Computer Methods and Programs in Biomedicine 185 (2020) 105153

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Texture analysis and multiple-instance learning for the classification of malignant lymphomas

Marco Lippi[a,j,k,*], Stefania Gianotti[a], Angelo Fama[c], Massimiliano Casali[d], Elisa Barbolini[i], Angela Ferrari[c], Federica Fioroni[b], Mauro Iori[b], Stefano Luminari[c,f], Massimo Menga[g], Francesco Merli[c], Valeria Trojani[h], Annibale Versari[d], Magda Zanelli[e], Marco Bertolini[b]

[a] Department of Sciences and Methods for Engineering, University of Modena and Reggio Emilia, Italy
[b] Medical Physics, Azienda USL-IRCCS di Reggio Emilia, Italy
[c] Hematology, Azienda USL-IRCCS di Reggio Emilia, Italy
[d] Nuclear Medicine, Azienda USL-IRCCS di Reggio Emilia, Italy
[e] Pathology Unit, Azienda USL-IRCCS di Reggio Emilia, Italy
[f] Surgical, Medical and Dental Department of Morphological Sciences related to Transplant, Oncology and Regenerative Medicine, University of Modena and Reggio Emilia, Italy
[g] Nuclear Medicine, ASUITS, Trieste, Italy
[h] School of Specialization in Health Physics, University of Bologna, Italy
[i] Gr.A.D.E. Onlus Foundation, Reggio Emilia, Italy
[j] Artificial Intelligence Research and Innovation center, University of Modena and Reggio Emilia, Italy
[k] InterMech Center, University of Modena and Reggio Emilia, Italy

ARTICLE INFO

Article history:
Received 13 February 2019
Revised 16 October 2019
Accepted 23 October 2019

Keywords:
Multiple-instance learning
Texture analysis
Malignant lymphomas

ABSTRACT

Background and objectives: Malignant lymphomas are cancers of the immune system and are characterized by enlarged lymph nodes that typically spread across many different sites. Many different histological subtypes exist, whose diagnosis is typically based on sampling (biopsy) of a single tumor site, whereas total body examinations with computed tomography and positron emission tomography, though not diagnostic, are able to provide a comprehensive picture of the patient. In this work, we exploit a data-driven approach based on multiple-instance learning algorithms and texture analysis features extracted from positron emission tomography, to predict differential diagnosis of the main malignant lymphomas subtypes.

Methods: We exploit a multiple-instance learning setting where support vector machines and random forests are used as classifiers both at the level of single VOIs (instances) and at the level of patients (bags). We present results on two datasets comprising patients that suffer from four different types of malignant lymphomas, namely diffuse large B cell lymphoma, follicular lymphoma, Hodgkin's lymphoma, and mantle cell lymphoma.

Results: Despite the complexity of the task, experimental results show that, with sufficient data samples, some cancer subtypes, such as the Hodgkin's lymphoma, can be identified from texture information: in particular, we achieve a 97.0% of sensitivity (recall) and a 94.1% of predictive positive value (precision) on a dataset that consists in 60 patients.

Conclusions: The presented study indicates that texture analysis features extracted from positron emission tomography, combined with multiple-instance machine learning algorithms, can be discriminating for different malignant lymphomas subtypes.

* Corresponding author.
E-mail addresses: marco.lippi@unimore.it (M. Lippi), ste.gianotti93@gmail.com (S. Gianotti), angelo.fama@ausl.re.it (A. Fama), massimiliano.casali@ausl.re.it (M. Casali), elisa.barbolini@ausl.re.it (E. Barbolini), angela.ferrari@ausl.re.it (A. Ferrari), federica.fioroni@ausl.re.it (F. Fioroni), mauro.iori@ausl.re.it (M. Iori), stefano.luminari@unimore.it (S. Luminari), massimo.menga@asuits.sanita.fvg.it (M. Menga), francesco.merli@ausl.re.it (F. Merli), valeria.trojani@studio.unibo.it (V. Trojani), annibale.versari@ausl.re.it (A. Versari), magda.zanelli@ausl.re.it (M. Zanelli), marco.bertolini@ausl.re.it (M. Bertolini).

https://doi.org/10.1016/j.cmpb.2019.105153
0169-2607/

1. Introduction

In the last decade, machine learning and artificial intelligence have produced stunning results in many domains [1]. Health-care systems have also been strongly affected by this process, as clinical data are now produced and stored at an unprecedented rate: this has enabled the rapid development of a novel research field named radiomics [2], where data analytics is applied to medical data, and in particular to imaging data.

In this paper, we exploit this kind of approach in the diagnostic phase of malignant lymphomas (ML), heterogenous cancers originating from the immune system. ML are classified into several subtypes based on their pathologic and immunologic features. Heterogeneity of ML is not only seen between ML subtypes but also within each subtype [3]. This is the case, for example, of grading and transformed areas in follicular lymphomas (FL) and other indolent lymphomas, cell of origin for diffuse large B cell lymphomas (DLBCL), and blastoid features in mantle cell lymphomas (MCL). Of note lymphoma subtype and intrapatient heterogeneity are major drivers of patients' outcome [3]. ML diagnosis and subtype definition are usually based on the sampling (biopsy) of a single tumor site, typically the easiest to biopsy lymph node, that however does not necessarily provide a full characterization of the ML features. Conversely, total body examinations such as computer tomography (CT) and fluorodeoxyglucose positron emission tomography (FDG-PET) scans, though not diagnostic, provide a comprehensive picture of the patient, characterizing multiple sites with a single exam.

