Journal of Theoretical and Applied Computer Science
ISSN 2299-2634 (printed), 2300-5653 (online)

Vol. 9, No. 1, 2015, pp. 41–55
http://www.jtacs.org

Using topological data analysis for diagnosis pulmonary embolism

Matteo Rucco[1], Emanuela Merelli[1], Damir Herman[2], Devi Ramanan[2], Tanya Petrossian, Lorenzo Falsetti[3], Cinzia Nitti[3], Aldo Salvi[3]

[1] University of Camerino, School of Science and Technology, Italy, [2] Ayasdi, Inc., USA, [3] Internal and Subintensive Medicine of Ospedali Riuniti - Ancona, IT

{matteo.rucco, emanuela.merelli}@gmail.com,
{damir, devi.ramanan}@ayasdi.com, Tanya.Petrossian@gmail.com,
{lorenzo.falsetti, c.nitti, a.salvi}@ospedaliriuniti.marche.it

Abstract: *Pulmonary Embolism (PE) is a common and potentially lethal condition. Most patients die within the first few hours from the event. Despite diagnostic advances, delays and underdiagnosis in PE are common. Moreover, many investigations pursued in the suspect of PE result negative and no more than 10% of the pulmonary angio-CT scan performed to confirm PE confirm the suspected diagnosis. To increase the diagnostic performance in PE, current diagnostic work-up of patients with suspected acute pulmonary embolism usually starts with the assessment of clinical pretest probability using plasma d-Dimer measurement and clinical prediction rules. One of the most validated and widely used clinical decision rules are the Wells and Geneva Revised scores. However, both indices have limitations. We aimed to develop a new clinical prediction rule (CPR) for PE based on a new approach for features selection based on topological concepts and artificial neural network. Filter or wrapper methods for features reduction cannot be applied to our dataset: the application of these algorithms can only be performed on datasets without missing data. Alternatively, eliminating rows with null values in the dataset would reduce the sample size significantly and result in a covariance matrix that is singular. Instead, we applied Topological data analysis (TDA) to overcome the hurdle of processing datasets with null values missing data. A topological network was developed using the Ayasdi-Iris software (Ayasdi, Inc., Palo Alto). The PE patient topology identified two flares in the pathological group and hence two distinct clusters of PE patient populations. Additionally, the topological network detected several sub-groups among healthy patients that likely are affected with non-PE diseases. To be diagnosed properly even though they are not affected by PE, in a next study we will introduce also the survival curves for the patients. TDA was further utilized to identify key features which are best associated as diagnostic factors for PE and used this information to define the input space for a back-propagation artificial neural network (BP-ANN). It is shown that the area under curve (AUC) of BP-ANN is greater than the AUCs of the scores (Wells and revised Geneva) used among physicians. The results demonstrate topological data analysis and the BP-ANN, when used in combination, can produce better predictive models than Wells or revised Geneva scores system for the analyzed cohort. The new CPR can help physicians to predict the probability of PE.*

Keywords: *Clinical Prediction Rule (CPR), Pulmonary Embolism, Topological Data Analysis, Artificial Neural Network (ANN), Computer Aided Diagnosis (CAD)*

1. Introduction

Several statistical and machine learning techniques have been proposed in the literature to deal with output of implicit or explicit rules and good classification performance [1, 2]. Most available techniques, such as linear discriminant approaches, multilayer perceptrons or

support vector machines, are able to achieve a good degree of provisional accuracy but these methods lack accuracy sufficient for the implementation of computer-aided diagnosis (CAD) for pulmonary embolism diagnosis. Different studies have been developed for CAD system predicting development of PE in patients. Tang et al used data from the Shangai Xin Hua Hospital, Tourassi et al and Patil S used the data collected from the collaborative study of the *Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED)* and built neural network [3, 4]. To improve the performance of a CAD we built a new system based on topological data analysis and statistical approach for features selection to define the input space for the artificial neural network. The result of the CAD must be interpreted as a new CPR that might be used to assign a probability of an occurrence of PE [5].

1.1. Pulmonary embolism

Pulmonary embolism (PE) is a relatively common and potentially lethal condition, affecting a proportion between 0.04% and 0.09% of the general population [6, 7], and ranging, in most of the cohorts of outpatients with suspected PE, between 8% and 12% [8, 9]. Among patients who die of PE, the largest part is observed in the first few hours from the acute event [10]. Despite diagnostic advances, delays in PE diagnosis are common and represent an important issue [11]. As a cause of sudden death, PE is second only to arrhythmic death. Among survivors, recurrent embolism and death can be prevented with prompt diagnosis and therapy. Current clinical guidelines suggest to perform a first-level, clinical stratification based upon patient's history, clinical findings and, in some cases, physician's judgment. For this purpose, several clinical prediction rulers (CPR) have been suggested and validated, and are currently in use in common clinical practice [8, 9, 12]. Subsequent management of the patient relies mainly on this stratification. Patients at high-risk for PE should immediately undergo a computed tomography pulmonary angiography (CTPA), while patients with intermediate or low pretest probability should be tested for high-sensitive dDimer assay [12]. Only patients with increased dDimer levels should be further investigated with second level examinations. CTPA is currently the most widely accepted imaging method recommended to confirm a suspect diagnosis of PE. However, increasing evidences and its direct and indirect costs suggest a limitation in its use: CTPA has been associated to increased risk of secondary cancer from radiation exposure [13] and contrast-induced nephropathy [14]. Another interesting point is that current use of 64-slices detectors has increased the frequency of subsegmental PE diagnoses: this finding, in the absence of a documented deep vein thrombosis (DVT), may cause clinical uncertainty, and lead to unnecessary therapies [15]. Echocardiography is currently recommended for shocked patients with high suspect of PE. For haemodinamically stable patients it is currently recommended only to better stratify the prognosis by detecting right ventricle dysfunction (RVD) [12]. Biomarkers such as BNP and troponin are used to assess disease severity and associated to right ventricle overload or damage, but are currently not included in diagnostic algorithms.

