Learning a confidence score and the latent space of a new Supervised Autoencoder for diagnosis and prognosis in clinical metabolomic studies

Michel Barlaud (barlaud@i3s.unice.fr)
Université Côte d'Azur (UCA)
David Chardin
Université Côte d'Azur (UCA)
Cyprien Gille
Université Côte d'Azur (UCA)
Thierry Pourcher
Université Côte d'Azur (UCA)

Research Article

Keywords:

Posted Date: March 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1460785/v1

RESEARCH

Learning a confidence score and the latent space of a new Supervised Autoencoder for diagnosis and prognosis in clinical metabolomic studies.

David Chardin [1,3], Cyprien Gille [2], Thierry Pourcher [1] and Michel Barlaud [2*]

Full list of author information is available at the end of the article

Abstract

Background: Presently, there is a wide variety of classification methods and deep neural networks approaches in bioinformatics. Deep neural networks have proven their effectiveness for classification tasks, and have outperformed classical methods, but they suffer from a lack of interpretability. Therefore, these innovative methods are not appropriate for decision support systems in healthcare. The algorithm should provide the main pieces of information allowing computed diagnosis and prognosis for the final decision by the clinician.

To address this lack of interpretability, we used a new supervised autoencoder (SAE). The main advantage of our supervised autoencoder is its ability to provide a diagnosis with a confidence score for each patient thanks to a softmax classifier, and a meaningful latent space visualization. This confidence score and visual evaluation are crucial for clinical interpretability. Moreover, we used a new efficient feature selection method with a structured constraint for biologically interpretable results.

Experimental results on three metabolomics datasets of clinical samples illustrate the effectiveness of our confidence score diagnostic method:
Results: The supervised autoencoder provides an accurate localization of the patients in the latent space, and an efficient confidence score. Experiments show that the SAE outperforms classical methods (PLS-DA, Random Forests, SVM, and neural networks (NN)). Furthermore, the metabolites selected by the SAE were found to be relevant.

Conclusion: In this paper, we have proposed a new efficient method for diagnosis or prognosis support using clinical metabolomics analyses.

1 Introduction

Deep neural networks have proven their effectiveness in bioinformatics for classification and feature selection [1, 2, 3, 4, 5]. They have also been used recently in metabolomic studies [6, 7, 8, 9, 10].

Autoencoders were introduced within the field of neural networks decades ago, their most efficient application at the time being dimensionality reduction [11, 12]. Autoencoders have been used for denoising different types of data [13], and for lossy image coding

[14], to extract relevant features. Classical stacked autoencoders [13] were used recently in metabolomic studies [15].

An autoencoder is a discriminative model that maps feature points from a high dimensional space to labels in a low dimensional latent space [16, 17]. One of the main advantages of the autoencoder is this projection of the data in the low dimensional latent space. Let us recall that when a model properly learns to construct a latent space, it identifies general features which are relevant for predicting classes.

These autoencoder models include variational autoencoders (VAEs) [18, 19, 20, 21]. VAE networks encourage the latent space to fit a prior distribution, like a Gaussian, which can alter the accuracy of the model. In order to cope with this issue, some recent papers have proposed latent spaces with more complex distributions (e.g. mixtures of Gaussians [22, 23, 24]) on the latent vectors, but they are non-adaptive and unfortunately may not match the specific data distribution. In this work, we relaxed the parametric distribution assumption on the latent space to learn a non-parametric data distribution of clusters [25]. Our network encourages the latent space to fit a distribution learned with the clustering labels rather than a parametric prior.

Recent metabolomic methods using mass spectrometers allow fast and high-resolution detection of massive amounts of metabolites. Therefore, liquid chromatography coupled with high Resolution Mass Spectrometry (LC-MS/MS) based untargeted metabolomics is a very promising omic approach for fundamental research in biology as well as for clinical research applications. A metabolomic analysis is a very sensitive way of revealing biomarkers in physiological or pathological states [26, 27, 28, 29], that could be particularly useful for personalized medicine [30, 31].

In this study, we described in section 2 the SAE method using structured constraints. In section 3 we provide results of experiments on three metabolomic databases. In section 4 we present discussed on these experimental results. Finally, section 5 concludes the paper and provides some perspectives.

2 Method: A New supervised Autoencoder (SAE) framework

2.1 Criterion

Projecting the samples in the lower dimension latent space is crucial to separate them accurately. Herein we propose to use a neural network autoencoder framework.

Let us recall that the encoder part of the autoencoder maps feature points from a high dimensional space to a low dimensional latent space, and that the decoder maps feature points from that low dimensional space to a high dimensional space.

Figure 1 depicts the main constitutive blocks of our proposed approach. We have added to our SAE a ”soft max” block to compute the classification loss.

Let X be the dataset, as an $m \times d$ data matrix made of m line samples $x_1, \ldots, x_m$. Let $y_i = j, j \in [1...k]$ be the label, indicating that the sample x_i belongs to the j-th cluster. Let Z, be the latent space, $\widehat{X}$ the reconstructed data (Figure 1) and W the weights of the neural network.

