Computers in Biology and Medicine 148 (2022) 105916

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Application of data augmentation techniques towards metabolomics

Francisco J. Moreno-Barea [a,*], Leonardo Franco [a], David Elizondo [b], Martin Grootveld [c]

[a] Departamento de Lenguajes y Ciencias de la Computación, Escuela Técnica Superior de Ingeniería Informática, Universidad de Málaga, No. 35, Bulevar Louis Pasteur, Málaga, 29071, Spain

[b] School of Computer Science and Informatics, Faculty of Technology, De Montfort University, The Gateway, Leicester, LE1 9BH, United Kingdom

[c] Leicester School of Pharmacy, Faculty of Health and Life Sciences, De Montfort University, The Gateway, Leicester, LE1 9BH, United Kingdom

ARTICLE INFO

Keywords:
Data augmentation
Machine learning
Metabolomics
Niemann–Pick type C disease
Rare diseases

ABSTRACT

Niemann–Pick Class 1 (NPC1) disease is a rare and debilitating neurodegenerative lysosomal storage disease (LSD). Metabolomics datasets of NPC1 patients available to perform this type of analysis are often limited in the number of samples and severely unbalanced. 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 have been employed to create synthetic samples, i.e. the addition of noise, oversampling techniques and conditional generative adversarial networks. These techniques have been used to evaluate their predictive capacities on a set of urine samples donated by 13 untreated NPC1 disease and 47 heterozygous (parental) carrier control participants. Results on the prediction have also been obtained using different machine learning classification models and the partial least squares techniques. These results provide strong evidence for the ability of DA techniques to generate good quality synthetic data. Results acquired show increases in sensitivity of 20%–50%, an F_1 score of 6%–30%, and a predictive capacity of 0.3 (out of 1). Additionally, more conventional forms of multivariate data analysis have been employed. These have allowed the detection of unusual urinary metabolite profiles, and the identification of biomarkers through the use of synthetically augmented datasets. Results indicate that urinary branched-chain amino acids such as valine, 3-aminoisobutyrate and quinolinate, may be employable as valuable biomarkers for the diagnosis and prognostic monitoring of NPC1 disease.

1. Introduction

1.1. Disease and metabolomics

Niemann–Pick type C disease (NPC, OMIM 257220) is a very rare neurodegenerative lysosomal storage disease (LSD) caused by mutations in two genes NPC1 (95% of clinical cases) and NPC2 [1]. NPC is estimated to occur in 1 in 100,000–120,000 live births [2]. The true frequency of this disorder in the general population cannot be exactly determined because many cases go undiagnosed or misdiagnosed. NPC involves the altered lysosomal storage of sphingosine, and leads to a loss of lysosomal calcium ions, a process accompanied by the accumulation of unesterified cholesterol and glycosphingolipids [3,4], along with decreased acidic store calcium levels [5]. The NPC1 gene encodes a large transmembrane protein that resides in the limiting membrane of the lysosome, and is known as NPC1 protein [3]. NPC2 encodes a small soluble lysosomal protein that binds cholesterol within the lysosomallumen and is believed to transfer cholesterol to the NPC1 protein. Full details on how these proteins may complement each other to allow cholesterol to pass through the glycocalyx and the lysosome membrane are available in [3].

Usually, NPC disease presents in childhood with clumsiness, ataxia, learning difficulties, vertical gaze paralysis, and dysphagia, together with cataplexy, epilepsy, and hepatosplenomegaly. Respiratory dysfunction may be another clinical feature. Additionally, adult-onset illness may occur, and this may be associated with a neuropsychiatric presentation [1]. NPC disease also involves neuroinflammation, neuronal apoptosis, and oxidative stress within its pathological cascade [6].

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, and at the point of response after exposure to specific stimuli, treatments or exercise regimens. The non-invasive nature of metabolomics and the close link of this type of data with

* Corresponding author.
E-mail addresses: fjmoreno@lcc.uma.es (F.J. Moreno-Barea), lfranco@lcc.uma.es (L. Franco), elizondo@dmu.ac.uk (D. Elizondo), mgrootveld@dmu.ac.uk (M. Grootveld).

https://doi.org/10.1016/j.compbiomed.2022.105916
Received 26 April 2022; Received in revised form 11 July 2022; Accepted 23 July 2022
Available online 27 July 2022

the phenotype, make it an ideal tool for pharmaceutical and preventative health, amongst others. Metabolomics allow researchers to develop a deep understanding of how the system explored responds to such stimuli. In clinical studies, this has potential applications for monitoring patient diagnosis and prognosis, therapeutic management, and even drug development for conditions investigated, the latter approach involving the identification of abnormal activities in selected metabolomics pathways, so that any defective enzymes may be drug-targeted.

Metabolomics is also applicable to the discovery of biomarkers as a support for decision making. In metabolomics, biomarkers are small molecules, known as metabolites. As an example, 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. High-resolution proton (^{1}H) nuclear magnetic resonance (NMR) spectroscopy is an established analytical tool which has been extensively used for the purpose of probing the metabolic status of biofluids [7]. ^{1}H NMR spectroscopy is useful for metabolic profiling studies in view of its multicomponent analytical capacity and now very high sensitivity. The technology allows for accurate high-throughput simultaneous screening of more than 100 metabolites present in a biological sample. ^{1}H NMR profiles of urine samples such as those analysed in this work contain informative metabolites that can be easily analysed for the purpose of discovering new biomarkers.

Metabolomics is typically performed using specialised and highly reproducible multi-component analytical techniques. However, metabolomic datasets are often limited in the number of samples and heavily unbalanced. In this study, computational intelligence based data augmentation techniques are used to create more observations. Experimentation is then proposed for the purpose of detecting unusual metabolic patterns in patients with NPC1 disease via an analysis of the ^{1}H NMR profiles of their urine samples, using both the original and the augmented datasets. In order to further support these investigations, and also for comparative purposes, we further utilised more conventional forms of multivariate data analysis. These included both (unsupervised) principal component analysis (PCA) and (supervised) partial least squares-discriminatory analysis (PLS-DA).

To date, the majority of metabolomics research conducted on lysosomal storage diseases has been focused on NPC1 disease. Indeed, high-resolution ^{1}H NMR analysis of urine samples has successfully sought and identified potential biomarkers for NPC1 disease diagnosis. These incorporated branched-chain amino acids (BCAAs), N-acetylsugars, selected bile acids and 3-aminoisobutyrate [8], the latter metabolite originating from metabolic perturbations to either BCAA catabolism or thymine degradation pathways. These analyses further indicated that, along with the lysosome itself, the brain and liver were key site features involved, and this observation was fully congruous with NPC1 disease characteristics, including seizures and hepatomegaly features. AUROC values determined ranged from 0.81–0.91 for the most significant biomolecules, a PLS-DA strategy applied gave a significant Q^2 value of 0.56, with an accuracy of 0.93 [8].

1.2. Data augmentation

Data augmentation (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 [10–14]. The application of DA techniques to datasets that are not images, signals or time series, is, however, 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 related to other domains are involved. An example of this includes genomic or clinically-relevant metabolic data.

Other DA techniques are available to handle this type of dataset. These include: noise injection techniques [15,16] or the application of SMOTE techniques (synthetic minority oversampling technique) [17]. SMOTE is designed to deal with datasets containing unbalanced classifications. A more recent technique known as Generative Adversarial Networks (GANs) has been proposed to be suitable for the analysis of these types of datasets [18,19]. The main objective of this technique is to learn the distribution of original data and, based on this, to generate new samples. The GAN method produces a confrontation between two neural networks, a network known as a "generator" and another one known as the "discriminator". The generator network generates synthetic data and subsequently attempts to deceive the discriminator. The discriminator network, in turn, then attempts to discern whether data received from the generator network is true or false. GAN models have shown an impressive level of success in generating realistic images. Furthermore, 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 [20,21], also in some biomedical applications [22–25],

Considering all the above aspects, the main objectives of this work were: (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 the usefulness of these methods with small and non-spatial structured datasets in this research domain. Current research strategies, however, attempt to analyse the ability of these DA methods to replicate the information of the metabolites. This is performed by the use of conventional forms of multivariate data analysis, such as PCA (unsupervised) and PLS-DA (supervised) approaches for the purpose of detecting unusual metabolic patterns in samples with NPC1 disease for the purpose of distinguishing these profiles from those of the heterozygous carrier controls group.

The paper is structured as follows: Section 2 shows some work related to DA in bioinformatics. Section 3 introduces the Materials and Methods employed in this study. This includes the collection and processing of the clinical urine samples, the DA methods examined, and the implementations made to conduct the experiments. Section 4 presents the results divided into three subsections: the classification of results acquired, an analysis of the augmented dataset, and the effect of DA on the information related to metabolites. We finalise the investigation with a Discussion in Section 5, and relevant Conclusions in Section 6.

2. Related works

Most of the recent works that apply Data Augmentation (DA) in bioinformatics tasks are focused on the treatment of medical images with Generative Adversarial Networks (GANs). Frid-Adar et al. [26] applied GAN models to improve liver lesion classification. Han et al. [27] designed a GAN to generate MR images for brain tumour detection. Waheed et al. [11] generated chest X-ray (CXR) images for COVID-19 detection. There are extensive reviews of medical imaging DA techniques and deep learning [28].

Nevertheless, there is a clear idea that DA application in domains where the samples are not images or time series is a challenging task. The works on biomedical problems with omics data are scarce and recent, but they show that DA methods can be beneficial to increase the prediction performance. Among the DA studies with classical methods, Açıcı et al. [29] applied ADASYN for the generation of pseudo proteins and prediction of T4SS effector proteins. Beinecke and Heider [30] applied Gaussian noise, SMOTE and ADASYN methods to clinical datasets from the UCI ML Repository covering different medical fields. As with medical imaging, most recent works try to apply Deep Learning based models with unstructured data. Liu et al. [22] proposed a GAN model for the DA application with serum samples in cancer staging. Marouf et al. [23] used a GAN for the realistic generation of single-cell RNA-Seq data and the detection of marker genes. García-Ordás et al. [24] built

a Variational Autoencoder for the prediction of pima indians diabetes. Barile et al. [25] employed a Generative Adversarial Autoencoder for the generation of synthetic structural brain network with Multiple Sclerosis.

To the best of our knowledge, the paper is the first work that proposes the application of DA to generate new synthetic information to improve prediction performance with non-structured metabolomics datasets. Only a couple of works were found that apply DA without samples generation (sampling) for missing values in metabolomics studies [31], or a rescaling DA with 2D photoacoustic metabolomics signals to assess breast tumour progression [32].

3. Materials and methods

In this section, the materials, methods and techniques used for the application of Data Augmentation (DA) to a metabolomics dataset are featured. This Methodology section is organised as follows: it commences by describing the ^{1}H NMR measurements present in the studied problem and the raw ^{1}H NMR analytical data preprocessing stages employed. The noise addition methods implemented and the SMOTE technique are described below. The CGAN model is described in another sub-section, and the process that the model follows for the generation of synthetic samples is then explained. Subsequently, a description is made of the general process performed in order to perform the experiments and the application of DA to create synthetic samples that, together with the original ones, will be used for the formation of a classifier; finally a description of the technical implementation details, including the architecture and the parameters used in the models developed is presented.

3.1. Materials

3.1.1. ^{1}H NMR measurements

This study features a UK-based clinical cohort consisting of 13 untreated NPC1 patients (age range 2.7–30 years, 6 male/7 female) and 47 corresponding parental heterozygous carriers (age range 25–62 years, 19 male/28 female). For the NPC1 patient cohort, only patients not receiving any therapeutic agents were carefully selected in order to avoid 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 donated by these participants were stored at −80 °C. When ready for analysis, these were thawed and centrifuged to remove any cells and debris (5000 rpm for a period of 10 min), and 0.60 ml volumes of the supernatants derived therefrom were then thoroughly vortex mixed with 0.07 ml of deuterium oxide (^2H_2O). These sample mixtures were then transferred to 5 mm diameter NMR tubes.

