mathematics

Article

A Deep Learning Approach to Analyze NMR Spectra of SH-SY5Y Cells for Alzheimer's Disease Diagnosis

Filippo Costanti [1,2,*,†], Arian Kola [3,†], Franco Scarselli [1], Daniela Valensin [3] and Monica Bianchini [1]

[1] Department of Information Engineering and Mathematics, University of Siena, 53100 Siena, Italy
[2] Department of Information Engineering, University of Florence, 50139 Firenze, Italy
[3] Department of Biotechnology, Chemistry, and Pharmacy, University of Siena, 53100 Siena, Italy; arian.kola@unisi.it (A.K.)
[*] Correspondence: filippo.costanti@unifi.it; Tel.: +39-347-9880136
[†] These authors contributed equally to this work.

Abstract: The SH-SY5Y neuroblastoma cell line is often used as an in vitro model of neuronal function and is widely applied to study the molecular events leading to Alzheimer's disease (AD). Indeed, recently, basic research on SH-SY5Y cells has provided interesting insights for the discovery of new drugs and biomarkers for improved AD treatment and diagnosis. At the same time, untargeted NMR metabolomics is widely applied to metabolic profile analysis and screening for differential metabolites, to discover new biomarkers. In this paper, a compression technique based on convolutional autoencoders is proposed, which can perform a high dimensionality reduction in the spectral signal (up to more than 300 times), maintaining informative features (guaranteed by a reconstruction error always smaller than 5%). Moreover, before compression, an *ad hoc* preprocessing method was devised to remedy the scarcity of available data. The compressed spectral data were then used to train some SVM classifiers to distinguish diseased from healthy cells, achieving an accuracy close to 78%, a significantly better performance with respect to using standard PCA-compressed data.

Keywords: Alzheimer's disease; SH-SY5Y cells; nuclear magnetic resonance (NMR); convolutional autoencoders; embedding of NMR spectra

MSC: 68T07

check for updates

Citation: Costanti, F.; Kola, A.; Scarselli, F.; Valensin, D.; Bianchini, M. A Deep Learning Approach to Analyze NMR Spectra of SH-SY5Y Cells for Alzheimer's Disease Diagnosis. *Mathematics* **2023**, *11*, 2664. https://doi.org/10.3390/math11122664

Academic Editor: Jianping Gou

Received: 9 May 2023
Revised: 8 June 2023
Accepted: 10 June 2023
Published: 12 June 2023

1. Introduction

Neurodegenerative diseases (NDs) are severe neurological disorders affecting millions of people worldwide with no effective treatments so far. NDs are characterized by a complex network of pathological events leading to oxidative stress, mitochondrial dysfunction, metal ion dyshomeostasis, protein misfolding, and neuroinflammation [1]. All these factors are strongly interdependent and, although playing a key role in neurodegeneration etiology, their implications as causes or consequences have not yet been clarified. Alzheimer's disease (AD) accounts for 60–80% of dementia cases, which are treated with few medicines available to relieve the early symptoms. The lack of treatment has led researchers to evaluate as many molecules as possible in order to discover active ingredients [2,3]. These preliminary screenings were carried out using in vitro neuronal models. The most useful model for such experiments is the SH-SY5Y cell line, derived from human neuroblastoma [4,5]. Previous studies have shown a comparison between different types of neuronal models demonstrating the many advantages offered by SH-SY5Y cells: they are human-derived, do not require animal testing, they are very proliferative, and their cost of cultivation is low [6].

AD is characterized by a slow progression, making early diagnosis difficult. Moreover, the small memory problems that characterize the early stages of AD are often misidentified as a natural consequence of aging. For these reasons, in recent years, the scientific community has put a lot of effort into the discovery of reliable biomarkers for early AD diagnosis. In this scenario, NDs need to be fully investigated, combining multidisciplinary approaches for the characterization of the corresponding metabolic features. NMR metabolomics allows us to identify the entire metabolite spectrum in many biological fluids and cellular extract/medium, obtaining a metabolic fingerprint (metabolic profile) that characterizes the healthy or diseased state [7]. NMR spectroscopy, compared to other techniques, has the advantage of being rapid and not requiring a particular treatment of the samples [8].

Data produced by spectroscopic investigations are diversified and are usually combined with multivariate statistical analysis [9]. Indeed, one of the main problems of NMR-related tasks is the huge number of features per spectrum to be analyzed, while the available spectra often amount to small numbers. As a consequence, both dimensionality reduction and data augmentation are normally needed to automatically classify spectral data. The main technique used for the former task is Principal Component Analysis (PCA), which allows a linear dimensionality reduction. PCA is an unsupervised multivariate statistical framework based on projecting data along the most informative directions in the eigenvector space of the feature covariance matrix. While widely used, PCA may not be appropriate for complex and highly nonlinear data such as NMR spectra. In fact, recent applications of deep learning (DL) methods to NMR data opened a new reliable path to guarantee information-conservative feature selection [10–13].

Actually, the application of DL models to the early detection and automated classification of AD has recently gained considerable attention, as rapid progress in neuroimaging techniques has generated large-scale multimodal data [14,15]. Based on brain NMR images, for instance, the gray-to-white matter ratio can be evaluated [16], while some forms of degeneration can be revealed, to highlight the premature brain aging typical of AD. Indeed, in the brain of AD patients, the deposition of the amyloid protein and the death of neurons, located in the gray matter, are observed. However, radiological examinations also show damage to the white matter, that part of the brain that is instead mainly made up of myelin, a substance that surrounds the neurons, facilitating their communication. In addition to deep learning techniques applied to neuroimaging, models have been implemented that are capable of integrating flexible combinations of routinely collected clinical information, including demographics, medical history, neuropsychological testing, and functional assessments [17]. Such DL tools have been shown to often outperform the diagnostic accuracy of practicing neurologists and neuroradiologists. Finally, attention modules, able to produce saliency maps, can give rise to interpretable methods for cerebral images, showing that the detected disease-specific patterns track distinct patterns of degenerative changes throughout the brain and correspond closely with the presence of neuropathological lesions on autopsy. For the non-invasive prediction of Alzheimer's disease, DL techniques have also been applied to retinal images [18]. In fact, a reduction in the density of the capillary network around the center of the macula has been observed in AD patients. To early diagnose AD, retinal images were captured in different modalities (optical coherence tomography, optical coherence tomography-angiography, ultra-widefield retinal photography, and retinal autofluorescence) and integrated with patient data [19].

