Feature learning with Deep Neural Networks for MS-based clinical diagnosis

Khawla Seddiki[1,2], Philippe Saudemont[2], Frédéric Precioso[3], Nina Ogrinc[2], Maxence Wisztorski[2], Michel Salzet[2], Isabelle Fournier[2], Arnaud Droit[1]

1 Centre de Recherche du CHU de Québec–Université Laval, Québec City, Canada
2 Université Lille, Inserm U1192-Laboratoire Protéomique, Réponse Inflammatoire et Spectrométrie de Masse (PRISM), Villeneuve d'Ascq Cedex, France
3 Université Cote d'Azur, CNRS, I3S, Sophia Antipolis, France

Abstract

Rapid and accurate clinical diagnosis of pathological conditions remains highly challenging. An important part of the development includes building effective classification models. Some machine learning approaches have been investigated to achieve Mass Spectrometry (MS) data classification. However, these algorithms require time-consuming preprocessing steps to remove data artifacts, making their application unsuitable for real-time analysis. Convolutional Neural Networks (CNNs) have shown to perform well under such circumstances. However, their effectiveness drastically decreases when a small number of samples is available, which is a common situation in medicine.

In this study, we investigate first CNNs transfer learning for 1D MS data classification, then we develop a new cumulative learning method when transfer learning is not powerful enough. We hence propose to train the same model through several classification tasks over different small datasets from different biological contexts, thus accumulating MS knowledge in the resulting representation. Our results show that using the cumulative learning approach improves classification accuracy beyond 98% for canine cancer, human ovarian cancer, and pathogenic microorganisms biotypes in 1D clinical datasets. Our proposed approach is a promising strategy to improve classification accuracy when only a small number of samples is available as prospective cohorts.

1 Introduction

In many clinical situations, the speed, sensitivity, and reliability of diagnostic may improve patient uptake. For instance, rapid identification of cancer tissues has a crucial impact on decisions made during surgery [1]. A similar need exists in the treatment of infections, where accurate and rapid identification of microorganisms is important to ensure the most appropriate and effective treatment [2]. Mass spectrometry (MS) is particularly useful for such purposes since it provides non-targeted molecular information on the millisecond time scales [3, 4]. Its sensitivity and reproducibility are well established. In this context, SpiderMass is a new MS system designed for *in vivo* and real-time analysis. It has several applications including cancer diagnosis [5], bacteria biotyping [6], and other scale applications [7].

For cancer diagnostics and microbial pathogen identification, many popular classification machine learning models, such as Support Vector Machine (SVM) [8], Random Forest (RF) [9], and Linear Discriminant Analysis (LDA) [10] have been already used and compared for MS data classification [11–14]. However, these methods design for real-time applications becomes a highly complex task, since they must follow a workflow involving several interdependent preprocessing steps such as denoising, baseline correction, alignment, etc. To address real-time MS data classification for a prospective implementation in the SpiderMass, Convolutional Neural Networks (CNNs) represent an attractive approach offering various advantages [15] even on raw data. However, their classification efficiency trained using a small number of spectra drops rapidly [16], which is the case in many clinical situations, where the samples are accessible in limited amounts especially for rarer diseases. For such applications, transfer learning has emerged as an interesting approach [17]. It has proven useful in many engineering areas [18]. This has yet to be explored for 1D spectral data, since no 1D spectral dataset as large as the ImageNet database is available for instance [19]. The aim of this study was to build CNN-based classification models for 1D mass spectra classification using small clinical datasets generated for diagnosis of cancer or microbial infection.

2 Methods

2.1 Datasets

We evaluated our proposed approach on independent MS datasets (Table 1):