So far, however, no study has been conducted to understand how imaging features may support histologic diagnosis, and better report on the heterogeneity of ML in a single patient. This paper aims to employ texture analysis techniques to extract relevant features from the volumes of interest (VOIs) contained in diagnostic PET-scans, so that machine learning algorithms can be subsequently used to identify ML subtype. In this framework, machine learning approaches are capable of automatically inferring which are the most significant data samples and features for the categories to be discriminated. In addition, from the point of view of machine learning, the problem is particularly challenging, as it can be naturally framed into the so-called *multiple-instance learning* framework, where each entity to be classified (the patient) typically consists of a collection of instances (the VOIs) that concur to the determination of the category of the main entity. In this paper, we exploit two different instantiations of multiple-instance learning: (i) a first one where predictions are first made at the level of VOIs, and further aggregated into an outcome at the level of patients, and (ii) a second one where classification is performed directly on patients.

We will present an experimental evaluation conducted on two datasets collected from the Arcispedale Santa Maria Nuova in Reggio Emilia. A first dataset contained examples regarding four different ML subtypes, while the second dataset contained Hodgkin's lymphoma (HL) patients only. Our results will show that HL is indeed the category that is best recognized by the proposed approach, achieving over 90% of precision (or positive predicted value) and recall (or sensitivity). We believe the implementation of this approach to be a first step towards the creation of a diagnosis support system, that, in the future, could avoid to perform biopsy in several cases. All the datasets and the source code needed to reproduce our results have been made publicly available.

The main contributions of the paper can be summarized as follows: (1) we present the first study that exploits machine learning and texture analysis to classify ML subtype; (2) we propose a natural formalization of the problem as a multiple-instance learning task; (3) we conduct a thorough experimental evaluation of the approach on two datasets; (4) we illustrate how interpretable models can be used to assess which are the most relevant texture features.

The paper is structured as follows. Section 2 discusses related works, highlighting the novelty of our approach. Section 3 describes our methodology, introducing the problem of multiple-instance learning in more detail, and illustrating the radiomics pipeline exploited in our approach. Then, in Section 4 we present the datasets used in our evaluation process, whereas in Section 5 we describe experimental results across different settings. Finally, Section 6 concludes the paper by presenting future research directions.

2. Related works

The research field of radiomics attempts to combine techniques for texture feature extraction from medical images with machine learning approaches, for the construction of systems capable to support diagnosis, prognosis, and response to treatment.

Building a diagnosis support system for the classification of ML subtype is a highly challenging task, due to the inherent heterogeneicity of the disease across different patients, as well as within a single patient. Availability of total body digitalized images assessing morphology and metabolism of the disease provide unique opportunity to dissect complexity of ML (and other cancers). Most of the existing approaches rely on the manual segmentation of VOIs, and on the extraction of texture-based features, that have been widely studied in the literature. This research field has recently received a growing attention, but only a few studies have investigated the potential of expoiting machine learning algorithms in combination with texture analysis. Moreover, none of these approaches have addressed the problem as a multiple instance classification task.

As for the categorization of ML subtypes, promising results have been obtained for the problem of differentiating DLBCL and FL in magnetic resonance images [4] with a study conducted on 41 patients, exploiting statistical analysis to measure correlations between texture features and ML category. The study reports both specificity and sensitivity around 76%.

Another problem that has received considerable attention is the task of FL grading. In [5], texture analysis and Bayesian classifiers are used to differentiate across three different levels of aggressivity, whereas Otzan et al. [6] use machine learning classifiers such as support vector machines and k-nearest neighbors in combination with multi-scale feature analysis. In both study, an accuracy of around 80% is reported.

Recently, convolutional neural networks have been employed in [7] to classify hematoxylin and eosin stained histopathology slides belonging to three different ML subtypes (FL, MCL, chronic lymphocytic leukemia). A study on the characterization of stages of malignant lymphomas from whole-body diffusion-weighted MRI was proposed in [8], exploiting statistical analysis over texture features. Texture analysis conducted on a set of 41 patients affected by ML has also been employed to provide prognostic information, showing how computer tomography can complement FDG-PET [9].

Looking at slightly different tasks, in [10] discriminant analysis is used to discriminate centroblast from non-centroblast cells in FL images. Support vector machines and texture analysis were exploited in [11] for the task of differentiating primary central nervous system lymphoma and enhancing glioma.

With respect to the aforementioned research works, our approach is, to the best of our knowledge, the first to exploit machine learning algorithms, and in particular a multiple-instance learning framework, to discriminate across four different ML subtypes, using texture features extracted from FDG-PET images.