In this work, we address the problem of the early diagnosis of AD from a new perspective. A DL approach, an alternative to PCA and based on convolutional autoencoders (CAEs), is presented. A targeted procedure for NMR spectral data compression is also presented in order to deal with the scarcity of data. Apart from the novelty inherent in the application of DL methodologies to this problem, the results of the downstream classification phase, needed for discriminating healthy and diseased cells, are very promising. To the best of our knowledge, this is the first time that such an automatic protocol has been used for the analysis of NMR spectra of SH-SY5Y cells.

In summary, the main contributions of this work are the following:

1. An entire pipeline to process the NMR spectral data of the SH-SY5Y cell line which includes:

- A *biologically-informed* preprocessing;
- A DL approach based on CAEs, an alternative to PCA, to automatically compress NMR spectra, represented as vectors of intensity values.

2. A set of comparative experiments—showing the superiority of our DL compression method with respect to PCA—to classify healthy/diseased cells for the early diagnosis of AD.

The rest of the paper is organized as follows. Section 2 contains a description of the specific in vitro experiments and of the collected dataset, together with an explanation of the whole processing workflow. In Section 3, the obtained results are presented, showing the advantage of using a DL approach instead of PCA for data dimensionality reduction. Finally, Section 4 collects some conclusions and traces future perspectives.

2. Materials and Methods

In this section, we focus first on the wet lab experimental setting used to obtain our dataset, from the cell culture to the NMR spectroscopical analysis of the metabolite content of the cells. Then, the collected dataset of spectra is described. Finally, we present the DL method used to embed the NMR spectra and to classify them as related to healthy or diseased cells, as well as the evaluation metrics. The key steps of the approach used in this study are schematically represented in Figure 1.

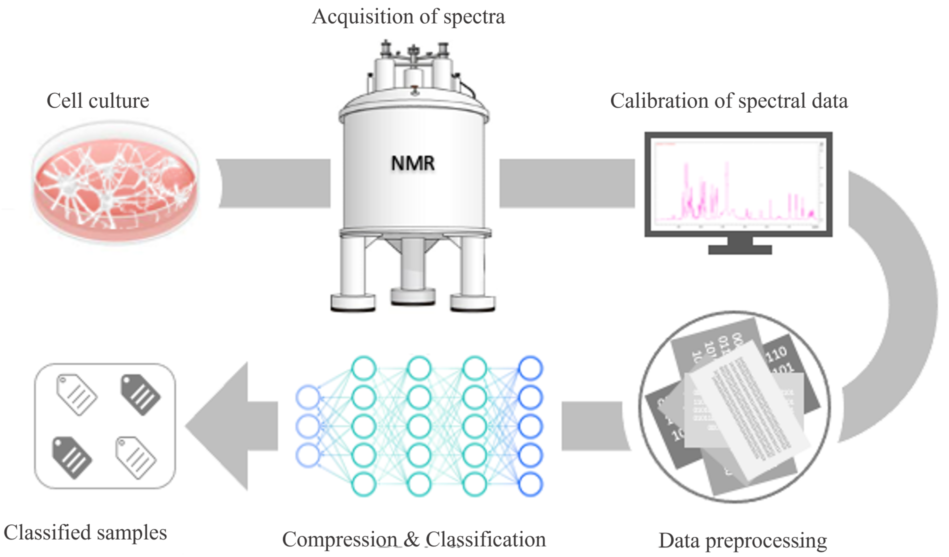


Figure 1. Processing workflow: From wet lab experiments to healthy/diseased cell classification.

2.1. NMR Experiments

The profile associated with a cell's health or disease state is characterized by a specific set of metabolites that can be determined using NMR spectroscopy, a powerful technique that can identify and quantify molecules in solutions. Indeed, NMR metabolomics studies have made it possible to map the content of metabolites in different biological fluids and/or cellular media in order to correlate specific NMR signatures with health or disease conditions [20,21]. Usually, NMR data are combined with statistical data that can classify NMR spectra based on metabolic profile.

The NMR samples analyzed in this work were obtained by collecting the cellular media of differentiated SH-SY5Y neuroblast-like cells treated with different concentrations of several compounds, including the peptide Amyloid β 1–42 and the galantamine and lycorine alkaloids. The cellular experiments were performed according to the procedure reported in [22]. The cellular vitality, determined by the Neutral Red Uptake (NRU) test, was strongly dependent on the type of applied treatment. Each cellular experiment was executed on biological triplicates. A total of 1 mL of the growth medium related to each experiment was collected for the metabolomics analysis and stored at −30 °C until the registration of 1H NMR spectra. Frozen samples were thawed at room temperature and shaken before use. A total of 60 µL of D_2O containing 1.1 mM TMSP was added to 540 µL of each sample. The mixtures were homogenized by vortexing for 30*s* and transferred into 5 mm NMR tubes for analysis. The pH of each sample was controlled and adjusted to the value of 8.0 for all the NMR tubes. All the 1H NMR spectra were recorded with a Bruker Avance III 600 MHz spectrometer (Bruker BioSpin, Billerica, MA, USA) operating at a controlled temperature (298 ± 0.1 K) and equipped with a 5 mm BBI probe with a z-axis gradient coil and an automatic tuning-matching (ATM). The spectra of the growth media were acquired with a 1D Nuclear Overhauser Enhancement Spectroscopy (NOESY) presaturation pulse sequence, 128 scans, 65,536 data points, a spectral width of 7184 Hz, and a relaxation delay of 4 s. The raw data were multiplied using exponential line broadening before applying the Fourier transform. Transformed spectra were automatically corrected for phase and baseline distortions and calibrated using TopSpin 3.6.4 (Bruker Biospin srl).