A Bruker Avance AV-600 spectrometer (Queen Mary University of London facility, London, UK) operating at a frequency of 600.13 MHz in quadrature detection mode and a probe temperature of 298 K was employed for the single-pulse ^{1}H NMR analysis of human urine samples, as described in [8]. The intense H_2O/HOD signal (δ = 4.80 ppm) was eliminated through gated decoupling during the delay between pulses. Chemical shift values were internally referenced to the methyl group resonances of acetate (s, δ = 1.920 ppm), alanine (d, δ = 1.487 ppm), lactate (d, δ = 1.330 ppm) and creatinine (>NCH3 s, δ = 3.030 ppm). Metabolite resonances present in spectra acquired were routinely assigned by a full consideration of chemical shift values, coupling patterns and coupling constants, and also from literature sources; these were then cross-checked with the *Human Metabolome Database (HMDB)* [33]. Amalgamations of one- (1D) and two-dimensional (2D) correlation (COSY) and total correlation (TOCSY) spectroscopic techniques were used to confirm these urinary assignments, along with the "spiking" of specimens with small μl volumes of approximately 5.00 mmol/L standard solutions of authentic biomolecules prepared in 0.10 M phosphate buffer (pH 7.00).

3.1.2. Raw ^{1}H NMR analytical data preprocessing stages

For the dataset, 33 variables were extracted from the 199 potential ^{1}H NMR metabolite predictor variables obtained previously. This feature extraction is performed following the conclusions obtained by Ruiz-Rodado et al. [8], in which an extensive multivariate analysis was performed to identify biomarkers in the ^{1}H NMR profiles. The urinary dataset matrix therefore consists of 60 spectra × 33 ^{1}H NMR-assigned metabolite predictor variables was generated through the employment of 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 "intelligent bucketing" processing sub-routine. All manipulations were performed using the ACD/Labs Spectrus Processor 2012 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 circumvented the splitting of signals across separate integral regions. A 0.04-ppm bucket width with a 50% looseness factor was selected for this purpose.

Spectral ISBs containing signals ascribable to ethanol (δ = 1.22–1.24 (t) and 3.63–3.67 ppm (q)), which were detectable in spectra acquired on a small number of the heterozygous carrier control urine specimens, and urea (broad, δ = 5.59–5.99 ppm) were removed from all the urinary ^{1}H NMR spectra acquired, as was that of the residual H_2O/HOD signal (δ = 4.65–5.16 ppm) via secondary irradiation described above. Prior to analysis, all sample ^{1}H NMR profiles were autoscaled column-(metabolite variable)-wise.

Outlier samples were removed from the original urinary ^{1}H NMR profile datasets following the inspection of both two- and three-dimensional PC3 *vs*. PC2 *vs*. PC1 plots. These included heterozygous carrier control classification samples in which paracetamol's glucuronide and sulphate metabolites were detected by ^{1}H NMR analysis (i.e. quite prominent aromatic doublet signals centred at δ = 7.13 and 7.35 ppm (glucuronide), and δ = 7.31 and 7.45 ppm (sulphate), along with their acetamido-NH-CO-CH_3 function resonances (δ = 2.15–2.17 ppm).

3.2. Methods

3.2.1. Addition of noise

To perform DA there are some methods with approaches which may be considered simple, such as resampling, shifting, flipping, clipping, or adding noise. In this study on the use of DA towards metabolomics datasets, a good initial approximation can be performed by using a simple method such as the addition of noise, based on a modification of original instances with an established degree of this factor. Although the approach of this method is simple, carrying out the design and applying a procedure based on the addition of noise can be extended, complicating its operation for adaptation towards different datasets, and with the ability to obtain effective results.

$$\tilde{x} = \min(\text{Max\_Val}, \max(\text{Min\_Val}, x + \text{RND}(-1.0, 1.0))) \quad (1)$$

The noise addition method designed performs an increase in the number of available samples based on the random selection of "training" samples. For each sample, a copy is made, the features of which are then modified to a maximal value of 25%. Eq. (1) mathematically describes the process of obtaining a new feature value $\tilde{x}$ from the original one x. If the attribute is chosen for modification, noise that arises from a random normal distribution (denoted "RND" in Eq. (1)) with a standard deviation/variance of 1.0, is added to the original value; the resulting "noisy" value must not exceed the limits established for its feature. Furthermore, this value is not excessively affected by any previous scaling of the data. In this manner, the new value must be greater than the minimum value present in the feature (Min_Val), a process which sets a lower bound. The value must also be less than the maximum value present in the feature (Max_Val) setting an upper

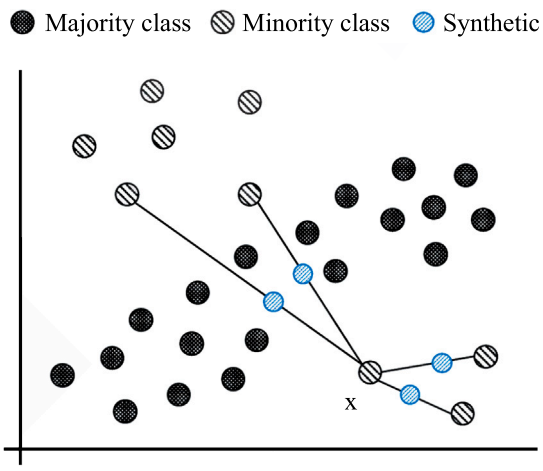

Fig. 1. Example of synthetic sample generation with the SMOTE technique. Synthetic samples are created with a random interpolation between samples of the minority class.

bound. A standard deviation of 1.0 is sufficient to create a sample that does not stray too far from the sample.

A variation of this noise addition method has also been designed. This method, abbreviated here as "noise bal", differs from the noise addition method described above, in that it performs the random selection only on the training samples of the minority class. Thus, only synthetic samples that belong to this class are created. The remainder of the method follows the same process as the standard noise addition approach.

3.2.2. SMOTE technique

Medical datasets, as well as metabolomics ones, may specifically present a characteristic that normally renders the task of prediction/classification difficult: large imbalances amongst the data classes. These sets usually show more samples of the control class than samples belonging to the class that indicates the presence of a particular disease. This is especially the case for biofluid samples collected from rare or very diseases such as the lysosomal storage disease NPC type 1 investigated here. An effective way to reverse this situation with DA is to apply an oversampling technique, such as the SMOTE technique (synthetic minority oversampling technique) [17].

In order to generate synthetic minority class data and balance the dataset, SMOTE uses a k-nearest neighbour algorithm on the minority class data, instead of random sampling with replacement. Thereby, different neighbours are randomly selected for each sample x, and a random interpolation is performed between these selected neighbours and sample x. Typically, this interpolation calculates the difference between 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 then adds this value to that of the original feature of sample x. Hence, this technique creates new synthetic samples that will be located within this space between neighbours and x. Fig. 1 shows the process followed by SMOTE for the creation of samples.

The disadvantages of the SMOTE algorithm application are connected to the random interpolation it performs, and its design as an oversampling technique. One of these disadvantages is the lack of control over the number of samples that will be generated from data, since oversampling aims to fully balance the dataset, i.e. it can only generate sufficient samples for this circumstance. Therefore, it is an ineffective method in well-balanced datasets, for which the number of samples to be generated in order to achieve balance is low in proportion when expressed relative to the number of samples involved. Another disadvantage derived from interpolation is the creation of synthetic samples that violate the geometry present in the dataset. An example of how SMOTE can lead to the creation of synthetic samples that do not follow the distributions of the original samples is also shown in Fig. 1.

3.2.3. Conditional generative adversarial network

The DA involved in the generation of realistic images has shown impressive success through the application of models known as Generative Adversary Networks (GAN) [18,19], which have a deep learning architecture. The objective of the GAN models is to learn the distribution of the original dataset in order to generate new samples from the learned distribution. With this aim, the standard GAN model has a structure divided into two neural networks, the *generator* and the *discriminator*. These two networks are trained simultaneously, yielding a confrontation between both so that they are able to learn from each other. In this context, the objective of the discriminator network (D) is to distinguish whether a sample arises from the set of 'real' data or is a generated sample, i.e. for the input sample x the discriminator estimates the probability of the sample belonging to the actual distribution or not. Notwithstanding, the generator network (G) takes as input a noisy random distribution z, and produces as output a distribution $G(z)$ assigned to the space of the real samples. The purpose of the generator is to create new 'synthetic' samples with features that approximate those present in the real samples. Hence, the discriminator network will not be able to distinguish these synthetic samples as samples not derived from the real distribution. Moreover, the generating process completely opposes the discriminating process, giving rise to the competitive environment noted above.

From the basic GAN, variants have been proposed that include variations in the network architecture, the loss function, or the inclusion of additional information. Some of these variants have demonstrated a valuable and effective performance in the generation of false images, being able to generate those that experts can consider 'real'. Amongst these, the most prominent variants may include the CycleGAN [34], the Conditional GAN [35], the Wasserstein GAN [36] and the Progressive Growing GAN [37].

Specifically, since a supervised task is performed in the present study, the model considered is the Conditional GAN (CGAN) [35]. This model is a variant of the standard GAN model in which the information concerning either a conditional parameter y, the sample label or other data information, is taken into account in the network, so that the generated samples directly present a label. In this manner, the latent space z and the condition y are passed onto the generator network as an input, y, that can be created randomly when training the model, and this can be controlled when generating synthetic samples. This condition (y) is also added to the input of the discriminator network, being the same as that employed to create a synthetic sample by the generator, or being the real label assigned to the real sample to be introduced in the discriminator.

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

The behaviour of the model can be observed in the GAN objective cost function (Eq. (2)), in which two parts that are identified within the competitive process may be distinguished. One is related to having an improved recognition of those samples that belong to the real distribution ($\mathbb{E}_x[\log D(x|y)]$), whereas another is related to an improved recognition of those samples that are generated by the generator ($\mathbb{E}_z[\log(1 - D(G(z|y)))]$). In this context, the capacity of the model to perceive whether the samples are real or false is expressed in Eq. (3), and the error of the model identified for the recognition of fake samples is modelled by Eq. (4). Additionally, the condition y in the computation of the objective cost function is the only variation with respect to that present in the standard GAN model.

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

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

The DA process performed with the CGAN model is shown in Fig. 2. Primarily, the noisy random distribution $\tilde{x}$ and a label y are introduced as input to the generator network G, which creates a synthetic sample,

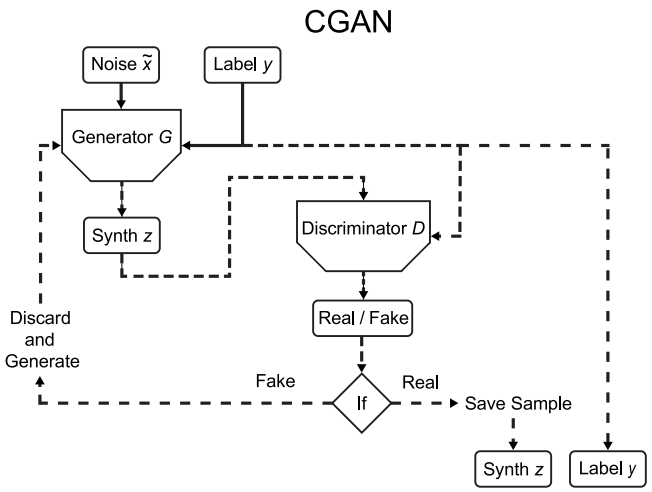

Fig. 2. Synthetic sample creation process for DA performed with a Conditional Generative Adversarial Network. Noise distribution $\tilde{x}$ and label y are inputted to the generator G, which creates a synthetic sample z. The discriminator D estimates whether it considers z as "real", in which case it is saved and assigned the label y. If z is considered false, the sample is discarded and the process is repeated.

z. This sample z, together with the label y, are passed to the discriminator network D that estimates the probability that the sample is indeed "real". If the synthetic sample z is considered as such, it is saved by assigning it to the same label class y. If the sample is considered fake, it is discarded and the method proceeds to create a new one.

3.3. Experiments

The details of the experiments conducted and the implementations of the models used in this study are provided in this sub-section. The set of parameters tested in the experiments arise from a selection based on previous experience, since the range of parameters should be limited in view of the large computational times involved.