	MS instrument	Dataset	Classes	# spectra	Description
Target datasets	Synapt G2-S Q-TOF (Waters, SpiderMass)	Canine cancer	Normal	482	Contained 1 normal and 11 heterogeneous sarcoma types as described previously [5]
			Myxosarcoma	60	
			Fibrosarcoma	404	
			Hemangiopericytoma	134	
			Malignant peripheral nerve tumor	60	
			Osteosarcoma	339	
			Undifferentiated pleomorphicsarcoma	376	
			Rhabdomyosarcoma	66	
			Splenic fibrohistiocytic nodules	63	
			Histiocytic sarcoma	105	
			Soft tissue sarcoma	69	
			Gastrointestinal stromal sarcoma	70	
			Total	**2228**	
	Hybrid quadrupole (QSTAR pulsar I)	Human ovarian cancer 1	Normal	95	Contained two classes of low-resolution spectra, normal and cancerous publicly available at `home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`
			Cancer	121	
			Total	**216**	
	Synapt G2-S Q-TOF (Waters, SpiderMass)	Microorganisms	Staphylococcus aureus	26	Contained a five human pathogen as described previously [6]
			E.coli D31	26	
			Pseudomonas aeruginosa	24	
			Enterococcus faecalis	19	
			Candida albicans	23	
			Total	**119**	
Source datasets	PBSII SELDI-TOF	Human ovarian cancer 2	Normal	91	Contained two classes of high-resolution spectra , normal and cancerous publicly available at `home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`
			Cancer	162	
			Total	**253**	
	Rapiflex MALDI-TOF (Bruker)	Rat brain	Gray matter	4635	Contained spectra of rat gray and white brain matter
			White matter	5465	
			Total	**10100**	
	Synapt G2-S Q-TOF (Waters, SpiderMass)	Beef liver	Positive mode	1372	Contained two types of spectra of healthy beef liver samples, one acquired in positive and the other in negative ion mode
			Negative mode	1265	
			Total	**2637**	

Table 1: Description of datasets

2.2 Experiments

All datasets were imported without undergoing any preprocessing step. Each dataset were linearly scaled between 0 and 1 and divided randomly into training, validation, and test with ratios of 60%, 20%, and 20%, respectively. Performance of trained classifiers was measured by global accuracy on test subsets averaged over 10 independent iterations. For each iteration a stratified 5-fold cross validation was used to maintain the original proportion of minority classes. A weighted loss function was used during the training for samples from under-represented classes.

Protocol for evaluating 2D-CNN architectures adapted to 1D We evaluated and compared the application of three prominent CNN architectures for classifying spectra in clinical datasets. The first of these was variant_LeNet contained two convolutional layers and two fully connected layers, adapted from [20], the second was LeNet_Liu included three convolutional layers and two fully connected layers [16], and the third was VGG9 with six convolutional layers and three fully connected layers adapted from [18]. We performed a grid search of several hyperparameters such as number of filters, kernels size, learning rate, etc. Using this approach, we expected to determine what model depth and hyper-parameters are optimal for classification of MS spectra, especially in the case of highly heterogeneous biological classes such as canine cancer types.

Protocol for evaluating transfer learning The three CNN architectures were trained on the large rat brain dataset with all weights initialized according to He normal distribution. The decision layers of the network were not useful, since the rat brain and clinical datasets were from different contexts. The convolutional weights were then frozen so that they would not be updated during back-propagation, the decision layers were removed, and the new specific decision layers dedicated to smaller clinical datasets were trained. Transfer learning from the rat brain dataset allowed the model to learn and detect generic representations of MS peaks. By freezing the lower CNN levels, we are assuming that the model has extracted the right patterns, and that only the high level is needed to take into account specific peak's features.

Protocol for evaluating cumulative learning Transfer learning in some cases may not be enough as an aid to classifying biologically similar materials using CNN models. This proximity is

reflected in a high degree of confusion between classes. This is typically the case when the biggest dataset which is supposed to be used to learn the pivotal data representation is not big enough. In addition, low-resolution or data heterogeneity can further complicate the classification task. We therefore propose two approaches to developing 1D CNN cumulative learning:

Scenario A The first step is to train CNN architectures on the rat brain dataset as described before for transfer learning. The model weights are then fine-tuned, the decision layers are removed, and new decision layers are trained with :

- the beef liver dataset, then its weights were frozen and new specific decision layers were added and trained using the canine cancer dataset.

- the human ovarian 2 dataset, then its weights were frozen and new specific decision layers were added and trained using the human ovarian 1 dataset.

Scenario B CNN architectures were trained on the rat brain and fine-tuned with the beef liver dataset as described in Scenario A, but instead of testing this model on the canine cancer dataset, an additional learning was added. Beef liver CNN weights were fine-tuned, decision layers were removed and new specific decision layers were added and trained using the microorganisms dataset, before freezing convolutional layer weighting and training new specific decision layers on the canine cancer dataset.

The resulting CNN model from Scenario B was tested with changes to the dimensionality of the output space (number of classes) and the activation function of the last fully connected layer on rat brain, beef liver and microorganisms datasets separately. The objective was to assess how much learning skill the final CNN gained or lost of MS knowledge through successive training.