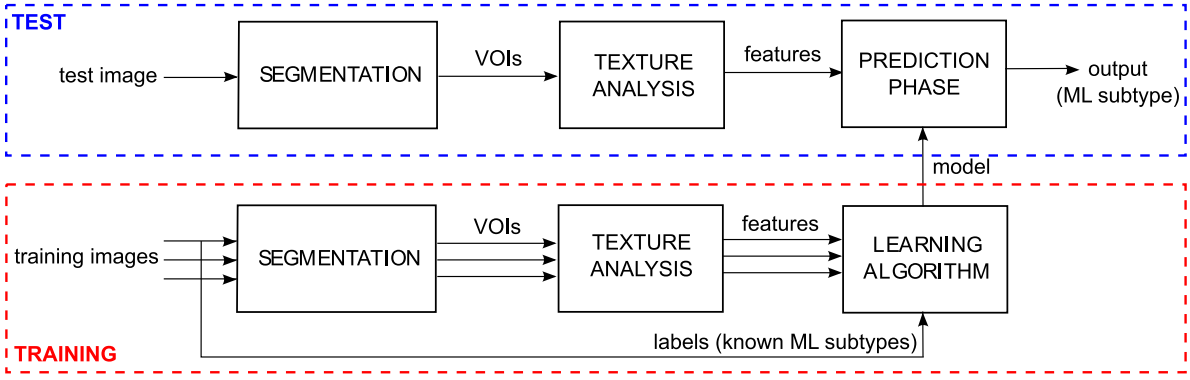

Fig. 1. Pipeline stages in our system, highlighting training (bottom) and test (top) phases.

3. Methods

3.1. Exploiting the radiomics pipeline

The system we implemented for the categorization of the ML subtype exploits a pipeline of stages that is typical of many applications in radiomics. As depicted in Fig. 1, the pipeline starts from raw images and the first stage consists in performing a segmentation of the VOIs, which in our case has been carried out manually by a nuclear medicine physician (more details in Section 4). Subsequently, texture analysis is performed on the extracted VOIs, so that features characterizing the tumors can be collected. Finally, a machine learning classifier is trained to learn a function that is capable to predict a desired outcome (in our case, the ML subtype) from the input features.

3.2. Texture analysis for ML feature extraction

Texture analysis has the goal to extract relevant characteristics from digital images, or from specific regions or volumes of interest within such images. The features that are extracted from medical images can be defined as shape-based, first-order, second-order, or higher-order [12]. Examples of shape-based features are volume and surface area. First-order features are typically obtained from the histogram of grey-level values obtained from the considered VOIs: these can be descriptive statistics such as mean or median value, minimum and maximum, skewness, kurtosis, etc. Second-order features are those that are usually referred to as *texture features*, since they take into account the spatial relationship between neighboring VOIs in an image, and thus they are capable to capture details regarding the heterogeneity of the lesions. These descriptors are typically computed through *parent matrices* such as the Gray Level Co-occurrence Matrix (GLCM) or the Gray Level Neighborhood Intensity-Difference Matrix (GLNIDM) [13]. An additional group of features that are specific of medical images is computed from the Standardized Uptake Value (SUV), that is a measure for the accumultion of radiopharmaceutical in the tissue. Examples of these features are its mean or maximum value within the considered VOI, or its peak within a region containing the maximum. In this work, we will use the texture features extracted with the CGITA software v1.4 [13], that has already been successfully used in other radiomics applications [14].

3.3. Multiple-instance learning (MIL)

From the point of view of machine learning, the classification of the ML subtype can be formulated as a multiple-instance learning (MIL) problem, which is a generalization of the supervised learning setting [15,16]. In such a framework, the examples to be classified consist of a collection (bag) of instances, and the label is typically attached to the bag rather than to each single instance. In our case, bags correspond to patients and single instances to VOIs.

More precisely, in a supervised MIL problem we are given a supervised dataset of n examples $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{n}$, where each example $X_i \in \mathcal{X}$ is a bag of k_i instances: $X_i = \{x_1^i, \ldots, x_{k_i}^i\}$. Although there are no restrictions on the nature of x_j^i instances, to simplify the notation we hereby treat them as p-dimensional vectors, thus $x_j^i \in \mathbb{R}^p$. The goal is to learn a classification function to predict the target y_i given the bag X_i. The classification can be produced either as the aggregation of the categorizations of single instances (instance-space, or IS), directly at the level of bags by embedding the set of instances into a single vector (embedded-space, or ES), or finally by exploiting a distance between bags (bag-space MIL). Both IS and ES approaches will be used in our experiments, thus we will describe them in more detail in the following subsections.

It is worth remarking a peculiarity of the problem of the diagnosis of malignant lymphomas: from the medical point of view, it is very often the case that all the instances in a single bag share the same lymphoma subtype. It is also possible – although very rare – that two different lymphoma subtypes co-exist in the same patient [17]. More generally, we also remark that the choice of the MIL paradigm is also supported by the large heterogeneity that is observed even within the same lymphoma subtype.

3.3.1. Instance-space MIL

In the instance-space paradigm, a classification function $f : \mathbb{R}^p \to \mathcal{Y}$ is learned at the level of instances. In this case, the underlying assumption is that the class of the bag is transferred to each instance within that bag, even though this fine-grained labeling could be potentially noisy. Given the classification of all the instances $\{x_1^i, \ldots, x_{k_i}^i\}$ in a bag X_i, an aggregation function is used in order to assign a label to the bag. The discrimination function F for a bag is thus computed as:

$$F(X_i) = \frac{f(x_1^i) \circ \ldots \circ f(x_{k_i}^i)}{Z} \tag{1}$$

where $\circ$ is the aggregation function and Z some (optional) normalization function. Typical choices assume that a bag is assigned to class C if the number of instances in the bag assigned to C is greater than a pre-determined threshold τ. The threshold can be absolute (a given number of instances) or relative (a given percentage of instances). According to the domain, different choices need just one positive instance to assign the positive label to the bag, or the majority of the instances. In general, several different solutions exist, and we refer the reader to the existing surveys on the topic for more details [15,16].

