Data augmentation techniques to improve metabolomic analysis in Niemann-Pick type C disease

Francisco J. Moreno-Barea[1]*[0000−0002−3887−9095], Leonardo Franco[1][0000−0003−0012−5914], David Elizondo[2][0000−0002−7398−5870], and Martin Grootveld[3][0000−0003−1502−3734]

[1] Escuela Técnica Superior de Ingeniería Informática
Universidad de Málaga, Málaga, Spain
{fjmoreno, lfranco}@lcc.uma.es
[2] School of Computer Science and Informatics, Faculty of Technology,
De Montfort University, The Gateway, Leicester, UK
[3] Leicester School of Pharmacy, Faculty of Health and Life Sciences,
De Montfort University, The Gateway, Leicester, UK
{elizondo, mgrootveld}@dmu.ac.uk

Abstract. Niemann-Pick Class 1 (NPC1) disease is a rare and neurodegenerative disease, and often metabolomics datasets of NPC1 patients are limited in the number of samples and severely imbalanced. In order to improve the predictive capability and identify new biomarkers in an NPC1 disease urinary dataset, data augmentation (DA) techniques based on computational intelligence are employed to create additional synthetic samples. This paper presents DA techniques, based on the addition of noise, on oversampling techniques and using conditional generative adversarial networks, to evaluate their predictive capacities on a set of Nuclear Magnetic Resonance (NMR) profiles of urine samples. Prediction results obtained show increases in sensitivity (30%) and in F_1 score (20%). In addition, multivariate data analysis and variable importance in projection scores have been applied. These analyses show the ability of the DA methods to replicate the information of the metabolites and determined that selected metabolites (such as 3-aminoisobutyrate, 3-hidroxivaleric, quinolinate and trimethylamine) may be valuable biomarkers for the diagnosis of NPC1 disease.

Keywords: Metabolomics · Data augmentation · Niemann-Pick type C · Bioinformatics · Lysosomal storage disease.

1 Introduction

Niemann–Pick type C disease (NPC, OMIM 257220) is a very rare neurodegenerative lysosomal storage disease caused by mutations in two genes NPC1 and NPC2 [22]. NPC affect approximately 1:100000 live births, although the NPC1 mutations account for 95% of cases observed. NPC involves the altered lysosomal

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_8

storage of sphingosine, and leads to a loss of lysosomal calcium ions, a process accompanied by the accumulation of unesterified cholesterol and glycosphingolipids [24, 19], along with decreased acidic store calcium levels [11]. Usually, NPC disease presents in childhood with clumsiness, ataxia, learning difficulties, vertical gaze paralysis, and dysphagia, together with cataplexy, epilepsy, and hepatosplenomegaly. Additionally, adult-onset illness may occur, and this may be associated with a neuropsychiatric presentation [22]. NPC disease also involves neuroinflammation, neuronal apoptosis, and oxidative stress [4].

For the diagnosis and prognostic monitoring of such diseases, metabolomics strategies are valuable because bioanalytical dataset systems can be analysed under pre-established conditions determined by the experimental design. The non-invasive nature of metabolomics and the close link of this type of data with the phenotype, make it an ideal tool for pharmaceutical and preventative health. Metabolomics is also applicable to the discovery of biomarkers, small molecules known as metabolites, as a support for decision making. Selected metabolites and their concentrations can be used to determine the status of different groups of samples based on their detection in control group samples, or in those collected from patients with a specified disease. Urine samples such as those analysed in this work contain informative metabolites that can be easily analysed for the purpose of discovering new biomarkers.

Notwithstanding, currently there is a clear lack of global, untargeted metabolomics studies focused on investigations of lysosomal storage diseases, with only a small number of studies being reported [21, 20, 18]. These studies justify the value offered by NMR-based metabolomics data analysis techniques, and the use of composites of both bioanalytical techniques and computational intelligence techniques is therefore further evolving and becoming more popular [13]. However, metabolomics datasets are often limited in the number of samples and heavily imbalanced. In this case, the lysosomal storage disorders are genetically-distinct and metabolically-related, rare inherited diseases. Because of this, the prior collection and the parental ethical consent required are often highly challenging hurdles to surmount. Additionally, obtaining a sufficient number of biofluid samples for NMR or other analyses adds to this complexity.

In this work, computational intelligence based Data Augmentation (DA) methods are used to generate more observations. DA has proven to be an effective technique to improve the performance of machine learning models, especially for applications related to problems involving datasets consisting of images [9], also in biomedical applications [6, 23, 14]. The application of DA techniques to datasets that are not images, signals or time series, is more complex. Experts find it easier to evaluate a generated image, being able to measure its quality and distinguish whether it is a 'synthetic' or a 'real' image. However, this type of evaluation conducted by human experts is not feasible when applications includes genomic or clinically-relevant metabolic data. DA techniques such as noise injection techniques [26, 17] or the application of SMOTE techniques (synthetic minority oversampling technique) [2] are available to handle this type of dataset. A more recent technique known as Generative Adversarial Networks (GANs) has been proposed to be suitable for the analysis of these types of datasets [8]. GAN models have shown an impressive level of success in generating realistic images, and recently, it has been shown that they can also be applied as a DA method for datasets without any type of spatial or temporal structure [5, 16], also in some biomedical applications [10, 12, 7]. To the best of the authors knowledge, there are no recent DA studies that show its application on metabolomics analysis.