The process followed to perform the experiments can be observed in Fig. 3. From the original dataset, a stratified division into training and test sets is performed in order to maintain total independence between the synthetic data generation process. This relates to the training of the classifier models and the evaluation of the classification metrics. In this context, the test set is not included in the data generation process, and is retained separately until the final testing for an external honest test of the accuracy takes place. A division of 60% of data for training and 40% for testing was established in view of the small number of samples present in the benchmark datasets. This proportion allows for a better evaluation of the test results. In addition, the entire process of dividing the data and its application is performed by a cross-validation procedure to provide an improved estimate of the results acquired. This reduces the random effects arising from the process. The data generation process employs the training set to create the desired number of synthetic samples. SMOTE then creates a sufficient number of samples for the set to become balanced. Each desired augmentation model follows the procedure described above in the previous sections. In this process, the quality of the samples generated can be affected by the training data provided; therefore, the implementation of a cross-validation strategy is considered important.

Once the synthetic samples have been created, a principal component analysis (PCA) is performed on the training dataset, obtaining the scores vectors and transforming all the data sub-sets (training, test and synthetic). This PCA strategy is applied prior to any experimentation performed on the dataset. This approach also considers the percentages of the variance explained by each principal component (PC), in order to select a sufficient number of components. Once a satisfactory number

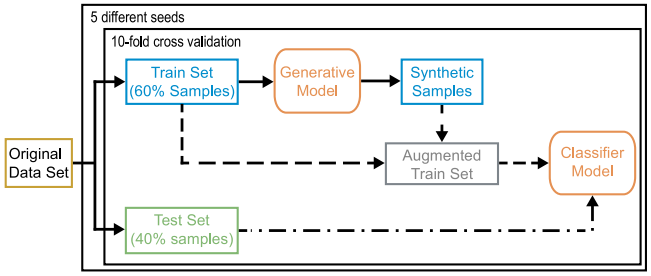

Fig. 3. Flow diagram of the whole experimentation process. A generative model is fed with samples from the training set, synthetic samples are created and added. Once the classifier is trained, predictions are evaluated using the Test set. The process is repeated with 5 different seeds in a 10-fold cross validation process.

Table 1
Characteristics of the generator and discriminator networks that compose the CGAN model for the generation process.

Layer	Output size	Batch-norm.	Activation
Generator			
Input = [z, y]	33 + 1	–	–
Fully connected	1024	Yes	ReLU
Fully connected	512	Yes	ReLU
Fully connected	256	Yes	ReLU
Fully connected	128	Yes	ReLU
Fully connected	33	No	Sigmoid
Discriminator			
Input = [x, y]	33 + 1	–	–
Fully connected	64	Yes	Leaky ReLU
Fully connected	128	Yes	Leaky ReLU
Fully connected	256	Yes	Leaky ReLU
Fully connected	512	Yes	Leaky ReLU
Fully connected	1	No	Sigmoid
Gen. optimiser	Adam (lr = 0.0004, beta1 = 0.5, beta2 = 0.9)		
Discrim. optimiser	Adam (lr = 0.0002, beta1 = 0.5, beta2 = 0.9)		
Leaky ReLU slope	0.2		
Backend	Tensorflow		

of PCs have been selected, in this case 14 PCs, the transformation of the sub-sets modifies the samples. In this manner, the samples present values that are represented by PC vectors instead of the original variable values. This process is performed to remove any high levels of correlation (multicollinearity) between the variables that are present in the metabolomics dataset. With the transformed sub-sets, synthetic data is added to training data, a process resulting in an augmented training set with which the classifier model is trained. Once the training is completed, prediction of the test set is conducted, obtaining the necessary metrics, which are then averaged for cross-validation. The entire process described was repeated with 5 different seeds featured in order to alleviate random effects.

Regarding the specific implementation of DA methods, the noise addition and SMOTE techniques do not present parameters or unique details of this procedure beyond those described in the respective sections featured in this report. The most important requirement is to indicate the implementation and architecture details of the CGAN model used for the generative process of synthetic samples, since a standard implementation process is not used, and its application depends on a series of different elements. Table 1 shows the characteristics of our CGAN process, which involves a distinction between the generator and discriminator networks, since they present different parameters and values. The columns in the Table show the type of layer, the output size of each one (which is identified by the number of neurons present in each layer), the use of batch normalisation, and the activation function used in the layer.

The neural network architecture used by the generator network consists of an input layer, four hidden layers and an output layer. The input layer is composed of the dimension of the noise vector, which is

Table 2
Test results acquired with logistic regression as classifier. Comb 1 is a combination of samples created with the CGAN and SMOTE strategies. Comb 2 is a combination of samples created with the CGAN and NOISE Bal approaches.

Model	DA %	Train	Accuracy	Specificity	Precision	Sensitivity	F_1 score	AUROC
None	None	36	88.42 ± *0.48*	**95.89** ± *0.56*	79.37 ± *0.62*	60.00 ± *1.41*	68.34 ± *1.04*	77.95 ± *0.71*
CGAN	50	54	88.08 ± *0.43*	93.58 ± *0.45*	73.36 ± *1.13*	67.20 ± *1.85*	70.15 ± *1.23*	80.39 ± *0.88*
NOISE	1000	396	**89.33** ± *0.78*	**95.89** ± *0.71*	**80.50** ± *2.54*	64.40 ± *2.48*	71.56 ± *2.32*	80.15 ± *1.30*
SMOTE	100	56	88.83 ± *0.52*	94.63 ± *0.35*	76.61 ± *1.36*	66.80 ± *1.36*	71.37 ± *1.25*	80.72 ± *0.81*
NOISE Bal	500	76	89.25 ± *0.64*	93.79 ± *0.68*	75.31 ± *2.03*	72.00 ± *0.89*	73.62 ± *0.96*	82.68 ± *0.67*
NOISE Bal	2000	196	88.58 ± *0.81*	91.16 ± *0.65*	70.11 ± *1.44*	**78.80** ± *1.85*	**74.20** ± *2.04*	**84.98** ± *1.15*
Comb 1	50	74	87.58 ± *0.46*	91.89 ± *0.49*	69.80 ± *0.82*	71.20 ± *2.58*	70.49 ± *1.84*	81.55 ± *1.18*
Comb 2	100	112	87.17 ± *0.31*	90.63 ± *0.35*	67.52 ± *0.81*	74.00 ± *0.63*	70.61 ± *0.92*	82.32 ± *0.36*

equivalent to the size of the dataset (33 variables), together with one that represents the classification label. The output layer presents a dimension equivalent to the original dataset. This network is responsible for creating synthetic data, and uses a sigmoidal activation function. In the hidden layers, the Rectified Linear Unit (ReLU) [38] function is used as the activation function. Batch normalisation is employed as a regularisation technique [39]. The architecture of the discriminator network also has an input layer, four hidden layers, and an output layer. The input layer has a dimension equivalent to the size of the dataset plus one for the class label, and the output layer has a single neuron, which decides whether the input sample is real or false. In the discriminator network, the hidden layers use the Leaky ReLU [40] activation function, which provides more stability in the classification task than the ReLU version employed in the generator network. The hidden layers of this network also use batch normalisation as a regularisation technique. The bottom of Table 1 shows other details of the model. Both networks use the Adam algorithm with adaptive learning rate [41], and the values used for the first and second moments (beta1 and beta2) are indicated. Further, the value of the slope of the Leaky ReLu activation function and the software backend used (Tensorflow [42]) are shown.

4. Results

The experimentation process proposed in Section 3.3 was followed, and an auto-scaled transformation (mean-centred followed by division by the standard deviation of each metabolite variable) of the raw ^{1}H NMR analytical dataset described in the Section 3.1.1 was performed.

4.1. Classification performance

The objective is the classification of the samples in one of the two groups: the heterozygous (parental) carrier control group or to the NPC1 disease one. Table 2 shows the test results obtained for the different methods and DA models applied, when a logistic regression approach was used as a classifier model in the experimentation. In this table, 'None' indicates the results obtained when the number of data samples is not increased. This was used for the results to show if the data augmentation improves the classification performance. In addition, the results obtained with two additional methods are included, where samples from two different models of DA are combined. 'Comb 1' refers to the results obtained with a combination between the samples created using the CGAN model and the SMOTE techniques, where as 'Comb 2' indicates the results obtained with a combination between the samples created using the CGAN model and the Noise Bal method, when a 500% level of sample creation with Noise Bal is involved.

The first column of Table 2 ('Percent') refers to the size of the synthetic data created compared to the original training set. Thus, a percentage of 100 indicates that as many samples are created as those in the training set. A percentage of 50 indicates that half the number of samples are created. Finally, the value 1000 indicates that the number of training samples is increased 10-fold. The 'Train' column indicates the number of samples used for training once the DA has been performed. The following columns show the values (± SE) obtained for each of the test metrics (boldface font indicates the best values).

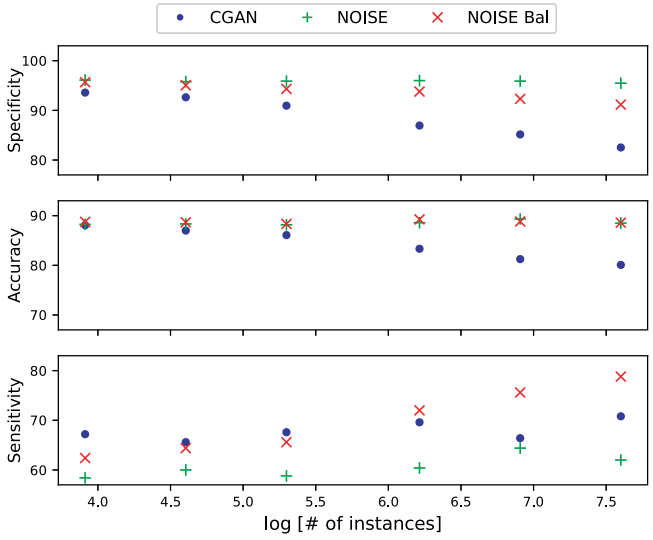

Fig. 4. Comparisons of the accuracy, specificity and sensitivity obtained using logistic regression as classifier with different DA models with respect to the log of the number of instances created. Dots represent the results obtained with the CGAN model, crosses those with the NOISE method, and cross-hairs represent results obtained by using the NOISE Bal method.

The Table also shows the accuracy level obtained. This indicates the proportion of the correct predictions (true positives and true negatives) amongst the total number of test samples. The specificity is also shown — this measures the number of true negatives with respect to the total number of negative patients. This is related to the ability of the classifier model to correctly reject heterozygous control patients. The sensitivity is also shown; this measures the number of true positives in relation to the total number of positive patients, and relates to the ability of the test to correctly detect patients with the disease.

Precision is also shown in Table 2, and indicates how good the classifier model is in predicting a sample as being positive. This is calculated as the number of true positives amongst the total of samples predicted as positive. The following column shows the results obtained for the F_1 score. This metric is the harmonic mean of the precision and sensitivity. It allows a more reliable measure of the performance of the classifier, and this is particularly the case in circumstances where sensitivity becomes more important. The last column shows the results obtained for the AUROC, and represents the ability of a classifier to distinguish between classes according to the relationship between sensitivity and 1 - specificity.

Results shown in Table 2 show that an improvement in test prediction accuracy is achieved with the addition of the noise method compared to that obtained with the dataset without augmentation ('None'). Using the Noise method with 1000 percent, an accuracy of 89.33% with a specificity of 95.89% is achieved. The latter assumes the highest value as being the same as that obtained with the benchmark dataset. The other outstanding method is that of the Noise Bal approach, with 2000%. This method reaches the highest sensitivity

Table 3
Results acquired with a random forest (RF) system used as a classifier. Comb 1 and Comb 2: described in Table 2.