1.1.1. Clinical prediction rules

The extreme aspecificity and the variety of the clinical presentation and symptoms and the slenderness of the clinical signs of PE are cause on a side of the proven underdiagnosis on the other of an excess of negative examinations performed. The positive CTPA rate is very low - less than 15% of the overall number of exams performed for suspected acute PE [16]. Recent studies have demonstrated the safety of rejecting the diagnosis of PE by the combination of

a low clinical probability, assessed by a CPR, and a normal quantitative d-Dimer test result, thereby decreasing the need for further diagnostic radiological imaging in up to 30% of patients [17]. The most widely used CPRs are the Wells rule (table 2) and the Geneva Revised score [10] The Wells score has three different risk categories, and more recently it has been simplified assuming two risk categories with a cutoff set to 4. Perrier et al., has been successful in generating a model based solely on objective parameters using the Geneva Revised score [12]. This CPR is easily standardizable and has been validated internally and externally, although less extensively than the Wells rule. Both scores appeared to have a comparable predictive value for PE. Regardless of the rule being performed, the proportion of patients with PE is around 10% for low probability, 30% for moderate probability and 65% for high clinical probability category. It is important to note that these scores have severe limitations. The Wells rule includes the physician's judgment of whether an alternative diagnosis is more likely than PE [11, 12] This criterion, which carries a major weight in the score, is subjective and cannot be standardized. Moreover, it has been suggested that the predictive value of the Wells rule is derived primarily from its subjective component [18]. The Geneve Revised score is based on 13 entirely objective variables, requires a blood gas analysis while breathing room air. Interestingly, these parameters have only been evaluated for patients in ED with a clinical performance result not superior to the Wells score.

Table 1. Revised Geneva Score

Variable	Points
Predisposing factors	
Age $>$ 65 years	+1
Previous Deep Vein Thrombosis (DVT) or Pulmonary Embolism (PE)	+3
Surgery or fracture within 1 month	+2
Active malignacy	+2
Symptons	
Unilateral lower limb pain	+3
Haemoptysis	+2
Clinical signs	
Heart rate	
75-94 beats/min	+3
$\geq$ 95 beats/min	+5
Pain on lower limb deep vein at palpation an unilateral oedema	+4
Clinical probability	**Total**
Low	0-3
Intermediate	4-10
High	$\geq$11

1.1.2. Topological data analysis

Topological Data Analysis (TDA) is a method to analyze multidimensional, complex data primarily driven by geometry. TDA is a result of over a decade of research in applications of pure mathematics to practical problems. The main idea of this approach is that the shape of the data in an abstract multidimensional space drives the analysis by exploring the parallelism of a large number of machine learning algorithms. The three fundamental concepts of TDA are independence of coordinate systems; insensitivity to deformation; and compressed representation [19]. A typical example of insensitivity to deformation would be writing in a different

Table 2. Wells Score

Variable	Points
Predisposing factors	
Previous Deep Vein Thrombosis (DVT) or Pulmonary Embolism (PE)	+1.5
Recent surgery or immobilization	+1.5
Cancer	+1
Symptons	
Haemoptysis	+1
Clinical Signs	
Heart rate $>$ 100 beats/min	+1.5
Clinical signs of DVT	+3
Clinical judgement	
Alternative diagnosis less likely than PE	+3
Clinical probability (3 levels)	**Total**
Low	0 - 1
Intermediate	2 - 6
High	≥ 7
Clinical probability (2 levels)	**Total**
PE unlikely	0 - 4
PE likely	$>$4

font as long as the underlying meaning is preserved, while compressed representation refers to approximating a complex shape such that of a circle with a hexagon. Using a mathematical concept of lenses [20], data can be projected onto a subspace suitable for visualization. The topological features of the subspace are then inspected with traditional statistical approaches such as Kolmogorov-Smirnov or t-test analysis. We applied TDA to the clinical data of patients for identifying different subgroups of patients. The analysis of these subgroups was used for extracting the features statistically relevant. This information was used for defining the input space of an artificial neural network.

1.1.3. Artificial neural network

Generally speaking, a network is formed by a set of nodes and edges between nodes. The expressiveness of a network can be increased, e.g. equipping each node with computational features and using the edges for representing the communication channels among the computational units.The interactions of nodes though the connections lead to a global behavior of the network, which cannot be observed in the isolated elements of the network. This global behavior is said to be emergent, meaning that the abilities of the network supersede the ones of its elements. Networks can be used as very powerful tool because many systems can be assembled into a network-based space, applications including proteins, computers, and communities [21]. Artificial neural network (ANN) uses nodes as artificial neurons, and an artificial neuron is a computational model inspired by the natural neurons. Neurons receive signals through synapses located on the dendrites or membrane of the neuron. When the signals received are strong enough (greater than a certain threshold), the neuron is activated and emits a signal though the axon. This can activate a cascade process. The complexity of real neurons is highly abstracted when modeling artificial neurons. These basically consist of inputs, which are multiplied by weights, and then computed by a mathematical function that determines the activation of the neuron. Another function computes the output of the artificial