The goal is to compute the weights W minimizing the total loss, which depends on both the

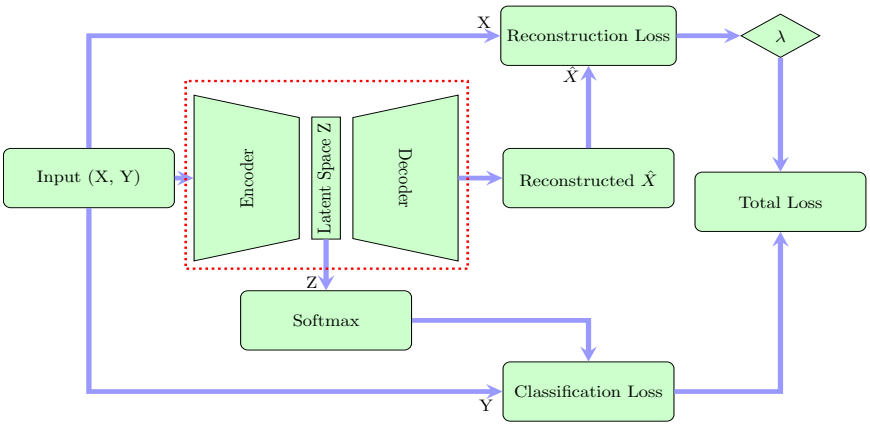

Figure 1 Supervised Autoencoder framework

classification loss and the reconstruction loss. Thus, we propose to minimize the following criterion to compute the weights W of the autoencoder (see [25] for details).

$$Loss(W) = \phi(Z, Y) + \lambda\psi(\widehat{X} - X) \text{ s.t. } \|W\|_1^1 \leq \eta. \quad (1)$$

Where $\phi(Z, Y)$ is the classification loss in the latent space and $\psi(\widehat{X} - X)$ is the reconstruction loss.

The parameter λ controls the weight of the reconstruction loss in the criterion. We used the Cross Entropy Loss for the classification loss function ϕ. We used the robust Smooth ℓ_1 (Huber) Loss [32] as the reconstruction loss function ψ. The dimension of the latent space is defined by the number of clusters.

2.2 Structured constrains, sparsity and feature selection

The basic idea for feature selection is to use a sparse regularizer that forces some coefficients to be zero. To achieve feature selection, classically, the Least Absolute Shrinkage and Selection Operator (LASSO) formulation [33, 34, 35, 36, 37] is used to add an ℓ_1 penalty term to the classification loss. However the LASSO is computationally expensive [34, 35]. Thus, we used a feature selection method by optimizing a criterion under constraints [38]. Let us recall that the classical ℓ_2 norm constraint does not induce any sparsity. Moreover the "group Lasso $\ell_{2,1}$ constraint" induces small sparsity [39] and the ℓ_1 constraint induces unstructured sparsity [40, 41]. Thus we used $\ell_{1,1}$ constrained regularization penalty $\|W\|_1^1 \leq \eta$ for feature selection [25].

2.2.1 Algorithm

We compute the $\ell_{1,1}$ constraint with the following algorithm: we first compute the radius t_i and then project the rows using the ℓ_1 adaptive constraint t_i.

Following the work developed by [42], which proposed a double descent algorithm, we replaced the thresholding by our $\ell_{1,1}$ projection and devised a new double descent algorithm (See Barlaud and Guyard 2020 [43]) as follows :

Algorithm 1 Projection on the $\ell_{1,1}$ norm—$proj_{\ell_1}(V, \eta)$ is the projection on the ℓ_1-ball of radius η, $\nabla\phi(W, M_0)$ is the masked gradient with binary mask M_0, f is the ADAM optimizer, γ is the learning rate

Input: W, γ, η
for $n = 1, \ldots, N(epochs)$ **do**
  $V \leftarrow f(W, \gamma, \nabla\phi(W))$
end for
$t := proj_{\ell_1}((\|v_i\|_1)_{i=1}^d, \eta)$
for $i = 1, \ldots, d$ **do**
  $w_i := proj_{\ell_1}(v_i, t_i)$
end for
Output: W, M_0
Input: W
for $n = 1, \ldots, N(epoch)$ **do**
  $W \leftarrow f(W, \gamma, \nabla\phi(W, M_0))$
end for
Output: W

3 Experimental results

3.1 Settings

3.1.1 Pytorch implementation of our supervised autoencoder

We implemented our sparse supervised autoencoder model in the Pytorch framework. The losses are averaged across observations for each batch. We chose the ADAM optimizer [44], as the standard optimizer in PyTorch. We used the Cross Entropy Loss for the classification loss and the Smooth ℓ_1 Loss (Huber Loss) for the reconstruction loss.

We used the Netbio SAE, a linear fully connected network (LFC), which has an input layer of d neurons, 1 hidden layer of 96 neurons followed by a ReLU activation function, and a latent layer of dimension 2 (the number of classes). The parameter η is determined by the maximum accuracy after cross-validation.

We compared the Netbio SAE with a classical linear fully connected Neural Network (NN) with the same structure.

We used the captum package [45] to compute the feature weights of the SAE.

We provide comparisons with a PLS-DA using 4 components, with Random Forests using 400 estimators and a maximum depth of 3 (using the Gini importance (GI) for feature ranking), and with a support vector classifier (SVM) with a linear kernel. For the SVM, we perform a cross-validation grid search to find the best regularization parameter C.

We provide the statistical evaluation (Accuracy, AUC, and F1 score) using a 4-fold cross validation process. We compare the performances of the different methods using the F1 Score. The F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Note that the F1 score is more relevant than accuracy, especially for unbalanced datasets.

The computation of the statistical metrics, the classifiers, the cross-validation function and the grid search were all provided by the scikit-learn machine learning python package. The python code is available on github: https://github.com/CyprienGille/Supervised-Autoencoder .

