






Data augmentation to improve molecular subtype prognosis prediction in breast cancer

Francisco J. Moreno-Barea^{1*} , José M. Jerez¹ , Nuria Ribelles² , Emilio Alba² , and Leonardo Franco¹ 

¹ Departamento de Lenguajes y Ciencias de la Computación, Escuela Técnica Superior de Ingeniería Informática, Universidad de Málaga, Málaga, Spain.

² Unidad de Gestión Clínica Intercentros de Oncología, Hospitales Universitarios Regional y Virgen de la Victoria, Málaga, Spain.

*{fjmoreno}@lcc.uma.es

Abstract. Breast cancer is a major public health problem, with 2.3M new cases diagnosed each year. Immunotherapy is an effective treatment for breast cancer depending on several factors like subtype of tumours or associated prognosis. However, the immune system's efficiency depends on the local microenvironment and requires region-specific trials with a reduced number of samples. To minimise this drawback and improve the accuracy of patient prognosis predictions, we explore several data augmentation methods, i.e. noise injection, oversampling techniques and generative adversarial networks. The experiment was conducted through a set of immune system gene expression samples donated by 165 breast cancer patients from the Málaga region. Results showed a 5% increase in AUC and a 23-36% increase in F_1 score for subtype prediction.

Keywords: Data augmentation · Breast Cancer · Cancer Prognosis · Data Mining · E-Health.

1 Introduction

Despite several advances, cancer remains a serious medical problem. Cancer is currently the second leading cause of death in developed countries, after major cardiovascular diseases. Breast cancer in particular is one of the world's major health problems, with a high number of cases diagnosed each year. It is the most common tumour in women, with an estimated 2.3 million cases worldwide in 2022 (11.7% of all diagnosed cancers). Fortunately, breast cancer has a lower mortality rate due to improvements in early detection and less aggressive therapies.

In recent decades, immunotherapy has emerged as a potent and less aggressive strategy for treating cancer, and countless studies have clearly provided a boost to understanding the effects of the immune system on tumour development and establishment [9]. One way to monitor the immune response of patients is to analyse the molecular variables encoded by the major histocompatibility complex (MHC). The MHC is a system of interrelated genes whose main role is to control the expression of cell surface molecules acting as markers of the immune

response. This paper examines a cohort of breast cancer patients from the region of Málaga, Spain. The dataset represents the MHC gene expression of patients associated with the molecular subtype of breast cancer [15] considered as the labelled class in a supervised learning analysis.

The disadvantage of applying machine learning (ML) methods to these regional bioinformatics datasets is the scarcity of data. The application of data augmentation (DA) methods, which allow the addition of synthetically generated samples, has become a relevant topic. DA has proven to be very effective for improving the performance of ML models, and recent results from generative artificial intelligence are demonstrating the potential of these models, even in the medicine and healthcare field [7,17]. But applying DA techniques to non-structured datasets is far more complex. DA techniques available to deal with this type of dataset include noise injection techniques [13], oversampling techniques [3], and recently the Generative Adversarial Networks (GAN) [5]. GAN models have shown an impressive level of success in generating synthetic samples, and recently they have shown good results as a DA method for datasets without spatial or temporal structure [11], also in the biomedical domain [8,4,12,6].

Considering all the above aspects, the main objective of this work is to apply state-of-the-art DA methods to a small MHC gene expression data set expecting an increase in the prediction performance of the prognosis associated with the molecular subtype of breast cancer patients.

2 Related Works

Works on the application of DA to bioinformatics problems have mainly focused on the treatment of medical images and tasks involving time series. However, DA with -omic data is recent and challenging, but recently works shows that DA methods can be beneficial to improve prediction performance. Among the DA studies using classical methods, Beinecke and Heider [2] applied Gaussian noise, SMOTE and ADASYN methods to clinical data from the UCI ML repository covering different medical fields. Related to the application of deep learning-based models to unstructured data, the work of Marouf et al. [8] used a GAN for realistic generation of single cell RNA-Seq data and detection of marker genes. García-Ordás et al. [4] built a variational autoencoder to predict diabetes in pima indians. Barile et al. [1] employed a Generative Adversarial Autoencoder for the generation of synthetic structural brain network with sclerosis.

To the best of our knowledge, this paper is the first work that proposes the application of DA to improve the predictive performance of prognosis associated with molecular subtypes of breast cancer using genomic data. Only a few papers applied DA to genomic cancer samples. Moreno-Barea et al. [12] used DA to improve the prediction of fixed-time events in cancer given 18 different cancer types from The TCGA database; Wei et al. [16] developed a GAN-based model to expand 12 cancer datasets, improving the accuracy of cancer diagnosis; and Gutta et al. [6] used a GAN model and image processing to improve survival prediction in breast cancer patients.