Considering all the above aspects, the main objectives of this work are: (1) to apply different state-of-the-art DA methods to a small size metabolomics dataset aimed at obtaining an increase in the prediction performance of urine samples belonging to NPC1 disease patients, in order to demonstrate their usefulness in this research domain; (2) to analyse the ability of these DA methods to replicate the information of the metabolites using conventional forms of multivariate data analysis, such as partial least squares - discriminatory analysis (PLS-DA).

2 Materials

This study presents a UK-based clinical cohort consisting of 13 untreated NPC1 patients and 47 corresponding parental heterozygous carriers. The selection process for the NPC1 patient cohort was carefully conducted to select only patients not receiving any therapeutic agents. This process avoids any complications arising from the presence of urinary ^{1}H NMR resonances attributable to such drugs and their metabolites in the urinary metabolite profiles explored. The data for this study was collected with informed consent and previously approved by the appropriate Research Ethics Committee (06/MRE02/85). Urine samples were collected, thawed and centrifuged to remove any cells and debris. The sample mixtures were then transferred to NMR tubes for in-depth analysis.

Single-pulse ^{1}H NMR analysis of human urine samples were obtained using a Bruker Avance AV-600 spectrometer (Queen Mary University of London facility, London, UK) operating at a frequency of 600.13 MHz, as described in [21]. The intense H_2O/HOD signal ($\delta = 4.80$ ppm) was suppressed via gated decoupling during the delay between pulses. Chemical shift values were internally referenced to the methyl group resonances of acetate (s, $\delta = 1.920$ ppm), alanine (d, $\delta = 1.487$ ppm), creatinine (>NCH3 s, $\delta = 3.030$ ppm) and lactate (d, $\delta = 1.330$ ppm). Through a complete consideration of chemical shift values, coupling patterns and coupling constants, the identities of metabolite resonances present in spectra acquired were routinely assigned. These assignments were cross-checked with the *Human Metabolome Database (HMDB)* [25] and confirmed by one- (1D) and two-dimensional (2D) correlation (COSY) and total correlation (TOCSY) spectroscopic techniques.

The urinary dataset matrix consists of 60 spectra × 33 ^{1}H NMR-assigned metabolite predictor variables. This dataset was generated using macro procedures for line broadening, zero filling, Fourier-transformation and phase and baseline corrections, together with the subsequent application of a separate macro for the “intelligently-selected bucketing” (ISB) processing sub-routine. All procedures were performed using the ACD/Labs Spectrus Processor 2012

DOI: 10.1007/978-3-031-08757-8_8

software package (ACD/Labs, Toronto, Ontario, Canada M5C 1T4). This ISB strategy ensured that all bucket edges featured did not coincide with ^{1}H NMR resonance maxima, and hence this approach avoided the splitting of signals across separate integral regions. Prior to data augmentation experiments, all sample ^{1}H NMR profiles were autoscaled column-(metabolite variable)-wise.

3 Data Augmentation Methods

Addition of Noise The first of the methods used in this work for data augmentation is a simple and straightforward one, that can be easily applied, and has the ability to lead to competent results. Specifically, the method randomly selects samples and modifies a maximum of a 25% of the features present in the data. The process of generating a new feature value $\tilde{x}$ from the original value x is mathematically described in Eq. 1. The noise value obtained from a random normal distribution (denoted "RND") with a standard deviation/variance of 1.0, is added to the original value for the chosen feature. The resulting "noisy" value is controlled so that it does not exceed the real limit values established for its feature (MIN_Value and MAX_Value). A standard deviation of 1.0 at the random normal distribution is sufficient to generate a sample that is not too far from the actual sample.

$$\tilde{x} = \min(\mathrm{MAX\_Value}, \max(\mathrm{MIN\_Value}, x + \mathrm{RND}(-1.0, 1.0))) \tag{1}$$

A variation of the addition of noise method described above has been designed for balancing purposes. The method abbreviated as "Noise Bal", differs from the standard method in that it applies the random selection only to samples belonging to the minority class. Therefore, it only modifies and generates synthetic samples that belong to the minority class in an oversampling process. The rest of the method follows the same noise addition process described before.