3.1.2 Diagnosis with confidence score

One of the main advantages of an autoencoder is the projection of the data in the latent space, which can easily be visualized if the latent space is of dimension 2. [1] Thanks to this, we propose a clinical diagnosis simulation: having trained a network on a database of patients, we can predict a diagnosis with a confidence score for new patients. To perform this simulation, we removed a patient from each of the k classes from the databases. We then trained the SAE on (n-k) patients and we fed the k "test" patients through the best net. We thus obtained a visualization of the projections of these new "test" patients in the latent space as well as their classification with a confidence score (see figures 3, 7 and 5).
The clinician then has a very accurate system to help with the diagnosis. Indeed, in addition to obtaining the confidence score for the diagnosis, the clinician can see where the patient is located among the others in the database and can highlight some of the peculiarities of the pathology that could require additional treatment.

3.2 Metabolomic profiling

The SAE was tested on three different metabolomic datasets : the "LUNG" , "BREAST", and "BRAIN" datasets.
The LUNG dataset was published by Mathe et al [46] and is available at MetaboLights. The BREAST dataset was kindly provided by Dr. Jan Budczies and can be found in the supplementary material of Badczies et al [47]. The BRAIN dataset was obtained through a study performed in our lab[1*].
The characteristics of the three metabolomic datasets are presented in Table 1. We chose to study these databases for their diversity both in terms of the number of features and number of patients, to test the robustness of our method on different types of databases.
The LUNG dataset includes a very large number of patients (1,005), with an equivalently large number of features (2,944), and 2 classes. The BREAST dataset includes a midsize number of patients (271), with a small number of features (161), and 2 classes. The BRAIN dataset includes a limited number of patients (88), with a much higher number of features (7,022), and 2 classes.
The LUNG and BREAST dataset were used without additional filtering. For the BRAIN dataset, metabolomic profiling was performed using MZmine (Version 2.29) [48]. Data obtained from positive and negative ionization modes were analysed separately. Results obtained with each polarity were combined. Only ions that were detected in all the samples after gap-filing were used.

Table 1 Overview of the datasets.

Dataset	No. of Samples	No. of features	Sample type
LUNG	1,005	2,944	Urine
BREAST	271	161	Tumor tissue
BRAIN	88	7,022	Glial tumor tissue

3.3 LUNG dataset

The LUNG dataset was provided by Mathe et al. [46]. This dataset includes metabolomic data concerning urine samples from 469 Non-Small Cell Lung Cancer (NSCLC) patients prior to treatment and 536 controls.

[1] If the latent space is of dimension $k > 2$, we can project the latent space on a 2D plot using a PCA.

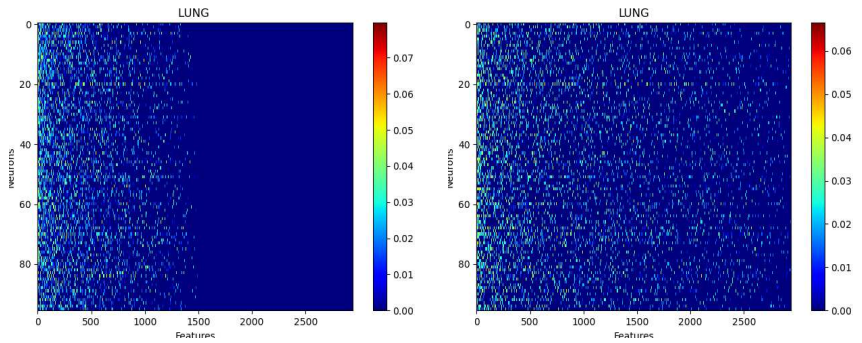

Figure 2 **LUNG** SAE Netbio Matrix: features versus hidden layer:Left with $\ell_{1,1}$ constraint,Right with ℓ_1 constraint

3.3.1 Statistical performances

Table 2 **LUNG** dataset: Accuracy using 3 seeds and 4-fold cross validation: comparison with PLS-DA, Random Forest, SVM and NN

Lung	SAE	PLS-DA	RF	SVM	NN
Accuracy %	81.22	76.56	72.47	76.26	78.27
AUC	80.98	76.85	74.46	78.37	77.94
F1 score	80.74	76.16	71.16	71.11	78.00

As shown in Table 2 our SAE outperformed PLS-DA, Random Forests, SVM and NN by 4.58, 9.58, 9.63 and 2.74% respectively for the F1 score. Note that we checked that increasing the number of trees for Random forests from 100 to 400 resulted in a small improvement in accuracy of only 1% while the computational cost increased by a factor of 3.

3.3.2 Feature selection using the $\ell_{1,1}$ structured constraint

Figure 2 shows the matrix ($d \times n$) of the network connections between the input layer (d feature neurons) and the hidden layer (n neurons). It shows the benefit of using the $\ell_{1,1}$ constraint: The $\ell_{1,1}$ constraint selects features while the constraint ℓ_1 selects only weights of features. All the following results are given with the $\ell_{1,1}$ constraint.

