Expert Systems With Applications 215 (2023) 119296

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Unsupervised neural networks as a support tool for pathology diagnosis in MALDI-MSI experiments: A case study on thyroid biopsies

Marco S. Nobile [a,b,h,i,1,*], Giulia Capitoli [c,b,1], Virgil Sowirono [i], Francesca Clerici [d], Isabella Piga [d], Kirsten van Abeelen [g], Fulvio Magni [d], Fabio Pagni [c,e], Stefania Galimberti [c,b], Paolo Cazzaniga [f,b,2], Daniela Besozzi [g,b,**,2]

[a] Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy
[b] Bicocca Bioinformatics Biostatistics and Bioimaging Centre - B4, University of Milano-Bicocca, Milano, Italy
[c] School of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy
[d] Proteomics and Metabolomics Unit, School of Medicine and Surgery, University of Milano-Bicocca, Vedano al Lambro, Italy
[e] Pathology, Department of Medicine and Surgery, University of Milano-Bicocca, San Gerardo Hospital, ASST, Monza, Italy
[f] Department of Human and Social Sciences, University of Bergamo, Bergamo, Italy
[g] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano, Italy
[h] Eindhoven Artificial Intelligence Systems Institute, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands
[i] Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands

ARTICLE INFO

Keywords:
Self-Organizing Maps
Unsupervised learning
MALDI-MSI
Mass spectrometry
Thyroid carcinoma
Precision medicine

ABSTRACT

Artificial intelligence is getting a foothold in medicine for disease screening and diagnosis. While typical machine learning methods require large labeled datasets for training and validation, their application is limited in clinical fields since ground truth information can hardly be obtained on a sizeable cohort of patients. Unsupervised neural networks – such as Self-Organizing Maps (SOMs) – represent an alternative approach to identifying hidden patterns in biomedical data. Here we investigate the feasibility of SOMs for the identification of malignant and non-malignant regions in liquid biopsies of thyroid nodules, on a patient-specific basis. MALDI-ToF (Matrix Assisted Laser Desorption Ionization - Time of Flight) mass spectrometry-imaging (MSI) was used to measure the spectral profile of bioptic samples. SOMs were then applied for the analysis of MALDI-MSI data of individual patients' samples, also testing various pre-processing and agglomerative clustering methods to investigate their impact on SOMs' discrimination efficacy. The final clustering was compared against the sample's probability to be malignant, hyperplastic or related to Hashimoto thyroiditis as quantified by multinomial regression with LASSO. Our results show that SOMs are effective in separating the areas of a sample containing benign cells from those containing malignant cells. Moreover, they allow to overlap the different areas of cytological glass slides with the corresponding proteomic profile image, and inspect the specific weight of every cellular component in bioptic samples. We envision that this approach could represent an effective means to assist pathologists in diagnostic tasks, avoiding the need to manually annotate cytological images and the effort in creating labeled datasets.

Abbreviations: SOM, Self-Organizing Maps; MALDI, Matrix Assisted Laser Desorption Ionization; ToF, Time of Flight; MSI, Mass Spectrometry Imaging; LASSO, Least Absolute Shrinkage and Selection Operator; AI, Artificial Intelligence; FNA, Fine Needle Aspiration; DESI, Desorption Electrospray Ionization; DSUUL, Discrimination of Spectra Using Unsupervised Learning; ROI, Region of Interest; H&E, Hematoxylin and Eosin; ANN, Artificial Neural Network; BMU, Best Matching Unit; PTC, Papillary Thyroid Carcinoma; HP, Hyperplastic; HT, Hashimoto Thyroiditis; NIFTP, Noninvasive Thyroid Neoplasm with Papillary-like Nuclear Features; TIC, Total Ion Current; MAD, Mean Absolute Deviation; GPU, Graphics Processing Unit; SIMD, Same Instruction Multiple Data

* Corresponding author at: Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy.
** Corresponding author at: Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano, Italy.
E-mail addresses: marco.nobile@unive.it (M.S. Nobile), giulia.capitoli@unimib.it (G. Capitoli), v.i.sowirono@student.tue.nl (V. Sowirono), francesca.clerici@unimib.it (F. Clerici), isabella.piga@unimib.it (I. Piga), kirsten.vanabeelen@unimib.it (K. van Abeelen), fulvio.magni@unimib.it (F. Magni), fabio.pagni@unimib.it (F. Pagni), stefania.galimberti@unimib.it (S. Galimberti), paolo.cazzaniga@unibg.it (P. Cazzaniga), daniela.besozzi@unimib.it (D. Besozzi).
[1] These authors contributed equally to this work.
[2] These senior authors contributed equally to this article.

https://doi.org/10.1016/j.eswa.2022.119296
Received 24 June 2022; Received in revised form 4 November 2022; Accepted 15 November 2022
Available online 22 November 2022
0957-4174/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Artificial Intelligence (AI) represents nowadays a promising means to aid clinicians in prognosis, diagnosis, treatment identification, and disease screening (Obermeyer & Topol, 2021). The most widespread supervised AI methods – e.g., machine learning (Rajkomar, Dean, & Kohane, 2019) and deep learning (Esteva et al., 2019) – require large amounts of labeled data for AI models' training and validation; though, adequate labeled datasets are not always available or demand a huge effort for their creation. To overcome this limit, unsupervised AI methods can be exploited to identify patterns in data, and to spontaneously learn the optimal separation of a dataset into clusters according to some measure of mutual similarity. In this context, neural networks like Self-Organizing Maps (SOMs) (Kohonen, 2012) have been extensively used as a tool for clustering (Günter & Bunke, 2002; Vesanto & Alhoniemi, 2000), complexity reduction (Wang, Delabie, Aasheim, Smeland, & Myklebost, 2002), anomaly detection (Tian, Azarian, & Pecht), and visualization of multi-dimensional numerical data (Vesanto, 1999) to assist and simplify its interpretation (Pourkia, Rahimi, & Baghaei, 2019). Practical application areas in which SOMs have been effectively applied include pattern recognition (Yamagutchi, Nagata, Truong, Pfurtscheller, & Inoue, 2007) and medical applications, such as clustering of gene microarray data of breast and prostate cancer cells (Hautaniemi et al., 2003; Markey, Lo, Tourassi, & Floy. Jr, 2003), integration of clinical and molecular information to the aim of classifying the risk of progression in bladder cancer (Borkowska et al., 2014), identification of patterns associated with the survival of patients with breast cancer (Shukla, Hagenbuchner, Win, & Yang, 2018), analysis of functional magnetic resonance imaging (Ngan, Yacoub, Auffermann, & Hu, 2002) and ophthalmological data (Henson, Spenceley, & Bull, 1997).

Here, we propose the application of SOMs for the analysis of Matrix-Assisted Laser Desorption Ionization (MALDI) Mass Spectrometry Imaging (MSI) data, which were generated to measure the spectral profiles of Fine Needle Aspiration (FNA) biopsies. FNA is a widely used procedure for the collection of specimens, employed in the diagnosis of benign and malignant lesions in the pre-surgical setting. In particular, we use MALDI-MSI data of FNA biopsies of thyroid nodules as a case study. Thyroid cancer can be diagnosed by detecting thyroid nodules, which are radiologically distinct from the normal types of tissue of the thyroid gland. Although the majority of such nodules – that are very common and most of the times incidentally detected during imaging procedures for other indications – are benign, approximately 7–15% of patients with thyroid nodules are affected by malignant thyroid carcinoma. Furthermore, around 20–30% of cases have an indeterminate for malignancy final report after biopsy and undergo surgery; however, after the thyroidectomy, 80% of these nodules are confirmed to be benign (Capitoli et al., 2022). The early identification of malignant nodules promoted by FNA, combined with the application of innovative technologies – such as MALDI-MSI – on cytological thyroid specimens thus represents a promising approach to better characterize and distinguish the molecular signature of different lesions.

Several computational approaches have been presented in the literature to analyze MALDI-MSI data. For instance, hierarchical clustering was exploited to highlight possible tumor areas within a tissue section (Deininger, Ebert, Futterer, Gerhard, & Rocken, 2008); machine learning techniques, such as support vector machines and random forests, were employed to differentiate cancer and non-cancer samples (Datta & DePadilla, 2006), and to perform data factorization and dimensionality reduction (Verbeeck, Caprioli, & Plas, 2020); convolutional neural networks were used in classification tasks of MALDI-MSI data (Seddiki et al., 2020). More recently, principal component analysis and t-distributed stochastic neighbor embedding were implemented in the tool M^{2}aia to realize a dimensionality reduction analysis of MSI data (Cordes et al., 2021); differently from the approach that is presented in this work, in M^{2}aia a peak picking of the spectra is performed, rather than considering them entirely.