3 Breast cancer cohort

According to studies of gene expression patterns associated with prognosis or metastatic risk [15], breast cancer can be divided into two main groups based on estrogen receptor positivity. A first group are low-grade neoplasms or estrogen receptor-positive tumours, also called luminal subtypes; and a second group are high-grade neoplasms or estrogen receptor-negative tumours, called non-luminal subtypes. The luminal subtypes are associated with low/medium grade and good/intermediate prognosis, whereas the non-luminal are associated with higher grade and poorer prognosis.

The cohort used to test the prognosis associated with molecular subtype of cancer was provided by the Carlos Haya Regional and the Virgen de la Victoria University Hospitals in Málaga, Spain. The data are part of a clinical study of a group of 165 breast cancer patients diagnosed between 2008 and 2013. The data set consists of patient gene expression (molecular variables encoded by the MHC) and cancer subtype. The division between the two groups according to estrogen receptor positivity was taken into account as classes. The most common molecular subtype in the dataset was luminal with 135 samples (82%), whereas 30 samples were non-luminal (18%).

4 Data augmentation methods

A simple but effective method to start testing DA is noise addition, technique based on a random modification of the original instances. To apply it, we did a random selection of samples and modified a maximum of 25% of the features [13]. Eq. 1 mathematically describes the process of obtaining a new feature value \tilde{x} from the original one x , where \min_V and \max_V are the actual limits of each feature. An oversampling noise addition method (“Noise bal”) is also applied, differing from the previous method in that it performs the random selection only on the training samples belonging to the minority class.

$$\tilde{x} = \min(\max_V, \max(\min_V, x + \text{RND}(-0.2, 0.2))) \quad (1)$$

SMOTE is a classic technique specifically designed for unbalanced data sets [3]. To generate synthetic minority class data and balance the classes, SMOTE uses a k-nearest neighbour algorithm on the minority class instead of random sampling with replacement. SMOTE performs an interpolation of the selected instance and its nearest neighbours, creating new instances for the minority class that are located in the region between the sample and its neighbours.

Currently, the generation of images has shown impressive success through the application of GAN models [5]. GANs attempt to learn the distribution of the original dataset in order to generate new samples from the learned distribution. The standard GAN model has a structure divided into two networks (Generator and Discriminator) trained simultaneously, so that they can learn from each other. In this context, the goal of the discriminator (D) is to distinguish whether a sample comes from the set of real data or is a generated sample. On the other

hand, the generator (G) produces as output a distribution assigned to the space of real samples with the purpose of presenting similar features.

Variations in network architecture, loss function or the inclusion of additional information have been proposed from the basic GAN model. Specifically, since a supervised task is performed in the study, the models considered are the Conditional GAN (CGAN) [10], the Auxiliary Classifier GAN (ACGAN) [14] and the Generative-Classifier GAN (ModCGAN) [11]. The CGAN model is a simple variant of the vanilla GAN model in which the information about the label of samples y is taken into account in D and G networks. The CGAN cost function (Eq. 2) contains two parts identified with the two networks involved in the competitive process. One related to the detection of samples which are in the real distribution, and the other involved in detecting those samples generated by G .

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}(x)} [\log D(x|y)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z|y)))] \quad (2)$$

The ACGAN model was also applied [14]. Like CGAN, ACGAN takes both the latent space and the class label as input to G . The main difference is that D receives only the sample as input, without the class label. The discriminator process still predicts whether the given sample is real or false, but also the class label of the sample. The architecture of D includes a single neural network model with two outputs: one output to distinguish the origin of the sample (real/false) and another to obtain the probability that the sample belongs to each class.

Finally, ModCGAN model was also considered [11]. This method is similar to ACGAN, but instead of integrating the classifier within the discriminator, uses externally a so-called generative classifier (GC). This GC is used to label the synthetic samples produced by the generator and discard them if they are of insufficient quality. A ModCGAN with ‘Balanced Multiclass’ (_BM) was also considered due to the unbalanced nature of the problem. This GAN-based modification uses two independent models, where each model is trained with an unbalanced set corresponding to each class. The purpose is to allow each generator to focus on one of the classes of the problem, always taking into account its differences from samples from the other class.