Table 3 Top 5 features on the **LUNG** dataset. From left to right: SAE, PLS-DA, Random Forest, SVM and NN

SAE	PLS-DA	Random Forest	SVM	NN
MZ 264.12	**MZ 264.12**	**MZ 264.12**	MZ 170.06	**MZ 264.12**
MZ 308.09	MZ 126.90	MZ 441.16	MZ 126.90	MZ 126.90
MZ 126.90	MZ 613.35	MZ 584.26	**MZ 264.12**	**MZ 308.09**
MZ 232.03	MZ 170.06	MZ 486.25	MZ 94.06	MZ 613.35
MZ 332.09	MZ 243.10	MZ 204.13	MZ 110.99	MZ 332.09

As shown in Table 3, all methods selected metabolite ”MZ 264.121”, which most likely corresponds to creatine riboside (expected m/z value in the positive mode: 264.1190). Note that the SVM selected metabolite ”MZ 264.121” at rank 3. Metabolite ”MZ 308.098”, which most likely corresponds to N-acetylneuraminic acid, was only selected by the SAE and the NN at rank 2 and 3, respectively. These metabolites were described by Mathé et al. [46] as the most important metabolites to discriminate between lung cancer patients and healthy individuals. Note that the author of RF proposes two measures for feature ranking, the variable importance (VI) and Gini importance (GI): a recent study showed that if predictors

are categorical, or real with multimodal Gaussian distributions, both measures are biased [49].

3.3.3 Diagnosis in the latent space with a confidence score

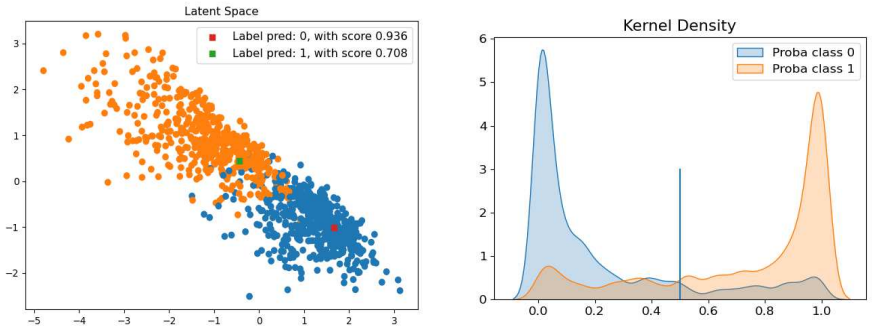

Figure 3 LUNG dataset. Right: Latent space, with test patients as squares. Left: Distribution using a Gaussian kernel

As shown in Figure 3, the two classes are mostly separated in the latent space of the SAE. Furthermore, the red and green squares show the location of the two random ”test” patients in the SAE's latent space. The red patient is at the heart of the class distribution and the green patient is close to the edge of the other class. This is important for the clinician's assessment of the result. Moreover, the distribution plot shows the perfect separability of the distributions calculated with the SAE, which means that most of the patients were accurately diagnosed with a high degree of confidence. The patient represented by the red square was classified in class 0 with a confidence score of 0.94 and the patient represented by the green square was labeled class 1 with a confidence score of 0.70. Both predicted labels were correct.

3.4 BREAST cancer dataset

The BREAST dataset was published by Dr Budczies et al from Charity Hospital Berlin [47]. 204 of these breast cancers were Estrogen Receptor positive (ER+) and 67 Estrogen Receptor Negative (ER-). Expression of the ER, as well as progesterone receptor and HER2 protein are commonly used to define the molecular subtypes of breast cancers. ER positive is a predictive marker for anti-hormonal therapy. The dataset includes a total of 162 metabolites were analyzed using untargeted Gas Chromatography followed by Time of Flight Mass spectrometry (GC–TOFMS). Here, we studied the performances of the supervised autoencoder framework to classify breast cancer patients by ER expression using the same clinical metabolomic dataset.

3.4.1 Statistical performances

As shown in Table 4 our SAE outperformed PLS-DA, Random Forests, SVM and NN by 9.16, 14.1, 9.11 and 2.23% respectively for the F1 score.

Table 4 BREAST dataset: Accuracy using 3 seeds and 4-fold cross validation: comparison with PLS-DA, Random Forest, Logistic Regression, SVM and NN

Breast	SAE	PLS-DA	RF	SVM	NN
Accuracy %	90.15	86.58	80.23	83.20	89.04
AUC %	84.88	83.07	88.02	77.64	80.34
F1 Score	85.17	76.01	71.07	76.06	82.94

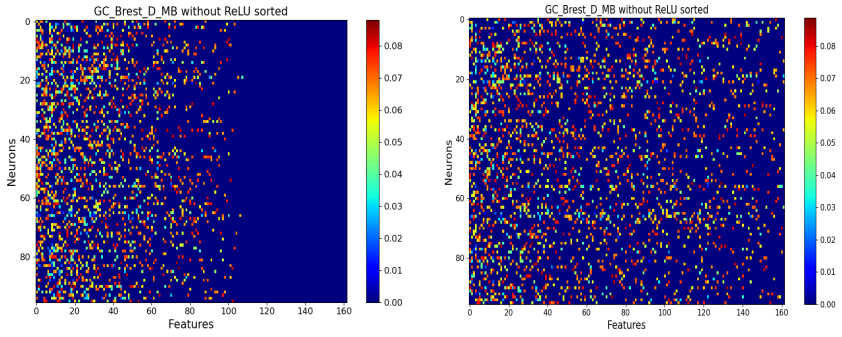

Figure 4 BREAST SAE Netbio Matrix: features versus hidden layer:Left with $\ell_{1,1}$ constraint,Right with ℓ_1 constraint

3.4.2 Feature selection using the $\ell_{1,1}$ structured constraint

Figure 4 shows the matrix ($d \times n$) of the network connections between the input layer (d feature-neurons) and the hidden layer (n neurons). It shows the benefit of using the $\ell_{1,1}$ constraint: The $\ell_{1,1}$ constraint selects features, while the constraint ℓ_1 selects only weights of features.