3.3.2. Embedded-space MIL

In the embedded-space MIL, a classification function $f : \hat{\mathcal{X}} \to \mathcal{Y}$ is learned from an embedded space $\hat{\mathcal{X}}$ onto which the original bags X_i are projected. This setting is more suitable in those cases where global information about the whole bag is useful in order to perform the classification, and local classifiers are not enough accurate. In general, the embedded space $\hat{\mathcal{X}}$ can be any space onto which a discriminant classifier can be applied. A typical choice is that of aggregating into such embedded space all the statistics of the single instances, such as the mean, minimum, maximum of each feature [15,16].

3.4. Support vector machines

In the MIL setting, any machine learning classifier can be used to learn the classification function. In our approach, we use linear support vector machines (SVMs), one of the mostly used machine learning approaches for its simplicity and efficiency. An SVM learns a function $f : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{X}$ is the input space, such as a vectorial space where each dimension represents a feature, and $\mathcal{Y}$ is the output space, that is the set of classes, or outcomes. In the context of SVMs, such function f is learnt by minimizing a loss function over a set of N given observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$.

When dealing with a binary classification task, i.e., when there are just two possible outcomes, a positive class and a negative class, in the *linear* formulation function f is computed as:

$$f(x) = \sum_{i=1}^{N} \alpha_i \langle x_i, x \rangle + b \tag{2}$$

where N is the number of training examples, α_i are the parameters to be learned, and $\langle \cdot, \cdot \rangle$ is the dot product between the input vectors, and it thus can be seen as a similarity measure between examples. Therefore, the resulting decision function is a linear hyperplane in the input space. Those examples for which the α_i coefficients are not equal to zero are called *support vectors*, since the discriminant function f only depends on them.

3.5. Random forests

As a further element of our experimental evaluation, we will employ also another machine learning classifier, named random forests (RFs) [18], that can be exploited to assess the importance of the features used in the classification process. RFs are an ensemble classifier, that is a collection of individual classifiers that are combined to obtain a global prediction.

In particular, an RF consists in multiple decision trees (DTs) [19], that are trees where a path from the root to the leaf is a specific classification rule, which can be seen as a conjunction of conditions over sets of features. For example, a path in the tree could specify that, if feature $f_7 > 0.7$ and feature $f_{12} < 2.3$ then the predicted class is positive. DTs are thus highly interpretable.

In a RF, a total of m different DTs are built, and grown to the largest extent possible. For the construction of each DT, a sample of n examples is selected at random, with replacement, where n is the size of the training set. When selecting the attribute to be inserted at a certain node in the tree, only a subset of all the features is tested. Given the result of the classification of each DT, a ranking is created, based on the number of votes obtained by each class, and the category that obtains the most votes is selected.

While DTs are highly sensitive to small changes in the training set, RFs are much more robust, as they leverage the contribution of many trees. Yet, differently from individual DTs, RFs do not produce interpretable classification rules. However, RFs allow to compute what is called feature *importance*, which is a score that takes into account the occurrence of each feature within the ensemble classifier. Importance is usually computed as the average reduction in weighted impurity of a feature across the collection of trees [18].

4. Data collection

In this section we describe the two datasets used in our experimental study, conducted at the Arcispedale Santa Maria Nuova, in Reggio Emilia. For all the patients, the histological diagnosis has been confirmed by an expert pathologist.

All the PET/CT scans collected in this study were performed on the same dedicated whole-body PET/CT scanner (Discovery STE16, GE Medical System) in three-dimensional mode (3D VUE Point HD algorithm with two iterations, 28 subsets, post-filter 5.5mm) corrected for attenuation. All patients fasted for at least 6 h before injection of the 18F-FDG tracers. The serum glucose level measured at the time of the injection was below 160 mg/dl in all patients. The examination was performed 60 min after intravenous administration of 3.7 MBq/kg of 18F-FDG using a standardized protocol. The image voxel size was 2.73 × 2.73 × 3.27 mm with a slice thickness of 3.27 mm without gap between slices. Matrix size was 256 × 256. In the assessment of PET-CT we used the Deauville five-point scale [20] that was defined for each case by one blinded nuclear medicine physician.

Only lymph nodes lesions (VOIs) were considered in this analysis. The VOIs were extracted by an experienced (5 years) nuclear medicine physician using a 40%-threshold of SUV_{max} (maximum SUV in the lesion) within a manually drawn volume.[1] The VOIs were independently checked by another nuclear medicine physician (10 years of experience). The texture features were extracted using Matlab CGITA software v. 1.4 [13]. SUV values were resampled in 64 discrete values using an absolute method (SUV range: 0–25) in order to reduce the impact of noise and size of matrices. The stability of features was studied in a previous work [21].

The 108 features computed by CGITA have then been reduced to 98, after removing nine features presenting a Kendall correlation coefficient larger than 0.999 with some other feature, and another feature[2] whose value was equal to zero in over 75% of the cases. All the datasets, the complete list of features, and the source code of our system are available in our repository at the following url: https://github.com/marcolippi83/MIL-lymphomas.

