A novel approach to learning through categorical variables applicable to the classification of solitary pulmonary nodule malignancy

Raquel Bosch-Romeu
CEU Cardinal Herrera University

Julian Librero
NavarraBiomed

Marina Senent-Valero
Hospital General Universitario de Alicante

Maria Carmen Sanfeliu-Alonso
CEU Cardinal Herrera University

Jose Maria Salinas-Serrano
Hospital Universitari Sant Joan d'Alacant

Jaume Forés-Martos
CEU Cardinal Herrera University

Beatriz Suay-Garcia
CEU Cardinal Herrera University

Joan Climent
CEU Cardinal Herrera University

Antonio Falco (✉ afalco@uchceu.es)
CEU Cardinal Herrera University

Maria Pastor-Valero
Miguel Hernandez University

Method Article

Keywords: Classification methods, Non-classical probabilities, Solitary Pulmonary Nodule

Posted Date: January 24th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2502360/v1

Additional Declarations: No competing interests reported.

METHODOLOGY

A novel approach to learning through categorical variables applicable to the classification of solitary pulmonary nodule malignancy

Raquel Bosch-Romeu[1,2], Julián Librero[3,8], Marina Senent-Valero[4,10], María Carmen Sanfeliu-Alonso[5], José María Salinas-Serrano[6], Jaume Forés-Martos[1,2], Beatriz Suay-García[1,2], Joan Climent[7], Antonio Falcó[1,2*] and María Pastor-Valero[4,9]

Abstract

Background: One of the main drawbacks in constructing a classification model is that some or all of the covariates are categorical variables. Classical methods either assign labels to each output of a categorical variable or are summarised measures (frequencies and percentages), which can be interpreted as probabilities.

Methods: We adopted a novel mathematical procedure to construct a classification model from categorical variables based on a non-classical probability approach. More specifically, we codified the variables following the categorical data representation from the Discriminant Correspondence Analysis before constructing a non-classical probability matrix system that represents an entangled system of dependent-independent variables. We then developed a disentangled procedure to obtain an empirical density function for each representative class (minimum of two classes). Finally, we constructed our classification model using the density functions.

Results: We applied the proposed procedure to build a classification model of the malignancy of Solitary Pulmonary Nodule (SPN) after five years of follow up using routine clinical data. First, with $2/3$ (270) of the sample of 404 patients with SPN, we constructed the classification model, and then validated it with the remaining $1/3(134)$ we validated it. We tested the procedure's stability by repeating the analysis randomly 1000 times. We obtained a model accuracy of 0.74, an F1 score of 0.58, a Cohen's Kappa value of 0.41 and a Matthews Correlation Coefficient of 0.45. Finally, the area under the ROC curve was 0.86.

Conclusion: The proposed procedure provides a machine learning classification model with an acceptable performance of a classification model of solitary pulmonary nodule malignancy constructed from routine clinical data and mainly composed of categorical variables. It provides an acceptable performance, which could be used by clinicians as a tool to classify SPN malignancy in routine clinical practice.

Keywords: Classification methods; Non-classical probabilities; Solitary Pulmonary Nodule

Background

The main aim of this article is to propose a procedure to be used in constructing a classification model by means of categorical variables. This model can be applied to construct health recommendations using previously anonymised routine clinical data from health centres and hospitals. Clinical data are usually collected as categorical variables, which is one of the main barriers to implementing a binary or multi-class classification system.

Chang *et al.* [1] and Krzanowski [2] were among the first researches to propose the use of continuous and dichotomous variables in this context. In their research, they implemented a classification model based on discriminant analysis involving two groups combining discrete and continuous variables previously established by Olkin and Tate [3]. Several other methodologies such as Logistic Regression and Neural

*Correspondence: afalco@uchceu.es
[1] Departamento de Matemáticas, Física y Ciencias Tecnológicas, Universidad Cardenal Herrera-CEU, CEU Universities, Alfara del Patriarca, Spain
Full list of author information is available at the end of the article

Networks use the above methodologies but in a different mathematical framework, i.e. they artificially convert categorical variables into continuous variables, when these models were only designed to be used with quantitative physical-based measures [4].