In previous approaches, MALDI-MSI data, obtained from FNA samples, was used to characterize the pathological state of patients. Conversely, in this work we propose a different approach to automatically discriminate different proteomic profiles within a FNA thyroid sample with the aim of identifying specific spectra footprint characteristic of morphological regions. We assess the feasibility of exploiting SOMs on proteomic profiles of thyroid FNA data. Namely, we evaluate the mass spectra clustering outcome and compare it to the corresponding morphological image. It is of great interest to evaluate the applicability of SOMs for the identification of clusters of different proteomic profiles as this approach can directly use the unlabeled liquid biopsy samples obtained with FNA, and has therefore the potential of supporting clinicians in cytological diagnosis.

Moving forward from the first results obtained using cytological smears or tissues taken from surgical procedures (Kurczyk et al., 2020; Mosele, Smith, Galli, Pagni, & Magni, 2017; Pietrowska et al., 2017), some previously studies investigated proteomics signatures on FNAs using MALDI-MSI technique (Capitoli et al., 2020, 2019), or metabolomic with desorption electrospray ionization mass spectrometry (DESI-MSI) technique (DeHoog et al., 2019). In both cases, based on the molecular profiles obtained from malignant thyroid carcinomas and benign thyroid tissues, classification models (e.g., LASSO (Tibshirani, 1996) and elastic net Zou & Hastie, 2005) were generated and used to predict a diagnosis based on the morphological composition of FNA material. It is worth highlighting that these methods often require the use of several pre-processing steps. Pre-processing filters out irrelevant and redundant information present in the data; however, it could potentially also remove relevant information that should have been included. In this context, an additional advantage in the use of SOMs is that they are typically applied to the original data with only minimal pre-processing, thus allowing for retaining as much information as possible. Moreover, SOMs aim at discovering patterns and then clustering the data accordingly, which is different from supervised statistical models that are used to identify specific "signals" based on which they are able to identify different entities, without taking into account the whole shape of a spectrum.

To sum up, in this work we present a novel framework, called DSUUL, for the automatic clustering of spectral profiles provided by MALDI-MSI data. We apply our framework to analyze FNA biopsies of thyroid nodules. Fig. 1 schematizes the workflow of mass spectrometry sample preparation and analysis, whose outcome represents the input for the machine learning strategies. LASSO (supervised) and DSUUL (unsupervised) learning approaches are depicted for comparison, to highlight the manual steps regarding the annotation of Regions of Interest (ROIs) in the sample and the labeling of the spectra, which are required to build a LASSO model. To the best of our knowledge, DSUUL represents the first attempt in using unsupervised learning as a possible complementary approach to routine FNA-based cytology. Since malignant thyroid samples can have different spectral footprints, we investigate the feasibility of unsupervised anomaly detection using SOMs. Various data pre-processing and clustering strategies are also used to investigate their impact on the discrimination efficacy of SOMs. Our results show that SOMs can be effective in automatically identifying and separating the areas of a sample containing benign cells from those containing malignant cells, thus avoiding the need for pathologists to manually annotate ROIs within the sample. Indeed, our method relies on mass spectra data analysis to determine cells' specific positions and different morphological areas, and it is able to extract the corresponding spectral footprints.

AI-based clinical decision support systems have gone through a fast development in the last decades, given their potentiality in improving the clinical workflow management, acting for cost containment, and providing automated diagnostic supplements (Ostropolets, Zhang, & Hripcsak, 2020; Sutton et al., 2020). By supporting pathologists with

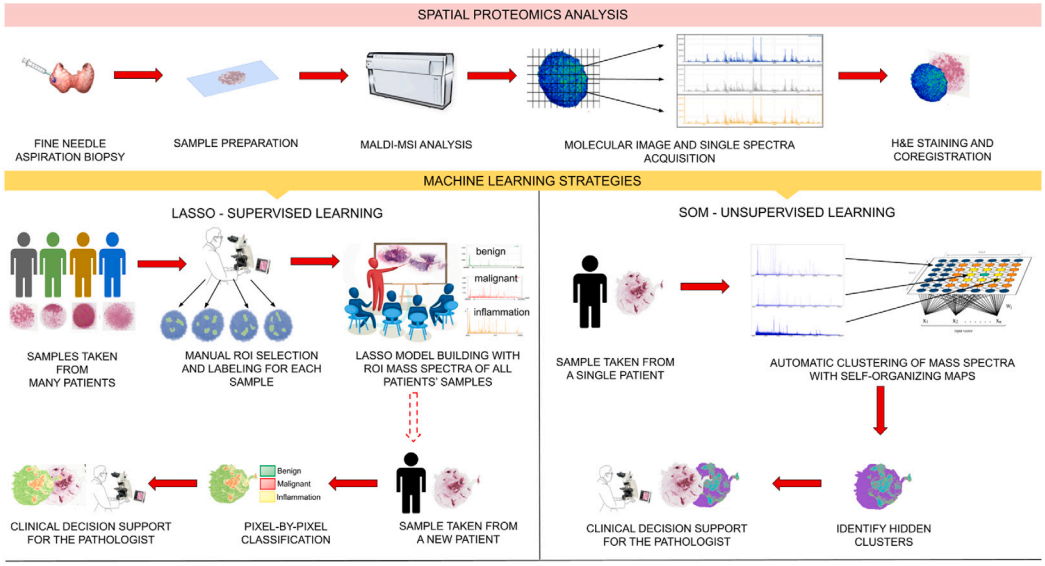

Fig. 1. *Top:* General workflow of the mass spectrometry sample preparation and analysis. FNA samples are collected and prepared to be analyzed with MALDI-MSI technique, resulting in the extraction of a single spectrum for each pixel of the sample. The H&E images are then coregistered with the molecular ones, combining the morphological information with the proteomic profile. *Bottom:* Strategies to analyze mass spectrometry data by means of supervised (LASSO) and unsupervised (SOM) learning models. The statistical LASSO model (*left side*) for the classification of thyroid nodules is constructed relying on mass spectra extracted from regions of interest (ROIs) manually selected from samples of many patients. Differently from LASSO, DSUUL (*right side*) takes as input sample data from a single patient, which is then automatically clustered by means of the SOM. The result of clustering is mapped to the sample coordinates in the H&E slide to provide the pathologist with an image highlighting the presence and separation of any region characterized by different cell types (benign, malignant, inflammatory) or noise.

computerized indications, diagnostic systems as DSUUL might help in confirming their clinical judgement or backtracking from the initial interpretation of the patient data. This outcome would become particularly advantageous in countries or remote regions where access to pathology experts of various medical specialties is not fully available, as it might reduce the use of more invasive or risky diagnostics exams. DSUUL could also give hints on the visual interpretability and the explainability of human diagnosis – two highly debated aspects related to trustworthy AI and ethical issues (Amann et al., 2022) – thanks to its possibility of inspecting specific areas of a cytology sample and comparing them with the corresponding proteomic profiles to better characterize the cellular components therein.

2. Methods and materials

2.1. Pathology

On a cohort of 8 patients US-guided FNAs were performed using a 25-gauge needle at the Department of Radiology, San Gerardo Hospital, Monza, Italy. One or two passes per nodule were executed and needle washing from every pass was sent for proteomics MALDI-MSI analysis following standard clinical procedure (Capitoli et al., 2022; Piga, Capitoli, Tettamanti, Denti, Smith, Chinello, Stella, Leni, Garancini, Galimberti, Magni, & Pagni, 2019b). Pathologists evaluated the corresponding Pap-stained smears for traditional morphological diagnosis certifying the existence of diagnostic criteria (e.g., the presence of benign and malignant thyrocytes clusters, a diffuse lymphocytic infiltrate and oncocytic changes of epithelial cells). The study was approved by the ASST Monza Ethical Board (Associazione Italiana Ricerca sul Cancro-AIRC-MFAG 2016 Id. 18445, HSG Ethical Board Committee approval October 2016, 27/10/2016). Appropriate informed consent was obtained from patients included in the study.