Standard Machine learning classification algorithms Conventional algorithms, namely SVM, RF, and LDA are not designed to classify MS spectra that have not been preprocessed. Spectra were corrected using sequential preprocessing of five steps: (1) Savitzky-Golay-Filter denoising, (2) baseline subtraction using the statistics-sensitive non-linear iterative peak-clipping, (3) normalization on the total ion count, (4) alignment using a cubic warping function, (5) and peaks detection using the median absolute deviation. Chi-square (χ^2) statistic was used to reduce data dimensionality before feeding to the classification algorithms.

3 Results

3.1 CNNs classification performance

As shown in Table 2, VGG9 was the best at canine cancer classification, while LeNet_Liu was the best at classifying ovarian dataset. Variant_LeNet was the best at classifying microorganisms, but accuracy suffers quickly from over-fitting when a deep architecture such as VGG9 was used. All three CNN architectures performed poorly, which was not surprising because of the low number of spectra used for the training.

Dataset	# classes	variant_LeNet	LeNet_Liu	VGG9
Canine cancer	12	0.66 ± 0.02	0.68 ± 0.08	**0.72** ± 0.02
Ovarian cancer 1	2	0.74 ± 0.01	**0.80** ± 0.00	0.53 ± 0.01
Microorganisms	5	**0.75** ± 0.02	0.73 ± 0.09	0.31 ± 0.06

Table 2: Overall classification accuracies using three CNN architectures. The best result is indicated in boldface

3.2 Transfer learning

As shown in Table 3, transfer learning clearly improved the classification accuracy of the three small datasets compared to models trained from scratch. These results suggest that training a CNN model with extracted spectral features transferred even from an unrelated field is better than training it with spectral features learned from scratch on a small dataset. Although improvements are still needed for canine and

Dataset	# classes	variant_LeNet	LeNet_Liu	VGG9
Canine cancer	12	0.81 ± 0.00 22%	**0.88** ± 0.03 29%	0.87 ± 0.01 20%
Ovarian cancer 1	2	0.80 ± 0.01 8%	**0.83** ± 0.02 3%	0.81 ± 0.02 52%
Microorganisms	5	**0.99** ± 0.00 30%	**0.99** ± 0.00 35%	0.81 ± 0.02 158%

Table 3: Overall classification accuracies using three CNN architectures with transfer learning. The improvement in performance is expressed as a percentage

ovarian cancer datasets.

3.3 Cumulative learning

Two scenarios were tested: (A) training on intermediate beef liver and then on canine cancer dataset or training on intermediate human ovarian 2 and then on human ovarian 1 cancer dataset; (B) training on beef liver, then on microorganisms and lastly on canine cancer dataset.

As shown in Table 4, Scenario A improved the classification accuracy considerably for canine cancer relative to learning from scratch and slightly relative to transfer learning, the best improvements was obtained with LeNet_Liu architecture. Scenario B provided a slight additional improvement over Scenario A, and the greatest accuracy was achieved with VGG9 architecture. For the human ovarian 1 dataset, accuracy was improved to 0.99 by scenario A with LeNet_Liu architecture.

Dataset	Protocol	variant_LeNet	LeNet_Liu	VGG9
Canine cancer	Scenario A	0.86 ± 0.02 30%* 6%**	**0.95** ± 0.033 39%* 8%**	0.93 ± 0.02 27%* 6%**
	Scenario B	0.90 ± 0.01 37%* 12%** 5%***	0.97 ± 0.00 41%* 10%** 2%***	**0.98** ± 0.00 35%* 12%** 6%***
Ovarian cancer 1	Scenario A	0.95 ± 0.00 28%* 18%**	**0.99** ± 0.00 24%* 19%**	0.96 ± 0.01 81%* 18%**

Table 4: Overall classification accuracies by the three CNN architectures; percent improvement to learning from scratch*, transfer learning**, and Scenario A***

These results show that in contrast with the previously observed lack of sensitivity and specificity of low-resolution datasets for cancer diagnosis [21], the CNN cumulative model was up to the task without any need for spectral preprocessing steps.

Classification accuracy obtained by CNN from scratch on data used for the training (rat brain and beef liver) and with transfer learning for microorganisms (Table 3) was equal to 0.99. Testing the best resulting cumulative representation of VGG9 from Scenario B on rat brain, beef liver and microorganisms datasets separately did not show improvement of the classification accuracy from 0.99. This indicates that the model accumulates knowledge through the successive training phases without any losses and suggests that a "universal" representation of MS classification might exist.