4.1. Dataset A: multiple lymphoma subtypes

In a first dataset, 36 patients were retrospectively included: 9 patients for each type of considered lymphomas (DLBCL, FL, HL and MCL). The number of VOIs per each patient varied from 1 to 37, being dependent on the lymphoma type. In the whole dataset, 349 VOIs were studied: 66 for DLBCL patients, 86 for FL patients, 53 for HL and 144 for MCL. The distribution of the number of VOIs across the four lymphoma subtypes is represented in Fig. 2 (left). As well known, the MCL subtype typically exhibits many lesions, whereas the HL subtype shows on average the minimum number of VOIs per patient. Fig. 2 (right) instead shows how VOI regions are distributed across the different ML subtypes: even in this case, we can notice how regions in HL are much more homogeneous, mostly appearing in the mediastinum, latero-cortical region, collarbone and collarbone pit. Although clearly the limited size of the dataset could lead to overfitting, and the information about regions and number of VOIs per patient is thus not discriminant per se, nevertheless it can be an important additional feature for the classification of the ML subtype.

[1] We used IntelliSpace Portal, Philips, Eindhoven, the Netherlands.

[2] Texture Feature Coding (TFC) homogeneity.

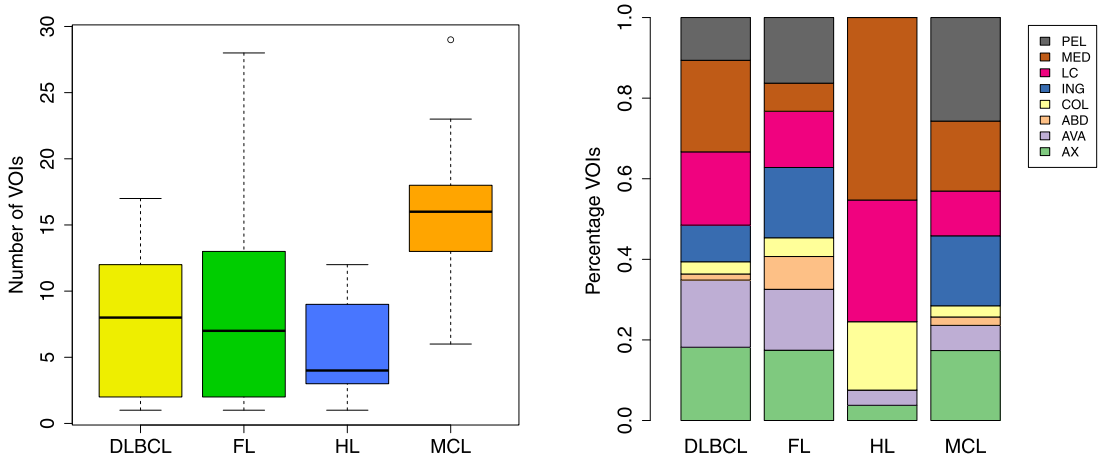

Fig. 2. Boxplot representing the distribution of the number of VOIs per patient (left) and barplot depicting the percentage of VOIs per region (right) across the four lymphoma subtypes in Dataset A. In the barplot on the right, the region abbreviations are: axillary (AX), abdominal vascular axis (AVA), abdominal (ABD), collarbone and collarbone pit (COL), inguinal (ING), latero-cortical (LC), mediastinum (MED), pelvic (PEL).

4.2. Dataset B: Hodgkin's lymphoma

In a second dataset, 24 patients affected by HL were retrospectively included. The number of VOIs per each patient ranged from 1 to 6 for a total of 78 VOIs. This second population of patients was chosen as an internal validation set for our model.

5. Results

5.1. Experimental setup

We now describe the experimental results conducted on the two datasets described in Section 4. In all our experiments, we used an SVM with linear kernel as the machine learning classifier, both in the instance-based and in the embedding-based setting. In a final, additional experiment, we also used RFs in order to assess the relevance of the texture features. To evaluate our approach, we employed a standard leave-one-out (LOO) procedure, where each patient was used, in turn, as the test set, and all the other patients constituted the training set. Clearly, in the instance-based setting, all the instances of a patient were assigned either to the training or to the test set. In order to perform model selection on the regularization hyper-parameter of SVM, for each fold of the LOO evaluation, an inner LOO procedure was applied on the training data only. This is a standard cross-validation procedure in machine learning, that is strongly encouraged in order to assess the robustness of the approaches in PET/CT image characterization with texture analysis [22].

In the embedded-space (ES) setting, for the embedded vector we exploited the minimum, maximum, and mean value of each feature, then the number of VOIs, and finally, where explicitly stated, also the histogram of frequencies of the VOI regions. For the instance-space (IS) setting, we simply used the texture features, and a one-hot encoding of the region, whereas the class of each VOI was inherited from the patient.

In order to measure the performance of our systems, we adopted standard classification metrics. For a given positive class (i.e., ML subtype) we define the True Positives (TP) as the number of correctly classified examples for that class, whereas the False Positives (FP) represent the number of examples predicted as positive, which are actually negative, and the False Negatives (FN) are the missed examples of positive class. Given these figures, we can define precision (or positive predictive value) $P = \frac{TP}{TP+FP}$ as the false positive ratio, the recall (or sensitivity) $R = \frac{TP}{TP+FN}$ as the false negative ratio, and the $F_1 = \frac{2PR}{P+R}$ as the harmonic mean between precision and recall. For completeness, we also report accuracy A as the total number of correctly classified examples, including negative cases. We remark that, in imbalanced datasets, it can be easy to achieve a high accuracy if only correctly detecting the most frequent class (which, in our case, would be the negative one). For this reason, we will mainly consider the other metrics in our evaluation.