2.2. Dataset Preparation

The collected dataset is composed of 94 spectra of metabolites from healthy and diseased cells—with 55 healthy and 39 diseased patterns. Moreover, a *percentage vitality* is attached to each sample, representing the cell health status. According to the literature [23,24], a sample is considered cytotoxic if the percentage vitality value is smaller than 70%. Hence, the vitality threshold is set to 70%.

Data are represented as sequences of ppm (parts per million) intensities, with each spectrum described by 262,144 intensity values (see Figure 2a, for an example). The first preprocessing step consists in filtering the spectra, setting zero values in the sequence elements corresponding to intensities between 4.6 and 5 ppm—which are consistent with water molecules—and greater than 9 or less than −0.5 ppm, corresponding to the head and the tail of the spectra, respectively. While the head and the tail must be filtered out because they collect noisy data, the water region has a very high signal coming from the hydrogen atoms, spanning all of the metabolite spectra. Excluding the first and the last zero values, each sequence is constituted by 220,000 intensity values. Finally, the intensity values are normalized in the range $[0,1]$ (see Figure 2b).

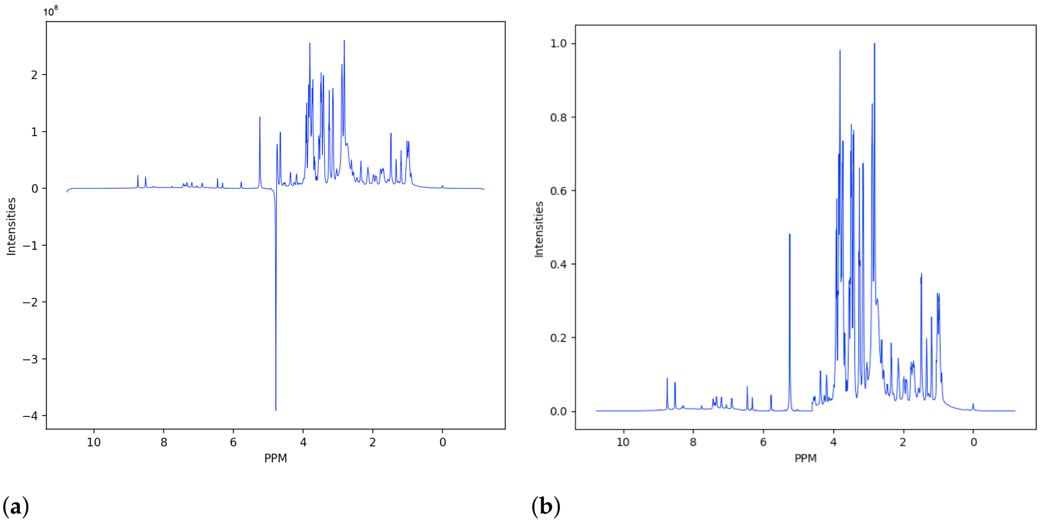


Figure 2. NMR spectral data before and after preprocessing: (**a**) An example of an NMR spectrum; (**b**) The same spectrum after preprocessing.

In order to obtain an adequate number of examples to train the DL model used for data compression, each original spectrum was split into 44 sub-spectra, with lengths of

5000. In practice, a dataset of 4136 samples was obtained, 80% of which (3308 samples) was used for training while the remaining 20% (828) constituted the test set.

2.3. Processing Pipeline

The considered task is a binary classification problem made difficult by both the scarcity of data and their high dimension, being the whole available dataset constituted by a set of 94 vectors, each consisting of 220,000 elements, with values in [0,1]. For this reason, our approach is based on two main phases: the reduction in the size of the feature vectors and their classification. Reducing the size of the spectra decreases the computational cost (in terms of the number of parameters to optimize) of the classification model. Dimensionality reduction is commonly performed using linear methods, such as PCA or LDA; however, in this document, autoencoders (AEs) have been chosen because they are nonlinear and learnable, while PCA is used as a baseline. Instead, some support vector machines (SVM), with different kernels, have been applied for classification.

2.3.1. Autoencoders

AEs [25,26] are unsupervised (actually, *self-supervised*) models able to learn compressed representations, or *embeddings*, of the input data in their inner layer. They have been used in various applications, e.g., for dimensionality reduction [27], anomaly detection [28,29], and sample generation [30–32]. Their structure consists of two modules, an *encoder* and a *decoder*. The encoder performs a dimensionality reduction in the input, producing its *latent space representation*, while the decoder re-expands the latent representation into the original dimension, so that, using the input data as targets, a reconstruction of the input can be obtained. The autoencoder architecture is shown in Figure 3. Usually, both the encoder and the decoder are MultiLayer Perceptrons (MLPs). However, in this work, we have also employed one-dimensional convolutional autoencoders (1D-CAEs), where the two modules are realized by a pool of one-dimensional convolutional layers (due to the sequential nature of the input data).

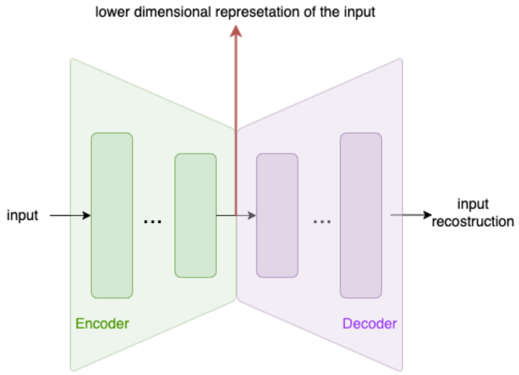


Figure 3. The autoencoder structure.

A similar model was proposed by Wang et al. [33] to compress ECG signals. Nevertheless, to the best of our knowledge, this is the first time that AEs have been applied for NMR spectral data embedding.

In our experiments, we have employed both MLP-AEs and 1D-CAEs, changing the encoder—and, consequently, the decoder—architecture with respect to the number of layers, of neurons per layer, and, for CAEs, the filter and kernel size. In particular, for the encoder of MLP-AEs, 2 or 3 dense layers with *ReLU* activation are considered, the first one composed of 5000 neurons, while the number of neurons in the last layer varies in $\{16, 32, 64\}$. The middle layer (when present) has a fixed dimension equal to 1024. It is worth noting that, if 16 neurons are used for representing each sub-spectrum in the latent space, then each spectrum is embedded into a vector of dimension 704 (16×44). The

1D-CAE architectures instead have 3, 4, or 5 layers in the encoder, with kernel size 7 and stride 5 for the first 4 layers, and kernel size 7 and stride 2 for the last layer. By changing the number of filters, compressed spectra with 4 or 8 channels were obtained. Table 1 collects the hyperparameters that define all the AE architectures used in the experiments. The values for the convolutional hyperparameters were set based on a grid search procedure aimed at optimizing the compression performance.