SMOTE Technique In clinical cohorts of rare diseases, it is easier to have more control samples available than samples from patients that present the disease. Therefore, medical datasets, as well as metabolomics datasets, are often imbalanced. The traditional oversampling method to reverse this situation by applying DA is SMOTE (Synthetic Minority Oversampling Technique) [2]. SMOTE uses a k-nearest neighbour algorithm on the minority class, rather than random sampling with replacement. In this way, the algorithm performs an interpolation between each sample x and its selected neighbours. The interpolation computes the difference between the sample x and each of the neighbours in the feature space, multiplies the difference of each feature by a random normalisation between 0 and 1, and adds this value to the feature of original sample x. This interpolation results in the synthetic samples generated by SMOTE being located within the space between the selected neighbours and the sample x. One disadvantage of the SMOTE algorithm application is the lack of control over the number of samples to generate. This technique is ineffective on well-balanced

DOI: 10.1007/978-3-031-08757-8_8

datasets, since oversampling aims to create a fully balance augmented dataset. Another disadvantage, derived from the interpolation process, is the creation of synthetic samples that do not follow the distribution of the original dataset.

Conditional GAN The DA application of deep learning models known as Generative Adversarial Networks (GAN) [8] has shown an impressive success in the generation of realistic images. Specifically, the model considered in this work is the Conditional GAN (CGAN) [15], since a supervised task is performed. The CGAN model is a variant of the vanilla GAN model in which the information contained in the sample label y is taken into account. The generation of synthetic samples using GAN models occurs by learning the distribution of the original dataset. With this aim, GAN models have a structure divided into two neural networks trained simultaneously, the *generator* and the *discriminator*, yielding a confrontation between both so that they are able to learn from each other. In this manner, the objective of the discriminator network (D) is to estimate the probability of the sample arises from the real distribution or is a generated sample. However, the purpose of the generator network (G), which takes as input a noisy random distribution z and the condition y, is to produce a distribution $G(z)$ (synthetic sample) with features that approximate those present in the real samples. Therefore, the generator intends that the discriminator cannot distinguish these synthetic samples from the real ones.

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

The two behaviours described above within the competitive process can be distinguished in the CGAN objective cost function (Eq. 2). One part is related to achieving a better recognition of those samples that belong to the real distribution, while the other is related to achieving a better recognition of those synthetic samples created by the generator network. Thereby, the discriminator network is updated based on the error associated with the ability to perceive whether the samples are real or false, expressed in Eq. 3; and the generator network is updated from the error identified for false sample recognition, modelled by Eq. 4.

$$\max_D \; \mathbb{E}_{x\sim p_{data}(x)}[\log D(x|y)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z|y)))] \tag{3}$$

$$\min_G \mathbb{E}_{z\sim p_z(z)}[\log(D(G(z|y)))] \tag{4}$$

4 Experiments and Results

The experimentation process followed is shown in Fig. 1. A stratified division of the original dataset into training and test sets is performed, allowing to maintain total independence between the synthetic data generation process and the evaluation of the experiments. Due to the reduced number of samples present in the

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_8

benchmark dataset, a split of 60% of samples for training and 40% for testing is conducted. Depending on the DA method, the synthetic data generation process uses the training set to create the desired number of samples, following the procedures described in the previous section. Before the augmentation process, a principal component analysis (PCA) is performed on the training dataset. This process removes any high level of correlation (multicollinearity) between the variables of the metabolomics dataset. In this way, the score vectors are obtained and the training, test and synthetic datasets are transformed. Samples values are represented by their principal components instead of the original values from metabolomic variables.

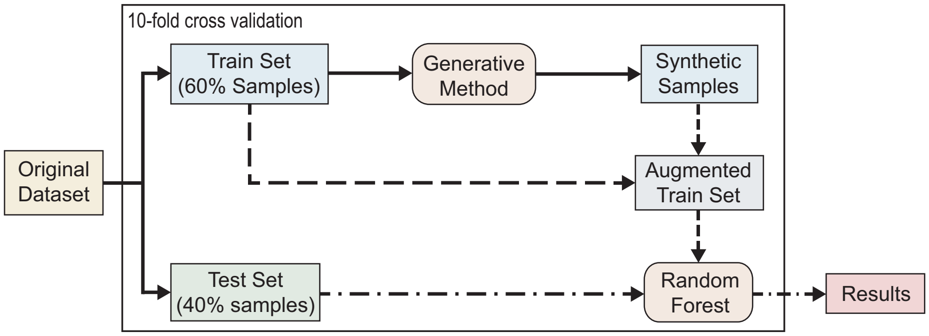

Fig. 1. Flow diagram of the whole experimentation process.

After this transformation process, the augmented training set is formed by adding synthetic generated data to the training one. A Random Forest system [1] is used for classifying the samples. Random Forest are essentially an ensemble of decision trees and establishes the outcome based on the individual tree predictions. The python scikit-learn package is used, with 1001 trees, bootstrap samples to build each tree, Gini impurity for tree splitting and 6 predictors selected at each node. The classifier model employs the augmented set to perform the training and the prediction is measured using the test set data. The entire process of data division, generation, training and evaluation is carried out through a cross-validation procedure, employed for obtaining a better estimate of the metrics, in order to avoid small dataset sampling biases.