5.2. Dataset A: multiple lymphoma subtypes

We first run experiments on dataset A, thus considering four lymphoma subtypes: DLBCL, FL, HL, MCL (see Section 4.1). For each subtype, we defined a binary classification task, where the goal is to discriminate that subtype (positive class) from the others (negative class). We chose to exploit four binary classification tasks instead of a single multi-class problem because these four subtypes are not the only existing lymphoma subtypes, thus a multi-class formulation would have implicitly made the (strong) assumption of knowing that the patient necessarily belongs to one of the four subtypes.

Table 1 presents the results obtained on this dataset, whereas Table 2 reports the confusion matrices for the best method for each ML subtype. We compare the results of the ES and IS settings, with or without the region information (R rows in Table 1) and, finally, we report also the performance when small VOIs are discarded[3] (ℓ rows in Table 1). First of all, the results confirm that the proposed approach is very effective in identifying the HL class, for which both precision and recall for patients are larger than 90% when region information is used, and only large VOIs are considered. For DLBCL and MCL performance are much lower, although far above a random baseline, as it can be observed from the confusion matrices shown in Table 2, achieving in both cases an F_1 score larger than 60%. The FL class is instead the most difficult to detect, although the ES approach is capable to identify few positive cases, without any false positive. As a further confirmation, by analyzing in more detail the errors of each classifier, indeed we observed that the large majority of the wrongly classified patients (over 50% of the cases) belong to the FL class. Conversely, again considering the ES case, no MCL patient is wrongly classified as affected by one

[3] The features computed for small VOIs are much more sensible to changes in the segmentation process. For this work, we consider a VOI to be small if its SUV statistics tumor volume is less than threshold value 2.6.

Table 1
We compare the performance on VOIs and patients (accuracy A, precision P, recall R, and F_1) for the embedded-space (ES) and instance-space (IS) classifiers, on each of the four binary classification problems, defined by the lymphoma subtype. Besides texture analysis rows with R also exploit information about region. Subscript ℓ indicates that large VOIs only are considered. Best results for each metric are highlighted in bold.

Subtype	Method	VOIs				Patients			
		A	P	R	F_1	A	P	R	F_1
DLBCL	ES	–	–	–	–	0.778	0.545	0.667	0.600
	ES + R	–	–	–	–	0.806	0.600	0.667	0.632
	ES_ℓ + R	–	–	–	–	0.778	0.571	0.444	0.500
	IS	0.725	0.317	0.394	0.351	0.667	0.364	0.444	0.400
	IS + R	0.765	**0.407**	**0.530**	**0.461**	0.806	0.583	**0.778**	**0.667**
	IS_ℓ + R	**0.800**	0.379	0.512	0.436	**0.833**	**0.714**	0.556	0.625
FL	ES	–	–	–	–	**0.833**	**1.000**	0.333	**0.500**
	ES + R	–	–	–	–	0.778	**1.000**	0.111	0.200
	ES_ℓ + R	–	–	–	–	0.750	0.000	0.000	0.000
	IS	0.504	0.182	0.291	0.224	0.457	0.143	0.222	0.174
	IS + R	0.553	0.292	**0.570**	0.386	0.528	0.250	0.444	0.320
	IS_ℓ + R	**0.565**	**0.316**	0.560	**0.404**	0.639	0.357	**0.556**	0.435
HL	ES	–	–	–	–	0.917	0.875	0.778	0.824
	ES + R	–	–	–	–	0.917	0.875	0.778	0.824
	ES_ℓ + R	–	–	–	–	**0.944**	**0.889**	**0.889**	**0.889**
	IS	0.728	0.294	0.566	0.387	0.722	0.462	0.667	0.545
	IS + R	0.791	0.384	**0.623**	0.475	0.861	0.750	0.667	0.706
	IS_ℓ + R	**0.818**	**0.419**	0.619	**0.500**	0.833	0.714	0.556	0.625
MCL	ES	–	–	–	–	0.556	0.360	**1.000**	0.529
	ES + R	–	–	–	–	0.556	0.360	**1.000**	0.529
	ES_ℓ + R	–	–	–	–	0.611	0.391	**1.000**	0.563
	IS	**0.662**	**0.610**	0.500	**0.550**	**0.806**	**0.600**	0.667	**0.632**
	IS + R	0.625	0.550	**0.500**	0.524	0.722	0.455	0.556	0.500
	IS_ℓ + R	0.593	0.540	0.488	0.513	0.694	0.429	0.667	0.522

Table 2
Confusion matrices on patients for each binary classification problem on the 36-patients dataset. Results are obtained with a leave-one-patient-out cross validation. For each subtype, we show the results obtained with the best configuration in terms of F_1 in Table 1.

DLBCL	0	1
0	22	5
1	2	7

FL	0	1
0	27	0
1	6	3

HL	0	1
0	26	1
1	1	8

MCL	0	1
0	23	4
1	3	6

of the other lymphoma subtypes, being always correctly detected as a negative case (when the positive class is DLBCL, FL, or HL). Another general observation is that the information about region typically improves the performance, except for the MCL category, which is in fact the one for which region distribution is the most heterogeneous, and the largest number of VOIs per patient is typically observed.

It is worth highlighting that the performance at the level of single VOIs are quite low for the IS approach, but they are substantially better when predictions are aggregated at the level of patients. This is not surprising, since predicting the class of individual instances is a much harder task than predicting the class of the patient. This is also the reason why the ES approach, which addresses the problem directly at the level of patients, typically performs better than the IS approach.