Table 1. Hyperparameters defining the AE architectures used in the experiments.

	Block Name	# of Layer	# of Neurons per Layer	Size of Embedded Spectrum
MLP	704_emb 2_lay	2	[5000, 16]	704
	1408_emb 2_lay	2	[5000, 32]	1408
	2816_emb 2_lay	2	[5000, 64]	2816
	704_emb 3_lay	3	[5000, 1024, 16]	704
	1408_emb 3_lay	3	[5000, 1024, 32]	1408
	2816_emb 3_lay	3	[5000, 1024, 64]	2816
	Block Name	**# of Layer**	**# of Output Channels**	**Size of Embedded Spectrum**
1D_CNN	4_ch 3_lay	3	4	[1760, 4]
	4_ch 4_lay	4	4	[352, 4]
	4_ch 5_lay	5	4	[176, 4]
	8_ch 3_lay	3	8	[1760, 8]
	8_ch 4_lay	4	8	[352, 8]
	8_ch 5_lay	5	8	[176, 8]

Finally, Figure 4a shows, as an example, the best-performing 1D-CAE architecture, composed of 4 layers of 1D convolutional blocks with a decreasing number of filters per layer, from 32 to 4, kernel dimension equal to 7, and stride equal to 5. Figure 4b depicts how the outputs of the 1D-CAEs are combined to reconstruct the full compressed spectrum from the sub-spectra. Indeed, each 220,000-component array is subdivided into 44 subvectors with 5000 entries, each fed to the relative autoassociator. After training, the 44 latent representations of the subvectors, obtained from the last layer of the encoders, are considered and concatenated to form the compressed representation of an entire spectrum.

2.3.2. PCA

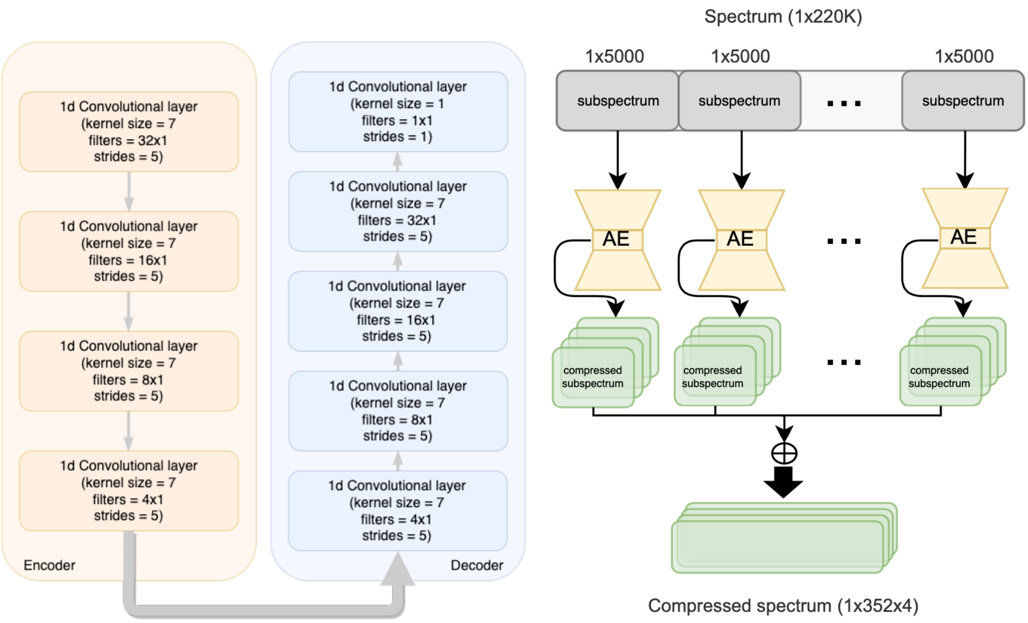
Concerning PCA, we selected 2 or 6 principal components. Considering the second case, more than 90% of the variance is captured, as shown in Figure 5a. For ease of visualization, in Figure 5b, the representation of the data with respect to the first two principal components is shown. As expected, the data distribution of healthy and diseased cells overlaps significantly.

2.3.3. SVM Classifiers

After compression, with AEs or PCA, reduced-dimensional data are presented to an SVM classifier. Memory efficiency is one of the key advantages of SVMs, as they use a subset of training points as part of the decision function. Moreover, SVMs tend to be an optimized algorithm with small datasets collecting high dimensional samples. Three types of kernels have been tested; linear, polynomial, and RBF, whose parameters were selected based on a grid search.

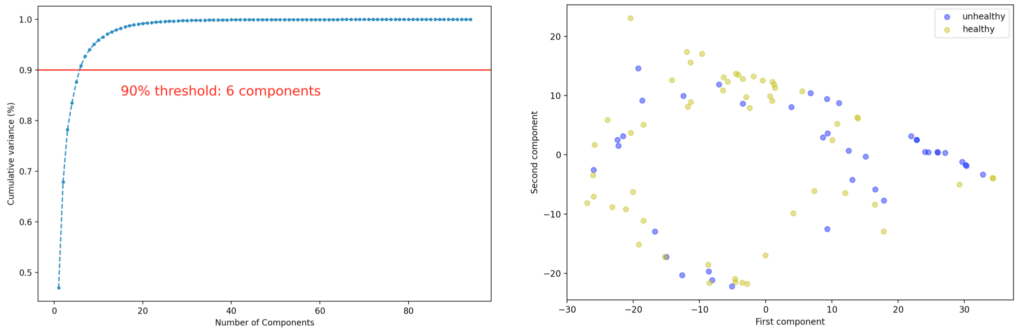
2.3.4. Evaluation Metrics

The quality of compressed data has been evaluated using three common metrics [34], namely Compression Ratio (CR), Root Mean Squared Error (RMSE), and Percentage Root mean squared Difference (PRD).