Table 5 Top 5 features on the **BREAST** dataset. From left to right: SAE, PLS-DA, Random Forest, SVM and NN

SAE	PLS-DA	Random Forest	SVM	NN
beta-alanine	beta-alanine	beta-alanine	3-phosphoglycerate	beta-alanine
xanthine	xanthine	xanthine	beta-alanine	xanthine
uracil	nicotinamide	glutamic acid	uracil	2-hydroxyglutaric
glutamic acid	isothreonic acid	idonic acid NIST	taurine	uracil
2-hydroxyglutaric acid	creatinine	uracil	2-ketoadipic acid	glutamic acid

As shown in Table 5, the SAE and the NN selected the same top five metabolites (beta-alanine, xanthine, uracil, glutamic acid). These metabolites have been already identified for significantly different concentrations in ER breast compared to ER+ breast cancer in the original metabolomics paper[47]. Increased concentrations of glutamic acid and 2-hydroxyglutaric acid indicate higher glutaminolysis, a key feature of metabolic changes in cancer cells. As shown in Budczies et al [47], increased concentrations of uracil, xanthine and beta-alanine levels are related to higher hexokinase 3, xanthine dehydrogenase and 4-aminobutyrate aminotransferase expressions, respectively.

3.4.3 Prognosis in the latent space with confidence score

Figure 5 (left), shows the accurate separation of the two classes in the latent space of the SAE. The red and green squares show the location of the two random ”test” patients in the SAE’s latent space. The patient represented by the red square was classified in class 0 with a confidence score of 0.55 and the patient represented by the green square was labeled class 1 with a confidence score of 0.80. Both predictions are correct. Figure 5 (right) shows the separability of the distributions calculated with the SAE.

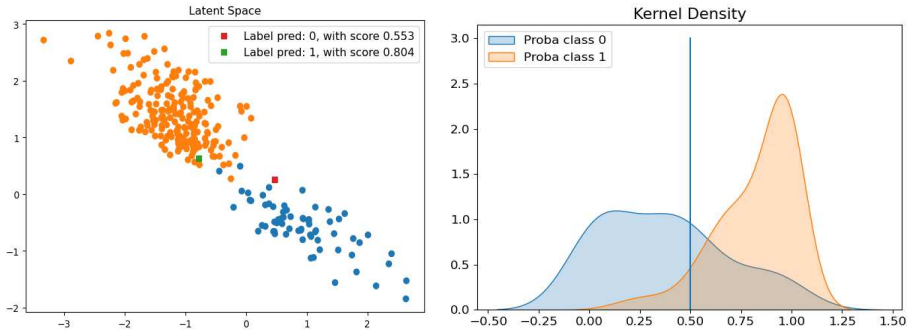

Figure 5 BREAST dataset. Left: latent space of the SAE. Right: Distribution using a Gaussian Kernel

3.5 BRAIN cancer dataset

The BRAIN dataset samples were retrospectively collected from two declared biobanks from the Central Pathology Laboratory of the Hospital of Nice and from the Center of Biological Resources of Montpellier (Plateforme CRB-CHUM). For each participant, the IDH mutational status was assessed using immunohistochemistry and pyrosequencing for immunonegative cases. This cohort is composed of 88 patients, 38 patients with mutated IDH and 50 with wild-type IDH. The dataset includes 7,022 features.

3.5.1 Statistical performances

Table 6 BRAIN dataset Accuracy using 3 seeds and 4-fold cross validation: comparison with PLS-DA, Random Forest , SVM and NN

Brain	SAE	PLS-DA	RF	SVM	NN
Accuracy %	92.80	84.84	86.73	87.12	75.75
AUC %	93.29	85.37	89.5	87.52	74.85
F1 score	92.66	83.88	88.05	86.51	74.19

Table 6 shows that, despite the small number of patients, the supervised autoencoder outperformed PLS-DA, Random Forest, SVM and NN by 8.78, 4.61, 6.15 and 18.47% respectively for the F1 score. For this base with few patients the performance of NNs collapses as reported in the literature.

3.5.2 Feature selection using the $\ell_{1,1}$ structured constraint

Figure 6 shows the matrix ($d \times n$) of the network connections between the input layer (d feature-neurons) and the hidden layer (n neurons). It shows the benefit of using the $\ell_{1,1}$ constraint: The $\ell_{1,1}$ constraint selects features, while the constraint ℓ_1 selects only weights of features.

As expected, the top features selected by each method (shown in Table 7) correspond mainly to different isotopes and adducts of 2-hydroxyglutarate (marked in bold). The features selected using the SAE were all different adducts of this specific product of IDH-mutated cells. Indeed, POS_MZ132.03 and POS_MZ131.03 correspond to the [M+H-H2O]+ adduct of 2-hydroxyglutarate with one 13C isotope for the first ion. POS_MZ171.02 is the [M+Na]+ adduct, NEG_MZ147.02 is the [M-H]- and POS_MZ86.03 is the [M+Na+H]2+ adduct. NEG_MZ148.03 is the [M-H]- adduct of 2-hydroxyglutarate with one 13C isotope.