Model	DA %	Train	Accuracy	Specificity	Precision	Sensitivity	F_1 score	AUROC
None	None	36	85.00 ± *0.47*	**96.95** ± *0.68*	77.34 ± *1.88*	39.60 ± *2.90*	52.38 ± *2.53*	67.58 ± *1.25*
CGAN	100	72	86.58 ± *0.83*	93.68 ± *0.96*	71.29 ± *1.90*	59.60 ± *2.14*	64.92 ± *2.00*	77.04 ± *1.01*
NOISE	1000	396	86.17 ± *0.42*	96.63 ± *0.69*	**78.38** ± *3.19*	46.40 ± *4.07*	58.29 ± *3.36*	71.22 ± *1.73*
SMOTE	100	20	**86.67** ± *0.85*	93.16 ± *1.55*	70.45 ± *4.03*	62.00 ± *3.38*	65.96 ± *1.65*	77.18 ± *1.23*
NOISE Bal	500	76	86.58 ± *1.35*	90.84 ± *1.46*	66.92 ± *2.82*	70.40 ± *1.94*	**68.62** ± *2.15*	**80.00** ± *1.40*
NOISE Bal	2000	196	78.92 ± *0.92*	79.26 ± *1.38*	49.62 ± *1.96*	**77.60** ± *2.15*	60.53 ± *1.16*	77.97 ± *0.89*
Comb 1	100	92	83.17 ± *1.05*	85.68 ± *1.62*	57.50 ± *2.35*	73.60 ± *1.41*	64.56 ± *0.64*	79.64 ± *0.46*
Comb 2	50	94	85.33 ± *0.96*	88.63 ± *1.17*	62.76 ± *1.64*	72.80 ± *2.48*	67.41 ± *1.69*	78.41 ± *1.26*

Table 4
Test results with a support vector machine system used as a classifier. Comb 1 and Comb 2: described in Table 2.

Model	DA %	Train	Accuracy	Specificity	Precision	Sensitivity	F_1 score	AUROC
None	None	36	83.17 ± *0.60*	**99.37** ± *0.31*	**90.00** ± *2.18*	21.60 ± *2.04*	38.84 ± *2.92*	60.48 ± *1.11*
CGAN	200	108	87.25 ± *0.75*	93.26 ± *0.51*	71.56 ± *1.97*	64.40 ± *2.48*	67.79 ± *2.83*	78.83 ± *1.35*
NOISE	100	72	88.25 ± *0.45*	96.84 ± *0.27*	82.25 ± *2.14*	55.60 ± *2.28*	66.35 ± *2.10*	69.05 ± *1.10*
SMOTE	100	20	89.25 ± *0.46*	95.68 ± *0.48*	79.80 ± *2.01*	64.80 ± *2.06*	71.52 ± *1.46*	80.24 ± *0.97*
NOISE Bal	500	76	**89.50** ± *0.65*	94.53 ± *0.74*	77.19 ± *2.19*	70.40 ± *1.85*	**73.64** ± *1.24*	83.51 ± *0.95*
NOISE Bal	2000	196	87.25 ± *0.67*	89.50 ± *0.43*	66.11 ± *1.15*	**79.60** ± *2.14*	72.23 ± *1.48*	**84.43** ± *1.18*
Comb 1	100	92	89.00 ± *0.48*	93.05 ± *0.51*	73.59 ± *0.96*	73.60 ± *1.67*	73.60 ± *0.94*	81.39 ± *0.83*
Comb 2	50	94	88.50 ± *0.43*	93.59 ± *0.26*	73.93 ± *1.25*	69.20 ± *1.62*	71.49 ± *1.39*	82.53 ± *0.85*

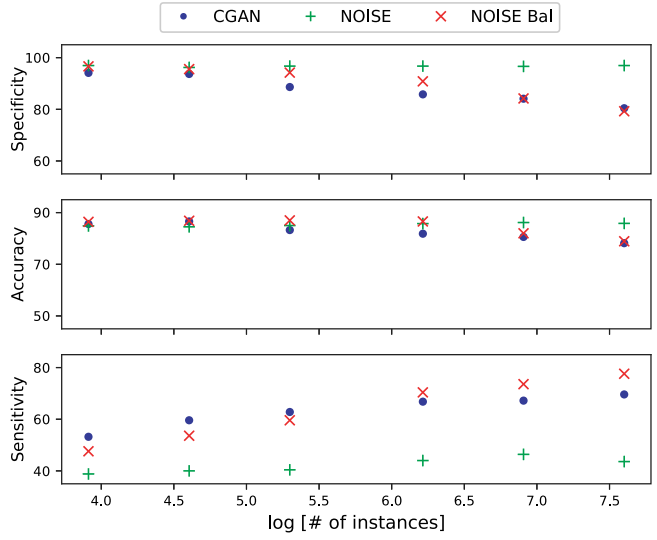

Fig. 5. Comparisons of the accuracy, specificity and sensitivity obtained using a random forest model as a classifier with different DA models with respect to the log of the number of instances created. Symbol codes: as Fig. 4.

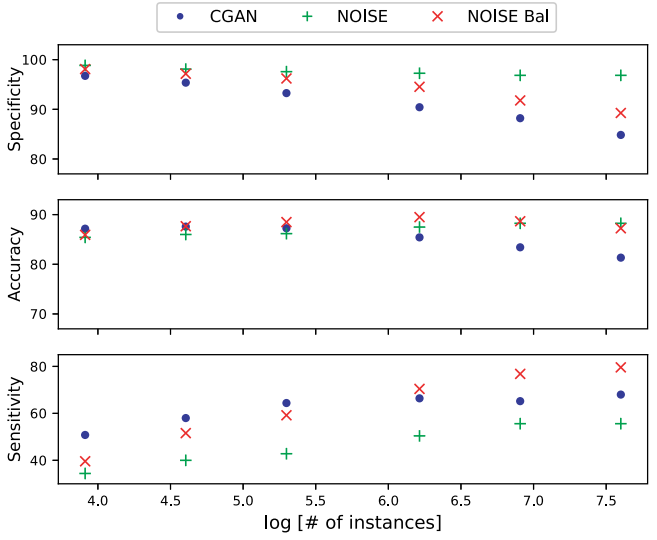

Fig. 6. Comparisons of the accuracy, specificity and sensitivity obtained using support vector machine system as a classifier with different DA models with respect to the log of the number of instances created. Symbol codes: as Fig. 4.

values (78.8%), F_1 score (74.2%) and AUROC (85.0%). These values show a substantial improvement compared to analysis of the dataset without augmentation.

Further analyses of the influence of the size of the datasets on the results obtained with different DA methods are presented in Fig. 4. Here, three test metrics (accuracy, specificity, and sensitivity) obtained with three DA methods (CGAN, Noise, and Noise Bal), *versus* the number of instances on a logarithmic scale on the abscissa axis, are presented. Taking into account the different models, for the Noise and Noise Bal methods, the accuracy of the results obtained are approximately stable with respect to the increase in the number of instances created. Furthermore, the specificity values are stable with the Noise method, and decrease slightly for Noise Bal. The sensitivity values increase slightly with Noise and there is clearly a significant positive correlation between sensitivity and the number of instances created with this analysis strategy. 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.

The previous results were obtained using a logistic regression model classifier. For comparison, a random forest (RF) system [43] was also implemented. The results obtained with the different methods and DA models applied are shown in Table 3. This Table displays the percentage of synthetic data created, and the test values obtained for the accuracy, specificity, sensitivity, precision and the F_1 score. In this case, the highest accuracy prediction value is obtained by increasing the set with the synthetic samples created with SMOTE (86.67%). Regarding the sensitivity and F_1 score metrics, they are still attained with the Noise Bal method. The results show a 77.6% sensitivity value using 2000% data size increment and a 68.62% F_1 score value using 500% increment. This demonstrates a substantial improvement over the results obtained when no data augmentation method is applied (39.6% sensitivity, 52.38% F_1 score).

Fig. 5 shows the influence that the size of the datasets has on the accuracy, specificity and sensitivity results obtained with the CGAN, Noise and Noise Bal methods. In these experiments, a random forest system is again employed, and results therefrom are compared to the number of instances on a logarithmic scale on the abscissa axis. The

results are similar to those obtained when using a logistic regression strategy. Emphasis was given to the cases of CGAN and Noise Bal with a negative correlation between the number of instances created and the gain in specificity, and also the precision of the prediction. In contrast, a positive correlation for the gain of sensitivity was considered.

A support vector machine model was also used as a classification system. The results obtained for the different methods and DA models applied are shown in the Table 4. This table shows the percentage of synthetic data created. It also shows the test values obtained for the accuracy, specificity, sensitivity, precision and F_1 score. The results in Table 4 show an improvement in the accuracy and sensitivity of the prediction with the Noise Bal method over and above that obtained with the dataset without augmentation. Using the Noise Bal method with 500%, a precision of 89.50% and an F_1 score of 73.64% were achieved. Training using only the original dataset, however, yielded the highest values of specificity (99.37%) and precision (90%). Noise Bal with 2000% also was very effective, achieving the highest sensitivity value of 79.6% and AUROC value of 84.4%.

The influence of the dataset size on the results can be viewed in Fig. 6. Indeed, this Figure shows the precision, specificity and recovery of the results obtained with the CGAN, Noise and Noise Bal methods. The results acquired are based on a support vector machine model and are compared to the number of instances on a logarithmic scale on the abscissa axis. This Figure shows similar results to the previous ones in the relationship between specificity and size of the dataset. Similar results for the increase with the CGAN strategy can also be observed. The results with both methods of noise addition show a positive correlation between the number of instances created and the sensitivity gain.

Considering the overall results of the different classifiers examined, an accuracy of 85.5% and a F_1 score of 53.2% were obtained without an augmentation process. The overall accuracy and F_1 score increases to 88.4% and 72.0% respectively using the Noise Bal method and 500% DA, and increases to 87.0% and 69.8% using Comb 2 (combination method CGAN + NOISE Bal).

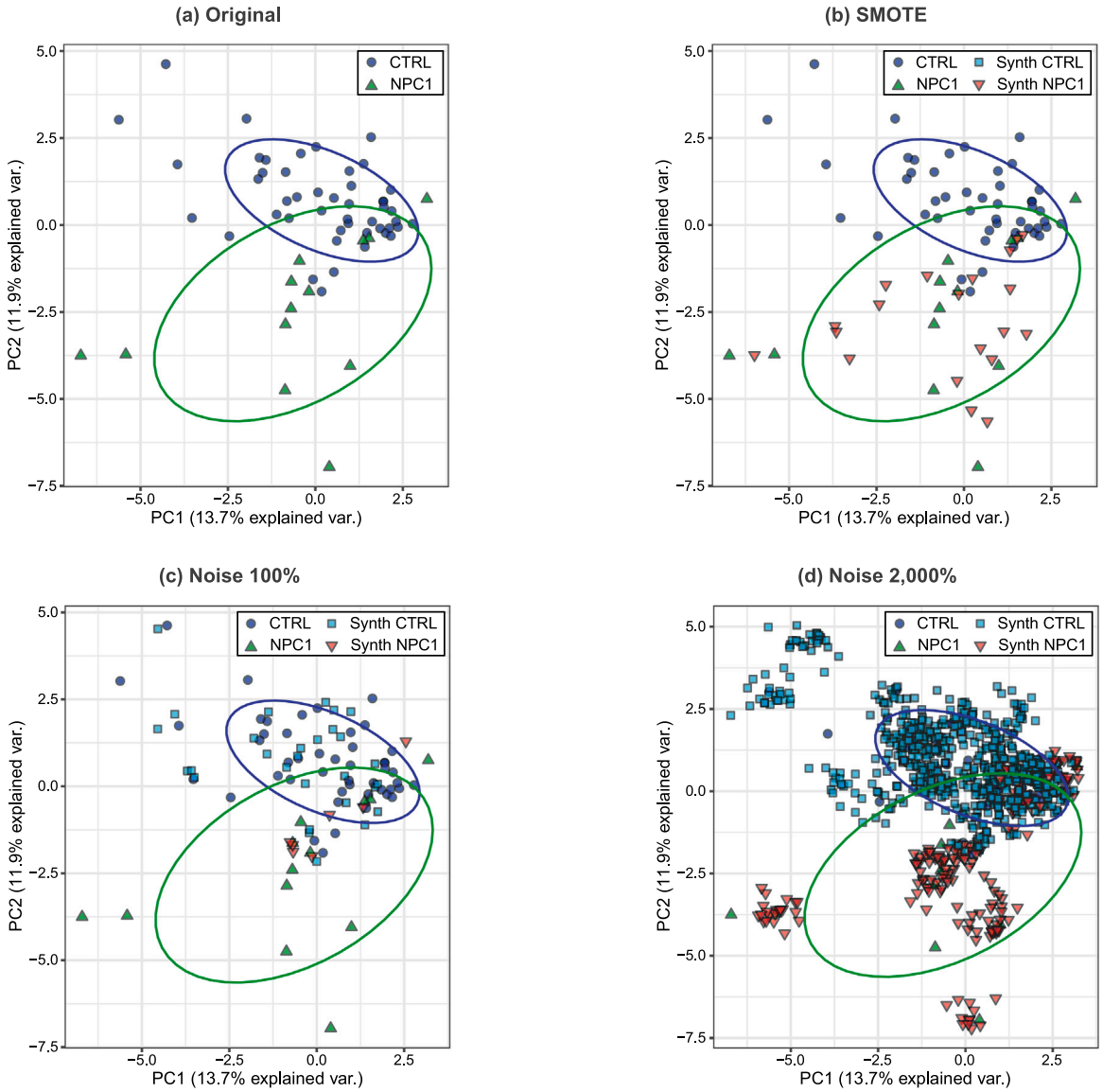

Fig. 7. PCA scores plot with: the original NPC1 disease dataset (a); the samples augmented with SMOTE (b); the samples augmented via the addition of noise with percentage of samples generated 100% (c) and 2000% (d). Colour codes: green triangles, NPC1 disease urine; dark blue circles, heterozygous carrier control urine; red inverse triangles, generated NPC1 disease urine samples; light blue squares, generated heterozygous carrier control urine samples.