neuron (sometimes in dependence of a certain threshold). Weights can also be negative, so we can say that the signal is inhibited by the negative weight. Depending on the weights, the computation of the neuron will be different. By adjusting the weights of an artificial neuron we can obtain the output we want for specific inputs. If we scale an ANN to hundreds or thousands of neurons, it is both complicated and labor intensive to manually discover all the necessary weights. However, by identifying algorithms to adjust the weights of the ANN in order to obtain the desired output from the network. This process of adjusting weights is called learning or training [22, 23].The number of ANN types and uses is very high; hence, there are hundreds of different models considered as ANNs. The differences among them can be related to functions, the accepted values, the topology of network, and/or the learning algorithms, etc. We applied backpropagation algorithm to perform a layered feed-forward ANNs. This allows us to organize the artificial neurons in layers to send their signals forward, and propagate errors backwards. The network receives inputs by neurons in the input layer, and the output of the network is given by the neurons on an output layer. There may be one or more intermediate hidden layers. The backpropagation algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. The benefit of using backpropagation algorithm is that this error is reduced until the ANN learns the training data. The training begins with random weights, and the goal is to adjust them so that the error will be minimal. Further mathematical explanations are detail in [24, 25].

1.2. Statistical concepts

1.2.1. Kolmogorov-Smirnov test

In statistics, the Kolmogorov-Smirnov test (K-S test) is a nonparametric test for the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K-S test), or to compare two samples (two-sample K-S test). The Kolmogorov-Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution (in the two-sample case) or that the sample is drawn from the reference distribution (in the one-sample case). In each case, the distributions considered under the null hypothesis are continuous distributions but are otherwise unrestricted. The two-sample KS test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples [26].

1.2.2. Receiver operating characteristic

Both Artificial Neural Networks and the two scores (Wells and Revised Geneva) were evaluated by using receiver-operating characteristic (ROC) analysis. Medical tests play a vital role in modern medicine not only for confirming the presence of a disease but also to rule out the disease in individual patient. A test with two outcome categories such as *test* + and *test* - is known as dichotomous, whereas more than two categories such as positive, indeterminate and negative called polytomous test. The validity of a dichotomous test compared with the gold

standard is determined by *sensitivity* and *specificity* [27]. The *Receiver Operating Characteristic (ROC)* curve is the plot that displays the full picture of trade-off between the sensitivity (true positive rate) and (1-specificity) (false positive rate) across a series of inherent validity of a diagnostic test. This curve is useful in:
— evaluating the discriminatory ability of a test to correctly pick up diseased and non-diseased subjects;
— finding optimal cut-off point to least misclassify diseased and non-diseased subjects;
— comparing efficiency of two or more medical tests for assessing the same disease;
— comparing two or more observers measuring the same test
Total area under ROC curve is a single index for measuring the performance a test. The larger AUC – Area Under Curve – the better is overall performance of the medical test to correctly identify diseased and non-diseased subjects. Equal AUCs of two test represent similar overall performance of tests but this does not necessariy mean that both the curves are identical [28].

1.2.3. Jaccard similarity coefficient

The Jaccard similarity coefficient is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between sample sets, and it is defined as [29]:

$$J = \frac{|MD \cap AD|}{|MD \cup AD|} \tag{1}$$

where:
— *Medical Doctor diagnosis* – MD.
— *ANN diagnosis* – AD.

2. Materials and methods

2.1. Patients and patients' dataset preparation

From a cohort of 1500 patients accepted in the department of Intensive and Subintesive medicine from 2009 to 2012 at *Ospedali Riuniti di Ancona*, a total of 987 patients of average age 69 years were included (see Figure 1). The inclusion criteria for this study were that the patients at least were recorded with Wells score, Revised Geneva score, d-Dimer and blood gas PO_2, the cutoff of d-Dimer was 230ng/ml. For each patients 26 variables were collected (see table 3). In these patients, the diagnosis of PE and the exclusion of PE were made by angio-CT. Among these, 793 had PE and 147 did not have PE. Characteristics of the history, objective data from the physical examination and the outcome of d-Dimer analysis were tabulated for each patient. To improve the interpretability of our data the logarithm transformation has been applied to d-Dimer and WBC (white blood count). In the present work we analyzed only outpatients. The patients' dataset was processed using topological data analysis with Ayasdi-Iris specifications. Ayasdi-Iris ready data files are composed of rows and columns. The calculations are performed on a row-by-row basis and the results elucidate the columns (features) that best define and explain particular groups of rows. The Ayasdi-Iris data format is a matrix with unique column headers and unique row identifiers. Column headers must uniquely identify the columns. In contrast, the column that uniquely identifies the rows may be in any column position, it does not have to be the first or the last column.