(**a**) 1D-CAE architecture. (**b**) Compressed sub-spectra recombination.

Figure 4. (**a**) Details of the layers composing the best performing 1D-CAE; (**b**) The combined architecture used to compress the sub-spectra and the reconstruction of the full spectrum.



(**a**) Cumulative variance plot. (**b**) Plot of the first two principal components.

Figure 5. PCA cumulative variance and selection of the number of components with a 90% threshold; based on the first two principal components (yellow/purple dots represent healthy/diseased cells, with vitality ≥ or <70%, respectively) the two distributions overlap significantly.

In particular, CR is the ratio between uncompressed and compressed data size:

$$CR = \frac{N}{M}, \tag{1}$$

where N is the length of uncompressed data and M is the dimension of the compressed patterns. This ratio estimates the compression efficiency—higher CR means higher compression—so it is desirable to be as high as possible.

RMSE is widely used to estimate the variance between reconstructed and original signals:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(S_0(i) - S_r(i))^2}{N}}, \tag{2}$$

where S_0 and S_r are the original and reconstructed signal, respectively. RMSE is a measure of the loss of information due to compression. While lossless compression algorithms should ideally ensure fully informative compressed data, some small losses can still occur due to quantization errors.

Finally, PRD is given by:

$$PRD = \sqrt{\frac{\sum_{i=1}^{N} (S_0(i) - S_r(i))^2}{\sum_{i=1}^{N} S_0^2(i)}}, \tag{3}$$

which is a measure of the percentage of information loss.

On the other hand, classification performance has been evaluated based on accuracy, precision, and recall, defined as:

$$accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{4}$$

$$precision = \frac{T_p}{T_p + F_p} \tag{5}$$

$$recall = \frac{T_p}{T_p + F_n} \tag{6}$$

where T_p, T_n, F_p, and F_n represent the true positive, true negative, false positive, and false negative predicted values, respectively.

3. Experimental Results

In this section, we present the comparative performance of different AE-based models in order to select the most accurate architecture to compress NMR spectra. As a subsequent verification of the quality of the compression procedure, the compressed spectra were then classified as descriptive of healthy or diseased cells.

Table 2 summarizes the compression performance of all the tested autoencoders. In the table, the model name explains the AE structure, e.g., *704_emb 2_lay* means that the MLP-AE has two layers for both the encoder and the decoder, while the whole latent space has dimension 704 (with 16 neurons for each compressed sub-spectrum embedding). Concerning 1D-CAEs, *4_ch* means that the compressed spectrum has four channels to be concatenated to evaluate the embedding length, while *3_lay* is the number of the encoder layers. Using the layers with stride 5 reduces the data dimensionality by a factor of 5^3, producing a compressed sub-spectrum with dimension $5000/5^3 = 40$, and a complete compressed spectrum described by a vector consisting of $40 \times 44 = 1760$ entries. Finally, the decoder has one more 1D convolutional layer to produce an output with the same dimension as the input.

As can be seen from Table 2, the reconstruction error is always small, while the PRD maintains under 1%; moreover, for comparable compression efficiency values (CR column), the best RMSE and PRD come from the 1D-CAE models. Actually, the CR metric is always very high, reaching a maximum value of 312.5 for the 1D-CAE with four channels and five layers (embedding dimension equal to 176) and for the MLP-AEs with both two and three layers and 16 neurons in the last layer (embedding dimension equal to 704). However, the minimum for RMSE and PRD is obtained with the 1D-CAE with eight channels and three layers (RMSE = 8×10^{-4} T/Hz and PRD = 0.02%), where the value of CR is comparatively small (in line with our expectations).

The classification task has been carried out by an SVM with an RBF kernel, which was chosen after also testing linear and polynomial kernels. A grid search approach was used in order to find the best kernel parameters ($\gamma = 0.1$ and $C = 1$). For performance evaluations, due to the very small number of available examples, a leave-one-out technique was implemented. Each experiment consists of 94 training/test runs, each of which reserves a single pattern for testing, while the other 93 are used in the training phase. The held-out test example is different at every run. The results obtained for each AE model are synthesized also in Table 2.

The accuracy of all models is about 76% and reaches its maximum value for the MLP-AE with two layers and embedding dimension equal to 2816, and for the 1D-CAE with four channels and four layers (embedding dimension equal to 1408), where it exceeds 77%. For these two models, the precision remains the highest (73% and 73.6%, respectively), while the recall never falls below 96% (98.2% and 96.4%, respectively). These results—which surpass those obtained with the PCA calculated on the original NMR spectra—indicate that the AE-based compression allows the extraction of features of better quality, which can be efficiently used for classification, (see Table 3). Indeed, while the 6D-PCA maintains slightly lower accuracy values, this cannot be said for the 2D-PCA, with which about 7% of accuracy is lost. Concerning the two best AE-based models, the 1D-CAE is absolutely preferable due to the significantly lower number of parameters (9897 against 2,5654,224) and because of its shorter embedding length (1408 against 2816 components).

Table 2. Autoencoder compression performance and SVM classification performance on AE models. In bold there are the bests results.