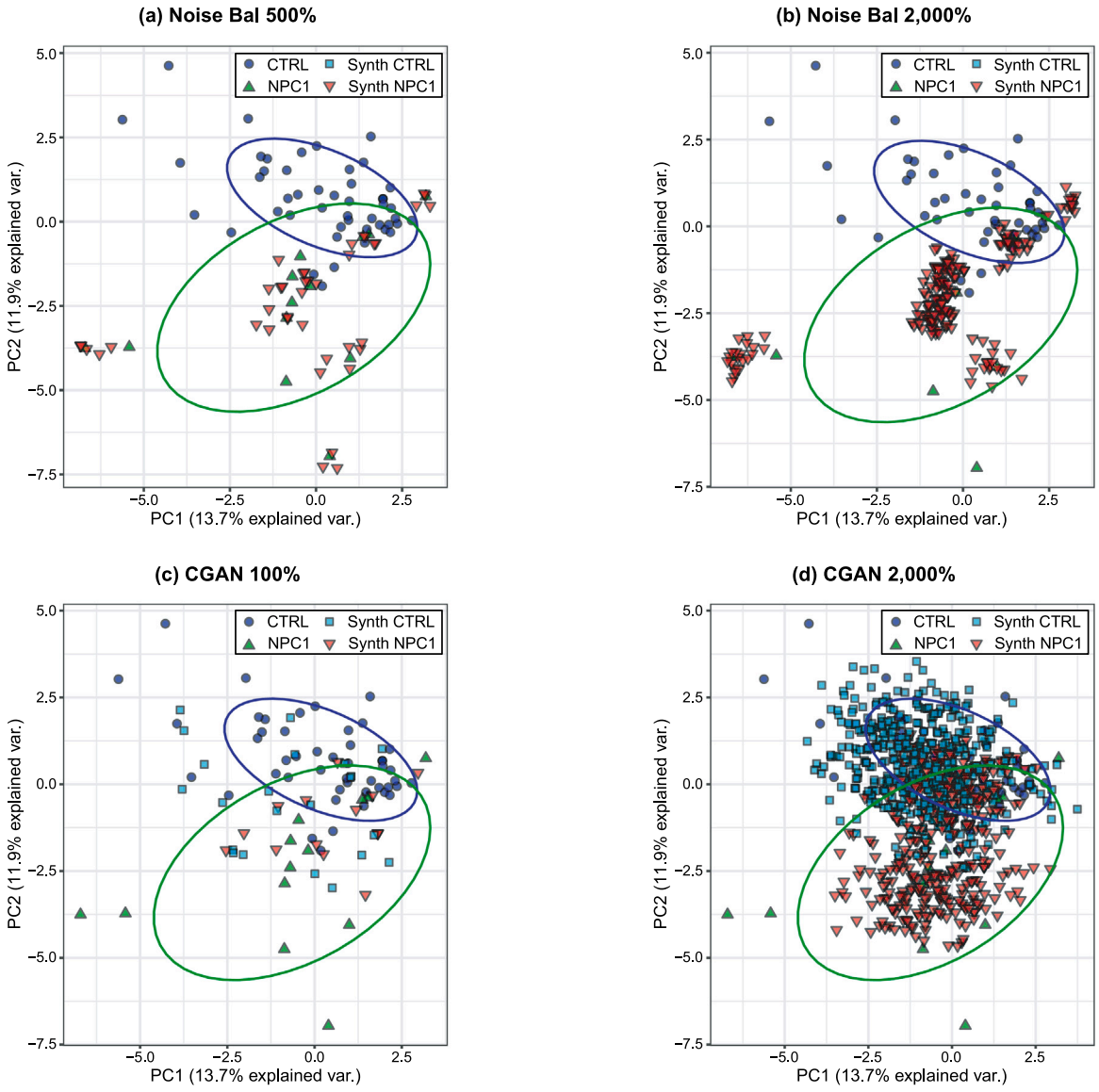

Fig. 8. PCA scores plot with the samples augmented via the addition of noise in a balance process with percentage of samples generated 500% (a) and 2000% (b); the samples augmented with the CGAN model with percentages 100% (c) and 2000% (d). Colour codes: as Fig. 7.

4.2. Augmented datasets analysis

In order to analyse how the DA methods used in the experimentation were able to replicate the information present in the metabolomics dataset, a main principal component analysis (PCA) and a partial least squares — discriminatory analysis (PLS-DA) were conducted. Through these methods, the configuration of the samples can be visualised in a two-dimensional space (i.e. component 2 *vs* component 1). This provided a means to check the distribution of the synthetic samples created, and compare this with the distribution of the original samples.

Fig. 7(a) shows the PCA model results obtained by using the original NPC1 dataset. This reveals that there were two significant clusters of the two sets of "disease state" classifications, whereas the cluster belonging to the heterozygous carrier group (dark blue circles) appeared as a compact cluster. On the contrary, the cluster conformed by the samples of NPC1 disease urine (green triangles), was more dispersed. It can also be noted that there is an area where both clusters converge.

The results of PCA scores after adding the samples generated with SMOTE are shown in Fig. 7(b). Here, it can be clearly seen how the creation of samples through SMOTE works. The synthetic samples (red inverse triangles) are distributed along the distribution of the original samples of NPC1 disease urine. These synthetic samples were generated by interpolation of the original samples.

Fig. 7 shows the distribution of the synthetic samples created by using the Noise method with respect to the original samples. Percentages of 100 (c) and of 2000 (d) were used respectively. Fig. 8 shows the distribution of the samples created by the Noise Bal method. Percentages of 500 (a) and 2000 (b) were used respectively. The difference between both methods of noise addition can be fully appreciated in this analysis, which creates Noise Bal-only samples for the minority class. Since the basic operation used by both methods is the same regarding modifications of the original samples, it can be observed how the synthetic samples are grouped around the original samples that they modify. Moreover, the synthetic samples form only small clusters.

The sample creation of the above DA methods clearly differs from the CGAN method. Fig. 8 shows the results obtained with the use of PCA scores and the distribution of the synthetic samples created by CGAN with respect to the original samples. Percentages of 100 (c) and 2000 (d) were used respectively. With a greater number of samples, the behaviour of this model can be better observed. Most of the samples are concentrated in the dispersion areas of the original data. This produces an admixture of samples from both groups in the originally present

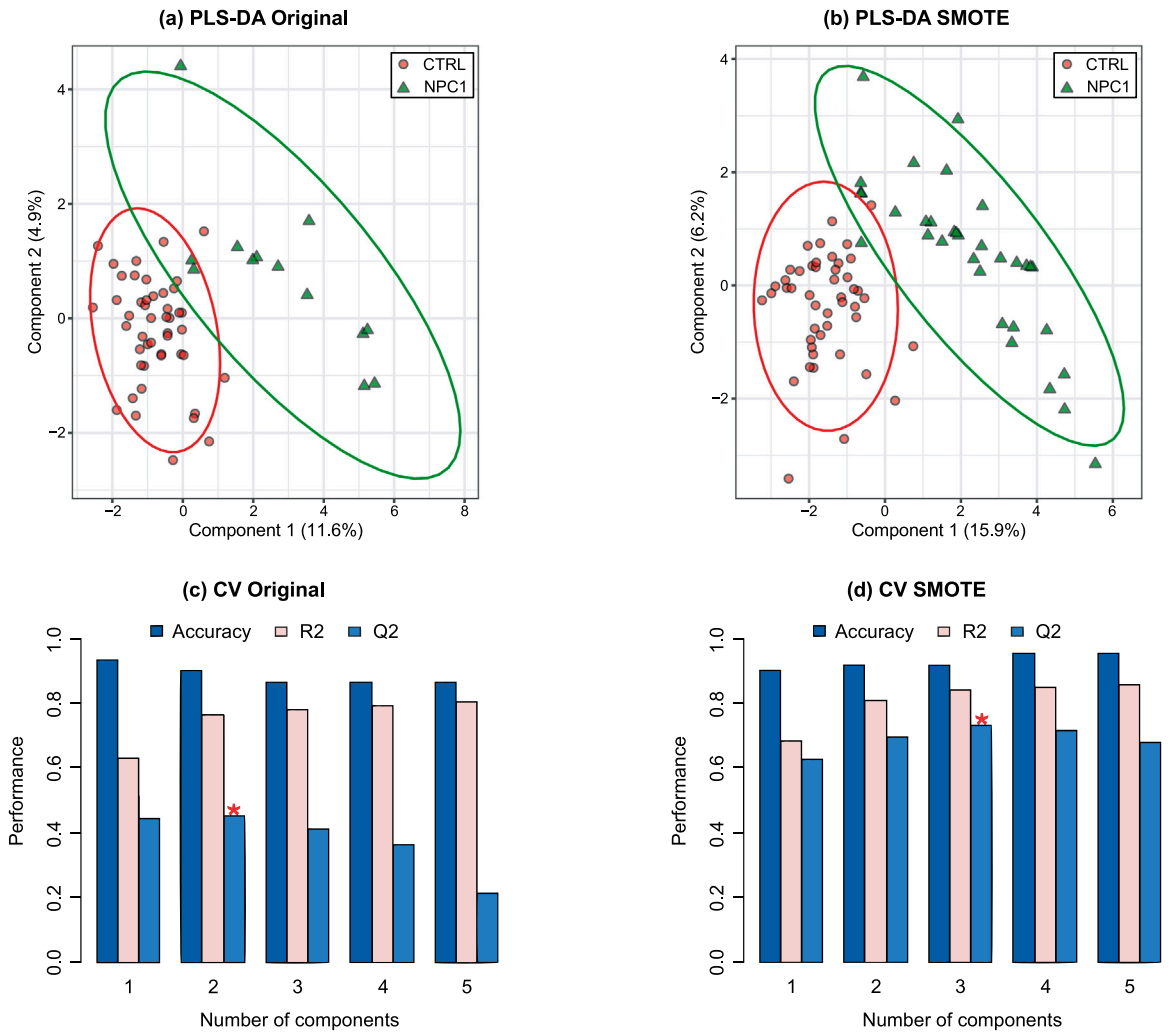

Fig. 9. PLS-DA component 2 *versus* component 1 scores plot (a) and cross-validation results for the original dataset (c), and PLS-DA scores plot (b) and cross-validation results (d) acquired on the augmented set with SMOTE applied. Colour codes: green triangles, NPC1 disease urine; red circles, heterozygous carrier control urine.

convergence zone. This model does not create any samples that stray excessively from these clusters, and this is probably accountable by the "discriminator" network of the model classifying them as false, and hence they are discarded.