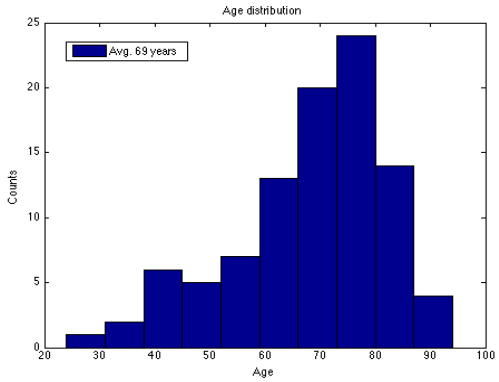

Figure 1. Age distribution: average 69 years

Table 3. Dataset

	Features	Clinical Significance
1	ID	Patient's identifier
2	Age	With the increase of the age, increase the incidence
3	Number Predictive Factors	Absolute number of predictive factors
4	Number Risk Factors	Absolute number of risk factors
5	Previous DVT	A previous DVT / PE is a risk factor repeated infringement DVT / PE
6	Palpitations	Aspecific symptom. If it implies a tachycardia could be associated with DVT / PE
7	Coughs	Symptom very nonspecific but frequently present in patients with DVT / PE
8	d-Dimer	A value of d-Dimer $< 230ng/ml$ is associated with a low risk / absent of DVT / PE. A very high value is associated with a high risk of DVT / PE
9	Systolic Pressure (PAS)	A low PAS is present in patients with DVT / PE and hemodynamic shock
10	Diastolic Pressure (PAD)	In cardiogenic shock with DVT / PE is low, sometimes undetectable. By itself has no value despite of the PAS
11	Heart Rate (FC)	In the patient with TVP / EP tachycardia is often found
12	Mean Pulmonary Artery Pressure (PAPS)	It is one of the criteria of right ventricular dysfunction. It can be normal in the case of EP low entity.
13	White Blood Cells Counter (WBC)	The value increases with inflammatory forms (pneumonia, etc ...) that can be confused with DVT / PE
14	Cancer at diagnosis	It is a risk factor for DVT / EP recognized
15	Troponin	It is a marker of myocardial infarction or heart failure and can be confused with DVT / PE
16	Shockindex	It is the ratio between PAS and FC, if it is greater than 1 is indicative of shock
17	Cancer	It is a risk factor for DVT / EP recognized
18	Right Ventricular Dsfunction (RVD)	Right ventricular overload in the course of DVT / PE
19	Score Wells	
20	Score Revised Geneva	
21	Score Wicki	
22	Dyspnea	Main symptom in DVT / PE
23	Chest pain	Chest pain is present in myocardial infarction, in pleural effusion, in the high DVT / PE
24	pCO2	Associated with low pO2 may be suggestive of DVT / PE
25	pO2	Associated with low pCO2 may be suggestive of DVT / PE
26	pH	In DVT / EP pH is usually normal
27	Final Diagnosis	Final physicians' diagnosis

2.2. Classical approaches for feature selection

Filter or wrapper methods for features reduction can not be applied to our dataset: the application of these algorithms can be done on a dataset without missing data. The elimination of rows with null values from our dataset would reduce extremely the number of samples and the covariance matrix is singular (ill conditioned) [30].

2.3. Topological data analysis: features selection

From the patients' dataset we selected only *d-Dimer, Revised Geneva score and Wells score* for the analysis and Variance Normalized Euclidean as metric function (equation 2). The Variance Normalized Euclidean metric is utilized when data are comprised of disparate quantities that are not directly comparable. The filters were L-Infinity centrality, which assigns to each point the distance to the point most distant from it, and *final diagnosis*. L-infinity centrality, is defined for a data point x to be the maximum distance from x to any other data point in the set. Large values of this function correspond to points that are far from the center of the data set (equation 3). It is also useful to use some filters that are not geometric, in that they do depend on a coordinate representation, such as coordinates in principal component or projection pursuit analysis. The parameters of the configuration were: resolution 60, gain 3.0x, equalized on (see Figure 2) [19].

$$D(h_1, h_2) = \sqrt{\sum_{i=1}^{d} \frac{(h_1(i) - h_2(i))^2}{\sigma_i^2}} \tag{2}$$

$$f(x) = \max_{y \in X} d(x, y) \tag{3}$$

2.4. Artificial neural network architecture

The neural network used in our study had a three-layer, feed-forward architecture and was trained by using the back-propagation algorithm with the sigmoid activation function. According to this learning scheme, the network tries to adjust its weights so that for every training input it can produce the desidered output. During the training phase, the network is presented with pairs of input-output patterns: supervised learning. It has been shown that this technique minimizes the mean squared error (MSE) between the decider and the actual network output following an iterative gradient search technique. The input of the network is a subset of the patient's dataset formed with all the patients but only with selected features by TDA (see table 4). Specifically our network had a hidden layer with 3 nodes, and an output layer with a single decision node. The network was trained to output 1 if PE was present and 0 if not. In this study, the network's output was interpreted as the probability of PE being present. In addition, input data were scaled to -1 and +1. The learning rate was selected to be 0.5 and the momentum coefficient to be 0.9. The network is stressed by following a k-fold cross validation (with k=10) between the training and the testing sets [31].

2.5. Softwares

2.5.1. Ayasdi-Iris

Topological data analysis has been performed using the Ayasdi-Iris software [32], properties of Ayasdi Inc. Ayasdi-Iris is based on the *Mapper* algorithm [33]. *Mapper* performs a

Table 4. Extracted features

	Features	Clinical Significance
1	Age	With the increase of the age, increase the incidence
2	d-Dimer	A value of d-Dimer $< 230ng/ml$ is associated with a low risk / absent of DVT / PE. A very high value is associated with a high risk of DVT / PE
3	Systolic Pressure (PAS)	A low PAS is present in patients with DVT / PE and hemodynamic shock
4	Heart Rate (FC)	In the patient with TVP / EP tachycardia is often found
5	Mean Pulmonary Artery Pressure (PAPS)	It is one of the criteria of right ventricular dysfunction. It can be normal in the case of EP low entity.
6	Shockindex	It is the ratio between PAS and FC, if it is greater than 1 is indicative of shock
7	Score Revised Geneva	
8	Score Wells	
9	pO2	Associated with low pCO2 may be suggestive of DVT / PE

topological simplification of a higher dimensional space by deriving a network representation of the input data space. Each node in the network is a cluster and is built by a metric-based clusterization approach (e.g. single linkage, etc. . .). The input data space is analyzed by computing n-dimensional mathematical functions, the so-called lenses. Two nodes are connected when corresponding clusters have common lenses values. Different lenses emphasizing different aspects of the dataset and different networks will be generated. Lens executes the division of data points in overlapping bins. Through the bins, the data are clustered. Each cluster is represented by a node, default node size is proportional to the number of data points in the cluster. The number of points in each bins and the size of the overlap between clusters can be adjusted by trigging the *resolution* and *gain* parameters.