3.4 Performance of standard algorithms applied to preprocessed datasets

As shown in Table 5, RF outperformed the other methods in canine cancer and microorganisms classification, while LDA was best for ovarian classification. Performance of RF and LDA appears not comparable to that of CNNs. In addition, RF and LDA require more time to carry out the necessary preprocessing steps and to determine the optimal hyper-parameters since datasets had different artifacts and therefore required different preprocessing strategies.

Dataset	# classes	SVM	RF	LDA
Canine cancer	12	0.53 ± 0.22	**0.75** ± 0.21	0.723 ± 0.025
Ovarian cancer 1	2	0.60 ± 0.05	0.88 ± 0.03	**0.97** ± 0.00
Microorganisms	5	0.88 ± 0.00	**0.98** ± 0.01	0.65 ± 0.03

Table 5: Overall classification accuracies by SVM, RF, and LDA

4 Discussion

This study shows for the first time the use of cumulative learning for 1D spectrum classification of datasets generated in vastly different biological contexts, on different organisms, acquired by a variety of instruments and technologies at different resolutions. Our CNN model was designed by accumulating mass spectral knowledge through multiple training steps on small datasets. It provided a viable alternative when transfer learning was inadequate, as was the case for low-resolution, heterogeneous MS data, or when the source domain dataset was not large enough. The novelty is that the model can be pre-trained on a dataset containing only two output categories and yet predict 2, 5 and even 12 outputs, that are unlikely to share common features. The final model accumulates MS knowledge through the successive training phases without any losses which suggests that a "universal" representation of MS classification might exist. In addition, CNNs appear to offer a unified solution for classification without the need for time-consuming preprocessing of spectra making the model adapted for a real-time analysis.

References

1. Zhang, J. *et al.* Nondestructive tissue analysis for ex vivo and in vivo cancer diagnosis using a handheld mass spectrometry system. *Science translational medicine* **9,** eaan3968 (2017).
2. Kumar, A. *et al.* Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical care medicine* **34,** 1589–1596 (2006).
3. Takats, Z. *et al.* in *Advances in cancer research* 231–256 (Elsevier, 2017).
4. Strittmatter, N. *et al.* Characterization and identification of clinically relevant microorganisms using rapid evaporative ionization mass spectrometry. *Analytical chemistry* **86,** 6555–6562 (2014).
5. Saudemont, P. *et al.* Real-time molecular diagnosis of tumors using water-assisted laser desorption/ionization mass spectrometry technology. *Cancer cell* **34,** 840–851 (2018).
6. Fatou, B. *et al.* Real time human micro-organisms biotyping based on Water-Assisted Laser Desorption/Ionization. *The EuroBiotech Journal* **3** (2019).
7. Fatou, B. *et al.* Real time and in vivo pharmaceutical and environmental studies with SpiderMass instrument. *Journal of biotechnology* **281,** 61–66 (2018).
8. Cortes, C. *et al. Support-vector networks* **3,** 273–297 (Springer, 1995).
9. Breiman, L. Random forests. *Machine learning* **45,** 5–32 (2001).
10. Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics* **7,** 179–188 (1936).
11. Wu, B. *et al.* Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* **19,** 1636–1643 (2003).
12. Gredell, D. A. *et al.* Comparison of Machine Learning Algorithms for Predictive Modeling of Beef Attributes Using Rapid Evaporative Ionization Mass Spectrometry (REIMS) Data. *Scientific reports* **9,** 5721 (2019).
13. Datta, S. *et al.* Feature selection and machine learning with mass spectrometry data for distinguishing cancer and non-cancer samples. *Statistical Methodology* **3,** 79–92 (2006).
14. Vervier, K. *et al.* Benchmark of structured machine learning methods for microbial identification from mass-spectrometry data. *arXiv preprint arXiv:1506.07251* (2015).
15. LeCun, Y. *et al.* Deep learning. *nature* **521,** 436 (2015).
16. Liu, J. *et al.* Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *Analyst* **142,** 4067–4074 (2017).
17. Pan, S. J. *et al.* A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22,** 1345–1359 (2009).
18. Lu, J. *et al.* Transfer learning using computational intelligence: a survey. *Knowledge-Based Systems* **80,** 14–23 (2015).
19. Krizhevsky, A. *et al. Imagenet classification with deep convolutional neural networks* in *Advances in neural information processing systems* (2012), 1097–1105.
20. LeCun, Y. *et al.* Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86,** 2278–2324 (1998).
21. Conrads, T. P. *et al.* High-resolution serum proteomic features for ovarian cancer detection. *Endocrine-related cancer* **11,** 163–178 (2004).