Results derived from the PLS-DA component 2 *versus* component 1 scores plots, and the cross-validation, provide results that support previous analyses. These analyses were conducted using *MetaboAnalyst v4.0* software (University of Alberta and National Research Council, National Nanotechnology Institute (NINT), Edmonton, AB, Canada) [44]. PLS-DA provides metrics that indicated the predictive power of the model [45]. These metrics were accuracy, goodness of fit (R^2) and predictive capacity (Q^2). During the PLS-DA, a cross-validation was performed and the predictive data was then compared with the original data. The prediction error in all samples is summed (predicted residual sum of squares or PRESS). To calculate the Q^2 value, PRESS is divided by the initial sum of squares and subtracted from 1 to match the scale of R^2.

Fig. 9(a) also reveals two significant groups in the original data for the two sets of "disease state" classifications, with a small area of convergence between the two. Fig. 9(c) shows the results of accuracy, R^2 and Q^2 obtained with the original dataset, with regard to the number of components used. The optimal number of components selected for the dataset without augmentation is 2 components, obtaining an accuracy of 0.873, an R^2 value of 0.781, and a Q^2 value of 0.457.

In a similar manner to the representation by PCA of the synthetic samples created with SMOTE, Fig. 9(b) shows how the samples created with this technique are distributed throughout the 'real' cluster, from the interpolation process. Fig. 9(d) shows the cross-validation results with respect to the number of components used, in this case the dataset augmented with the synthetic samples created with SMOTE. This model obtained the most effective results with 3 components, with an accuracy of 0.930, an R^2 value of 0.840, and a Q^2 index of 0.743. These results confirm that this model gave rise to a significant improvement over the original, non-augmented, dataset.

Fig. 10(a) shows the PLS-DA component 2 *versus* component 1 scores plot when the augmented dataset with the CGAN synthetic samples is used (using a percentage value of 50%). The analysis reveals a substantial change compared to that deduced from the original dataset. In this case, the analysis of the components has been modified, and this significantly changes the distribution shown. Although the dispersion and angle of the distributions are dissimilar, it is still possible to differentiate both clusters, with a larger convergence zone.

Fig. 10(c) shows the cross-validation results with respect to the number of components used (the augmented dataset with the synthetic samples created with CGAN model was employed). The optimal number of components selected was two, obtaining an accuracy of 0.880, and R^2 and Q^2 values of 0.620 and 0.398 respectively for a model with two components. The change induced in the analysis of the components and the larger convergence zone, may be the cause of the poor performance in predictive capacity.

Fig. 10(b) shows the PLS-DA results when the augmented dataset generated by the Noise Bal method was used (with a percentage value

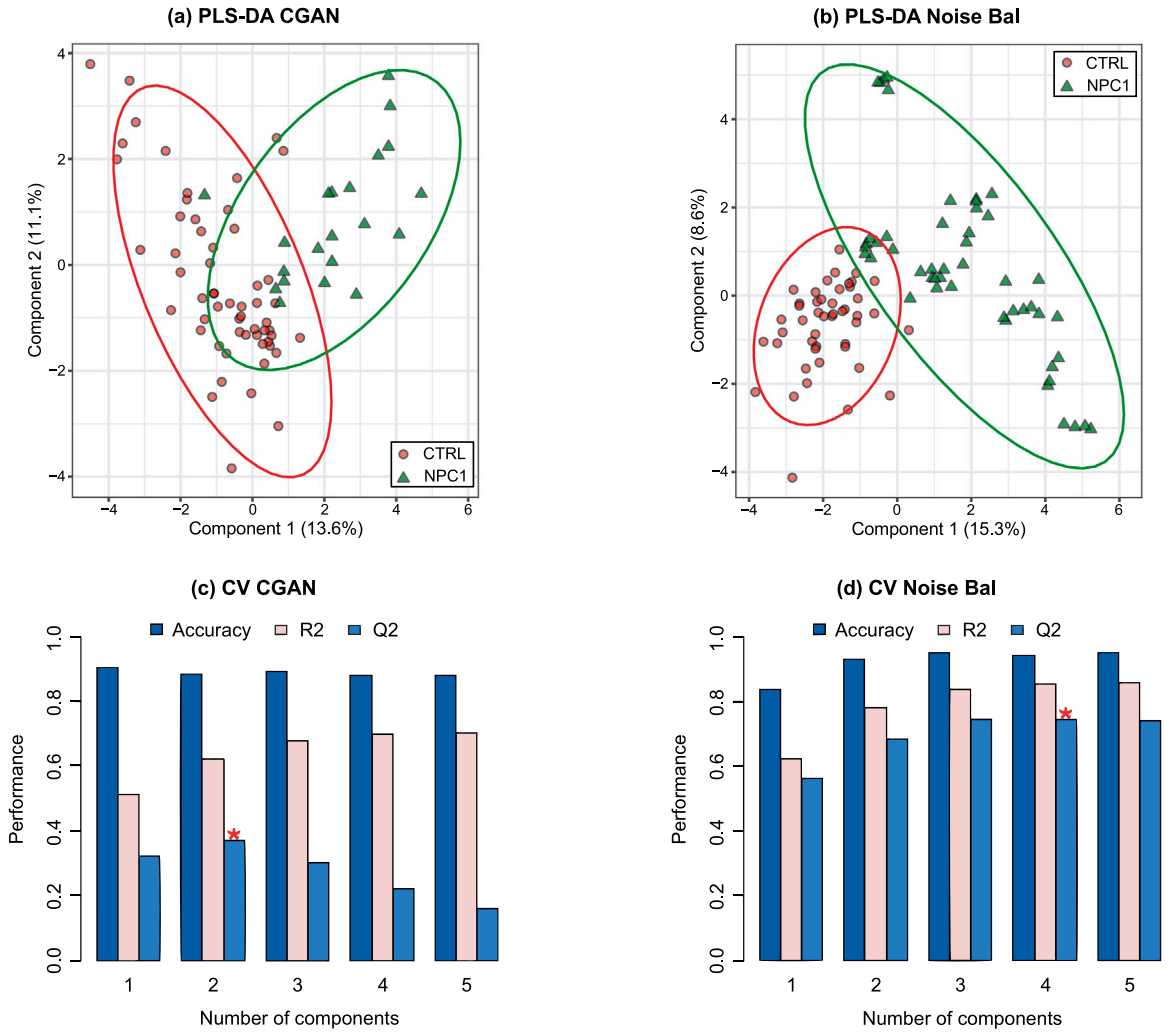

Fig. 10. PLS-DA scores plot (a) and cross-validation results (c) obtained with the augmented set generated with CGAN model, and PLS-DA scores plot (b) and cross-validation results (d) acquired with the Noise Bal strategy. Colour codes: as Fig. 9.

of 500%). The results are similar to those obtained with the original dataset, and with an augmented dataset produced by the SMOTE strategy. They are also similar to the results obtained with PCA. The samples generated NPC1 disease class samples accumulate around the original ones. The model obtains the optimal results with 4 components, with an accuracy of 0.960, an R^2 value of 0.852, and a Q^2 value of 0.749. These results significantly improve the results obtained on the dataset without augmentation, and also with the CGAN augmentation approach. They also improve the results obtained with the SMOTE augmentation method, but only to a small extent.

4.3. Metabolite importance

The variable importance in projection (VIP) scores with respect to component 1 were obtained using the PLS-DA technique. A VIP score is a measure of the contribution of an individual variable in the PLS-DA model [46]. This metric is calculated as the weighted sum of squares of the PLS-DA weights. This analysis indicates the importance of each metabolite in the process of differentiating between the clusters formed by the heterozygous carrier (control) group and NPC1 disease urine samples. This identifies the metabolites that may be employable as valuable biomarkers.

Fig. 11(a) shows the VIP scores obtained from PLS-DA applied to the original dataset. Values for the top 14 metabolites are shown. The coloured boxes on the right-hand side indicate the relative concentrations of the corresponding metabolite in each group under study. The most prominent metabolite were 3-aminoisobutyric acid (3-aminoisobutyrate at the normal pH values of human urine), with a VIP value equal to 1.89 (values >1.00 are considered significant).

Fig. 11(b) shows results from the VIP analysis with the augmented dataset using the synthetic samples created with the SMOTE method. The most prominent metabolites were trimethylamine with a VIP value of 1.81, and 3-hydroxyisovaleric acid with a VIP value of 1.69. The metabolite 3-aminoisobutyric acid occupied position 5 in this analysis.

The VIP analysis with the augmented dataset using the synthetic samples created with Noise Bal reflects a result similar to that obtained with SMOTE. Indeed, Fig. 11(c) revealed that the highest scoring metabolites with this augmented set were trimethylamine with a VIP value of 1.68, and 3-hydroxyisovaleric acid with a VIP value of 1.86. The metabolite 3-aminoisobutyric acid occupied position 4.

Fig. 11(d) shows the VIP analysis with the augmented dataset using the synthetic samples created with the CGAN model. In this case, the results present more differences than those observed for the alternative models. Indeed, the most prominent metabolites were found to be trigonelline with a VIP value of 1.79, and pyroglutamic acid with a VIP value of 1.70. The metabolite 3-aminoisobutyric acid occupied position 8 in this analysis.

The results of VIP scores shown in all the above Figures are summarised in Table 5. This table shows the VIP values obtained for each of the top 14 marker metabolites using the different augmented datasets. The results show that there are only 4 metabolites that are not included

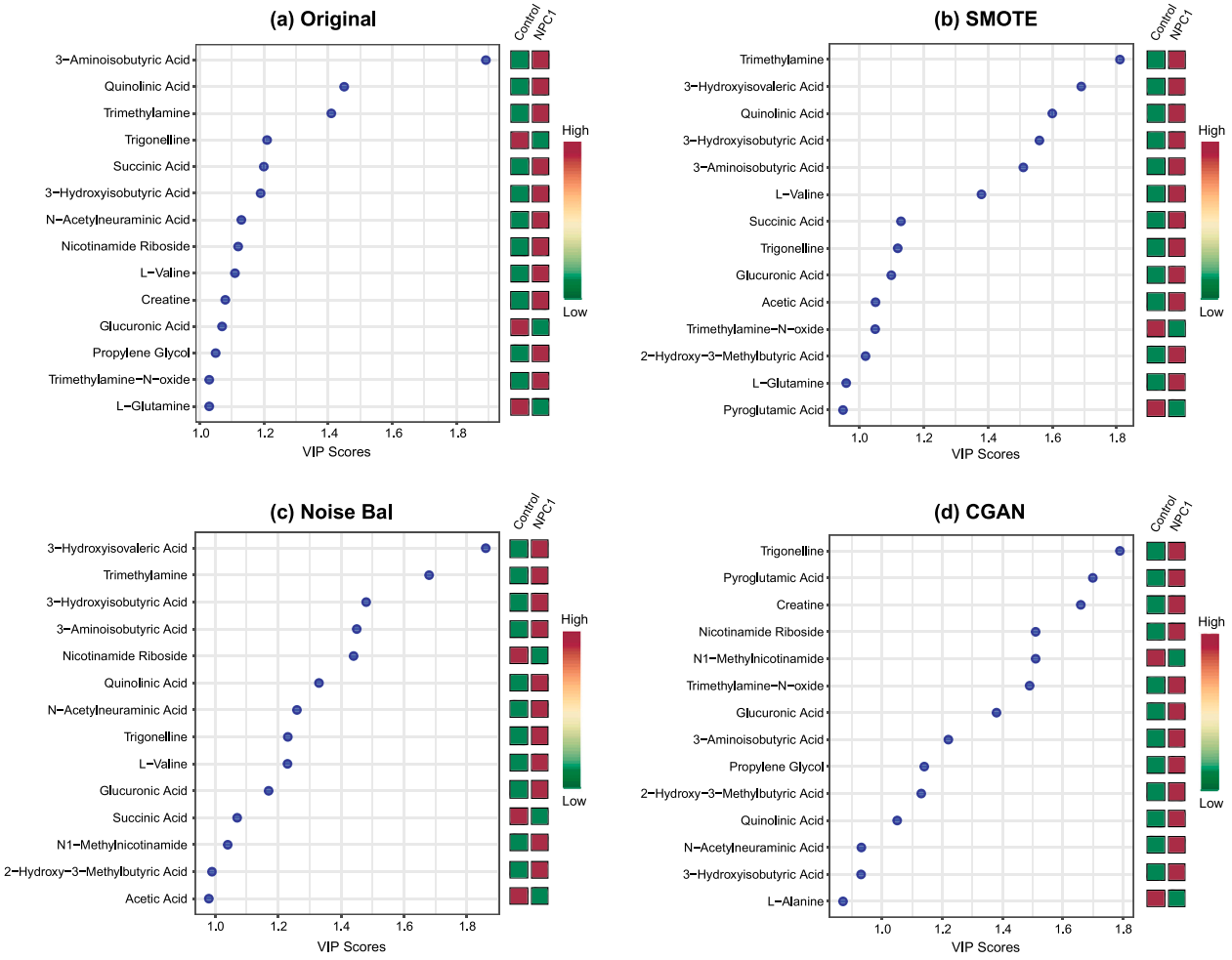

Fig. 11. PLS-DA variable importance parameter (VIP) values derived from the application of PLS-DA to: (a) the original dataset; (b) the augmented set using SMOTE; (b) the augmented set with Noise Bal method (using 500% of DA); (d) the augmented set with CGAN model. Colour codes indicate the relative concentrations of metabolites featured.