4.1 Classification Performance

Test results obtained from the application of the different DA methods are shown in Table 1. To show whether the application of DA improves the classification performance, the results are compared with those obtained with the original non-augmented dataset, indicated in the table as 'None'. In order to be more exhaustive in the study, the results obtained with the combination of samples from two different DA models are included. 'Comb 1' refers to the results obtained with a combination between the CGAN model and the SMOTE method, and

DOI: 10.1007/978-3-031-08757-8_8

'Comb 2' refers to the results obtained with a combination between the CGAN model and the Noise Bal strategy, when a 500% augmentation level is applied with Noise Bal.

Table 1. Test results acquired with a random forest system using each DA method and the percentage of augmentation applied.

Aug. Model	Percent	Accuracy	Specificity	Sensitivity	F_1 score
None	None	85.42 ± *1.0*	97.37 ± *0.6*	40.10 ± *5.6*	53.33 ± *2.4*
CGAN	500	85.83 ± *1.1*	90.53 ± *1.3*	67.99 ± *4.5*	63.47 ± *3.9*
NOISE	1000	86.25 ± *1.1*	**97.89** ± *0.4*	42.02 ± *5.0*	49.27 ± *5.6*
SMOTE	100	87.92 ± *1.0*	94.74 ± *0.4*	61.93 ± *4.8*	64.49 ± *4.1*
NOISE Bal	500	**89.17** ± *1.0*	94.21 ± *0.9*	70.12 ± *3.7*	**71.79** ± *2.8*
NOISE Bal	2000	80.83 ± *1.6*	82.11 ± *2.5*	**76.25** ± *2.9*	63.87 ± *1.9*
Comb 1	100	85.63 ± *1.0*	90.26 ± *1.2*	68.01 ± *4.0*	64.67 ± *3.0*
Comb 2	100	84.79 ± *1.4*	87.63 ± *1.4*	73.76 ± *3.9*	66.68 ± *3.1*

The 'Percent' column of Table 1 indicates the amount of synthetic data generated compared to the original set. Thus, if the number of generated samples is the same as the training set, a percentage of 100 is reported; and if the number of training samples is multiplied by 10, a percentage of 1000 is reported. The remaining columns show the values (± 'between-validation performance' SE) obtained for each of the test metrics. The test metrics showed are the accuracy, specificity, sensitivity and F_1 score obtained. The F_1 score is the harmonic mean of the precision and sensitivity (Eq. 5), and allows a reliable measure of the prediction performance achieved in problems where sensitivity is more important.

$$\mathrm{F}_1 \text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{sensitivity}}{\text{precision} + \text{sensitivity}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

Results in Table 1 show that an improvement in test prediction accuracy, sensitivity and F_1 is achieved with almost all the DA methods compared to the values obtained with the non-augmented dataset ('None'). Using the Noise Bal method with 1000% DA, the highest accuracy (89.17%) and F_1 values (71.78%) are obtained. The Noise Bal approach with 2000% reaches the highest sensitivity values (78.8%), but with a lower accuracy (80.83%) and F_1 score (63.87%). These values show a substantial improvement compared to analysis of the dataset without augmentation.

Additional analyses were performed reviewing the impact of the number of samples generated with different DA methods on the test results obtained. Three test metrics (accuracy, specificity and sensitivity) obtained with three different DA methods (CGAN, Noise and Noise Bal) *versus* the number of instances on a logarithmic scale, are presented in Fig. 2. The results obtained with the CGAN model indicate a negative correlation between the number of instances created and the specificity and accuracy gain of the prediction. However, a positive correlation for the sensitivity gain was also observed. For the Noise Bal method, there

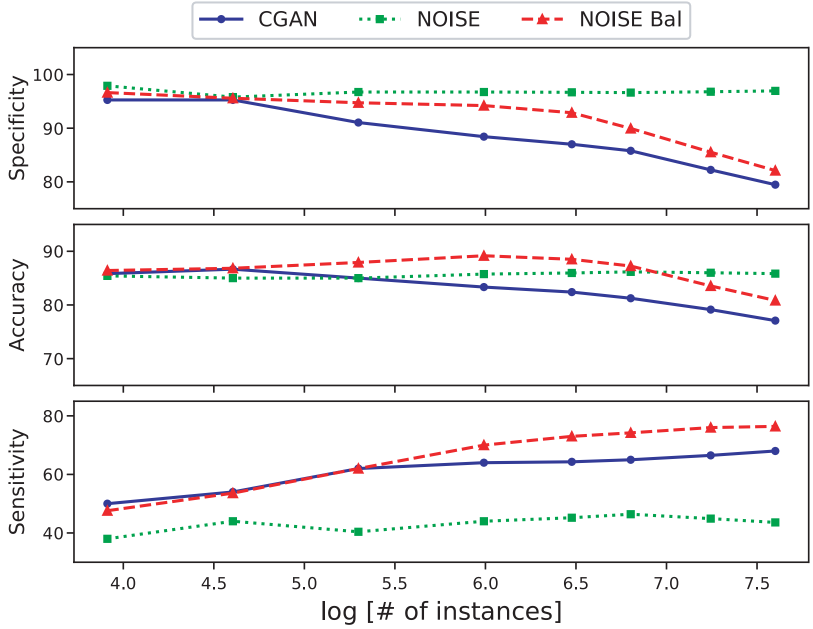

Fig. 2. Comparisons of the specificity, accuracy and sensitivity obtained with different DA methods *versus* the logarithm of the number of instances generated. Dots (solid) represent the results obtained with the CGAN model, squares (dotted) those with the NOISE method, and triangles (dashed) those with the NOISE Bal method.

is clearly a significant positive correlation between sensitivity and the number of instances created with this strategy. The specificity obtained decreases slightly until the abscissa axis reaches a value of 6.5, when it presents a significant negative correlation. The influence of specificity on the accuracy gain is noticeable, since they decreases at the same time when a large number of samples are generated. Finally, the values obtained for the metrics are approximately stable with respect to the number of instances generated with the standard Noise method.