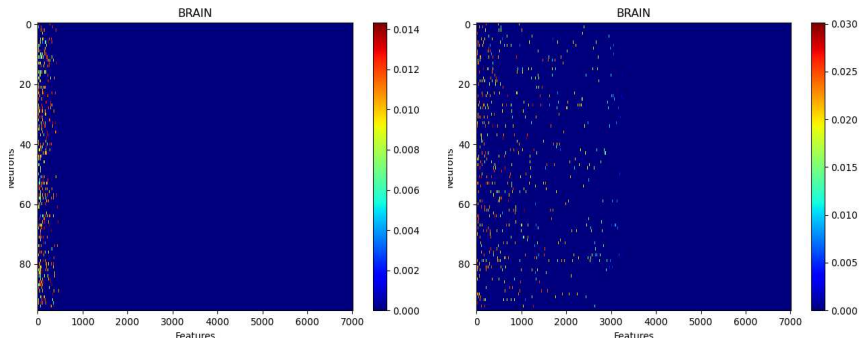

Figure 6 **BRAIN** SAE Netbio Matrix: features versus hidden layer: Left with $\ell_{1,1}$ constraint, Right with ℓ_1 constraint

Table 7 **BRAIN** dataset with 7,022 features : Top 8 features selected by the SAE, PLS-DA, Random Forests, SVM and NN

SAE	PLS-DA	RF	SVM	NN
NEG MZ147.02	**POS MZ131.03**	**NEG MZ148**	**POS MZ131**	**NEG MZ148.03**
POS MZ132.0374	**POS MZ132.03**	**POS MZ86**	**POS MZ132**	**NEG MZ147.02**
POS MZ171.02	**NEG MZ147.02**	NEG MZ215	**NEG MZ147**	**POS MZ132.03**
NEG MZ148.03	**NEG MZ148.03**	**NEG MZ147**	**NEG MZ148**	**POS MZ85.02**
POS MZ132.0375	**POS MZ171.02**	**POS MZ132**	**POS MZ171**	**POS MZ173.03**
POS MZ85.02	**POS MZ132.03**	**POS MZ149**	**POS MZ132**	**POS MZ171.02**
POS MZ86.03	POS MZ85.02	POS MZ132	**POS MZ173**	POS MZ132.52
POS MZ149	POS MZ132.52	**POS MZ171**	**POS MZ132**	POS MZ131.03

POS_MZ173.03 is the [M+Na]+ adduct with two 13C isotope. Finally, POS_MZ149.04 is the [M+H]+ adduct ion of 2-hydroxyglutarate.

3.5.3 Diagnosis in the latent space with confidence score

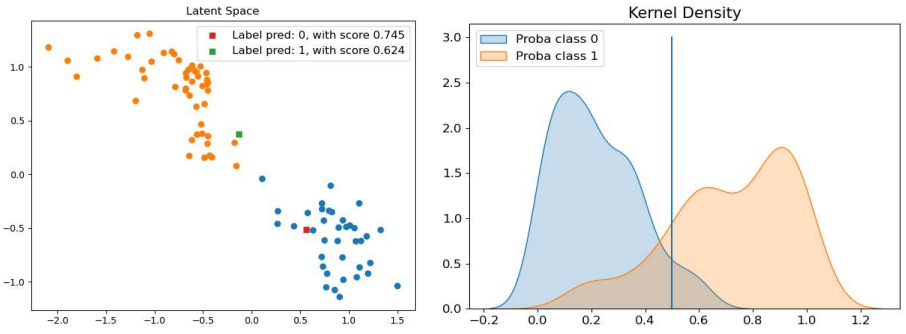

Figure 7 **BRAIN** dataset. Left: latent space of the SAE, Red and green squares are ”test” patients. Right: Distribution using a Gaussian kernel

Figure 7 (left), shows the perfect separation of the two classes in the latent space of the SAE. Furthermore, the red and green squares show the location of the two random ”test” patients in the SAE’s latent space. The patient represented by the red square was classified in class 0 with a confidence score of 0.75 and the patient represented by the green square was labeled class 1 with a confidence score of 0.62. Both predictions were correct. Figure

7 (right) shows the peak separability of the distributions calculated with the SAE, which means that most patients will have a good prediction with a high degree of confidence score.

4 Discussion

Our results led us to the following conclusions :

- Unlike classical autoencoders that assume a latent space modeling [23, 24], we have used a supervised autoencoder (SAE). The real distributions of many datasets, including metabolomics datasets, are far from multi-gaussian mixtures. Therefore, using a non-parametric autoencoder seems more appropriate than forcing a multi-gaussian distribution upon the data.
- Regardless of data size and feature space dimensions, the SAE outperforms all other methods (PLS-DA, Random Forests, SVM and NN). Note that the NN also outperformed classical methods (PLS-DA, Random Forests and SVM), except on small databases.
- The trained SAE can already be used to accurately classify new patients (see figures 3, 5 and 7).
- The SAE provides high-level distribution visualization of the patients to be diagnosed in the latent space, as well as their classification confidence score, which is crucial for any computerized decision support tool based on clinical metabolomic data. Indeed, these two features can greatly help a clinician in routine patient care. For instance, if a patient is visualized at the edge of the group, a clinician may want to take extra caution - a nuance only provided by the SAE.
- Furthermore, one should note that metabolomics is a very promising approach, particularly adapted to routine clinical practice, since metabolomics analyses are fast and relatively inexpensive.
- Interestingly, the SAE combined with a structured projection provides efficient feature selection (Tables 3, 5 and 7). Better yet, we have verified that the selected features in the LUNG, BREAST and BRAIN datasets were known to be biologically relevant metabolites. Efficient feature selection adds interpretability to the model which is crucial for metabolomic studies in biological research or clinical trials.