2.5.2. RULEX 2.0

The ANN was implemented by the Rulex software suite [34]. The Rulex software, developed and commercialized by Impara srl, is an integrated suite for extracting knowledge from data through statistical and machine learning techniques. An intuitive graphical interface allows to easily apply standard and advanced algorithms for analyzing any dataset of interest, providing the solution to classification, regression and clustering problems. To build classifiers we used a number of graphical components provided by Rulex 2.0. We utilized Visualization and editing components to visualize and export the confusion matrix, the training and validation sets, the results of the classifier, to access statistical data (e.g. Covering, Error and Relevance).

2.6. MedCalc

MedCalc 9.5.0.0 - MedCalc Software, Mariakerke, Belgium - is a statistical software package designed for the biomedical sciences. It has an integrated spreadsheet for data input and can import files in several formats. The software includes all the basic parametric and non-parametric statistical procedures and graphs, survival analysis, test evaluation methods and meta-analysis and sample size calculations [35].

3. Results and discussion

3.1. Topological data analysis results

Topological data analysis was applied on the PE dataset to generate the network in figure (see 2) using Ayasdi-Iris (Ayasdi, Inc, Palo Alto). We color in red the pathological patients and in blue the healthy group. Ayasdi-Iris highlighted two flares and a longer tail in the pathological group. From the comparison of the two flares we selected a sub-set of features characterized with high Kolmogorov- Smirnov and low p-value (see table 4). The information extracted from the comparison of the two flares is that there is a sub-group of patients who is characterized by the recurrence risk to be affected by PE. Ayasdi-Iris detected some sub-groups among healthy patients, which present other diseases to be diagnosed properly even though they are not affected by PE.

3.2. Comparison of AUC: Wells, Geneva and ANN

From the patients' dataset we random selected 152 patients: 101 pathological and 51 healthy and we studied the discriminatory ability of the system. Tables 5 and 6 5 show the performance of three classifier: Wells score, revised Geneva score and the ANN model. Any of the three AUC was statistically significant different. The comparative study of three AUCs has been made with MedCalc. The AUCs of Revised Geneva and Wells score obtained from the analysed dataset are directly comparable with the results previously published in literature by other groups [36]. Figure (see 5 and 4) showed the ROC curves of three classifiers, they were evaluated by leave-one-out method [37]. The difference between the AUC of the new classifier both with revised Geneva's AUC and Wells' AUC was statistically significant at the 95% confidence level, $P - value < 0.0001$ and $P - value = 0.0025$ respectively. Instead, the difference between revised Geneva's AUC and Wells' AUC is not statistically relevant: $P - value = 0.1456$. Thus, the network's discriminant power is significantly. The Jaccard coefficient for the ANN is $J = 0.88$ as every informatics system also the ANN-classifier is affected by the round problem, i.e. the probability value of occurrence of PE in range [0.45; 0.5) is rounded to 0.5 then the patient is classified as pathological. At the state of the art we left to the physician the final judgment over this situations [38, 39, 40, 41].

Table 5. Performance of classifiers

Comparison of classifier			
	AUC	Standard Error (SE)	95% Confidence Interval
ANN	0,891	0,0383	0,838 to 0,899
Revised Geneva	0,5533	0,0485	0,420 to 0,610
Wells	0,7454	0,0473	0,538 to 0,753

4. Conclusion

The study derives a new CPR for pulmonary embolism and evaluating patients' cohort characteristic with a topological approach. The new CPR has been obtained training an artificial neural network on the input formed by a set of features selected by Ayasdi-Iris. Ayasdi-Iris extracted new knowledge from the patients' dataset by the application of an innovative approach for data analysis : *Topological Data Analysis*. Our results show that the feature selection strategy is beneficial for the performance improvement of an ANN trained