	Model	CR	RMSE (T/Hz)	PRD	Accuracy	Precision	Recall
MLP	704_emb 2_lay	**312.5**	0.05	0.01	0.766	0.726	0.964
	1408_emb 2_lay	156.25	0.04	0.007	0.766	0.72	**0.982**
	2816_emb 2_lay	78.125	0.033	0.005	**0.777**	0.73	**0.982**
	704_emb 3_lay	**312.5**	0.02	0.004	0.755	0.716	0.964
	1408_emb 3_lay	156.25	0.022	0.003	0.755	0.716	0.964
	2816_emb 3_lay	78.125	0.02	0.003	0.766	0.72	0.981
1D_CNN	4_ch 3_lay	31.25	0.0015	0.0004	0.755	0.716	0.963
	4_ch 4_lay	156.25	0.009	0.003	**0.777**	**0.736**	0.964
	4_ch 5_lay	**312.5**	0.016	0.005	0.745	0.712	0.945
	8_ch 3_lay	15.625	**0.0008**	**0.0002**	0.745	0.707	0.964
	8_ch 4_lay	78.125	0.005	0.001	0.766	0.726	0.964
	8_ch 5_lay	156.25	0.009	0.003	0.745	0.701	**0.982**

Let us notice that, based on the values of both precision and recall, the classifier shows better performance in the recognition of healthy cells (positive patterns), with a recall greater than 0.94 for all models. This is not astonishing since the dataset is imbalanced, with the positive class containing around 60% of available data. Indeed, for imbalanced classification problems, the majority class is typically referred to as the negative outcome, and the minority class is typically referred to as the positive outcome. Actually, the precision calculated on the diseased class for the 1D-CAE architecture of Table 3 is equal to 0.91, assessing that the patterns that the classifier recognizes as belonging to this class are almost always spectra of diseased cells.

Finally, let us spend a few words on the interpretability of the compressed spectra. In the case of 1D-CAEs, the compressed representation does not preserve the number of ppm related to each peak, though the sequential order and intensity values of the spectra are mostly conserved. The MLP-AEs, on the other hand, only roughly maintain the original trend. These different behaviors could be due to the local nature of the convolution operation, which guarantees a sort of spatial consistency. Figure 6 shows an example of a spectrum (after preprocessing) and its compressed representations with both the best 1D-CAE and MLP-AE.

Table 3. Best results obtained with the SVM classifier, changing the compression method. In bold there are the bests results.

	Model	Accuracy	Precision	Recall
whole spectra		0.713	0.671	**1**
2D–PCA		0.702	0.675	0.945
6D–PCA		0.755	0.726	0.964
MLP–AE	2816_emb, 2_lay	**0.777**	0.710	0.982
1D–CAE	4_ch, 4_lay	**0.777**	**0.736**	0.964

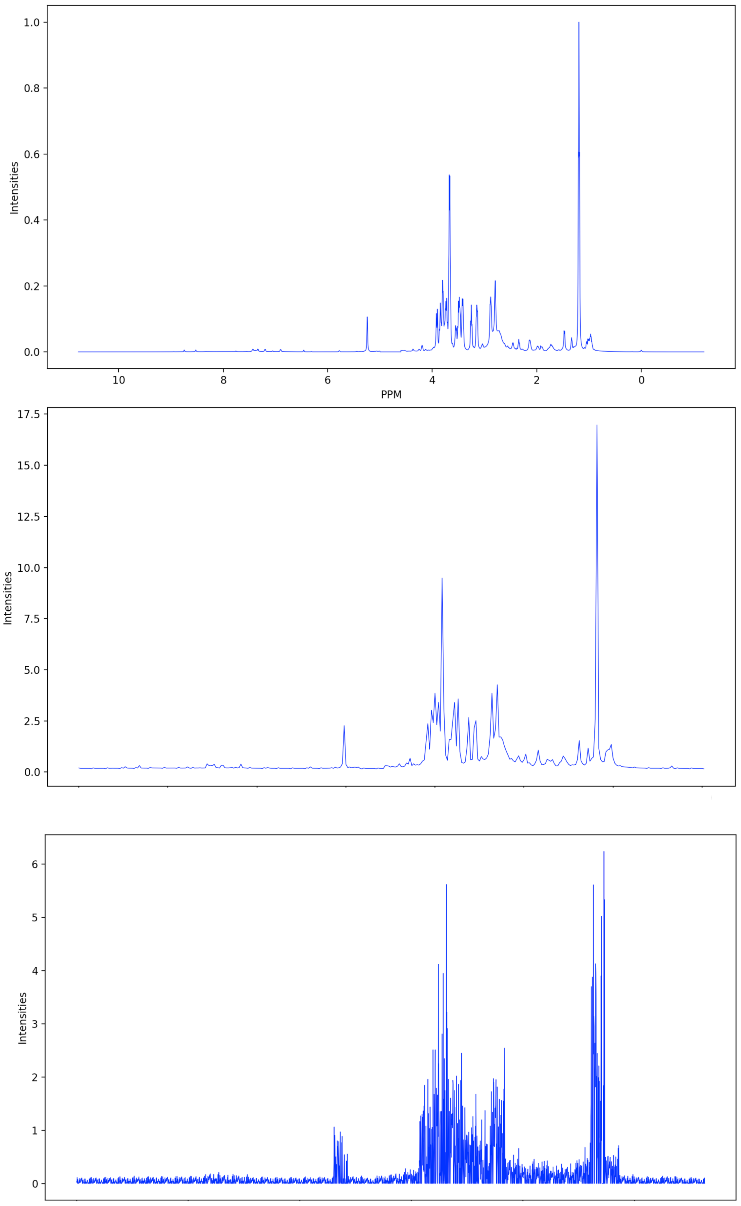


Figure 6. A spectrum from the dataset after preprocessing (on the **top**) and the same sample after compression by the best 1D-CAE (in the **middle**) and MLP-AE (at the **bottom**). The label of the x-axis is ppm for the top picture, while the other two graphical representations describe the obtained intensity values in the latent space.

4. Conclusions

In vitro cell-based metabolomics studies, often combined with other -omics, have found widespread use in many research areas, including the analysis of drug effect, action, and toxicology, tumor cell characterization, and signature extraction from cells. Indeed, cell signatures can constitute biomarkers usable for the early diagnosis of widespread pathologies, such as Alzheimer's disease. In fact, the common goal of metabolomics studies is to understand and decipher the influence and involvement of metabolism in biological effects and mechanisms and integrate this information into metabolic maps. However, analyzing the metabolic phenotype of cells using NMR spectroscopy and artificial intelligence tools is a branch of research still in its embryonic phase. In this paper, we proposed a whole pipeline to process NMR spectra in order to automatically establish if they correspond to a signature of a healthy or diseased cell. Identification of metabolite biomarkers depends on determining the latent variables responsible for class separation (healthy vs. diseased). PCA identifies the largest variations in the NMR data, but the latent variables responsible for class separation may not be in the direction of the largest variation. Instead, based on DL techniques, we have proved that high-dimensional NMR data can be embedded in compact representations — in *ad hoc* latent spaces — which can be used to represent the spectra in an information-conservative way, usable for automatic classification. Although the experimental results, based on the small dataset at our disposal, are preliminary, they are nonetheless promising and show the ability of compressed data to be used in a downstream classification task to distinguish between healthy and diseased neuronal cells, in the early diagnosis of neurodegenerative diseases. The significantly reduced dimension of embedded spectra, obtained with almost no information loss, would allow us to drastically reduce both the computational cost and memory storage, opening the door to the intensive use of automatic processing techniques for NMR data (not only related to the metabolic cell profile).