2.2. MALDI-MSI

Needle washing biopsies from thyroid FNAs were collected after which samples were prepared with a previously optimized protocol to provide reproducible results, and finally they were transferred as a cytospin spot onto ITO glass slides (Piga et al., 2019a, 2019b). MALDI-ToF-MSI was performed using an ultrafleXtreme MALDI-ToF (Bruker Daltonik GmbH, Bremen, Germany) in positive-ion linear mode, using 300 laser shots per spot, with a laser focus setting of 3 medium (diameter of 50 μm) and a pixel size of 50 × 50 μm. Protein Calibration Standard I (Bruker Daltonics, Billerica, MA, USA), which contained a mixture of standard proteins within the mass range of 5,730 to 16,950 Da, was used for external calibration (mass accuracy ±30 ppm). Spectra were recorded within the 3,000–20,000 m/z range. Data acquisition and visualization were performed using the Bruker software packages (flexControl 3.4, flexImaging 5.0). After the analysis, the slides were stained with hematoxylin and eosin (H&E), digitally scanned using a ScanScope CS digital scanner (Aperio, Park Center Dr., Vista, CA, USA), and images were coregistered to the MSI datasets in flexImaging for the integration of proteomic and morphological data (Capitoli et al., 2020).

2.3. Statistical model

Statistical analysis of proteomic data was performed for each patient on all the single spectra of the imzML MALDI-MSI files (pixel-by-pixel). A previously published multinomial regression with a LASSO regularization method (Tibshirani, 1996) was used to quantify the probability of the sample being malignant, hyperplastic or Hashimoto thyroiditis (Capitoli et al., 2020). For each pixel the probabilities to belong to the three aforementioned classes were calculated. The highest of the three indices obtained from each single spectrum was used to classify the corresponding pixel. Data preprocessing (`MALDIquant` package Gibb & Strimmer, 2012) and statistical analyses (`glmnet` package Friedman, Hastie, & Tibshirani, 2010) were performed using the open-source R software version 3.6.0.

2.4. Self-organizing maps

Self-Organizing Maps (SOMs) – also known as Kohonen maps or Kohonen neural networks – are a class of artificial neural networks (ANNs) that are trained in an unsupervised learning fashion (Kohonen, 2012). A SOM does not require any labeled data as the algorithm learns by observation instead of learning by examples. This feature

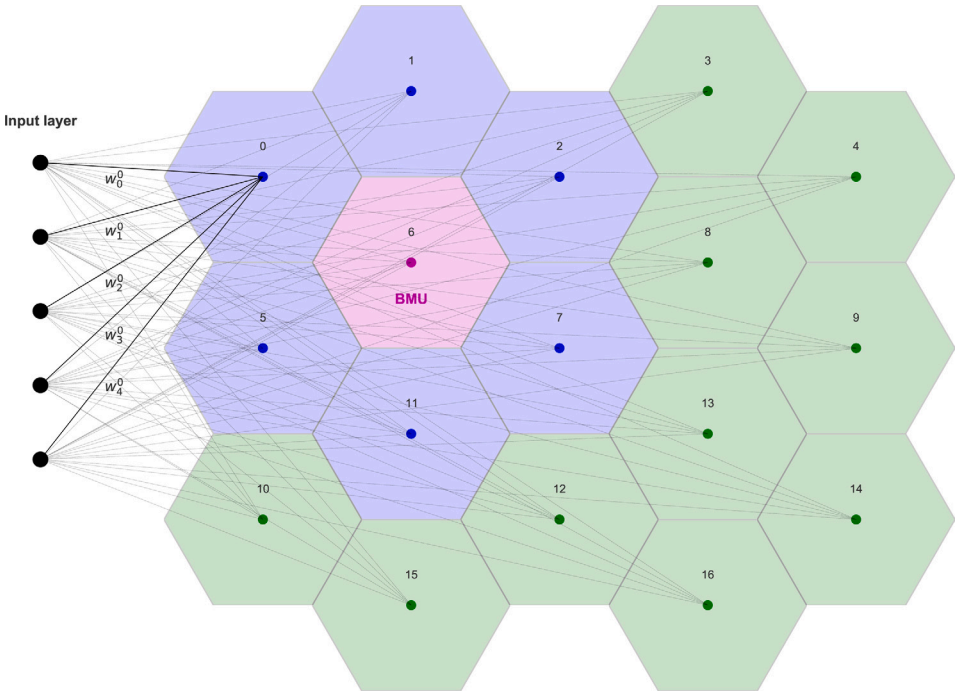

Fig. 2. Architectural scheme of a Self-Organizing Map. All neurons in the input layer (black dots) are connected to the neurons, called units, in the output layer (colored dots); bold lines represent the connections between the input layer neurons and unit #0, whose weights are denoted by $w_0^0, w_1^0, \dots, w_{m-1}^0$ (here, $m = 5$). The magenta neuron denotes the Best Matching Unit (BMU). The neighborhood of the BMU is denoted by blue neurons. The green neurons do not belong to the BMU's neighborhood, and thus they are not updated in the current iteration.

allows SOMs to discover hidden patterns within the data: the main goal of SOMs is to transform a complex high-dimensional input space into a two-dimensional discrete map, whilst preserving any existing topological relationships in the data (Asan & Ercan, 2012).

A SOM is conceptually composed of two layers of neurons: the input layer and the output layer. Since all neurons of the input layer are connected to all neurons in the output layer, a SOM is a completely connected ANN, as shown in Fig. 2. The input layer receives the data samples presented to the SOM for the training process, which are encoded as m-dimensional vectors whose elements represent the features of the dataset. The output layer is composed of interconnected neurons, named units, which are usually organized in a $u \times v$ lattice, for some user-defined integer numbers u and v. Each unit is connected to its neighboring units, using a rectangular or hexagonal neighborhood (see an example in Fig. 2). At the end of the learning process, the output layer will provide a low-dimensional representation of the input data.

A weight vector $\mathbf{w}_i = (w_0^i, \dots, w_{m-1}^i)$, $\mathbf{w}_i \in \mathbb{R}^m$, is associated with each connection among the input neurons and the ith output neuron. In the case of MALDI-MSI spectra analysis presented here, $m = 8,000$ since the histograms representing the spectra are composed of $8,000$ values. The initial weights of the SOM can be randomly generated using one of the existing initialization methods (Attik, Bougrain, & Alexandre, 2005), e.g., by means of random values taken from the input data, or random samples extracted from the subspace defined by the first two eigenvectors identified by principal components analysis on the dataset. According to Akinduko, Mirkes, and Gorban (2016), a purely stochastic initialization of the weights outperforms all other methods in the case of non-linear datasets; this approach was chosen as the default in this work.

The peculiar learning process exploited by SOMs is known as *competitive learning* (Asan & Ercan, 2012), whereby each input vector that is fed to the SOM is simultaneously processed by all units in the output layer. The units compete and the output layer neuron whose weight vector is most similar to the input vector is declared the winner; this neuron takes the name of Best Matching Unit (BMU, represented by the magenta hexagon in Fig. 2).

The BMU is identified according to a user-specified similarity measure, e.g., Euclidean, absolute value, or cosine distances (Wan, Vidavsky, & Gross, 2002). (Stein & Scott, 1994) showed that, in the context of mass spectrometry data, cosine similarity better differentiates very similar spectra than alternative methods, so that it was selected as similarity metric in this work. The cosine similarity (also known as dot-product distance) between two spectra $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^m$ is calculated as follows:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \cos(\theta) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|}, \tag{1}$$

where $\cdot$ denotes the dot-product and θ is the angle between the two vectors defined by the spectra in the m-dimensional hyperspace.

As soon as a BMU is identified, that neuron begins to interact with its neighbors; the rationale behind this is that, to the aim of preserving the topology of the data, nearby locations in the output layer (i.e., the topological neighborhoods) are supposed to share similar properties. Thus, not only the BMU but also the neurons within its neighborhood are activated (the blue hexagons in Fig. 2), so that they can learn from the same input vector. During this phase, the weights of both the BMU and its neighbors are updated and "pulled" towards the input vector, with a strength that is proportional to the topological proximity to the BMU. The weights update during the tth iteration is based on two hyper-parameters: the learning rate $\alpha(t)$, and the neighborhood size $\sigma(t)$.

The learning rate $\alpha(t)$ determines the rate of change of the weight vectors. It is generally based on a decay function that makes its value gradually decreasing in the interval $(0, 1)$ as a function of the iteration step t. The decay function can be linear, exponential, geometrical, inversely proportional, or user-defined. The following formula shows how the weight vector $\mathbf{w}_b$ of the BMU is updated using the learning rate:

$$\mathbf{w}_b(t+1) = \mathbf{w}_b(t) + \alpha(t)\big(\mathbf{x}(t) - \mathbf{w}_b(t)\big), \tag{2}$$

where $\mathbf{x}$ is the current sample shown to the SOM.