4.2 Augmented Datasets Analysis

An important objective is to analyse how the DA methods were able to replicate the metabolomic information present in the dataset. The configuration of the augmented samples can be visualised in a two-dimensional space (component 2 *vs* component 1) through a partial least squares - discriminatory analysis (PLS-DA), using *MetaboAnalyst v4.0* software (University of Alberta and National Research Council, National Nanotechnology Institute (NINT), Edmonton, AB, Canada) [3]. This provided a means to check the information contained in the augmented dataset and compare this with the information in the original samples, analyzing how the distribution and clustering is affected.

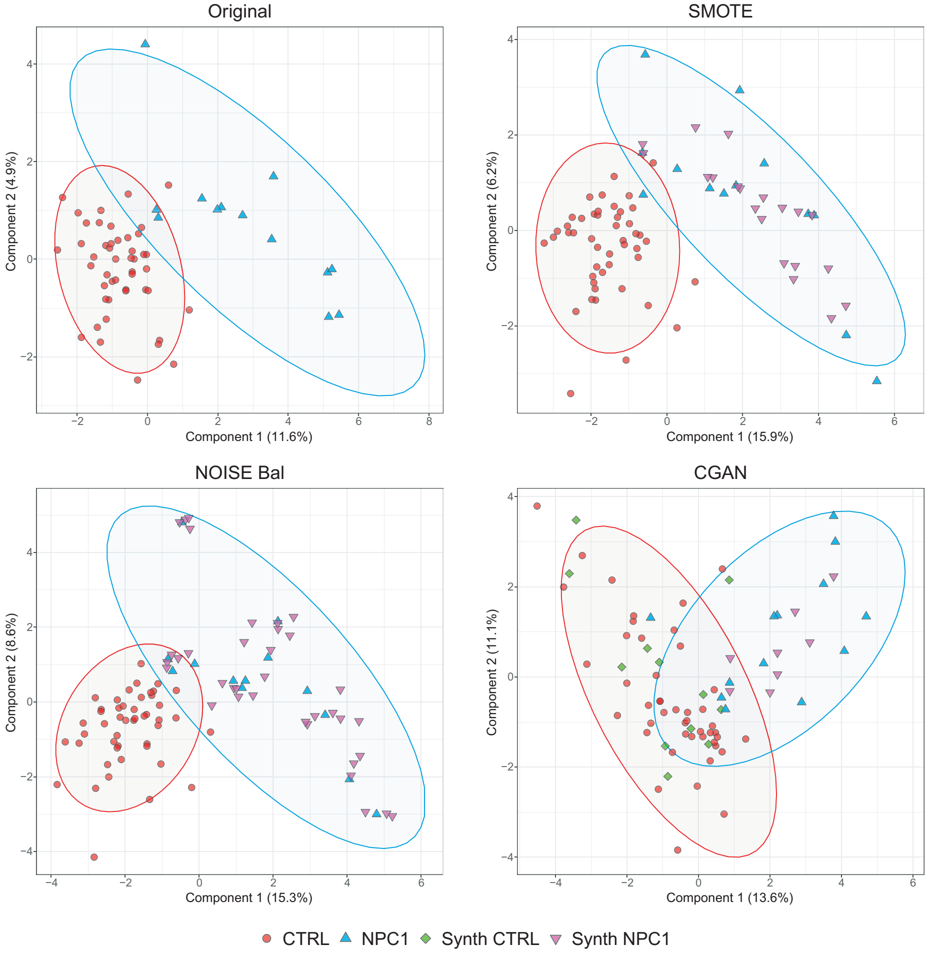

Fig. 3. PLS-DA component 2 *versus* component 1 scores plot for the original dataset, and the augmented datasets with SMOTE, the Noise Bal method and CGAN model. Colour codes: red circles, heterozygous carrier control urine; green diamond, synthetic control; blue triangles, NPC1 disease urine; purple inverse triangles, synthetic NPC1.