Table 5
PLS-DA VIP values obtained with the different DA methods (brackets represent the relative rank position of each metabolite variable for each method applied).

Metabolite	Original	SMOTE	Noise Bal	CGAN
3-Aminoisobutyric Acid	1.89 (1)	1.51 (5)	1.45 (4)	1.22 (8)
Quinolinic Acid	1.45 (2)	1.60 (3)	1.33 (6)	1.05 (11)
Trimethylamine	1.41 (3)	1.81 (1)	1.68 (2)	0.79 (16)
Trigonelline	1.21 (4)	1.12 (8)	1.23 (8)	1.79 (1)
Succinic Acid	1.20 (5)	1.13 (7)	1.07 (11)	0.70 (21)
3-Hydroxyisobutyric Acid	1.19 (6)	1.56 (4)	1.48 (3)	0.93 (13)
N-Acetylneuraminic Acid	1.13 (7)	0.93 (15)	1.26 (7)	0.93 (12)
Nicotinamide Riboside	1.12 (8)	0.68 (20)	1.44 (5)	1.51 (4)
L-Valine	1.11 (9)	1.38 (6)	1.23 (9)	0.66 (22)
Creatine	1.08 (10)	0.78 (18)	0.65 (22)	1.66 (3)
Glucuronic Acid	1.07 (11)	1.10 (9)	1.17 (10)	1.38 (7)
Propylene Glycol	1.05 (12)	0.43 (28)	0.40 (27)	1.14 (9)
L-Glutamine	1.03 (13)	0.96 (13)	0.77 (19)	0.70 (20)
Trimethylamine-N-oxide	1.03 (14)	1.05 (11)	0.93 (16)	1.49 (6)
2-H-3-Methylbutyric Acid	1.02 (15)	1.02 (12)	0.99 (13)	1.13 (10)
Acetic Acid	1.00 (16)	1.05 (10)	0.98 (14)	0.77 (17)
3-Hydroxyisovaleric Acid	0.99 (17)	1.69 (2)	1.86 (1)	0.58 (25)
N1-Methylnicotinamide	0.99 (18)	0.86 (17)	1.04 (12)	1.51 (5)
Pyroglutamic Acid	0.84 (19)	0.95 (14)	0.89 (17)	1.70 (2)
L-Alanine	0.43 (28)	0.16 (30)	0.24 (31)	0.87 (14)

with the original dataset, and which are present in the analysis with application of the SMOTE and Noise Bal methods.

Especially, the results obtained with the SMOTE and Noise Bal approaches are similar, and there are certain metabolites that occupy equivalent ranking positions. Amongst these, the following metabolites should be highlighted: trimethylamine (positions 1 and 2); 3-hydroxyisovaleric acid (pos. 2 and 1); 3-hydroxyisobutyric acid (pos. 4 and 3); 3-aminoisobutyric acid (pos. 5 and 4); and trigonelline (pos. 8 and 8). The similarity between the VIP scores obtained with these DA methods may arise from the fact that both are oversampling methods. The greater number of samples synthesised for the minority class (NPC1 disease) influenced the analysis significantly, which indicates a greater relevance of these metabolites to separate this group from the heterozygous carriers.

Notwithstanding, Table 5 shows the analysis with the augmented set using the CGAN model. This presents 5 differing metabolites from those deduced with the original dataset. The difference of the pyroglutamic acid metabolite at position 2 with CGAN, and position 19 with original dataset, should be noted. Similarly, the difference between the trimethylamine metabolite in position 16 with the CGAN approach, and position 3 with respect to the original dataset, should be highlighted.

5. Discussion

The strategies employed here clearly offer much potential regarding the metabolomics analysis of imbalanced datasets. These datasets predominantly comprise smaller or much smaller numbers of sample-donating participants recruited to the diseased group. This is particularly the case for diseases which are rare, and these include the NPC1 condition which is examined in this work in some detail. Indeed, NPC diseases caused by NPC1 and NPC2 mutations affect approximately

1:100,000 live births [2], although the NPC1 mutations account for 95% of cases observed. Overall, lysosomal storage disorders represent a series of > 41 genetically-distinct and metabolically-related, inherited diseases. Indeed, the prevalence of these diseases varied substantially. These values being between 1 per 57,000 and 1 per 4.2 million live births for Gaucher and sialidosis diseases respectively [47]. In such cases, 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. Therefore, the Data Augmentation (DA) approaches outlined here offer great potential in such cases. There are several relatively new studies showing the usefulness of using DA in biomedical problems with non-structured datasets (not images, signals or time series), but to the best of the authors knowledge, none before have shown its application and effect on metabolomics datasets.

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 [48–51]. These studies justify the value offered by NMR-based metabolomics data analysis techniques. Notably, the roles of branched-chain amino acids (BCAAs), the recognition of 3-aminoisobutyrate as a catabolite of BCAA degradation, and the molecular nature of excreted N-acetylated metabolites in NPC diseases was not properly deciphered until more global, untargeted ^{1}H NMR-based metabolomics approaches were employed for this purpose [8,49]. This global approach is slowly but surely developing. The use of composites of both bioanalytical techniques and multivariate statistical and computational intelligence techniques for their solution, is therefore further evolving and becoming more popular [52].

In the present study, the effect of using different state-of-the-art techniques for DA on the prediction performance that can be obtained when metabolomic NPC1 urinary dataset is considered. The results shown in Table 2 can be considered the most relevant results of this work in the improvement of test prediction. Indeed, they exhibit the most effective results obtained. These indicate that the DA can lead to an increase in predictive accuracy, when a logistic regression approach was used as a classifier model. The augmented dataset with the addition of noise reaches approximately a 1% improvement in accuracy compared to the analysis performed on the original dataset. However, given the features of the problem, accuracy is not the most representative metric. Since the dataset is very unbalanced, with 47 samples corresponding to parental heterozygous carriers (controls), and 13 samples corresponding to untreated NPC1 patients (diseased). Thus, it is easy to obtain a high value of accuracy simply by predicting all the control samples alone. Furthermore, the ability of the model not to predict disease samples from the participants' urinary metabolite profiles as if they were control samples implies more value than the accuracy obtained by the model. It is more important that the largest number of patients with the disease be diagnosed as such and not as control patients. Therefore, the most representative prediction metrics for this problem are sensitivity and F_1 score. Table 2 shows that when performing DA, the balanced noise addition method (Noise Bal) and 2000% DA, an approximate 19% improvement in sensitivity and a 6% improvement in F_1 score were obtained.

In order to analyse the effect of DA when using alternative classifier models, the same analysis with random forest (Table 3) and support vector machine systems (Table 4) was performed. The results acquired showed that the DA method with the most valuable test results is Noise Bal and 500% DA. This method gave rise to a ca. 30% improvement in sensitivity, and 16% in F_1 score, when using random forest. With the support vector machine system, however, the model yielded approximately a 49% improvement in sensitivity, along with a 35% improvement in F_1 score.

Results also demonstrate that the best performing method is Noise Bal and 500% DA. This method leads to the highest improvements using random forest and support vector machine systems as classifiers, and the second best improvement was obtained when using logistic regression (12% improvement in sensitivity and 5% in F_1 score). The obtained PLS-DA cross-validation results also support this decision (cf. Fig. 10(d)). The augmented set with involving data created using the Noise Bal model and 500% DA provide approximately a 9% improvement in accuracy, and a 0.3 (out of 1) improvement in predictive capacity (Q^2).

In order to determine the ability of DA methods to replicate metabolic information, an analysis was performed using conventional forms of multivariate data analysis. The PCA and PLS-DA results for component 2 *versus* component 1 clearly show the differences in the generation of data by each method applied. The noise addition methods implemented generate synthetic samples that are grouped around the original samples that they modify. The distance and dispersion of the synthetic samples with respect to the original samples depends on the applied random Gaussian error and the percentage of modified variables. Hence, these methods have a clear capacity to replicate the information if the applied error is not excessive. Indeed, using the SMOTE technique, the replication capacity is based on the interpolation of real samples. In this case, with the original dataset, two separate clusters with a small convergence zone were observed. This fact avoids the disadvantages that creation with the SMOTE method can present.

The results obtained from the analysis of the data augmented with the CGAN model are more interesting, however. These results show the ability of this model to replicate information that fits the distribution of the 'real' samples. However, a disadvantage can also be concluded from these results, and this arises from the small number of samples and their original configuration. Given the internal function of the discriminator network, samples that are far from the core of the distribution ("outliers") are considered as 'false' samples. This implies that when the generator tries to create a synthetic sample close to these outliers, the discriminator discards them. The PLS-DA results are affected by the synthetic samples created by this model; this is clearly shown by the component 2 *versus* component 1 scores plot analysis (Fig. 10) and the VIP value results (Table 5).

A previously conducted study of the dataset analysed, and which involved a series of contemporary multivariate metabolomics analysis techniques, and without any form of DA or expansion strategies, revealed a series of biomarkers which were valuable for distinguishing between the urinary ^{1}H NMR profiles of NPC1 patients and their heterozygous (parental) healthy controls. These included the branched-chain amino acid valine, 3-aminoisobutyrate, quinolinate, succinate, trigonelline, 2-hydroxy-3-methylbuyrate and L-alanine, and those selected and their relative importance rankings were found to be similar to those reported here (Table 5). Similarly, computational intelligence analysis approaches involving genetic algorithm (GA) techniques provided evidence to support these observations. With the exception of one metabolite, trigonelline (a caffeine metabolite), along with some further nicotinate and nicotinamide metabolism pathway intermediates, all of these biomarkers were upregulated in the urinary profiles of NPC1 patients. These studies were conducted using a range of row-wise normalisation preprocessing approaches, including creatinine-, constant sum- and median-normalisation.

From the current study, and the results available in [8], it appears that urinary BCAAs such as valine, the BCAA and thymine degradation product 3-aminoisobutyrate, and perhaps also quinolinate, may be employable as valuable biomarkers for the diagnosis and perhaps even prognostic monitoring of NPC1 disease. However, there is a primary requirement for their validation, and it will be necessary to achieve this through a comparison of these urinary biomarker levels in both untreated and NPC1-selective drug-treated patients. Two moderately successful therapies employed for controlling this disease and attenuating its activity is the orally-administered glucosylceramide synthase inhibitor miglustat, and more recently also the cholesterol-encapsulating agent 2-hydroxypropyl-β-cyclodextrin.

6. Conclusions

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 20%–50%. This, however, depends on the machine learning model used, and increases in the predictive capacity of 0.3 (out of 1) were observed for such models. The establishment of these DA techniques allows the identification of a series of urinary metabolomics biomarkers which will serve to provide valuable information on the diagnosis, pathogenesis, status, and monitoring of the severity of patients with NPC1 disease. In relation to this, several future studies are planned which employ the application of current methods to other metabolomics datasets acquired on the analysis of biofluid samples collected from patients with other diseases, with differential levels of metabolomics information. The aim of such studies should be focused on achieving an improvement in disease diagnosis, along with the search for new metabolomics information and clinically-acceptable biomarkers of interest.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the support from MINECO (Spain) through grants TIN2017-88728-C2-1-R and PID2020-116898RB-I00 (MICINN), from Universidad de Málaga *y* Junta de Andalucía through grant UMA20-FEDERJA-045, and from Instituto de Investigación Biomédica de Málaga – IBIMA (all including FEDER funds). Funding for open access charge: Universidad de Málaga / CBUA .