5 Experiments and Results

In the experimental procedure followed, a Principal Component Analysis (PCA) was first performed on the original dataset to obtain the component score vectors and transform the dataset from binary to continuous variables. After the PCA transformation, a stratified cross-validation procedure was performed. Due to the small number of non-luminal samples in the dataset, a 60% split was used for training and 40% for testing. The data generation process used the training set to generate the desired number of synthetic samples, except for SMOTE. The main generative models were tested with different levels of DA percentages. The synthetic samples generated were added to the training data. Prediction was performed on the test set to obtain the evaluation metrics, which were averaged on a cross-validation scheme that was repeated 10 times with different seeds.

Table 1. Test results obtained for the binary problem with the different DA methods using a RF and a SVM system as classifiers.

Method	Perc.	Accuracy	Sensitivity	Specificity	F ₁ score	AUC
RF						
Original	None	.8121 ± .002	.3083 ± .012	.9241 ± .004	.3652 ± .009	.7643 ± .005
CGAN	750	.8341 ± .003	.4133 ± .019	.9276 ± .004	.4624 ± .015	.7930 ± .005
ACGAN	500	.8335 ± .003	.3917 ± .014	.9317 ± .004	.4465 ± .012	.7851 ± .006
ModCGAN	750	.8364 ± .002	.4117 ± .016	.9307 ± .003	.4672 ± .011	.7857 ± .005
ModCGAN_BM	750	.8370 ± .003	.4500 ± .016	.9230 ± .004	.4903 ± .012	.8015 ± .006
NOISE	50	.8302 ± .003	.3250 ± .014	.9424 ± .003	.3978 ± .014	.7674 ± .005
NOISE Bal	1000	.8433 ± .002	.6500 ± .014	.8863 ± .003	.5988 ± .007	.8107 ± .005
SMOTE	None	.8380 ± .002	.5825 ± .016	.8948 ± .003	.5597 ± .009	.8040 ± .004
SVM						
Original	None	.8177 ± .002	.2358 ± .021	.9470 ± .004	.2723 ± .019	.7920 ± .005
CGAN	500	.8423 ± .002	.5092 ± .019	.9163 ± .004	.5291 ± .012	.8225 ± .005
ACGAN	500	.8452 ± .002	.4767 ± .017	.9270 ± .003	.5125 ± .012	.8301 ± .004
ModCGAN	200	.8427 ± .003	.4917 ± .012	.9207 ± .005	.5204 ± .009	.8296 ± .004
ModCGAN_BM	750	.8488 ± .002	.5383 ± .013	.9180 ± .003	.5535 ± .010	.8358 ± .005
NOISE	1000	.8188 ± .003	.2483 ± .013	.9456 ± .003	.3187 ± .014	.7592 ± .006
NOISE Bal	750	.8490 ± .002	.7262 ± .015	.8763 ± .003	.6360 ± .008	.8405 ± .006
SMOTE	None	.8321 ± .003	.6600 ± .015	.8704 ± .005	.5877 ± .008	.7978 ± .007

Table 1 shows the test results obtained for the different methods and DA models applied, using Random Forest (RF) and Support Vector Machines (SVM) as classifier models. In the table, “Original” indicates that no DA method is applied, serving as a reference level. The second column (‘Perc.’) refers to the DA percentage applied, noting that SMOTE only produces balanced classes. The remaining columns show the values (\pm ‘between-validation performance’ SE) obtained for each of the test metrics: accuracy, sensitivity, specificity, F₁ score and Area Under Curve ROC (AUC). The good/interm. prognosis subtype (majority) is considered the negative class to measure sensitivity and specificity.

The results show an improvement in prediction performance for most evaluation metrics with the Noise Bal method for both classifier models, compared to the performance obtained with the non-augmented data set. Specifically, for the RF system, augmentation using the Noise Bal strategy with 1000% DA achieves an improvement of 3.0% in accuracy, 4.6% in AUC and 23.4% in F₁ score with respect to the reference levels with the original set with no DA. On the other hand, using the Noise Bal method with 750% and the SVM classifier, similar values are obtained. With regard to the GAN-based deep generative models, the evaluation metrics also indicate an improvement in the performance of the classifiers compared to the reference set. Among them, the ModCGAN model with the sample balancing modification stands out, obtaining the highest values for the different metrics, except specificity. Accuracy and AUC values obtained with ModCGAN_BM are similar to those obtained with Noise Bal, although the latter obtains a better value for sensitivity, resulting in higher values for the rest of the metrics, especially F₁ score.