In future perspective, if a larger dataset becomes available, we will also consider spectral images (similar to the graphs shown in the figures) [35], which can collect richer information than sequences of intensity signals, also allowing the use of more complex convolutional architectures, presumably capable of achieving better compression/classification performance.

Author Contributions: Conceptualization, M.B. and F.C.; methodology, F.C., A.K., D.V., F.S. and M.B.; software, F.C.; validation, M.B. and F.C.; formal analysis, F.C.; investigation, F.C. and A.K.; resources, F.C., D.V. and A.K.; data curation, D.V. and A.K.; writing—original draft preparation, F.C. and A.K.; writing—review and editing, M.B. and D.V.; visualization, F.C. and A.K.; supervision, M.B., D.V. and F.S.; project administration, F.C., A.K., D.V., F.S. and M.B.; funding acquisition, D.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Publicly available dataset was analyzed in this study. This data can be found here: https://github.com/filippocostanti/NMR_compression, accessed on 8 May 2023.

Acknowledgments: Consorzio Interuniversitario Risonanze Magnetiche di Metallo Proteine (CIRMMP) is acknowledged for scholarship support. Paolo Andreini is acknowledged for helpful discussion on CNNs and autoencoders.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AD	Alzheimer's Disease
AE	Autoencoder
ATM	Automatic Tuning-Matching
CAE	Convolutional Autoencoder

CR	Compression Ratio
DL	Deep Learning
ECG	ElectroCardioGram
LDA	Linear Discriminant Analysis
MLP	MultiLayer Perceptron
MLP-AE	MultiLayer Perceptron Autoencoder
ND	Neurodegenerative Disease
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser Enhancement SpectroscopY
NRU	Neutral Red Uptake
PCA	Principal Component Analysis
PRD	Percentage Root-mean-square Difference
RBF	Radial Basis Function
RMSE	Root Mean Square Error
SVM	Support Vector Machine
TMSP	TriMethylSilylPropanoic acid
1D-CAE	1-Dimensional Convolutional Autoencoder
*n*D-PCA	*n*-Dimensional PCA

References

1. Kozlowski, H.; Luczkowski, M.; Remelli, M.; Valensin, D. Copper, zinc and iron in neurodegenerative diseases (Alzheimer's, Parkinson's and prion diseases). *Coord. Chem. Rev.* **2012**, *256*, 2129–2141. [CrossRef]
2. Stern, N.; Gacs, A.; Tátrai, E.; Flachner, B.; Hajdú, I.; Dobi, K.; Bágyi, I.; Dormán, G.; Lőrincz, Z.; Dual, S.C. Inhibitors of AChE and BACE–1 for Reducing Aβ in Alzheimer's Disease: From In Silico to In Vivo. *Int. J. Mol. Sci.* **2022**, *23*, 13098. [CrossRef]
3. Nagu, P.; Pathan, A.K.A.; Mehta, V. Screening of Herbal Molecules for The Management of Alzheimer's Disorder: In Silico and In Vitro Approaches. *Appl. Biol. Res.* **2022**, *24*, 255–272. [CrossRef]
4. Biedler, J.L.; Roffler-Tarlov, S.; Schachner, M.; Freedman, L.S. Multiple Neurotransmitter Synthesis by Human Neuroblastoma Cell Lines and Clones. *Cancer Res.* **1978**, *11*, 3751–3757.
5. Luo, Y.; Zhou, S.; Haeiwa, H.; Takeda, R.; Okazaki, K.; Sekita, M.; Yamamoto, T.; Yamano, M.; Sakamoto, K. Role of amber extract in protecting SH-SY5Y cells against amyloid beta 1-42-induced neurotoxicity. *Biomed. Pharmacother.* **2021**, *141*, 111804. [CrossRef]
6. Bell, M.; Zempel, H. A human cell model for TAU sorting and vulnerability. *Rev. Neurosci.* **2022**, *33*, 1–15. [CrossRef] [PubMed]
7. Huang, K.; Thomas, N.; Gooley, P.R.; Armstrong, C.W. Systematic Review of NMR–Based Metabolomics Practices in Human Disease Research. *Hum. Metab.* **2022**, *12*, 963. [CrossRef]
8. da Silva, G.H.R.; Mendes, L.F.; de Carvalho, F.V.; de Paula, E.; Duarte, I.F. Comparative Metabolomics Study of the Impact of Articaine and Lidocaine on the Metabolism of SH-SY5Y Neuronal Cells. *Hum. Metab.* **2022**, *12*, 581. [CrossRef]
9. Paris, D.; Melck, D.; Longo, A.; Napoletano, S.; Carotenuto, G.; Nicolais, L.; Motta, A.; Vitale, E. Metabolic response of SH-SY5Y cells to gold nanoparticles by NMR–based metabolomics analyses. *Biomed. Phys. Eng. Express* **2016**, *2*, 045003. [CrossRef]
10. Corsaro, C.; Vasi, S.; Neri, F.; Mezzasalma, A.M.; Neri, G.; Fazio, E. NMR in Metabolomics: From Conventional Statistics to Machine Learning and Neural Network Approaches. *Appl. Sci.* **2022**, *12*, 2824. [CrossRef]
11. Klukowski, P.; Riek, R.; Güntert, P. Rapid protein assignments and structures from raw NMR spectra with the deep learning technique ARTINA. *Nat. Commun.* **2022**, *13*, 6151. [CrossRef]
12. Klukowski, P.; Augoff, M.; Zięba, M.; Drwal, M.; Gonczarek, A.; Walczak, M.J. NMRNet: A deep learning approach to automated peak picking of protein NMR spectra. *Bioinformatics* **2018**, *34*, 2590–2597. [CrossRef]
13. Luo, J.; Zeng, Q.; Wu, K.; Lin, Y. Fast reconstruction of non–uniform sampling multidimensional NMR spectroscopy via a deep neural network. *J. Magn. Reson.* **2020**, *317*, 106772. [CrossRef] [PubMed]
14. Jo, T.; Nho, K.; Saykin, A.J. Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Front. Aging Neurosci.* **2019**, *11*, 220. [CrossRef]
15. Liu, S.; Masurkar, A.V.; Rusinek, H.; Chen, J.; Zhang, B.; Zhu, W.; Fernandez-Granda, C.; Razavian, N. Generalizable deep learning model for early Alzheimer's disease detection from structural MRIs. *Sci. Rep.* **2022**, *12*, 17106. [CrossRef] [PubMed]
16. Rossi, A.; Vannuccini, G.; Andreini, P.; Bonechi, S.; Giacomini, G.; Scarselli, F.; Bianchini, M. Analysis of brain NMR images for age estimation with deep learning. *Procedia Comput. Sci.* **2019**, *159*, 981–989. [CrossRef]
17. Qiu, S.; Miller, M.I.; Joshi, P.S.; Lee, J.C.; Xue, C.; Ni, Y.; Wang, Y.; Anda-Duran, I.D.; Hwang, P.H.; Cramer, J.A.; et al. Multimodal deep learning for Alzheimer's disease dementia assessment. *Nat. Commun.* **2022**, *13*, 3404. [CrossRef]
18. Wisely, C.E.; Wang, D.; Henao, R.; Grewal, D.S.; Thompson, A.C.; Robbins, C.B.; Yoon, S.P.; Soundararajan, S.; Polascik, B.W.; Burke, J.R.; et al. Convolutional neural network to identify symptomatic Alzheimer's disease using multimodal retinal imaging. *Br. J. Ophthalmol.* **2022**, *106*, 388–395. [CrossRef]
19. Cheung, C.Y.; Ran, A.R.; Wang, S.; Chan, V.T.T.; Sham, K.; Hilal, S.; Venketasubramanian, N.; Cheng, C.; Sabanayagam, C.; Tham, Y.C.; et al. A deep learning model for detection of Alzheimer's disease based on retinal photographs: A retrospective, multicentre case–control study. *Lancet—Digit. Health* **2022**, *4*, E806–E815. [CrossRef]