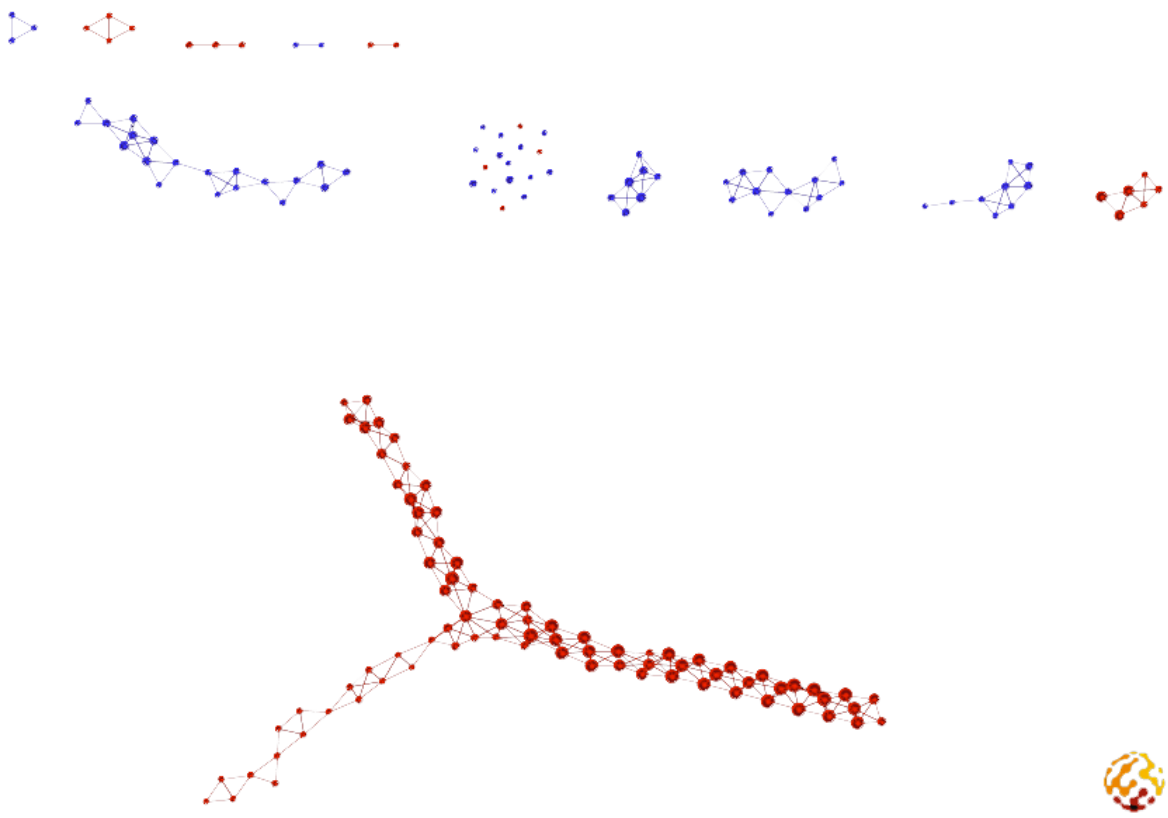

Figure 2. Ayasdi-Iris highlighted two flares in the pathological group. That means there are two cluster of patients in the PE population (red groups), also Ayasdi-Iris detected some sub-groups among healthy patients, which present other diseases to be diagnosed properly even though they are not affected by PE (blue groups).

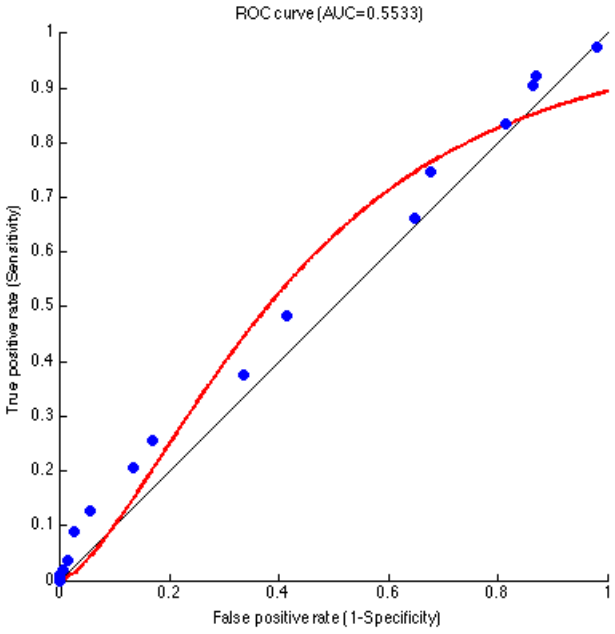

Figure 3. ROC curve of Revised Geneva score.

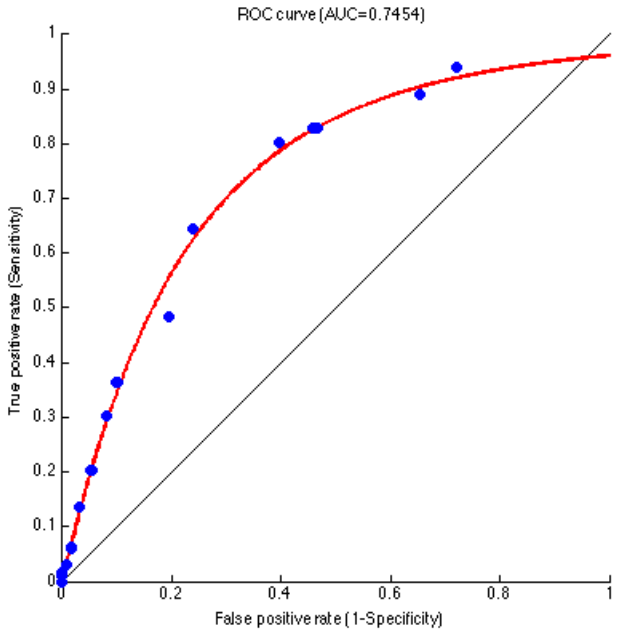

Figure 4. ROC curve of Wells score.