The neighborhood size $\sigma(t)$ determines to which extent the BMU influences the activation of the neighbor neurons at iteration t; stated



Fig. 3. The four phases of DSUUL workflow: (1) data pre-processing, (2) optimization of SOMs' hyper-parameters, (3) unsupervised learning and data discrimination, (4) outcome for clinical support.

otherwise, it represents the width or the radius of the neighborhood at each iteration. The rate of modification of the weight vectors in the neighborhood decreases according to a decay function that takes into account both the distance of the neighbor neuron from the BMU and the number of iterations. This decay function is also called the neighborhood function and can either be discrete or continuous. The most widespread neighborhood function is the Gaussian function:

$$h_{bi}(t) = \exp\left(-\frac{d_{bi}^2}{2\sigma(t)^2}\right), \tag{3}$$

where d_{bi} denotes the lateral distance between the BMU (denoted by b) and the excited neuron i. In the specific case of the BMU (i.e., neuron $i = b$), the value of the function $h_{bb}(t)$ is equal to 1.

After both the neighborhood and its neighborhood function are defined, the weight vectors of the units in the neighborhood are updated using the following formula:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \alpha(t)h_{bi}(t)\big(\mathbf{x}(t) - \mathbf{w}_i(t)\big). \tag{4}$$

The learning process is iterated until a termination criterion is met, e.g., a fixed number of iterations $t_{\max}$ is reached, or the change of the weights is smaller than some user-defined threshold value. In this work, we use the former halting criterion.

At the end of the learning process, each sample in the dataset will be recognized and mapped by a specific BMU. Similar samples in the dataset will be mapped to BMUs that are topologically close in the network, and characterized by similar weights. By clustering the units according to their similarity, it is possible to aggregate similar samples to the aim of discriminating one or more classes that are present (though hidden) in the dataset.

3. Results

3.1. DSUUL: automatic discrimination of malignant spectra with SOMs

DSUUL is a novel framework that supports the pathology diagnosis by exploiting unsupervised machine learning. Specifically, DSUUL automatically separates MALDI spectra by means of SOMs; the main phases of its functioning are schematized in Fig. 3.

DSUUL is applied here to mass spectra obtained from the MALDI-MSI analysis of the biopsies taken from each individual patient in the cohort. In particular, each MALDI-MSI analysis extracted a number of spectra ranging from $13,055$ to $20,421$.

Phase 1. The first phase of DSUUL consists in the pre-processing of spectra; in order to retain as much information as possible for the discrimination process, this phase consists only of normalization (total ion current, TIC method) and peak alignment (mean absolute deviation (MAD) noise estimation method and a half window width equal to 5. The TIC method divides each intensity by the sum of all intensities in the mass spectrum, resulting in the same intensity range across spectra and consequently allowing the comparison among spectra within the sample as well as spectra from other samples. This pre-processing step corrects for slight differences in m/z-values so that the same proteins can be identified among spectra: the method applies a cubic warping function to match each peak's m/z to the nearest peak's m/z of a mean spectrum acting as a reference within a given tolerance. Peak alignment is essential for the application of SOMs, since the underlying algorithm uses a distance measure to compute the similarity between the input vector and the output neuron; a misalignment of the peaks would lead to faulty similarity matching and inaccurate results. The pre-processing phase is performed by using the R `MALDIquant` library (Gibb & Strimmer, 2012).

Phase 2. The second phase of DSUUL consists in the optimization of SOM's hyper-parameters to maximize the discrimination performance. We automatically determine the size of the map, i.e., the values u and v that correspond to the number of neurons in the rows and columns of the output layer, respectively. The choice of the map size is particularly critical, because it influences the accuracy and the generalization capability of the SOM and, in turn, of DSUUL. Although there are no strict rules or best practices for selecting the map size, in this work we adopted a widespread heuristic presented by Vesanto and Alhoniemi (2000), which states that the overall number of neurons can be calculated as $uv = 5\sqrt{k}$, where k represents the number of input vectors in the dataset. As default setting, DSUUL exploits a hexagonal topology with a Gaussian neighborhood, as suggested by Asan and Ercan (2012), and performs $t_{\max} = 10,000$ iterations. The other hyper-parameters – notably, σ and α values – are automatically determined in a preliminary phase by exploiting the `hyperopt` Python library (Bergstra, Komer, Eliasmith, Yamins, & Cox, 2015). DSUUL executes 200 runs to determine the best settings for these hyper-parameters. The SOM was implemented with the `minisom` library (Vettigli, 2018) version 2.2.7. The imzML files produced by MALDI experiments were imported with the `pyimzml` library, version 1.2.6.

Phase 3. The third phase of DSUUL consists of actual unsupervised learning with the SOM. When the training is completed, the neurons are clustered by means of the agglomerative clustering algorithm implemented in `scikit-learn` (Pedregosa et al., 2011). Specifically, the clustering algorithm aims at discriminating the spectra in the following four classes:

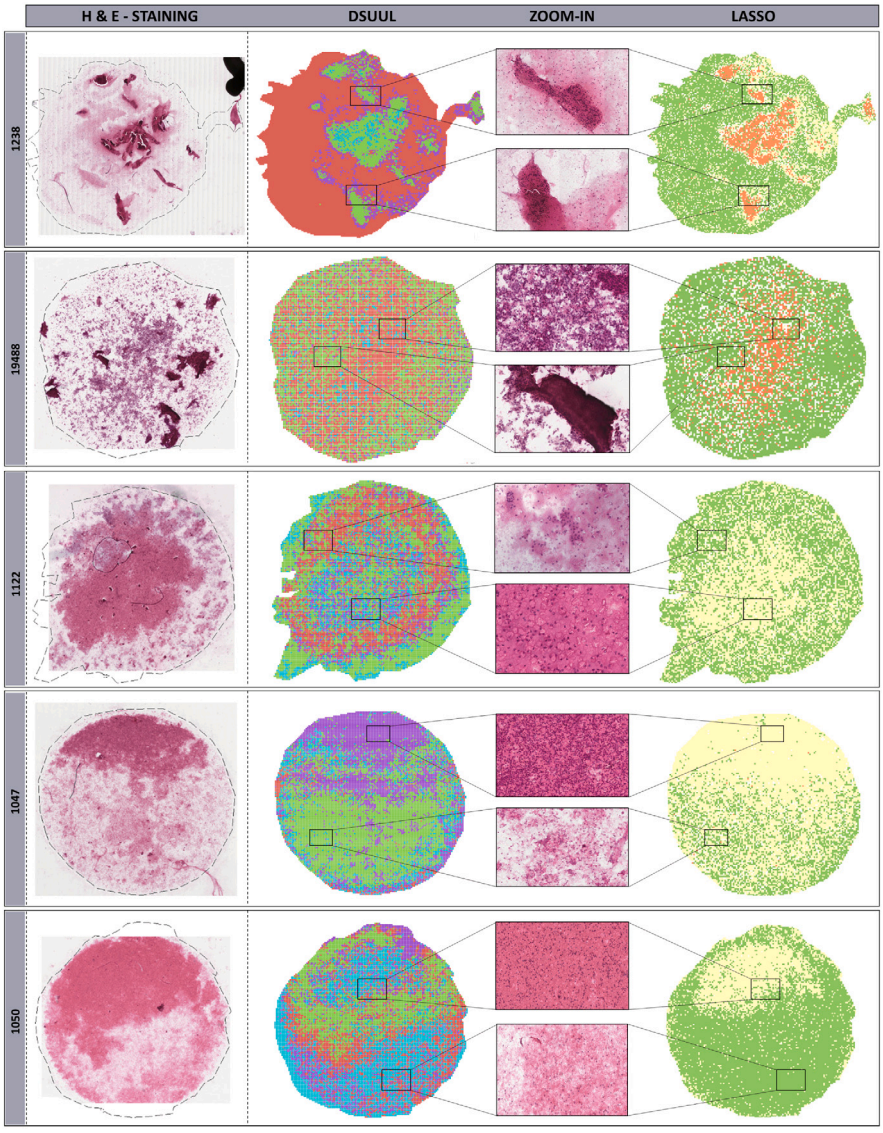

Fig. 4. Comparison between DSUUL and LASSO model: correct clustering of different cell populations in various samples. 1st column: H&E staining images; 2nd column: DSUUL clustering results; 3rd column: zoom-in of the morphological images; 4th column: results of the classification model (LASSO). For the LASSO model, the green, yellow, red, and white pixels correspond to epithelial cells, inflammatory background, malignant cells and empty/noise spectra, respectively.

