happens in two parts

- Part 1: The Discriminator is trained while the generator's parameters are fixed. the network only forward propagates. The Discriminator is trained on a positive sample x from the real data set. while the generator generates a negative sample G(z).
- Part 2: The Generator is trained while discriminator's parameters are fixed. The generated data is fed into the discriminator, The error is calculated between the output of the discriminator D (G(z)) and the sample label, and the parameters of generator are updated using the error of backpropagation algorithm.

The steps above are repeated until approaching Pdata(x) = Pg(x) . one the generator is well balanced

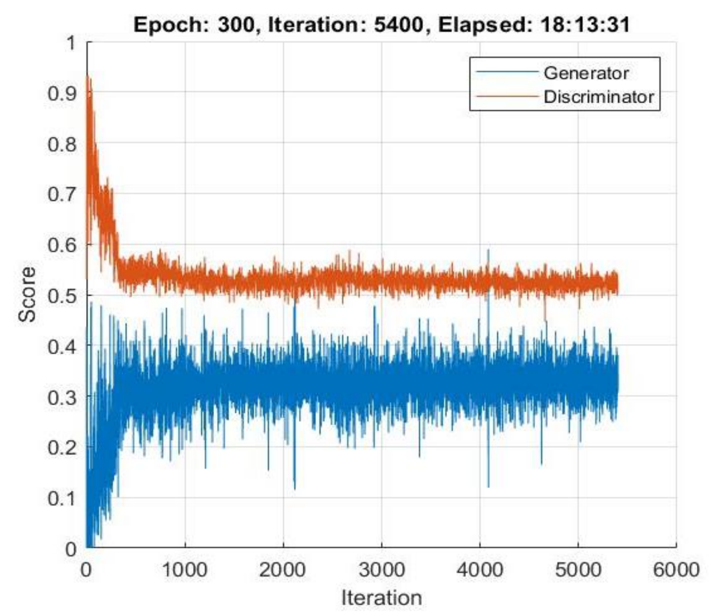


Figure 2. Progress of generator and discriminator during GAN training

GAN data augmentation for Imbalanced data

We used a pretrained Resnet on an open-source dataset [7] which achieved 95.2% accuracy, 60.8% sensitivity, 99.25% specificity and 90.48% PPV
We noticed the test dataset has an uneven distribution of images (125 positive images vs. 1066 negative Images)
We then used GAN to generate new positive images and balance the dataset. We ran several tests to find the best number of images to add to the dataset, with the right balance of improved specificity without losing too much in accuracy.
The new approach achieved 93.2% of accuracy, 89.0% of sensitivity, 97.37% of specificity ,97.13% of PPV and 89.9% of NPV.

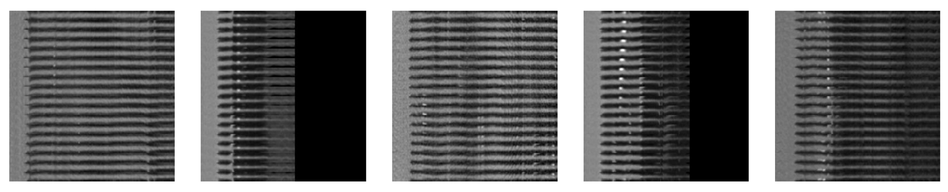
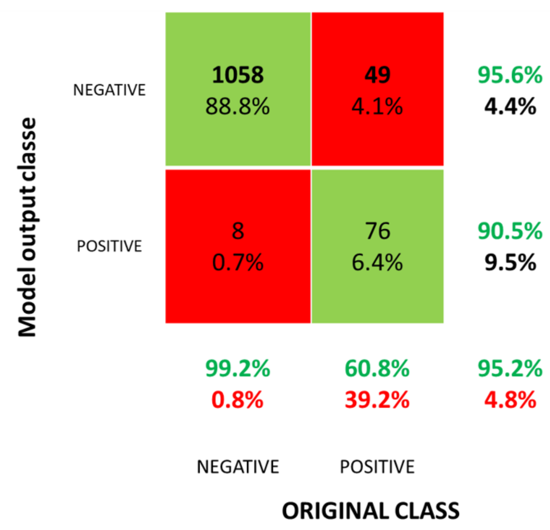


Figure 3. A sample of positive images from original dataset

Figure 4. A sample of positive images generated by the GAN

Discussion

- Intrinsic imbalance is a direct result of the nature of the dataspace, it is common in medical datasets as it reflects the prevalence of a disease in real life. It can affect the performance of deep learning algorithms.
- The performance of our model on the original dataset was comparable to Candemir et al. [8] who has achieved 90.9% accuracy, 68.9% Sensitivity, 93.6% specificity, and 58.8% PPV, and 96.1% NPV.