5 Conclusion

In this paper we have proposed a new and efficient diagnosis method for metabolomics datasets, based on the representation of data on the latent space of a new supervised autoencoder (SAE). In clinical applications, our method provides a diagnosis score to each patient's predicted class. Moreover, from a statistical point of view (Accuracy, AUC, F1 score) our SAE outperformed PLS-DA, Random Forest, SVM, and NN while selecting biologically relevant features.

6 Declarations

Ethics approval and consent to participate

LUNG dataset

The LUNG dataset was pulished by Mathe et al [] and available at MetaboLights database.

BREAST dataset

The BREAST dataset was kindly provided by Dr. Jan Budczies and could be find in the supplementary material of Badczies et al [47].

BRAIN dataset

The samples from the BRAIN dataset were retrospectively collected from two declared biobanks. The authors confirm that all methods were carried out in accordance with relevant guidelines and regulations.

The authors confirm that informed consent was obtained from all subjects.

The authors confirm that this retrospective study has been approved by the institutional ethics committees of the University Hospital of Nice and the University Hospital of Montpellier.

Consent for publication

Not applicable.

Availability of data and materials

We implemented the diagnostic and statistical analysis code with python. Functions and scripts are freely available at https://github.com/CyprienGille/Supervised-Autoencoder .

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been supported by the French government, through UCA-JEDI Investment in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01. Equipment for this study was purchased through grants from the Recherche en matières de Sûreté Nucléaire et Radioprotection program from the French National Research Agency and the Conseil Départemental 06.

Acknowledgments

The authors thank Dr Jean-Marie Guigonis (Bernard Rossi facility) for the LC-MS analyses. The authors thank Pr Fanny Burel-Vandenbos and Valerie Rigau for providing the samples of the Brain dataset.

Authors' information

[1]Transporters in imaging and Radiotherapy in Oncology (TIRO), Direction de la Recherche Fondamentale (DRF), Institut des sciences du vivant Fréderic Joliot, Commissariat à l'Energie Atomique et aux énergies alternatives (CEA), Université Côte d'Azur (UCA), Nice, France.

[2]Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis (I3S), Université Côte d'Azur (UCA), Centre de Recherche Scientifique (CNRS), Sophia Antipolis, France.

[3] Centre Antoine Lacassagne, Université Côte d'Azur (UCA), Nice, France.

*Correspondence to barlaud@i3s.unice.fr

Author details

References

1. Song, Y., Zheng, S., Li, L., Zhang, X., Zhang, X., Huang, Z., Chen, J., Wang, R., Zhao, H., Zha, Y., Shen, J., Chong, Y., Yang, Y.: Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2021)
2. Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G.-Z.: Deep learning for health informatics. IEEE Journal of Biomedical and Health Informatics **21**(1), 4–21 (2017)
3. Min, S., Lee, B., Yoon, S.: Deep learning in bioinformatics. Briefings in Bioinformatics **18**(5), 851–869 (2016)
4. Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., Tao, Y., Guo, Y., Ni, X., Shi, T.: Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Frontiers in Genetics **9**, 477 (2018)
5. Sen, P., Lamichhane, S., Mathema, V.B., McGlinchey, A., Dickens, A.M., Khoomrung, S., Orešič, M.: Deep learning meets metabolomics: a methodological perspective. Briefings in Bioinformatics (2020)
6. Alakwaa, F., Chaudhary, K., Garmire, L.: Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. Journal of Proteome Research, **17**, 337–347 (2018)
7. Bradley, W., Robert, P.: Multivariate analysis in metabolomics. Current Metabolomics **1**, 92–107 (2013)
8. Asakura, P., Date, Y., Kikuchi, J.: Application of ensemble deep neural network to metabolomics studies. Analytica Chimica Acta **1037**, 92–107 (2018)
9. Mendez, K., Broadhurst, D., Reinke, S.: Application of artificial neural networks in metabolomics: A historical perspective. Metabolomics **15** (2019)
10. Sen, P., Lamichhane, S., Mathema, V.B., McGlinchey, A., Dickens, A.M., Khoomrung, S., Orešič, M.: Deep learning meets metabolomics: a methodological perspective. Briefings in Bioinformatics **22**(2), 1531–1542 (2020)
11. Hinton, Z.R. Geoffrey: Autoencoders, minimum description length and helmholtz free energy. In: Advances in Neural Information Processing Systems, pp. 3–10 (1994)
12. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning vol. 1. MIT press, ??? (2016)
13. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**, 3371–3408 (2010)
14. Theis, L., Shi, W., Cunningham, A., Huszár, F.: Lossy image compression with compressive autoencoders. arXiv stat.ML /1703.00395 (2017)
15. Xiaojing, F., Xiye, W., Mingyang, J., Zhili, P., Shicheng, Q.: An improved stacked autoencoder for metabolomic data classification. Hindawi Computational Intelligence and Neuroscience **2021** (2021)
16. Hinton, G., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
17. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
18. Kingma, D., Welling, M.: Auto-encoding variational bayes. International Conference on Learning Representation (2014)
19. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Xing, E.P., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning, vol. 32, pp. 1278–1286. PMLR, ??? (2014)
20. Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O.: Ladder variational autoencoders. arXiv stat.ML /1602.02282 (2016)
21. Kingma, D., Mohamed, S., Jimenez Rezende, D., Welling, M.: Semi-supervised learning with deep generative models. Neural Information Processing Systems conference, 3581–3589 (2014)
22. Dilokthanakul, N., Mediano, P.A.M., Garnelo, M., Lee, M.C.H., Salimbeni, H., Arulkumaran, K., Shanahan, M.: Deep unsupervised clustering with gaussian mixture variational autoencoders (2016). 1611.02648
23. Emdadi, A., Eslahchi, C.: Auto-HMM-LMF: feature selection based method for prediction of drug response via autoencoder and hidden Markov model. BMC Bioinformatics (2021)
24. Liu, D., Huang, Y., Nie, W., Zhang, J., Deng, L.: SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost. BMC Bioinformatics **22** (2021)
25. Barlaud, M., Guyard, F.: Learning a sparse generative non-parametric supervised autoencoder. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, TORONTO , Canada (2021)