Figure 3 (top left) shows the PLS-DA results obtained by using the original NPC1 dataset. This reveals two significant groups for the samples that correspond to the possible classes of "disease state", with an area where both clusters converge. The cluster belonging to the control group appeared as a compact cluster, while the cluster conformed by the NPC1 disease samples was more dispersed. The results of PLS-DA after adding the samples generated with SMOTE are shown in Fig. 3 (top right). Here, it can be clearly seen how the creation of samples through SMOTE works. The synthetic samples are distributed throughout the 'real' NPC1 disease cluster, from the interpolation process. In this case, the small convergence zone between clusters avoids the SMOTE method disad-

DOI: 10.1007/978-3-031-08757-8_8

vantages. Figure 3 (bottom left) shows the distribution of the augmented dataset created by using the Noise Bal method with respect to the original samples.The distribution of the samples is similar to the distribution observed with the original dataset, and with the augmented dataset produced by the SMOTE. Through the noise injection process, the synthetic samples belonging to the NPC1 disease class are found grouped around the original samples that they modify.

The PLS-DA scores plot when using the augmented dataset with the CGAN model is shown in Figure 3 (bottom right). Contrary to previous DA methods, CGAN is not an oversampling technique, so the model creates samples belonging to both classes. The generated samples modify the dispersion and angle presented by the component analysis, causing the control group less compact. Although it is still possible to differentiate both groups, with a larger convergence zone. The generated synthetic samples fit satisfactorily the distribution of the 'real' samples for both groups, with some of them generated in the convergence zone.

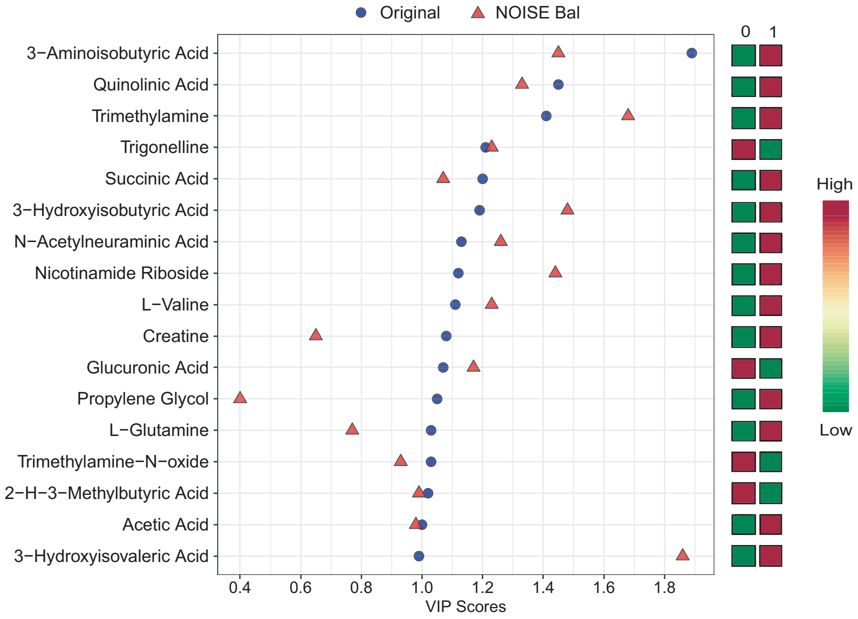

Fig. 4. Variable importance parameter (VIP) scores obtained from the PLS-DA applied to the original dataset and the application to the Noise Bal augmented dataset. Colour codes: blue circles, original VIP; red triangles, Noise Bal augmentation VIP.

The variable importance in projection (VIP) scores with respect to component 1 were also obtained using PLS-DA. The VIP scores allow to analyse the differences between the heterozygous carrier group and NPC1 disease urine samples, measuring the importance of each metabolite in the differentiation process. Figure 4 shows the VIP scores for the top 17 metabolites obtained from the original dataset and a comparison with the VIP scores obtained from the augmented dataset with balanced addition of noise. The coloured boxes indicate the relative concentrations of the corresponding metabolite in each "disease state" group on the original dataset. Considering that the values >1.00 are significant, the most outstanding metabolite of the analysis in the original set was 3-aminoisobutyric acid with a VIP score equal to 1.89. With the Noise Bal augmentation, the most prominent metabolites were trimethylamine with a VIP score of 1.68 and 3-hydroxyisovaleric acid with a VIP score of 1.86. These values are higher than those obtained for the original dataset. Conversely, the 3-aminoisobutyric acid obtained a VIP score of 1.45 and propylene glycol obtained a VIP score of 0.4 with the Noise Bal dataset, values lower than the original ones. Notwithstanding, the analysis reveals a total of nine metabolites that obtain similar VIP values.

Table 2. PLS-DA VIP scores obtained using augmented datasets with the SMOTE, Noise Bal method and CGAN model for the top 17 metabolites with original dataset.