1. benign cells/epithelial cells;
2. malignant cells/malignant thyrocytes;
3. any other cells/inflammatory background;
4. noise/empty spectra.

Empty profiles regard spectra containing several empty bins (i.e., intensity values equal to 0). These spectra are caused by points in the sample where the MALDI-MSI instrument was unable to extract proteins due to the absence of cells in that particular area of the sample. The noisy profiles refer to spectra rich in signal, but consisting of very low and similar intensity values due to instrumental background noise.

Phase 4. At the end of the clustering process, all units are assigned to one of the four aforementioned classes. Since all samples in the dataset are associated with a specific BMU, each sample is transitively associated with the same class of the unit. This association is mapped to the sample coordinates to visually represent the discrimination of pixels in the four classes. This representation is aimed at providing the pathologist with an immediate perception of the presence and separation of regions, if any, characterized by different cell types, most notably benign or malignant thyrocytes.

3.2. DSUUL analysis outcome

The dataset analyzed in this work consists of multiple FNA biopsy samples of thyroid nodules. The outcome of the proteomic MALDI-MSI analysis of these samples was processed by DSUUL to assess its capability in distinguishing different molecular signatures.

Figs. 4 and 5 show the results achieved by DSUUL, together with a pixel-by-pixel comparison with the results obtained with the LASSO model (Capitoli et al., 2020). The two figures include samples taken from 8 patients – with ID 268, 993, 1047, 1050, 1084ev, 1122, 1238, 19488 – who show various lesions, as highlighted by the presence of different types of entities within the specimens (first column, H&E staining images). In particular, the biological heterogeneity investigated in the analyses presented here includes 2 Papillary Thyroid Carcinoma (PTC), 1 Hyperplastic (HP), 3 Hashimoto Thyroiditis (HT), and 2 Noninvasive Thyroid Neoplasm with Papillary-like Nuclear Features (NIFTP).

In general, the outcome of DSUUL is in good agreement with both the cytological diagnosis and the MALDI-MSI data, with only a few exceptions. In the case of patient 1238 (NIFTP), shown in the first row of Fig. 4, specific groups of spectra – corresponding to epithelial cells

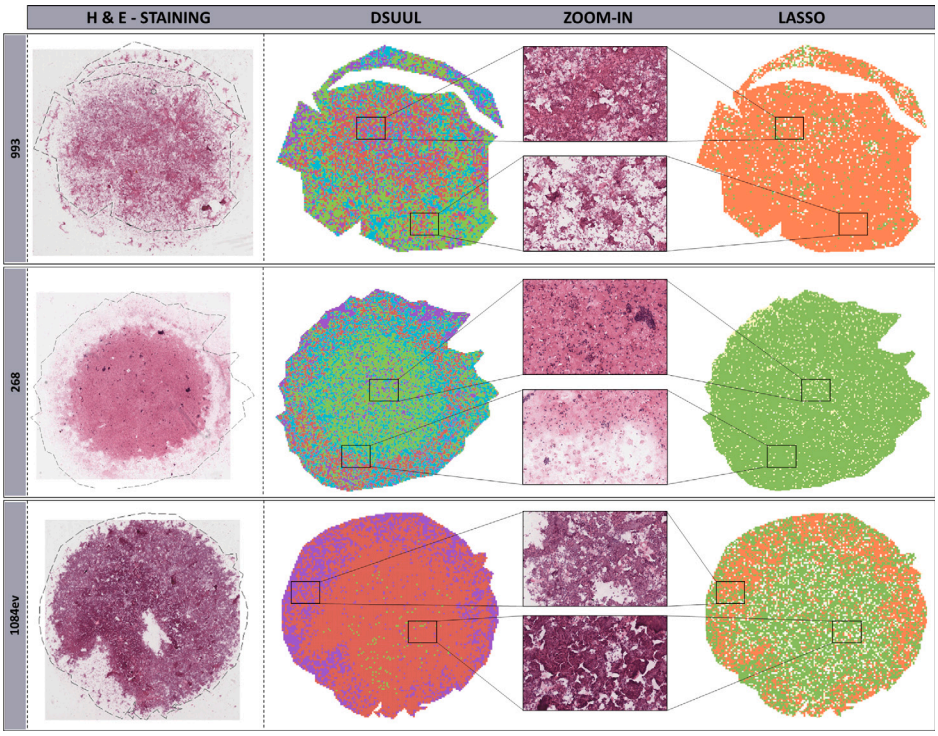

Fig. 5. Comparison between DSUUL and LASSO model: representative cases of homogeneous samples. Due to the homogeneous distribution of cellular entities, DSUUL is not able to differentiate clusters of different types of spectra but only to differentiate them based on the cellular concentration or quality of samples. 1st column: H&E staining images; 2nd column: DSUUL clustering results; 3rd column: zoom-in of the morphological images; 4th column: results of the classification model (LASSO). For the LASSO model, the green, yellow, red, and white pixels correspond to epithelial cells, inflammatory background, malignant cells and empty/noise spectra, respectively.

(red), lymphocyte/oncocytic background (purple), and malignant cells (green/cyan) – were correctly separated in different clusters according to the classification obtained with the LASSO model. These results also highlight the ability of proteomic MALDI-MSI analysis to generate specific molecular signatures representing different entities. It is worth noting that DSUUL performed a blind clinical assessment of cytological specimens, showing an excellent agreement with the results provided by a supervised classifier (Capitoli et al., 2020). The different proteomic NIFTP profiles were also distinguishable in patient 19488, in which the purple cluster corresponds to the presence of specific "signals of alert" for malignancy that can be identified by the pathologist in the pixel-by-pixel classification (Piga et al., 2020), while the red cluster represents colloidocistic areas, as confirmed in the zoom-in of the H&E image (reported in the second row of Fig. 4). While the red cluster may appear sparse at first observation, it is in agreement with the morphological image as smaller regions of thyrocytes are present throughout the sample. In the field of thyroid tumors, NIFTP lesions represent a heterogeneous group of nodules that often require surgical treatment. These lesions can be diagnosed as NIFTP only after thyroidectomy with the histological evaluation. While the detection of possible differences at the level of proteomic spectra of thyroid FNAs is a very challenging issue (Canini et al., 2019; Piga et al., 2020), DSUUL proved its capability in correctly distinguishing lymphocytes from epithelial cells based on specific signals, as illustrated in Fig. 4, where lymphocytes background (cyan and violet clusters) and epithelial cells (red and green clusters) were localized in agreement with H&E.

The same level of accuracy in the results holds for the sample of patient 1122, which is characterized by a mixture of epithelial and inflammatory background, and for the samples of patients 1047 and 1050, which show an undeniable evidence of homogeneous lymphocytes in the higher part of the samples (see the last three rows of Fig. 4). These specific signals, localized in an area with an abundance of lymphocytes and epithelial cells, were previously investigated (Capitoli et al., 2020). The zoom-in of the H&E image of patient 1122 presents areas rich in inflammatory background (e.g., granulocytes, oncocytes) in contrast to the poor background within the better defined cluster of benign thyrocytes. Overall, these results were found in rather good agreement with the presence of epithelial and lymphocytes cells based on the pixel-by-pixel classification elaboration of the MALDI-MSI data.

Patients 993 and 268 (HP and PTC, respectively) are representative cases of homogeneous samples in terms of their morphology (see Fig. 5), which represent a complex and challenging scenario for the automatic identification of proteomics profiles of different entities. The obtained results highlight the need for further investigation for this specific case. In homogeneous samples, DSUUL was not able to differentiate specific spectra profiles even in the presence of different types of cells; on the contrary, it identified clusters according to the amount of cells, thus separating highly populated areas from those with a paucity of cells. It is worth noting that this is a consequence of the proteomic MALDI-MSI analysis, executed with an *ad hoc* spatial resolution (> 20 μm, as in the case of this study) that does not allow for extracting spectra profiles at a single-cell level. A different aspect is shown in the sample of patient 1084ev (last row in Fig. 5), which is characterized by an homogeneous presence of malignant cells that the LASSO model was able to detect only in the contour of the sample, due to the fact that the center is characterized by the presence of a dry area that did not allow for extracting enough molecules to obtain specific spectra of malignant cells. The same clustering result was achieved by DSUUL, which identified the presence of two main different clusters (red and purple, Fig. 5).