We hereby remark that the presented results are obtained on a relatively small set of patients, which makes the task very challenging but at the same time also prone to overfitting. For this reason, we avoided using information regarding patients, such as sex, age, weight, or height: the considered sample would have not been large enough to be significant for the whole population. Nevertheless, even with such a small amount of data, results are far beyond a random prediction for all the four considered subtypes. For HL, in particular, both precision and recall larger than 90% are achieved with a dataset of just 60 patients. All these figures confirm the great potential behind this research direction.

Table 3
Performance achieved on the 60-patients dataset for HL prediction task, with the embedding-space approach (ES) and the instance-space approach (IS), respectively. Results compare accuracy A, precision P, recall R and F_1. Besides texture analysis rows with R also exploit information about region. Best results for each metric are highlighted in bold. Subscript ℓ indicates that large VOIs only are considered: in this case, the $(^*)$ superscript indicates that one patient is not included (having just one small VOI).

Subtype	Method	VOIs				Patients			
		A	P	R	F_1	A	P	R	F_1
HL	ES	–	–	–	–	0.883	0.906	0.879	0.892
	ES + R	–	–	–	–	0.883	0.842	**0.970**	0.901
	$\text{ES}_\ell^{(*)}$ + R	–	–	–	–	0.881	0.857	0.938	0.896
	IS	0.799	0.645	0.763	0.699	0.850	0.875	0.848	0.862
	IS + R	0.843	0.703	0.847	0.768	**0.950**	**0.941**	**0.970**	**0.955**
	$\text{IS}_\ell^{(*)}$ + R	**0.851**	**0.704**	**0.888**	**0.785**	0.915	0.909	0.938	0.923

5.3. Dataset B: Hodgkin's lymphoma

As a second testbed for our approach, we considered Dataset B too, thus only focusing on HL. As a first experiment, we trained our model on Dataset A (36 patient) and used Dataset B as a test set only. The ES model using all the VOIs, and exploiting region information too, wrongly classified 5 patients out of 24 whereas a model trained without small VOIs – which corresponds to ES_ℓ + R row in Table 1, that is the best performing model for HL – instead correctly predicted 22 patients out of 23 as positives (one patient could not be classified, as it had only one small VOI[4]).

Furthermore, we also performed a LOO validation by merging Datasets A and B, thus obtaining a total of 60 patients. Results are reported in Table 3, showing that, by increasing the number of examples, performance greatly improves, achieving $F_1 > 0.85$ in all the settings, with a maximum of 0.955 for IS with region information. Performance on single VOIs improves as well, reaching $F_1 = 0.785$. These results confirm that the HL subtype can be identified with remarkable accuracy.

5.4. Feature importance

As a further experiment, we also tested an RF classifier, so as to measure the importance of the considered features. While on Dataset A the performance of RFs resulted to be significantly lower than that of SVMs, for the binary task of HL identification on the union of Datasets A and B, performance were satisfactory, although not as good as those achieved by SVMs. In particular, the RF achieved $F_1 = 0.845$, resulting from $P = 0.789$ and $R = 0.909$.[5]

Therefore, we could use the RF to compute the importance of features (as explained in Section 3.5). When ranking all the features by their importance score, we found the five most important features to be the entropy, number nonuniformity and small-number emphasis from the neighborhood gray-level dependence matrix, and the complexity and strength from the neighborhood intensity difference matrix [13]. When training a linear SVM on the union of Datasets A and B to detect HL with the ES setting, using only these five features, we achieved a remarkable 0.773 value for F_1, which could be improved up to 0.901 when including also information about regions. We believe this to be a very important step towards creating an interpretable system, since from a detailed analysis of the features, and from the results obtained with small feature sets, it could be possible to derive classification rules (e.g., single decision trees) that are understandable for humans.

[4] This is the reason for which we indicate the results with a $(^*)$ symbol in Table 3.
[5] We observed negligible differences across multiple runs of the RF classifier.

Table 4
Confusion matrix on single VOIs (left) and on patients (right) for the binary classification of HL, on the 60-patients dataset, using the instance-based approach, exploiting both texture features and region information.

	0	1
0	249	47
1	20	111

	0	1
0	25	2
1	1	32

We aim to address this issue in our future research, since a larger dataset would be necessary to assess the generalization capabilities of such rules, and to prevent overfitting (Table 4).

6. Discussion

In this work, we addressed the task of predicting the subtype of ML from texture features, using multiple-instance learning with support vector machines. Experimental results show the great potential of the approach, in particular for what concerns the detection of the Hodgkin's lymphoma, where precision and recall larger than 90% are achieved on a dataset of just 60 patients. An analysis of the importance of features conducted with random forests allows to identify the most relevant texture features for the considered task. To summarize, the proposed approach indicates that texture features extracted from FDG-PET, coupled with machine learning algorithms, are highly discriminative of the ML subtype. This is the first study of this kind, conducted to discriminate across four different ML subtypes, exploiting multiple-instance learning. Although no direct comparison can be made in terms of results achieved with respect to related works – as no previous method addressed the same task – the performance achieved in our experimental study are in line with those achieved in the literature for similar tasks.