References

[1] M.T. Vanier, Niemann-pick disease type C, Orphanet J. Rare Dis. 5 (1) (2010) 1–18, http://dx.doi.org/10.1186/1750-1172-5-16.

[2] T. Geberhiwot, A. Moro, A. Dardis, U. Ramaswami, S. Sirrs, M.P. Marfa, M.T. Vanier, M. Walterfang, S. Bolton, C. Dawson, et al., Consensus clinical management guidelines for Niemann-Pick disease type C, Orphanet J. Rare Dis. 13 (1) (2018) 1–19, http://dx.doi.org/10.1016/j.eswa.2017.09.030.

[3] M.B. Winkler, R.T. Kidmose, M. Szomek, K. Thaysen, S. Rawson, S.P. Muench, D. Wüstner, B.P. Pedersen, Structural insight into eukaryotic sterol transport through Niemann-Pick type C proteins, Cell 179 (2) (2019) 485–497, http://dx.doi.org/10.1016/j.cell.2019.08.038.

[4] F.M. Platt, A. d'Azzo, B.L. Davidson, E.F. Neufeld, C.J. Tifft, Lysosomal storage diseases, Nat. Rev. Dis. Primers 4 (1) (2018) 1–25, http://dx.doi.org/10.1038/s41572-018-0025-4.

[5] E. Lloyd-Evans, A.J. Morgan, X. He, D.A. Smith, E. Elliot-Smith, D.J. Sillence, G.C. Churchill, E.H. Schuchman, A. Galione, F.M. Platt, Niemann-Pick disease type C1 is a sphingosine storage disease that causes deregulation of lysosomal calcium, Nature Med. 14 (11) (2008) 1247, http://dx.doi.org/10.1038/nm.1876.

[6] A. Cougnoux, C. Cluzeau, S. Mitra, R. Li, I. Williams, K. Burkert, X. Xu, C. Wassif, W. Zheng, F. Porter, Necroptosis in Niemann–Pick disease, type C1: A potential therapeutic target, Cell Death Dis. 7 (3) (2016) e2147, http://dx.doi.org/10.1038/cddis.2016.16.

[7] M. Grootveld, C.J.L. Silwood, 1H NMR analysis as a diagnostic probe for human saliva, Biochem. Biophys. Res. Commun. 329 (1) (2005) 1–5, http://dx.doi.org/10.1016/j.bbrc.2005.01.112.

[8] V. Ruiz-Rodado, R. Marcos Luque-Baena, D. te Vruchte, F. Probert, R. H Lachmann, C. J Hendriksz, J. E Wraith, J. Imrie, D. Elizondo, D. Sillence, 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, Curr. Metabol. 2 (2) (2014) 88–121.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, http://dx.doi.org/10.1109/cvpr.2016.90.

[10] V. Sandfort, K. Yan, P.J. Pickhardt, R.M. Summers, Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks, Sci. Rep. 9 (1) (2019) 1–9, http://dx.doi.org/10.1038/s41598-019-52737-x.

[11] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, P.R. Pinheiro, CovidGAN: Data augmentation using auxiliary classifier gan for improved covid-19 detection, IEEE Access 8 (2020) 91916–91923, http://dx.doi.org/10.1109/access.2020.2994762.

[12] T.K. Yoo, J.Y. Choi, H.K. Kim, A generative adversarial network approach to predicting postoperative appearance after orbital decompression surgery for thyroid eye disease, Comput. Biol. Med. 118 (2020) 103628, http://dx.doi.org/10.1016/j.compbiomed.2020.103628.

[13] A. Marzullo, S. Moccia, M. Catellani, F. Calimeri, E. De Momi, Towards realistic laparoscopic image generation using image-domain translation, Comput. Methods Programs Biomed. 200 (2021) 105834, http://dx.doi.org/10.1016/j.cmpb.2020.105834.

[14] Z. Liu, H. Zhao, X. Fang, D. Huo, Abdominal computed tomography localizer image generation: A deep learning approach, Comput. Methods Programs Biomed. (2021) 106575, http://dx.doi.org/10.1016/j.cmpb.2021.106575.

[15] R.M. Zur, Y. Jiang, L. Pesce, K. Drukker, Noise injection for training artificial neural networks: A comparison with weight decay and early stopping, Med. Phys. 36 (10) (2009) 4810–4818, http://dx.doi.org/10.1118/1.3213517.

[16] F.J. Moreno-Barea, F. Strazzera, J.M. Jerez, D. Urda, L. Franco, Forward noise adjustment scheme for data augmentation, in: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 2018, pp. 728–734, http://dx.doi.org/10.1109/ssci.2018.8628917.

[17] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357, http://dx.doi.org/10.1613/jair.953.

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

[19] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015, CoRR, arXiv:1511.06434.

[20] G. Douzas, F. Bacao, Effective data generation for imbalanced learning using conditional generative adversarial networks, Expert Syst. Appl. 91 (2018) 464–471, http://dx.doi.org/10.1016/j.eswa.2017.09.030.

[21] F.J. Moreno-Barea, J.M. Jerez, L. Franco, Improving classification accuracy using data augmentation on small data sets, Expert Syst. Appl. 161 (2020) 113696, http://dx.doi.org/10.1016/j.eswa.2020.113696.

[22] Y. Liu, Y. Zhou, X. Liu, F. Dong, C. Wang, Z. Wang, Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology, Engineering 5 (1) (2019) 156–163, http://dx.doi.org/10.1016/j.eng.2018.11.018.

[23] M. Marouf, P. Machart, V. Bansal, C. Kilian, D.S. Magruder, C.F. Krebs, S. Bonn, Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks, Nature Commun. 11 (1) (2020) 1–12, http://dx.doi.org/10.1038/s41467-019-14018-z.

[24] M.T. García-Ordás, C. Benavides, J.A. Benítez-Andrades, H. Alaiz-Moretón, I. García-Rodríguez, Diabetes detection using deep learning techniques with oversampling and feature augmentation, Comput. Methods Programs Biomed. 202 (2021) 105968, http://dx.doi.org/10.1016/j.cmpb.2021.105968.

[25] B. Barile, A. Marzullo, C. Stamile, F. Durand-Dubief, D. Sappey-Marinier, Data augmentation using generative adversarial neural networks on brain structural connectivity in multiple sclerosis, Comput. Methods Programs Biomed. 206 (2021) 106113, http://dx.doi.org/10.1016/j.cmpb.2021.106113.

[26] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, Neurocomputing 321 (2018) 321–331, http://dx.doi.org/10.1016/j.neucom.2018.09.013.

[27] C. Han, L. Rundo, R. Araki, Y. Furukawa, G. Mauri, H. Nakayama, H. Hayashi, Infinite brain MR images: PGGAN-based data augmentation for tumor detection, in: Neural Approaches to Dynamics of Signal Exchanges, Springer, 2020, pp. 291–303, http://dx.doi.org/10.1007/978-981-13-8950-4_27.

[28] Y. Chen, X.-H. Yang, Z. Wei, A.A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, Q. Guan, Generative adversarial networks in medical image augmentation: A review, Comput. Biol. Med. (2022) 105382, http://dx.doi.org/10.1016/j.compbiomed.2022.105382.

[29] K. Açıcı, T. Aşuroğlu, Ç.B. Erdaş, H. Oğul, T4SS effector protein prediction with deep learning, Data 4 (1) (2019) 45, http://dx.doi.org/10.3390/data4010045.

[30] J. Beinecke, D. Heider, Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making, BioData Min. 14 (1) (2021) 1–11, http://dx.doi.org/10.1186/s13040-021-00283-6.

[31] J. Shah, G.N. Brock, J. Gaskins, Bayesmetab: Treatment of missing values in metabolomic studies using a Bayesian modeling approach, BMC Bioinformatics 20 (24) (2019) 1–13, http://dx.doi.org/10.1186/s12859-019-3250-2.

[32] J. Rodrigues, A. Amin, C.R. Raghushaker, S. Chandra, M.B. Joshi, K. Prasad, S. Rai, S.G. Nayak, S. Ray, K.K. Mahato, Exploring photoacoustic spectroscopy-based machine learning together with metabolomics to assess breast tumor progression in a xenograft model ex vivo, Lab. Invest. 101 (7) (2021) 952–965, http://dx.doi.org/10.1038/s41374-021-00597-3.

[33] D.S. Wishart, Y.D. Feunang, A. Marcu, A.C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, et al., HMDB 4.0: The human metabolome database for 2018, Nucleic Acids Res. 46 (D1) (2018) D608–D617.

[34] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232, http://dx.doi.org/10.1109/iccv.2017.244.

[35] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014, CoRR, arXiv:1411.1784.

[36] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, 2017, pp. 214–223.

[37] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, 2017, CoRR, arXiv:1710.10196.

[38] B. Xu, N. Wang, T. Chen, M. Li, Empirical evaluation of rectified activations in convolutional network, 2015, CoRR, arXiv:1505.00853.

[39] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015, pp. 448–456.

[40] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: International Conference on Machine Learning, 2013, pp. 3.

[41] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, CoRR, arXiv:1412.6980.

[42] M. Abadi, A. Agarwal, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, et al., TensorFlow: A system for large-scale machine learning, in: 12th USENIX Symposium on Operating Systems Design and Implementation, 2016, pp. 265–283.

[43] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/a:1010933404324.

[44] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D.S. Wishart, J. Xia, MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis, Nucleic Acids Res. 46 (1) (2018) 486–494, http://dx.doi.org/10.1093/nar/gky310.

[45] E. Szymańska, E. Saccenti, A.K. Smilde, J.A. Westerhuis, Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies, Metabolomics 8 (1) (2012) 3–16, http://dx.doi.org/10.1007/s11306-011-0330-3.

[46] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemom. 29 (10) (2015) 528–536, http://dx.doi.org/10.1002/cem.2736.

[47] P.J. Meikle, J.J. Hopwood, A.E. Clague, W.F. Carey, Prevalence of lysosomal storage disorders, JAMA 281 (3) (1999) 249–254, http://dx.doi.org/10.1001/jama.281.3.249.

[48] V. Ruiz-Rodado, E.-R. Nicoli, F. Probert, D.A. Smith, L. Morris, C.A. Wassif, F.M. Platt, M. Grootveld, 1H NMR-linked metabolomics analysis of liver from a mouse model of NP-C1 disease, J. Proteome Res. 15 (10) (2016) 3511–3527, http://dx.doi.org/10.1021/acs.jproteome.6b00238.

[49] F. Probert, V. Ruiz-Rodado, D. Te Vruchte, E.-R. Nicoli, T.D. Claridge, C.A. Wassif, N. Farhat, F.D. Porter, F.M. Platt, M. Grootveld, NMR analysis reveals significant differences in the plasma metabolic profiles of Niemann Pick C1 patients, heterozygous carriers, and healthy controls, Sci. Rep. 7 (1) (2017) 1–12.

[50] B.C. Percival, M. Gibson, P.B. Wilson, F.M. Platt, M. Grootveld, Metabolomic studies of lipid storage disorders, with special reference to Niemann-Pick type C disease: A critical review with future perspectives, Int. J. Mol. Sci. 21 (7) (2020) 2533, http://dx.doi.org/10.3390/ijms21072533.

[51] B.C. Percival, Y.L. Latour, C.J. Tifft, M. Grootveld, Rapid identification of new biomarkers for the classification of GM1 type 2 gangliosidosis using an unbiased 1H NMR-linked metabolomics strategy, Cells 10 (3) (2021) 572, http://dx.doi.org/10.3390/cells10030572.

[52] D.D. Marshall, R. Powers, Beyond the paradigm: Combining mass spectrometry and nuclear magnetic resonance for metabolomics, Prog. Nucl. Magn. Reson. Spectrosc. 100 (2017) 1–16, http://dx.doi.org/10.1016/j.pnmrs.2017.01.001.