Fig. 6 reports examples of spectra taken from different clusters. For what concerns patient 1238, we show the spectra representative of the two main clusters that were highlighted in Fig. 4: here, it is possible to observe the expression of signals intensities in different regions of the mass spectra associated with specific cellular morphological characterizations of malignant thyrocytes (green cluster) and lymphocytes (purple cluster). The plot referring to cluster 1 of patient 1238 is characterized by the hypothetical signals of malignancy reported in Piga

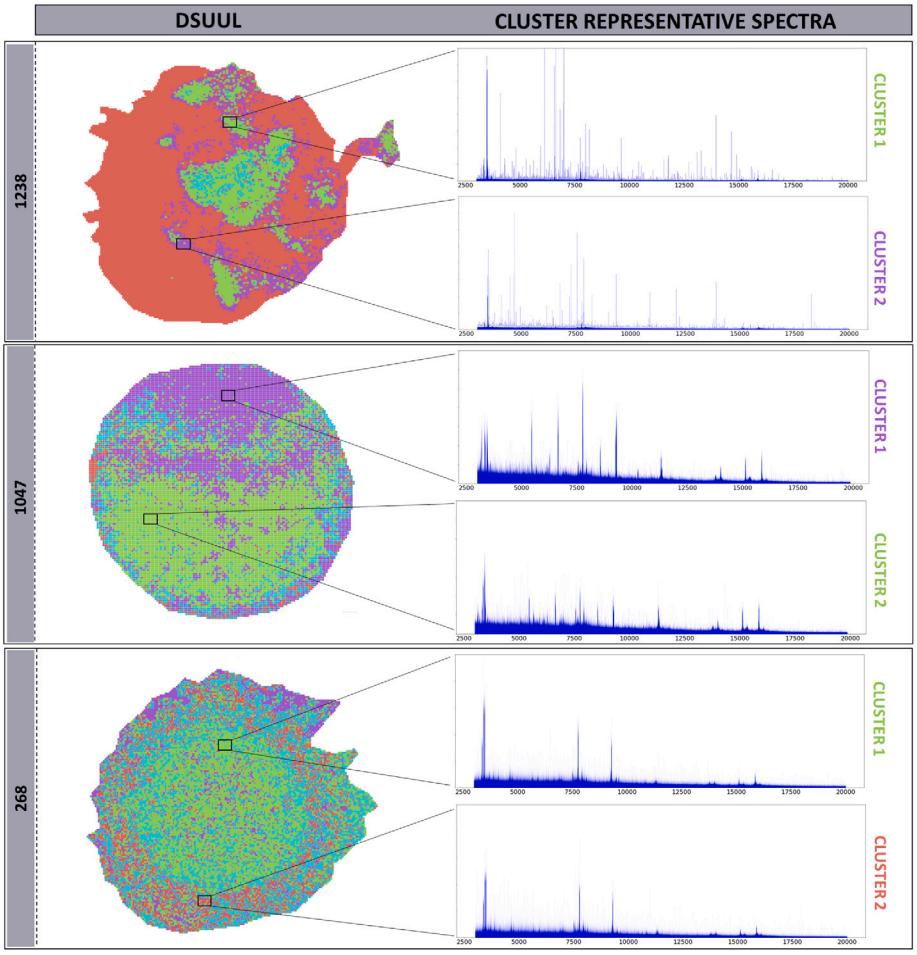

Fig. 6. Examples of clusters' representative spectra. In the case of patient 1238, two clusters are considerably different, representing the classes of malignant and benign/inflammatory cells. In the case of patient 1047, the spectra might seem similar but there are relevant differences due to the cellular relative amounts. In the case of patient 268, the spectra are too similar and DSUUL was not able to properly discriminate between the four classes.

et al. (2020) due to the fact that this sample falls in the pre-malignancy class (NIFTP); the spectrum of cluster 2 of patient 1238 represents inflammatory background, as already seen in the morphological zoom-in of Fig. 4. In addition, the spectra profile of cluster 2 of patient 1238 has signals in common with the spectra profile of cluster 1 of patient 1047, which is a typical case of Hashimoto thyroiditis, with a high presence of lymphocytes. Patient 1047 shows spectra profiles that differ in the intensity of the signals, due to the different cellular amount of inflammatory background present in the upper part of the sample. It is worth mentioning that even if the spectra of the two clusters of patient 1047 may look similar, from a proteomic point of view they are very dissimilar, representing different cellular amounts in the sample and also denoting the presence of different signals, as already reported in Capitoli et al. (2020). Conversely, patient 268 is one of the cases in which DSUUL was not able to properly discriminate between the four classes. The heterogeneity of cells in this sample, together with their homogeneity in the spatial distribution, led DSUUL to misidentify different clusters as a result of the data obtained from proteomic MALDI-MSI analysis executed with 50 μm laser diameter. This is reflected on the representative alpha blended spectra profile reported in Fig. 6, which appear to be very similar, thus bringing to the same conclusion from a proteomic point of view.

As a final test, we investigated the influence of different agglomerative clustering methods, implemented using `scikit-learn` (Pedregosa et al., 2011), on the outcome of DSUUL. As shown in Fig. 7, in almost all cases DSUUL achieved the same results, with the exception of K-means (for the particular case of patient 1047) and of Single linkage. The K-means algorithm clusters data by trying to separate samples in n-groups of equal variance. As already seen in Fig. 6, the majority of the data related to patient 1047 represent the molecular information of inflammatory background, characterized by similar spectra (with the same expressed proteins signals, with different intensities). Due to this fact, K-means included all the spectra in the same cluster, despite having different intensities of the same peaks (which underlines a morphological difference that was not detected). Instead, the Single linkage works on minimizing the distance between the closest observations of pairs of clusters, and it resulted to be the less robust approach when dealing with noisy data, as in the case of these analyses.

4. Discussions & conclusions

The present study introduces DSUUL, an original methodological approach that exploits the unsupervised learning ability of SOMs in (thyroid) cytopathology, by taking advantage of data measured with the MALDI-MSI technology. The capability of SOMs to identify clusters of different proteomic profiles has never been investigated before on thyroid liquid biopsy specimens. Here, we have shown their feasibility on samples characterized by four different proteomic profiles that correspond to the presence of benign and epithelial cells, malignant cell and thyrocytes, other cells and inflammatory background, or noise and empty spectra. An additional advantage of DSUUL is the possibility to overlap the different areas of the H&E stained slides and the proteomic profile image, which allows to verify the specific weight of every cellular component of FNA samples. This is particularly important

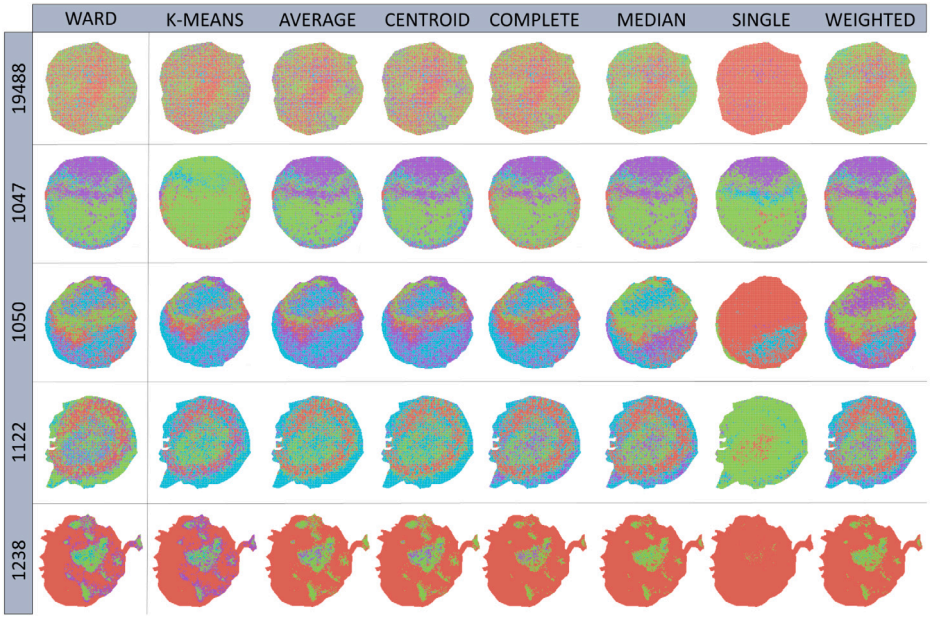

Fig. 7. Outcome of DSUUL exploiting different clustering methods. As default setting, DSUUL uses the Ward method.

since FNA provides challenging specimens in terms of heterogeneity, with inter-sparsed epithelial cells, macrophages, lymphocytes and/or malignant thyrocytes. Despite the technical challenges of this study, the coupled application of proteomics and imaging may help to elucidate key biomolecular events and pathways in oncogenic processes. DSUUL indeed represents an effective means to assist pathologists, since it automatically identifies morphological regions based only on experiment-specific information – that is, mass spectra – without any further action. This is in contrast to what happens in a typical diagnostic scenario where, generally, the ROIs in a given cytological image are visually inspected and annotated by the pathologist after the MALDI-MSI analysis. Since the cytological image and the corresponding molecular image might be affected by slight differences, which hinder their co-registration, the manual identification of the ROIs cannot ensure an exact correspondence between the cell morphology information and the related mass spectra data.