- When new GAN generated were added to balance the dataset, the same model achieved 93.2% of accuracy, 89.0% of sensitivity, 97.37% of specificity ,97.13% of PPV and 89.9% of NPV.
- The model got better at recognizing positive cases.

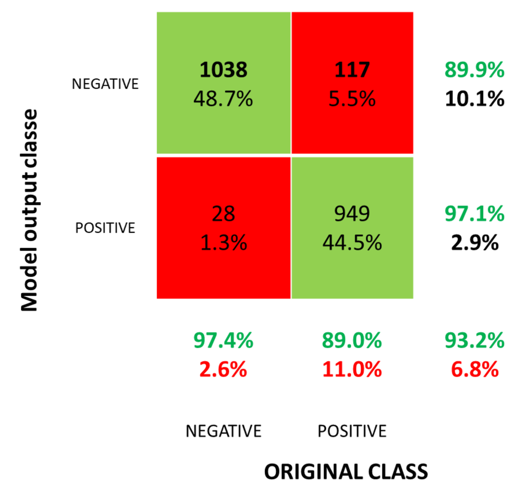


Table 1. the outcome of testing with generated images

Added images	Accuracy	sensitivity	specifity	PPV	NPV
Original	95.2	60.8	99.25	90.48	95.57
100 images	94.0	75.2	96.2	60.6	97.1
300 images	86.9	81.6	87.5	43.4	97.6
1000 images	93.2	89.0	97.4	97.1	89.9

Conclusions

GAN is a useful tool for data augmentation, and it worked well on our datatset to overcome the imbalanced data issue.

The results achieved are not only comparable to the state of the art, but also allow a fast and accurate automatic detection of atherosclerosis, which has important clinical implications.

For a more detailed description of the method and results, check the paper (or scan QR code below)

Contact

AMEL LAIDI
LIMOSE Laboratory, MHAMED BOUGARA UNIVERISTY
Email: a.laidi@univ-Boumerdes.dz
Linkedin: www.linkedin.com/in/amel-laidi-
Phone: +213662817554

References

1. A. Laidi, M. Ammar, M. E. H. Daho and S. Mahmoudi, "GAN Data Augmentation for Improved Automated Atherosclerosis Screening from Coronary CT Angiography," EAI Endorsed Transactions on Scalable Information Systems: Online First, 2022.
2. "World Health Organization," 11 June 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Accessed 23 May 2022].
3. O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester and others, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?," IEEE transactions on medical imaging, vol. 37, no. 11, pp. 2514--2525, 2018.
4. L. Engstrom, B. Tran, D. Tsipras, L. Schmidt and A. Madry, "A rotation and a translation suffice: Fooling cnns with simple transformations," arXiv, 2018.
5. A. Fernàndez, S. García, M. Galar, R. C. Prati, B. Krawczyk and F. Herrera, Learning from imbalanced data sets, vol. 10, Springer, 2018.
6. F. H. K. d. S. Tanaka and C. Aranha, "Data augmentation using GANs," arXiv preprint arXiv:1904.09135, 2019.
7. M. Demirer, V. Gupta, M. Bigelow, B. Erdal, L. Prevedello and R. White, "Image dataset for a CNN algorithm development to detect coronary atherosclerosis in coronary CT angiography," 2019.
8. S. Candemir, R. D. White, M. Demirer, V. Gupta, M. T. Bigelow, L. M. Prevedello and B. S. Erdal, "Automated coronary artery atherosclerosis detection and weakly supervised localization on coronary CT angiography with a deep 3-dimensional convolutional neural network," Computerized Medical Imaging and Graphics, vol. 83, p. 101721, 2020.