20. Gebregiworgis, T.; Powers, R. Application of NMR metabolomics to search for human disease biomarkers. *Comb. Chem. High Throughput Screen.* **2012**, *15*, 595–610. [CrossRef]
21. Song, Z.; Wang, H.; Yin, X.; Deng, P.; Jiang, W. Application of NMR metabolomics to search for human disease biomarkers in blood. *Clin. Chem. Lab. Med.* **2019**, *57*, 417–441. [CrossRef] [PubMed]
22. Kola, A.; Lamponi, S.; Currò, F.; Valensin, D. A Comparative Study between Lycorine and Galantamine Abilities to Interact with AMYLOID β and Reduce In Vitro Neurotoxicity. *Int. J. Mol. Sci.* **2023**, *24*, 2500. [CrossRef] [PubMed]
23. Cannella, V.; Altomare, R.; Leonardi, V.; Russotto, L.; Di Bella, S.; Mira, F.; Guercio, A. In Vitro biocompatibility evaluation of nine dermal fillers on L929 cell line. *Biomed. Res. Int.* **2020**, *2020*, 8676343. [CrossRef]
24. Cannella, V.; Altomare, R.; Chiaramonte, G.; Di Bella, S.; Mira, F.; Russotto, L.; Pisano, P.; Guercio, A. Cytotoxicity Evaluation of Endodontic Pins on L929 Cell Line. *BioMed Res. Int.* **2019**, *2019*, 3469525.
25. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*; MIT Press: Cambridge, MA, USA, 1986; pp. 318–362.
26. Bank, D.; Koenigstein, N.; Giryes, R. Autoencoders. *arXiv* **2020**, arXiv:2003.05991.
27. Ryu, S.; Choi, H.; Lee, H.; Kim, H. Convolutional Autoencoder Based Feature Extraction and Clustering for Customer Load Analysis. *IEEE Trans. Power Syst.* **2020**, *35*, 1048–1060. [CrossRef]
28. An J.; Cho, S. *Variational Autoencoder Based Anomaly Detection Using Reconstruction Probability*; Special Lecture on IE; SNU Data Mining Center: Seoul, Republic of Korea, 2015 ; pp. 1–18.
29. Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M.R.; Venkatesh, S.; Hengel, A.V.D. Memorizing normality to detect anomaly: Memory–augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1705–1714.
30. Xu, W.; Keshmiri, S.; Wang, G. Adversarially approximated autoencoder for image generation and manipulation. *IEEE Trans. Multimed.* **2019**, *21*, 2387–2396. [CrossRef]
31. Semeniuta, S.; Severyn, A.; Barth, E. A hybrid convolutional variational autoencoder for text generation. *arXiv* **2017**, arXiv:1702.02390.
32. Wan, Z.; Zhang, Y.; He, H. Variational autoencoder based synthetic data generation for imbalanced learning. In Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, USA, 27 November 27–1 December 2017; pp. 1–7.
33. Wang, F.; Ma, Q.; Liu, W.; Chang, S.; Wang, H.; He, J.; Huang, Q. A novel ECG signal compression method using spindle convolutional auto–encoder. *Comput. Methods Programs Biomed.* **2019**, *175*, 139–150. [CrossRef]
34. Tiwari, A.; Falk, T.H. Lossless electrocardiogram signal compression: A review of existing methods. *Biomed. Signal Process. Control* **2019**, *51*, 338–346. [CrossRef]
35. Chen, D.; Wang, Z.; Guo, D.; Orekhov, V.; Qu, X. Review and Prospect: Deep Learning in Nuclear Magnetic Resonance Spectroscopy. *arXiv* **2020**, arXiv:2001.04813.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.