The influence of the size of the generated dataset on the prediction results can be seen in Fig. 1. This presents values for accuracy, specificity and sensitivity obtained with CGAN, ModCGAN_BM, Noise and Noise Bal, against the number of instances on a logarithmic scale on the abscissa axis. The CGAN and

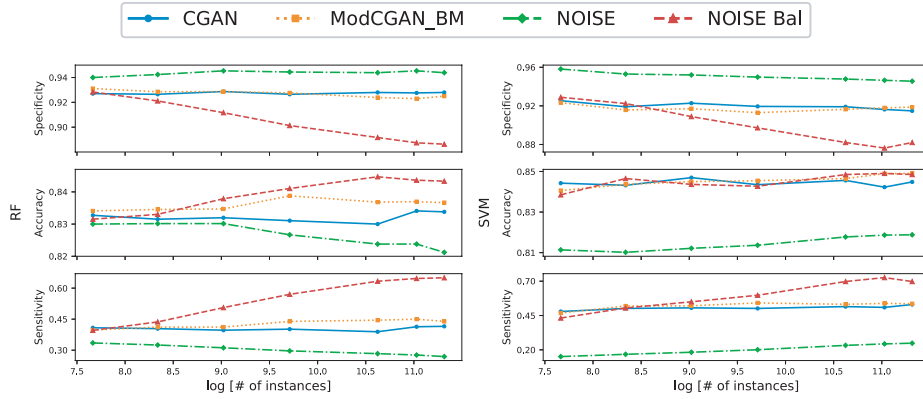


Fig. 1. Comparison of the accuracy, specificity and sensitivity obtained using different DA methods, in terms of the logarithm of the number of instances generated. Code lines: blue solid, CGAN; orange dots, ModCGAN with balancing; green dashes and dots, noise; red dashes, Noise Bal.

ModCGAN_BM results show stability in the values as the number of generated samples increases, with a slightly gain in sensitivity. On the other hand, the Noise Bal strategy shows a significant negative correlation for the gain in specificity and a positive correlation for sensitivity and accuracy. This fact explains the higher performance achieved by the Noise Bal technique in F_1 score.

After analysing the performance obtained by applying DA, it is necessary to analyse the quality of the synthetic generated data. A PCA was carried out to visualise the configuration of the samples in a two-dimensional space (PC1 vs. PC2). This enables comparison of the distribution of synthetic samples and original samples, analysis of the performance of DA methods, and explanation of the obtained classification performance. Figure 2 shows the distribution of synthetic samples created using the Noise Bal method and ModCGAN (with BM modification) model. These analyses reflect the inner workings of each DA method and model in sample generation. The Noise Bal method generates samples only for the minority (non-luminal) class, so it can be seen how the samples generated are around the original samples modified with noise, although since these are generated from component scores transformed data, the noise in this case generates more variability, i.e. samples further away.

As representative of the deep generative models, ModCGAN_BM has also generated majority class samples. It is important to note that the synthetic samples are adapted to the real distribution of the samples. When there are samples that can be considered as outliers, the Noise Bal method generates samples around them. However, the deep generative models, due to their internal discriminative operation, consider these outliers as data generated by the generator, which prevents the model from generating samples close to, or simply influenced by, these samples.

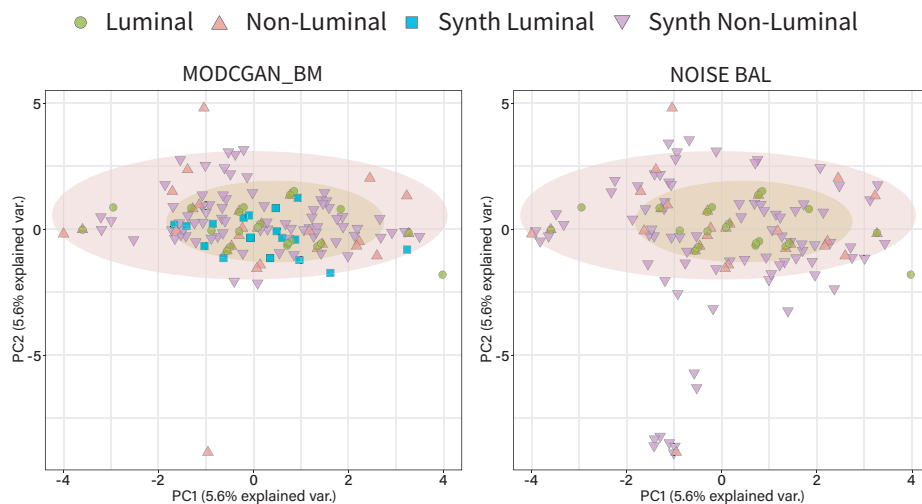


Fig. 2. PCA plot with the original samples and the samples generated by the Noise Bal and ModCGAN_BM methods. Colour codes: green circle, luminal; red triangle, non-luminal; blue square, synthetic luminal; purple triangle, synthetic non-luminal.