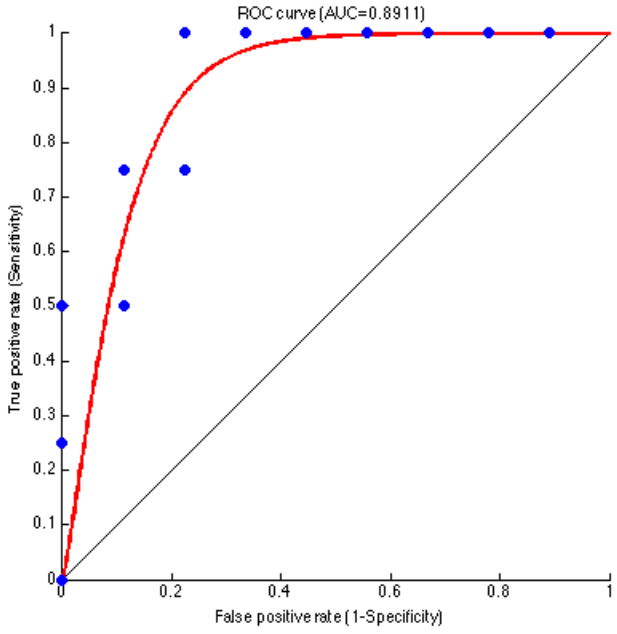

Figure 5. Comparison study of ROC curves: the AUC obtained with the new classifier is greater then the AUCs from the other two CPRs.

on the analyzed cohort. A three-layer neural network can be trained to successfully perform the diagnostic task. In conclusion a system based on Ayasdi-Iris and an ANN can form the basis of a CAD system to assist physicians with the right stratification of patients. At the moment the authors are involved in the development of a web tool for computing the CPR. In the future we will perform a validation of the system both increasing the number of patients in

the dataset and using different cohorts, we will perform a comparison study among artificial neural networks and other classification systems.

List of abbreviations: CPR: clinical predictive rule, CAD: computer aided detection, ANN: artificial neural network, TDA: topological data analysis, PE: pulmonary embolism, DVT: deep venous thrombosis, angio-CT: angiography computed tomography, K-S or KS: Kolmogorov-Smirnov test, ROC: receiver operating characteristic, FP: false positive, TN: true negative, FN: false negative, TP: true positive, DD: doctor diagnosis, AD: artificial neural network dianosis, AUC: area under curve, WBC: white blood count, MSE: mean sqared error.

Competing interests At the moment of the experiment, Damir Herman and Devi Ramanan were employees of Ayasdi, Inc. and hold stock in the company. All the authors declare that they have no competing interests.

Author's contributions Emanuela Merelli, Aldo Salvi, Lorenzo Falsetti and Cinzia Nitti and conceived the project, supervised the study. Lorenzo Falsetti wrote the clinical background. Matteo Rucco designed the experimental flowchart, performed the computer experiments and wrote the manuscript.

Acknowledgements We acknowledge the department of Internal and Subintensive medicine of Ospedali Riuniti di Ancona – Torrette for the patients' dataset, the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under the FET- Proactive grant agreement TOPDRIM (Topology-driven methods for multievel complex systems), number FP7-ICT-31812, Francesco Vaccarino and Giovanni Petri from ISI foundation for their introduction to algebraic topology and complex networks. Impara s.r.l. and Ayasdi Inc. for their software products. A special thanks is extended to Damir Herman for his support during the TDA.

References

[1] Cangelosi, D., Blengio, F., Versteeg, R., Eggert, A., Garaventa, A., Gambini, C., Conte, M., Eva, A., Muselli, M., Varesio, L.: Logic Learning Machine creates explicit and stable rules stratifying neuroblastoma patients. BMC Bioinformatics, 14(Suppl 7), p. S12, 2013.

[2] Kotsiantis, S. B., Zaharakis, I. D., Pintelas, P. E.: Machine learning: a review of classification and combining techniques. Artif. Intell. Rev., 26(3), pp. 159–190, 2006. ISSN 0269-2821.

[3] Tourassi, G., Floyd, C., Sostman, H., Coleman, R.: Acute pulmonary embolism: artificial neural network approach for diagnosis. Radiology, 189, pp. 555–558, 1993.

[4] Patil, S.: Neural network in the clinical diagnosis of acute pulmonary embolism. Chest, 1, pp. 1685–1689, 1993.

[5] Tang, L., Wang, L., Pan, S., Su, Y., Chen, Y.: A neural network to pulmonary embolism aided diagnosis with a feature selection approach. In: Biomedical Engineering and Informatics (BMEI), 2010 3rd International Conference on, volume 6, pp. 2255–2260. Oct.

[6] Tagalakis, V., Patenaude, V., Kahn, S., S, S.: Incidence of and Mortality from Venous Thromboembolism in a Real-world Population: The Q-VTE Study Cohort. Am J Med, 126(9), pp. 832.e13–21, 2013. Doi: 10.1016/j.amjmed.2013.02.024. Epub 2013 Jul 3. PubMed PMID: 23830539.

[7] Silverstein, M., Heit, J., Mohr, D., Petterson, T., O'Fallon, W., Melton, L. r.: Trends in the incidence of deep vein thrombosis and pulmonary embolism: a 25-year population-based study. Arch Intern Med., 158(6), pp. 585–93, 1998. PubMed PMID: 9521222.

[8] Wells, P., Anderson, D., Rodger, M., Stiell, I., Dreyer, J., Barnes, D., Forgie, M., Kovacs, G., Ward, J., Kovacs, M.: Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and d-dimer. Ann Intern Med., 135(2), pp. 98–107, 2001. PubMed PMID: 11453709.

[9] Wolf, S., McCubbin, T., Nordenholz, K., Naviaux, N., Haukoos, J.: Assessment of the pulmonary embolism rule-out criteria rule for evaluation of suspected pulmonary embolism in the emergency department. Am J Emerg Med, 26(2), pp. 181–5, 2008. Doi: 10.1016/j.ajem.2007.04.026. PubMed PMID: 18272098.