26. Yazdani, H., Cheng, L.L., Christiani, D.C., Yazdani, A.: Bounded fuzzy possibilistic method reveals information about lung cancer through analysis of metabolomics. IEEE/ACM Transactions on Computational Biology and Bioinformatics **17**(2), 526–535 (2020)

27. Liu, Y., Xu, X., Deng, L., Cheng, K.-K., Xu, J., Raftery, D., Dong, J.: A novel network modelling for metabolite set analysis: A case study on crc metabolomics. IEEE Access **8**, 106425–106436 (2020)

28. Banimustafa, A., Hardy, N.: A scientific knowledge discovery and data mining process model for metabolomics. IEEE Access **8**, 209964–210005 (2020)

29. Qi, Z., Voit, E.O.: Strategies for comparing metabolic profiles: Implications for the inference of biochemical mechanisms from metabolomics data. IEEE/ACM Transactions on Computational Biology and Bioinformatics **14**(6), 1434–1445 (2017)

30. Long, N.P., Nghi, T.D., Kang, Y.P., Anh, N.H., Kim, H.M., Park, S.K., Kwon, S.W.: Toward a Standardized Strategy of Clinical Metabolomics for the Advancement of Precision Medicine. Metabolites **10**(2), 51 (2020). doi:10.3390/metabo10020051. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. Accessed 2020-12-01

31. Cakmak, A., Celik, M.H.: Personalized metabolic analysis of diseases. IEEE/ACM Transactions on Computational Biology and Bioinformatics **18**(3), 1014–1025 (2021)

32. Huber, P.J.: Robust statistics. 1981

33. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288 (1996)

34. Hastie, T., Rosset, S., Tibshirani, R., Zhu, J.: The entire regularization path for the support vector machine. Journal of Machine Learning Research **5**, 1391–1415 (2004)

35. Friedman, J., Hastie, T., Tibshirani, R.: Regularization path for generalized linear models via coordinate descent. Journal of Statistical Software **33**, 1–122 (2010)

36. Hastie, T., Tibshirani, R., Wainwright, M.: Statistcal learning with sparsity: The lasso and generalizations. CRC Press (2015)

37. Li, J., Cheng, K., Wang, S., Morstatter, F., P. Trevino, R., Tang, J., Liu, H.: Feature selection: A data perspective. ACM Computing Surveys **50** (2016). doi:10.1145/3136625

38. Barlaud, M., Belhajali, W., Combettes, P., Fillatre, L.: Classification and regression using an outer approximation projection-gradient method, vol. 65, pp. 4635–4643 (2017)

39. Barlaud, M., Chambolle, A., Caillau, J.-B.: Classification and feature selection using a primal-dual method and projection on structured constraints. International Conference on Pattern Recognition, Milan (2020)

40. Condat, L.: Fast projection onto the simplex and the l1 ball. Mathematical Programming Series A **158**(1), 575–585 (2016)

41. Perez, G., Barlaud, M., Fillatre, L., Régin, J.-C.: A filtered bucket-clustering method for projection onto the simplex and the ℓ_1-ball. Mathematical Programming (2019)

42. Zhou, H., Lan, J., Liu, R., Yosinski, J.: Deconstructing lottery tickets: Zeros, signs, and the supermask. In: Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 3597–3607. Curran Associates, Inc., ??? (2019)

43. Barlaud, M., Guyard, F.: Learning sparse deep neural networks using efficient structured projections on convex constraints for green ai. International Conference on Pattern Recognition, Milan (2020)

44. Kingma, D., Ba, J.: a method for stochastic optimization. International Conference on Learning Representations, 1–13 (2015)

45. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Neural Information Processing Systems, Barcelone, Spain **30** (2017)

46. Mathé *et al*, E.: Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. Cancer research **74**(12), 3259–3270 (2014)

47. Budczies, J., Brockmöller, S., Müller, B., Barupal, D., Richter-Ehrenstein, C., Kleine-Tebbe, A., Griffin, J., Orešič, M., Dietel, M., Denkert, C., Fiehn, O.: Comparative metabolomics of estrogen receptor positive and estrogen receptor negative breast cancer: Alterations in glutamine and beta-alanine metabolism. Journal of Proteomics **94**, 279–288 (2013)

48. Pluskal, T., Castillo, S., Villar-Briones, A., Oresic, M.: MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC bioinformatics **11**, 395 (2010). doi:10.1186/1471-2105-11-395

49. Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**(10), 1340–1347 (2010)