The proposed system undergoes a pipeline of steps, which currently includes a manual segmentation of the volumes of interest. This is a time-consuming procedure, requiring experts to manually scan each image, and contour the relevant regions. As a future research direction, we aim to employ deep learning approaches such as convolutional neural networks, that have recently achieved significant results in many medical imaging applications, to directly extract features from the whole images, without the need to perform manual segmentation. Another interesting research direction would be that of building a machine learning system capable of differentiating healthy patients from those affected by any category of ML. Finally, further studies involving relational learning could

also be exploited to include also clinical and imaging-related data, with background knowledge given by experts.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Statement of ethical approval

The study was approved by the ethical committee of AUSL-IRCCS, Reggio Emilia (reference number 2016/0014693, June 13th, 2016).

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[2] R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: images are more than pictures, they are data, Radiology 278 (2) (2015) 563–577.
[3] S.H. Swerdlow, E. Campo, S.A. Pileri, N.L. Harris, H. Stein, R. Siebert, R. Advani, M. Ghielmini, G.A. Salles, A.D. Zelenetz, et al., The 2016 revision of the world health organization classification of lymphoid neoplasms, Blood 127 (20) (2016) 2375–2390.
[4] X. Wu, M. Sikiö, H. Pertovaara, R. Järvenpää, H. Eskola, P. Dastidar, P.-L. Kellokumpu-Lehtinen, Differentiation of diffuse large b-cell lymphoma from follicular lymphoma using texture analysis on conventional MR images at 3.0 tesla, Acad. Radiol. 23 (6) (2016) 696–703.
[5] O. Sertel, J. Kong, U.V. Catalyurek, G. Lozanski, J.H. Saltz, M.N. Gurcan, Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading, J. Signal Proces. Syst. 55 (1–3) (2009) 169.
[6] B. Oztan, H. Kong, M.N. Gürcan, B. Yener, Follicular lymphoma grading using cell-graphs and multi-scale feature analysis, in: Medical Imaging 2012: Computer-Aided Diagnosis, 8315, International Society for Optics and Photonics, 2012, p. 831516.
[7] N. Codella, M. Moradi, M. Matasar, T. Sveda-Mahmood, J.R. Smith, Lymphoma diagnosis in histopathology using a multi-stage visual learning approach, in: Medical Imaging 2016: Digital Pathology, 9791, International Society for Optics and Photonics, 2016, p. 97910H.
[8] K.N. De Paepe, F. De Keyzer, P. Wolter, O. Bechter, D. Dierickx, A. Janssens, G. Verhoef, R. Oyen, V. Vandecaveye, Improving lymph node characterization in staging malignant lymphoma using first-order ADC texture analysis from whole-body diffusion-weighted MRI, J. Magn. Reson. Imaging 48 (4) (2018) 897–906.
[9] B. Ganeshan, K. Miles, S. Babikir, R. Shortman, A. Afaq, K. Ardeshna, A. Groves, I. Kayani, Ct-based texture analysis potentially provides prognostic information complementary to interim FDG-pet for patients with Hodgkins and aggressive non-Hodgkins lymphomas, Eur. Radiol. 27 (3) (2017) 1012–1020.
[10] K. Belkacem-Boussaid, M. Pennell, G. Lozanski, A. Shanaah, M. Gurcan, Computer-aided classification of centroblast cells in follicular lymphoma, Anal. Quant. Cytol. Histol. 32 (5) (2010) 254.
[11] P. Alcaide-Leon, P. Dufort, A. Geraldo, L. Alshafai, P. Maralani, J. Spears, A. Bharatha, Differentiation of enhancing glioma and primary central nervous system lymphoma by texture-based machine learning., AJNR 38 (6) (2017) 1145.
[12] G. Castellano, L. Bonilha, L. Li, F. Cendes, Texture analysis of medical images, Clin. Radiol. 59 (12) (2004) 1061–1069.
[13] Y.-H.D. Fang, C.-Y. Lin, M.-J. Shih, H.-M. Wang, T.-Y. Ho, C.-T. Liao, T.-C. Yen, Development and evaluation of an open-source software package cgita for quantifying tumor heterogeneity with molecular images, BioMed Res. Int. 2014 (2014).
[14] G. Feliciani, F. Fioroni, E. Grassi, M. Bertolini, A. Rosca, G. Timon, M. Galaverni, C. Iotti, A. Versari, M. Iori, et al., Radiomic profiling of head and neck cancer: 18F-FDG pet texture analysis as predictor of patient survival, Contrast Media Mol. Imaging 2018 (2018).
[15] J. Foulds, E. Frank, A review of multi-instance learning assumptions, Knowl. Eng. Rev. 25 (1) (2010) 1–25.
[16] J. Amores, Multiple instance classification: review, taxonomy and comparative study, Artif. Intell. 201 (2013) 81–105.
[17] H. Kim, Composite lymphoma and related disorders., Am. J. Clin. Pathol. 99 (4) (1993) 445–451.
[18] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[19] J. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[20] M. Meignan, A. Gallamini, M. Meignan, A. Gallamini, C. Haioun, Report on the first international workshop on interim-pet scan in lymphoma, Leuk. Lymphoma 50 (8) (2009) 1257–1260.
[21] G. Feliciani, M. Bertolini, F. Fioroni, M. Iori, Texture analysis in ct and pet: a phantom study for features variability assessment, Phys. Med. 32 (2016) 77.
[22] M. Hatt, F. Tixier, L. Pierce, P.E. Kinahan, C.C. Le Rest, D. Visvikis, Characterization of pet/ct images using texture analysis: the past, the present any future? Eur. J. Nucl. Med. Mol. Imaging 44 (1) (2017) 151–165.