In our tests we observed that the performance of SOMs can be hampered in the case of unbalanced (e.g., homogeneous) datasets and in the presence of empty and/or noisy spectra. These two circumstances warrant further investigations to promptly detect these situations and provide clear feedback to the user. For example, in the case of homogeneous samples, information can instead be provided on the quality and subtypes of different cellular phenotypes within the benign or malignant classes. We plan to extend DSUUL to analyze and compare the identified clusters against reference profiles from existing proteomic libraries (benign, malignant and inflammatory spectra) through similarity measures. Following the promising agreement shown between SOM's clustering and the results obtained with the supervised classifier and the morphological image, we can assess the independent (unsupervised) identification of the clusters made by DSUUL. So doing, DSUUL will be able to (*i*) assist the pathologist in the complex decision making processes, with the aim of obtaining insights into the distribution of cellular entities in the sample, also allowing for reviewing the representative spectra of each cluster, and (*ii*) provide the pathologist with specific features to integrate the library with atypical tumor entities and broaden the spectrum of cases that can be compared to the new experimental data.

In order to carry out the discrimination of malignant cells in a personalized way, DSUUL systematically re-trains a SOM from scratch exploiting the patient's own MALDI-MSI data. However, due to the large number of samples and neurons in the SOM, both the training of the network and the generation of figures can be computationally expensive. In our tests, performed on a workstation equipped with an Intel Core i7-7700HQ CPU @ 2.80 GHz and 16 GB of RAM, the whole process required approximately 2 h. Nevertheless, many portions of DSUUL's code are intrinsically parallel (e.g., the determination of the BMU for each sample) so that they could be offloaded to a SIMD (same instruction, multiple data) co-processor, most notably the Graphics Processing Unit (GPU). As a future development, we aim at integrating in DSUUL the GPU-based SOM implemented in `CUDA-SOM` (Rundo et al., 2021), in order to strongly reduce the running time and highlight potential regions characterized by malignant cells within a few minutes.

Moreover, we plan to analyze the performance of DSUUL using different map sizes of the SOM, that is, different numbers of neurons, and alternative SOM models, such as dynamic and growing grids (Barbalho, Costa, Neto, & Netto, 2003), which might affect the clustering accuracy.

Finally, we are currently developing a user-friendly graphical user interface, designed to make the analysis and interpretation of DSUUL's results faster and easier. We envision a system where pathologists can drag and drop a imzML file directly into DSUUL and obtain, within a few minutes, an insight about the distribution of cellular entities in the sample, with the possibility of reviewing the representative spectra of each cluster, and rapidly identify and highlight groups of malignant cells. DSUUL can potentially complement clinicians in their cytological diagnosis of FNA samples.

CRediT authorship contribution statement

Marco S. Nobile: Conceptualization, Computational methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Project administration. **Giulia Capitoli:** Conceptualization, Computational methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization, Project administration. **Virgil Sowirono:** Maldi computational methodology, Software, Formal

analysis, Investigation, Data curation, Writing – original draft, Visualization. **Francesca Clerici:** MALDI-MSI methodology, Resources, Visualization. **Isabella Piga:** MALDI-MSI methodology, Resources. **Kirsten van Abeelen:** Writing – original draft. **Fulvio Magni:** Resources, Writing – review & editing, Supervision. **Fabio Pagni:** Resources, Writing – review & editing, Supervision, Funding acquisition. **Stefania Galimberti:** Resources, Writing – review & editing, Supervision. **Paolo Cazzaniga:** Writing – original draft, Supervision. **Daniela Besozzi:** Writing – original draft, Supervision.

Data availability

The data that has been used is confidential.

Acknowledgment

All authors approved the version of the manuscript to be published.

Funding

This research was funded by Regione Lombardia POR FESR 2014-2020, Call HUB Ricerca ed Innovazione: Immun-HUB, Regione Lombardia, regional law n. 9/2020, resolution n. 3776/2020: Programma degli interventi per la ripresa economica: sviluppo di nuovi accordi di collaborazione con le universitá per la ricerca, l'innovazione e il trasferimento tecnologico: NephropaThy, Associazione Italiana Ricerca sul Cancro Grant - AIRC-MFAG 2016 Id. 18445, and Ricerca Finalizzata GR-2019-12368592.

Statements of ethical approval

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of ASST Monza HSG (protocol code 18445 and date of approval 27/10/2016).

Informed consent statement

The study was carried out in accordance with the relevant guidelines and regulations. It was approved by the ASST Monza Ethical Board (Associazione Italiana Ricerca sul Cancro - AIRC-MFAG 2016 Id. 18445, HSG Ethical Board Committee approval October 2016, 27/10/2016), and study participants signed an informed consent.

References

Akinduko, A. A., Mirkes, E. M., & Gorban, A. N. (2016). SOM: Stochastic initialization versus principal components. *Information Science*, *364*, 213–221. http://dx.doi.org/10.1016/j.ins.2015.10.013.

Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., et al. (2022). To explain or not to explain?– artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, *1*, Article e0000016. http://dx.doi.org/10.1371/journal.pdig.0000016.

Asan, U., & Ercan, S. (2012). An introduction to self-organizing maps. *Computational Intelligence Systems in Industrial Engineering*, 295–315.

Attik, M., Bougrain, L., & Alexandre, F. (2005). Self-organizing map initialization. *International Conference on Artificial Neural Networks*, 357–362. http://dx.doi.org/10.1007/11550822_56.

Barbalho, J., Costa, J., Neto, A., & Netto, M. (2003). Hierarchical and dynamic SOM applied to image compression. In *Proceedings of the international joint conference on neural networks, vol. 1* (pp. 753–758). IEEE.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., & Cox, D. D. (2015). Hyperopt: a Python library for model selection and hyperparameter optimization. *Computer Science Discoveries*, *8*, Article 014008.

Borkowska, E. M., Kruk, A., Jedrzejczyk, A., Rozniecki, M., Jablonowski, Z., Traczyk, M., et al. (2014). Molecular subtyping of bladder cancer using Kohonen self-organizing maps. *Cancer Medicine*, *3*, 1225–1234. http://dx.doi.org/10.1002/cam4.217.

Canini, V., Leni, D., Pincelli, A. I., Scardilli, M., Garancini, M., Villa, C., et al. (2019). Clinical-pathological issues in thyroid pathology: study on the routine application of NIFTP diagnostic criteria. *Science Reports*, *9*, http://dx.doi.org/10.1038/s41598-019-49851-1.

Capitoli, G., Piga, I., Clerici, F., Brambilla, V., Mahajneh, A., Leni, D., et al. (2020). Analysis of Hashimoto's thyroiditis on fine needle aspiration samples by MALDI-Imaging. *Biochimica et Biophysica ACTA (BBA) - Proteins and Proteom*, *1868*, Article 140481.

Capitoli, G., Piga, I., Galimberti, S., Leni, D., Pincelli, A. I., Garancini, M., et al. (2019). MALDI-MSI as a Complementary Diagnostic Tool in Cytopathology: A Pilot Study for the Characterization of Thyroid Nodules. *Cancers*, *11*(1377), http://dx.doi.org/10.3390/cancers11091377.

Capitoli, G., Piga, I., L'Imperio, V., Clerici, F., Leni, D., Garancini, M., et al. (2022). Cytomolecular classification of thyroid nodules using fine-needle washes aspiration biopsies. *International Journal of Molecular Sciences*, *23*, http://dx.doi.org/10.3390/ijms23084156.

Cordes, J., Enzlein, T., Marsching, C., Hinze, M., Engelhardt, S., Hopf, C., et al. (2021). M2aia–interactive, fast, and memory-efficient analysis of 2D and 3D multimodal mass spectrometry imaging data. *GigaScience*, *10*, http://dx.doi.org/10.1093/gigascience/giab049, giab049.