6 Conclusion and Future Work

This paper presents the application of DA techniques as a way to improve the prognosis prediction for breast cancer associated with molecular subtype using a dataset of MHC gene expression profiles. A binary problem is addressed by performing a classification according to whether the subtype is associated with a good/intermediate prognosis (luminal) or a poor prognosis (non-luminal). The overall results obtained suggest and confirm that DA is quite effective when small and unbalanced data are used, leading to a high increase in predictive performance. Among the DA methods applied in the study, the Noise Bal method showed the best performance, leading to an increase in accuracy of 3.0%, with particular attention to the increase in sensitivity, resulting in an improvement of 4.5 – 5% in AUC and 23 – 36% in F_1 score. This resulted in an F_1 score of 64%, compared to 30% reference level for the original data, representing a major advance in the ability to predict early treatments in patients who may have a poor prognosis and require more aggressive therapies.

In future works, we would like to extend the present approach to a multiclass problem (prediction of each subtype) and we aim to utilise generative AI and Transformer-based GAN models to enhance the quality of generated data.

Acknowledgements. The authors acknowledge the support from the the Ministerio de Ciencia e Innovación (MICINN) under project PID2020-116898RB-I00, from the Universidad de Málaga through grant UMA20-FEDERJA-045, and from the Fundación General UMA and Pfizer S.L. (UMA-FGUMA-Pfizer).

References

1. Barile, B., Marzullo, A., Stamile, C., Durand-Dubief, F., Sappey-Marinier, D.: Data augmentation using generative adversarial neural networks on brain structural connectivity in multiple sclerosis. *Computer Methods and Programs in Biomedicine* **206**, 106113 (2021)
2. Beinecke, J., Heider, D.: Gaussian noise up-sampling is better suited than smote and adasyn for clinical decision making. *BioData Mining* **14**(1), 1–11 (2021)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16**, 321–357 (2002)
4. García-Ordás, M.T., Benavides, C., Benítez-Andrades, J.A., Alaiz-Moretón, H., García-Rodríguez, I.: Diabetes detection using deep learning techniques with oversampling and feature augmentation. *Computer Methods and Programs in Biomedicine* **202**, 105968 (2021)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: *Advances in Neural Information Processing Systems*. pp. 2672–2680 (2014)
6. Guttà, C., Morhard, C., Rehm, M.: Applying a gan-based classifier to improve transcriptome-based prognostication in breast cancer. *PLOS Computational Biology* **19**(4), e1011035 (2023)
7. He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., et al.: Transformers in medical image analysis. *Intelligent Medicine* **3**(1), 59–78 (2023)
8. Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D.S., et al.: Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nature communications* **11**(1), 1–12 (2020)
9. Martin, J.D., Cabral, H., Stylianopoulos, T., Jain, R.K.: Improving cancer immunotherapy using nanomedicines: progress, opportunities and challenges. *Nature Reviews Clinical Oncology* **17**(4), 251–266 (2020)
10. Mirza, M., Osindero, S.: Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014)
11. Moreno-Barea, F.J., Jerez, J.M., Franco, L.: Improving classification accuracy using data augmentation on small data sets. *Expert Systems with Applications* **161**, 113696 (2020)
12. Moreno-Barea, F.J., Jerez, J.M., Franco, L.: Gan-based data augmentation for prediction improvement using gene expression data in cancer. In: *International Conference on Computational Science*. pp. 28–42. Springer (2022)
13. Moreno-Barea, F.J., Strazzera, F., Jerez, J.M., Urda, D., Franco, L.: Forward Noise Adjustment Scheme for Data Augmentation. In: *IEEE Symposium Series on Computational Intelligence*. pp. 728–734 (2018)
14. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: *International conference on machine learning*. pp. 2642–2651 (2017)
15. Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al.: Molecular portraits of human breast tumours. *Nature* **406**(6797), 747–752 (2000)
16. Wei, K., Li, T., Huang, F., Chen, J., He, Z.: Cancer classification with data augmentation based on generative adversarial networks. *Frontiers of Computer Science* **16**, 1–11 (2022)
17. Zhang, P., Kamel Boulos, M.N.: Generative ai in medicine and healthcare: Promises, opportunities and challenges. *Future Internet* **15**(9), 286 (2023)