Metabolite	Original	SMOTE	Noise Bal	CGAN
3-Aminoisobutyric Acid	1.89	1.51	1.45	1.22
Quinolinic Acid	1.45	1.60	1.33	1.05
Trimethylamine	1.41	1.81	1.68	0.79
Trigonelline	1.21	1.12	1.23	1.79
Succinic Acid	1.20	1.13	1.07	0.70
3-Hydroxyisobutyric Acid	1.19	1.56	1.48	0.93
N-Acetylneuraminic Acid	1.13	0.93	1.26	0.93
Nicotinamide Riboside	1.12	0.68	1.44	1.51
L-Valine	1.11	1.38	1.23	0.66
Creatine	1.08	0.78	0.65	1.66
Glucuronic Acid	1.07	1.10	1.17	1.38
Propylene Glycol	1.05	0.43	0.40	1.14
L-Glutamine	1.03	0.96	0.77	0.70
Trimethylamine-N-oxide	1.03	1.05	0.93	1.49
2-H-3-Methylbutyric Acid	1.02	1.02	0.99	1.13
Acetic Acid	1.00	1.05	0.98	0.77
3-Hydroxyisovaleric Acid	0.99	1.69	1.86	0.58

The results obtained for each of the top 17 marker metabolites shown in Figure 4 are summarised in Table 2. The analysis reveals certain metabolites showing analogous VIP scores for the SMOTE and Noise Bal approaches. Amongst these, the following metabolites should be highlighted: trimethylamine, 3-hydroxyisovaleric acid, 3-hydroxyisobutyric acid, 3-aminoisobutyric acid, and trigonelline. Both methods (SMOTE and Noise Bal) are oversampling ones, thus increasing the number of samples for the NPC1 disease class. This significantly influenced the analysis, which indicates a greater relevance of these metabolites to separate this group from the heterozygous carriers. Regarding the analysis using the augmented set with CGAN, the results show fewer similarities. The most differentiating metabolites with respect to the original dataset are trimethylamine with a VIP score of 0.79, compared to the original value equal to 1.41; and creatine with a VIP score of 1.66, and a value of 1.08 with the original dataset.

5 Conclusions

The different state-of-the-art techniques for Data Augmentation (DA) employed in this work clearly offer much potential regarding the analysis of metabolomics datasets, as these predominantly comprise small numbers of sample-donating participants, as it is the case of the NPC1 data examined here.

The results shown in Table 1 indicate a great improvement of test prediction, with an increase in predictive accuracy. This renders the balanced addition of noise (Noise Bal) the best DA method for this purpose. The augmented dataset reaches approximately a 4% improvement in accuracy compared to the analysis performed on the original dataset. Since the dataset is quite imbalanced, predictive accuracy is not the most representative metric, as it is more important that the largest number of patients with the disease be diagnosed as such. Therefore, most representative prediction metrics for this type of imbalanced problem are sensitivity and F_1 score. Table 1 shows that when performing data augmentation with the Noise Bal method and 500% DA, an approximate 30% improvement in sensitivity and a 20% improvement in F_1 score can be obtained.

In order to determine the ability of DA methods to replicate metabolic information, a PLS-DA was performed. The SMOTE and Noise Bal method show a good capacity to replicate the information of the metabolites from samples representing NPC1 disease. The results obtained from the analysis of the CGAN augmentation show the ability of this model to replicate information that fits the distribution of the 'real' samples. However, because CGAN can generate samples for both classes in the convergence zone of the clusters, the PLS-DA results differ from the original one. Finally, the VIP scores results obtained revealed a series of biomarkers which may be valuable for distinguishing between the urinary ^{1}H NMR profiles of NPC1 patients and their heterozygous healthy controls. These included the branched-chain amino acid valine, 3-aminoisobutyrate, 3-hidroxivaleric, quinolinate and trimethylamine. The selected metabolites and their relative importance rankings were found to be similar to those reported in a previously conducted study of the dataset analysed, and without any form of DA strategies [21].

In conclusion, DA techniques constitute a suitable approach to increase the prediction performance of Niemann-Pick Class C1 (NPC1) disease activity in patients when analysing ^{1}H NMR urinary metabolic datasets. DA techniques are capable of generating good quality synthetic data that lead to an increase in sensitivity of 30%, allowing the identification of urinary metabolomics biomarkers which will serve on the diagnosis and monitoring of the severity of patients with NPC1 disease. Future research directions will focus on testing different machine learning algorithms analysing their robustness in the prediction of rare diseases.

Acknowledgements. The authors acknowledge the support from MICINN (Spain) through grant TIN2017-88728-C2-1-R and PID2020-116898RB-I00, from Universidad de Málaga y Junta de Andalucía through grant UMA20-FEDERJA-045, and from IBIMA (all including FEDER funds).