Datta, S., & DePadilla, L. M. (2006). Feature selection and machine learning with mass spectrometry data for distinguishing cancer and non-cancer samples. *Statistical Methodology*, *3*, 79–92. http://dx.doi.org/10.1016/j.stamet.2005.09.006.

DeHoog, R. J., Zhang, J., Alore, E., Lin, J. Q., Yu, W., Woody, S., et al. (2019). Preoperative metabolic classification of thyroid nodules using mass spectrometry imaging of fine-needle aspiration biopsies. *Proceedings of the National Academy of Sciences*, *116*, 21401–21408. http://dx.doi.org/10.1073/pnas.1911333116.

Deininger, S.-O., Ebert, M. P., Futterer, A., Gerhard, M., & Rocken, C. (2008). MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *Journal of Proteome Research*, *7*, 5230–5236. http://dx.doi.org/10.1021/pr8005777.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*, 24–29. http://dx.doi.org/10.1038/s41591-018-0316-z.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22. http://dx.doi.org/10.18637/jss.v033.i01.

Gibb, S., & Strimmer, K. (2012). MALDIquant: A versatile R package for the analysis of mass spectrometry data. *Bioinform*, *28*(2270–2271), http://dx.doi.org/10.1093/bioinformatics/bts447.

Günter, S., & Bunke, H. (2002). Self-organizing map for clustering in the graph domain. *Pattern Recognition Letters*, *23*, 405–417. http://dx.doi.org/10.1016/S0167-8655(01)00173-8.

Hautaniemi, S., Yli-Harja, O., Astola, J., Kauraniemi, P., Kallioniemi, A., Wolf, M., et al. (2003). Analysis and visualization of gene expression microarray data in human cancer using self-organizing maps. *Machine Learning*, *52*, 45–66. http://dx.doi.org/10.1023/A:1023941307670.

Henson, D., Spenceley, S. E., & Bull, D. (1997). Artificial neural network analysis of noisy visual field data in glaucoma. *Artificial Intelligence in Medicine*, *10*, 99–113. http://dx.doi.org/10.1016/S0933-3657(97)00388-6.

Kohonen, T. (2012). *Self-organizing maps, vol. 30*. Springer Science & Business Media.

Kurczyk, A., Gawin, M., Chekan, M., Wilk, A., Łakomiec, K., Mrukwa, G., et al. (2020). Classification of thyroid tumors based on mass spectrometry imaging of tissue microarrays; a single-pixel approach. *Journal of Molecular Science*, *21*(6289), http://dx.doi.org/10.3390/ijms21176289.

Markey, M. K., Lo, J. Y., Tourassi, G. D., & Floy. Jr, C. E. (2003). Self-organizing map for cluster analysis of a breast cancer database. *Artificial Intelligence in Medicine*, *27*, 113–127. http://dx.doi.org/10.1016/S0933-3657(03)00003-4.

Mosele, N., Smith, A., Galli, M., Pagni, F., & Magni, F. (2017). MALDI-MSI analysis of cytological smears: The study of thyroid cancer. *Methods in Molecular Biology*, *1618*, 37–47. http://dx.doi.org/10.1007/978-1-4939-7051-3_5.

Ngan, S.-C., Yacoub, E. S., Auffermann, W. F., & Hu, X. (2002). Node merging in Kohonen's self-organizing mapping of fMRI data. *Artificial Intelligence in Medicine*, *25*, 19–33. http://dx.doi.org/10.1016/S0933-3657(02)00006-4.

Obermeyer, Z., & Topol, E. J. (2021). Artificial intelligence bias, and patients' perspectives. *The Lancet*, *397*, 2038. http://dx.doi.org/10.1016/S0140-6736(21)01152-1.

Ostropolets, A., Zhang, L., & Hripcsak, G. (2020). A scoping review of clinical decision support tools that generate new knowledge to support decision making in real time. *The official journal of the American Medical Informatics Association*, *27*, 1968–1976. http://dx.doi.org/10.1093/jamia/ocaa200.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pietrowska, M., Diehl, H. C., Mrukwa, G., Kalinowska-Herok, M., Gawin, M., Chekan, M., et al. (2017). Molecular profiles of thyroid cancer subtypes: Classification based on features of tissue revealed by mass spectrometry imaging. *Biochimica et Biophysica ACTA (BBA) - Proteins and Proteom*, *1865*, 837–845. http://dx.doi.org/10.1016/j.bbapap.2016.10.006.

Piga, I., Capitoli, G., Clerici, F., Brambilla, V., Leni, D., Scardilli, M., et al. (2020). Molecular trait of follicular-patterned thyroid neoplasms defined by MALDI-imaging. *Biochimica et Biophysica ACTA (BBA) - Proteins and Proteom*, *1868*, Article 140511. http://dx.doi.org/10.1016/j.bbapap.2020.140511.

Piga, I., Capitoli, G., Denti, V., Tettamanti, S., Smith, A., Stella, M., et al. (2019a). The management of haemoglobin interference for the MALDI-MSI proteomics analysis of thyroid fine needle aspiration biopsies. *Analytical and Bioanalytical Chemistry*, *411*(5007–5012), http://dx.doi.org/10.1007/s00216-019-01908-w.

Piga, I., Capitoli, G., Tettamanti, S., Denti, V., Smith, A., Chinello, C., et al. (2019b). Feasibility study for the MALDI-MSI analysis of thyroid fine needle aspiration biopsies: evaluating the morphological and proteomic stability over time PROTEOM. *Clinical and Applied*, *13*, Article 1700170. http://dx.doi.org/10.1002/prca.201700170.

Pourkia, J., Rahimi, S., & Baghaei, K. T. (2019). Hospital data interpretation: A Self-Organizing Map approach. In *International fuzzy systems association world congress* (pp. 493–504). Springer, http://dx.doi.org/10.1007/978-3-030-21920-8_44.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*, 1347–1358. http://dx.doi.org/10.1056/NEJMra1814259.

Rundo, L., Tangherloni, A., Cazzaniga, P., Mistri, M., Galimberti, S., Woitek, R., et al. (2021). A CUDA-powered method for the feature extraction and unsupervised analysis of medical images. *The Journal of Supercomputer*, 1–18. http://dx.doi.org/10.1007/s11227-020-03565-8.

Seddiki, K., Saudemont, P., Precioso, F., Ogrinc, N., Wisztorski, M., Salzet, M., et al. (2020). Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification. *Nature Communications*, *11*, 1–11. http://dx.doi.org/10.1038/s41467-020-19354-z.

Shukla, N., Hagenbuchner, M., Win, K. T., & Yang, J. (2018). Breast cancer data analysis for survivability studies and prediction. *Computer Methods and Programs in Biomedicine*, *155*, 199–208. http://dx.doi.org/10.1016/j.cmpb.2017.12.011.

Stein, S. E., & Scott, D. R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, *5*, 859–866. http://dx.doi.org/10.1016/1044-0305(94)87009-8.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits risks, and strategies for success. *Npj Digital Medicine*, *3*, 1–10. http://dx.doi.org/10.1038/s41746-020-0221-y.

Tian, J., Azarian, M. H., & Pecht, M. Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm. In *Proceedings of the European conference of the prognostics and health management society* (pp. 1–9). Citeseer, http://dx.doi.org/10.36001/phme.2014.v2i1.1554.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(267–288), http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Verbeeck, N., Caprioli, R. M., & Plas, R. Van de. (2020). Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass Spectrometry Reviews*, *39*, 245–291. http://dx.doi.org/10.1002/mas.21602.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, *3*, 111–126. http://dx.doi.org/10.3233/IDA-1999-3203.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, *11*, 586–600. http://dx.doi.org/10.1109/72.846731.

Vettigli, G. (2018). MiniSom: minimalistic and NumPy-based implementation of the self organizing map. https://github.com/JustGlowing/minisom/.

Wan, K. X., Vidavsky, I., & Gross, M. L. (2002). Comparing similar spectra: from similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry*, *13*, 85–88. http://dx.doi.org/10.1016/S1044-0305(01)00327-0.

Wang, J., Delabie, J., Aasheim, H. C., Smeland, E., & Myklebost, O. (2002). Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. *BMC Biomedicine*, *3*, 1–9. http://dx.doi.org/10.1186/1471-2105-3-36.

Yamagutchi, T., Nagata, K., Truong, P. Q., Pfurtscheller, G., & Inoue, K. (2007). Pattern recognition of EEG signal during motor imagery by using SOM. In *Second International Conference on Innovative Computing, Information and Control (ICICIC 2007)* (p. 121). IEEE, http://dx.doi.org/10.1109/ICICIC.2007.447.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *67*, 301–320. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.