[10] Wood, K.: Major pulmonary embolism: review of a pathophysiologic approach to the golden hour of hemodynamically significant pulmonary embolism. Chest, 121(3), pp. 877–905, 2002.

[11] Kline, J., Hernandez-Nino, J., Jones, A., Rose, G., Norton, H., Camargo, C. J.: Prospective study of the clinical features and outcomes of emergency department patients with delayed diagnosis of pulmonary embolism. Acad Emerg Med., 14(7), pp. 592–8, 2007.

[12] Torbicki, A., Perrier, A., Konstantinides, S., Agnelli, G., Gali, N., Pruszczyk, P., Bengel, F., Brady, A., Ferreira, D., Janssens, U., Klepetko, W., Mayer, E., Remy-Jardin, M., Bassand, J.: ESC Committee for Practice Guidelines (CPG). Guidelines on the diagnosis and management of acute pulmonary embolism: the Task Force for the Diagnosis and Management of Acute Pulmonary Embolism of the European Society of Cardiology (ESC). Eur Heart J., 29(18), pp. 2276–315, 2008. Doi: 10.1093/eurheartj/ehn310. Epub 2008 Aug 30. PubMed PMID: 18757870.

[13] Einstein, A., Henzlova, M., Rajagopalan, S.: Estimating risk of cancer associated with radiation exposure from 64-slice computed tomography coronary angiography. JAMA, 298, p. 317323, 2007.

[14] Mitchell, A., Kline, J.: Contrast nephropathy following computed tomography angiography of the chest for pulmonary embolism in the emergency department. J Thromb Haemost, 5, pp. 50–54, 2007.

[15] Brunot, S., Corneloup, O., Latrabe, V., Montaudon, M., Laurent, F.: Reproducibility of multi-detector spiral computed tomography in detection of sub-segmental acute pulmonary embolism. Eur.Radiol., 15, pp. 2057–2063, 2005.

[16] Haap, M., Gatidis, S., Horger, M., Riessen, R., Lehnert, H., Haas, C.: Computed tomography angiography in patients with suspected pulmonary embolism-too often considered? Am J Emerg Med, 30(2), pp. 325–30, 2012.

[17] Beck, K., Holzschuh, J.: Schattauer GmbH - Verlag für Medizin. Naturwissenschaften - The Science of Nature, 2011.

[18] Sanchez, O., Trinquart, L., Caille, V., Couturaud, F., Pacouret, G., Meneveau, N., Verschuren, F., Roy, P.-M., Parent, F., Righini, M., et al.: Prognostic factors for pulmonary embolism: the prep study, a prospective multicenter cohort study. American journal of respiratory and critical care medicine, 181(2), pp. 168–173, 2010.

[19] Lum, P., Singh, G., Lehman, A., et al.: Extracting insights from the shape of complex data using topology. Scientific Reports - Nature, 3, pp. 561–577, 2013.

[20] Singh, G., Memoli, F., Carlsson, G.: Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In: Botsch, M., Pajarola, R., Chen, B., Zwicker, M. (eds.), Symposium on Point Based Graphics, pp. 91–100. Eurographics Association, 2007.

[21] Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. Reviews of modern physics, 74(1), p. 47, 2002.

[22] Lippmann, R.: An introduction to computing with neural nets. ASSP Magazine, IEEE, 4(2), pp. 4–22, Apr.

[23] Rosenblatt, F.: Principles of neurodynamics. Perceptrons and the theory of brain mechanisms. Spartan Books, Washington, 1962.
[24] Rojas, R.: Neural Networks: A Systematic Introduction. 1996.
[25] Kauffman, S. A.: The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, New York, 1993.
[26] Kolmogorov-Smirnov test.
[27] Hanley, J., McNeil, B.: The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology., 143(1), pp. 29–36, 1982.
[28] McClish, D. K.: Analyzing a Portion of the ROC Curve. Medical Decision Making, 9, pp. 190–195, 1989.
[29] Jaccard Coefficient.
[30] Lloyd, N. T., Bau, D.: NUMERICAL LINEAR ALGEBRA. pp. xii+361. Society for Industrial and Applied Mathematics, UK, iii edition, 1997.
[31] Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai, volume 14, pp. 1137–1145. 1995.
[32] Ayasdi-Iris.
[33] Singh, G., Mémoli, F., Carlsson, G. E.: Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In: SPBG, pp. 91–100. Citeseer, 2007.
[34] Rulex 2.0.
[35] MedCalc.
[36] Penaloza, A., Melot, C., Motte, S.: Comparison of the Wells score with the simplified revised Geneva score for assessing pretest probability of pulmonary embolism. Thromb Res, 127(2), pp. 81–4, 2011.
[37] Picard, R., Cook, D.: Cross-Validation of Regression Models. Journal of the American Statistical Association., 79(387), p. 575583, 1984.
[38] DeLong, E., DeLong, D., Clarke-Pearson, D.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 44, pp. 837–845, 1988.
[39] Griner, P., Mayewski, R., Mushlin, A., Greenland, P.: Selection and interpretation of diagnostic tests and procedures. Annals of Internal Medicine, 94, pp. 555–600, 1981.
[40] Metz, C.: Basic principles of ROC analysis. Seminars in Nuclear Medicine. Annals of Internal Medicine, 8, pp. 283–298, 1978.
[41] Zweig, M., Campbell, G.: Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry, 39, pp. 561–577, 1993.