References

1. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001). https://doi.org/10.1023/a:1010933404324
2. 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). https://doi.org/10.1613/jair.953
3. Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D.S., Xia, J.: MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic acids research **46**(W1), W486–W494 (2018). https://doi.org/10.1093/nar/gky310
4. Cougnoux, A., Cluzeau, C., Mitra, S., Li, R., Williams, I., Burkert, K., Xu, X., Wassif, C., Zheng, W., Porter, F.: Necroptosis in Niemann–Pick disease, type C1: a potential therapeutic target. Cell death & disease **7**(3), e2147–e2147 (2016). https://doi.org/10.1038/cddis.2016.16
5. Douzas, G., Bacao, F.: Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Systems with Applications **91**, 464–471 (1 2018). https://doi.org/10.1016/j.eswa.2017.09.030
6. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing **321**, 321–331 (12 2018). https://doi.org/10.1016/j.neucom.2018.09.013
7. 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). https://doi.org/10.1016/j.cmpb.2021.105968
8. 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. vol. 3, pp. 2672–2680 (2014). https://doi.org/10.1145/3422622
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/cvpr.2016.90
10. Liu, Y., Zhou, Y., Liu, X., Dong, F., Wang, C., Wang, Z.: Wasserstein GAN-Based Small-Sample Augmentation for New-Generation Artificial Intelligence: A Case Study of Cancer-Staging Data in Biology. Engineering **5**(1), 156–163 (2 2019). https://doi.org/10.1016/j.eng.2018.11.018
11. Lloyd-Evans, E., Morgan, A.J., He, X., Smith, D.A., Elliot-Smith, E., Sillence, D.J., Churchill, G.C., Schuchman, E.H., Galione, A., Platt, F.M.: Niemann-Pick disease type C1 is a sphingosine storage disease that causes deregulation of lysosomal calcium. Nature medicine **14**(11), 1247 (2008). https://doi.org/10.1038/nm.1876
12. Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D.S., Krebs, C.F., Bonn, S.: Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. Nature communications **11**(1), 1–12 (2020). https://doi.org/10.1038/s41467-019-14018-z
13. Marshall, D.D., Powers, R.: Beyond the paradigm: Combining mass spectrometry and nuclear magnetic resonance for metabolomics. Progress in nuclear magnetic resonance spectroscopy **100**, 1–16 (2017). https://doi.org/10.1016/j.pnmrs.2017.01.001

14. Marzullo, A., Moccia, S., Catellani, M., Calimeri, F., De Momi, E.: Towards realistic laparoscopic image generation using image-domain translation. Computer Methods and Programs in Biomedicine **200**, 105834 (2021). https://doi.org/10.1016/j.cmpb.2020.105834
15. Mirza, M., Osindero, S.: Conditional Generative Adversarial Nets. CoRR **abs/1411.1784** (11 2014), https://arxiv.org/abs/1411.1784
16. 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). https://doi.org/10.1016/j.eswa.2020.113696
17. 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 (IEEE SSCI 2018). pp. 728–734 (2018). https://doi.org/10.1109/ssci.2018.8628917
18. Percival, B.C., Latour, Y.L., Tifft, C.J., Grootveld, M.: Rapid Identification of New Biomarkers for the Classification of GM1 Type 2 Gangliosidosis Using an Unbiased 1H NMR-Linked Metabolomics Strategy. Cells **10**(3), 572 (2021). https://doi.org/10.3390/cells10030572
19. Platt, F.M., d'Azzo, A., Davidson, B.L., Neufeld, E.F., Tifft, C.J.: Lysosomal storage diseases. Nature reviews Disease primers **4**(1), 1–25 (2018). https://doi.org/10.1038/s41572-018-0025-4
20. Probert, F., Ruiz-Rodado, V., Te Vruchte, D., Nicoli, E.R., Claridge, T.D., Wassif, C.A., Farhat, N., Porter, F.D., Platt, F.M., Grootveld, M.: NMR analysis reveals significant differences in the plasma metabolic profiles of Niemann Pick C1 patients, heterozygous carriers, and healthy controls. Scientific reports **7**(1), 1–12 (2017). https://doi.org/10.1038/s41598-017-06264-2
21. Ruiz-Rodado, V., Marcos Luque-Baena, R., te Vruchte, D., Probert, F., H Lachmann, R., J Hendriksz, C., E Wraith, J., Imrie, J., Elizondo, D., Sillence, D., et al.: 1H NMR-linked urinary metabolic profiling of Niemann-Pick Class C1 (NPC1) disease: Identification of potential new biomarkers using correlated component regression (CCR) and genetic algorithm (GA) analysis strategies. Current Metabolomics **2**(2), 88–121 (2014). https://doi.org/10.2174/2213235X02666141112215616
22. Vanier, M.T.: Niemann-Pick disease type C. Orphanet journal of rare diseases **5**(1), 1–18 (2010). https://doi.org/10.1186/1750-1172-5-16
23. Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., Pinheiro, P.R.: Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection. Ieee Access **8**, 91916–91923 (2020). https://doi.org/10.1109/access.2020.2994762
24. Winkler, M.B., Kidmose, R.T., Szomek, M., Thaysen, K., Rawson, S., Muench, S.P., Wüstner, D., Pedersen, B.P.: Structural insight into eukaryotic sterol transport through Niemann-Pick type C proteins. Cell **179**(2), 485–497 (2019). https://doi.org/10.1016/j.cell.2019.08.038
25. Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., et al.: Hmdb 4.0: the human metabolome database for 2018. Nucleic acids research **46**(D1), D608–D617 (2018). https://doi.org/10.1093/nar/gkx1089
26. Zur, R.M., Jiang, Y., Pesce, L., Drukker, K.: Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. Medical physics **36**(10), 4810–4818 (2009). https://doi.org/10.1118/1.3213517

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_8