Classical methods – to work with categorical variables – either assign labels to each output of a variable or they summarise measures using frequencies and percentages, which can be interpreted as probabilities. Following this latter strategy, Discriminant Correspondence Analysis (DCA) [5] uses tables showing the frequency of each category of the variables in the different groups into which individuals can be classified. It is most commonly used to analyse data obtained through surveys. In such a context, each question corresponds to a variable and each possible answer to a category of that variable. DCA is a specific application of Correspondence Analysis (CA) and Discriminant Analysis (DA). The aim of the CA is to summarise the relationships between the variables, which are studied either in pairs or as a whole. On the other hand, the aim of DA is to categorise observations into different groups using continuous variables. DCA is primarily employed to classify observations but using categorical independent variables based on the geometry of a point cloud, which allows distances between groups and categories to be defined.

On this basis, our methodological challenge was to construct a classification model to help clinicians manage diagnoses based on a quantum probability framework.

The model will be expressed as a classification map (or oracle) denoted by ℓ. We adopted a novel strategy based on observed categorical variables to construct ℓ from a training data set. This oracle assigns an output category n output category $\mathbf{Y} = \ell(\mathbf{X})$ in a finite output space $\mathcal{Y}$ to a data vector $\mathbf{X}$, from an input space $\mathcal{X}$. In practice, In practice, clinicians will be able to query the oracle ℓ using collected input data $\mathbf{X}$ obtained from a new patient and get the answer $\ell(\mathbf{X}) \in \mathcal{Y}$.

For example, from a previous survey using hospital patients, we have a set of two categorical variables *Sex* and *Smoker* that take values in a set of modalities given by

$$\{Sex.Male, Sex.Female, Smoker.Never, Smoker.Current, Smoker.Former\}.$$

The patients are classified in two modalities, namely

$$\{Positive\ Diagnosis, Negative\ Diagnosis\},$$

which are obtained from each individual clinical history. Remember that classical approaches, like logistic regression, assume that

$$\mathcal{Y} \in \{-1\ (Negative\ Diagnose), 1\ (Positive\ Diagnose)\}.$$

Next, we can construct an input space $\mathcal{X}$, where each element is written in the form

$$\begin{aligned}\mathbf{X} &= \alpha_1 Sex.Male + \alpha_2 Sex.Female \\ &\quad + \alpha_3 Smoker.Never + \alpha_4 Smoker.Current + \alpha_5 Smoker.Former.\end{aligned}$$

Here each coefficient α_i takes the value 1 if its corresponding modality is active and 0 otherwise. Similarly, the output space $\mathcal{Y}$ is composed of elements written in the form

$$\mathbf{Y} = \beta \textit{Positive Diagnosis} + (1-\beta)\textit{Negative Diagnosis}$$

where $\beta = 0, 1$. Formally, we can represent both elements as vectors

$$\mathbf{X} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} \text{ and } \mathbf{Y} = \begin{bmatrix} \beta \\ 1-\beta \end{bmatrix},$$

composed of binary entries. Hence, the individual elements of the survey can be represented with the help of the following tensor product operation:

$$\mathbf{X} \otimes \mathbf{Y} := \mathbf{X}\,\mathbf{Y}^T = \begin{bmatrix} \beta\alpha_1 & (1-\beta)\alpha_1 \\ \beta\alpha_2 & (1-\beta)\alpha_2 \\ \beta\alpha_3 & (1-\beta)\alpha_3 \\ \beta\alpha_4 & (1-\beta)\alpha_4 \\ \beta\alpha_5 & (1-\beta)\alpha_5 \end{bmatrix}.$$

Here the superscript ^{T} denotes the transpose matrix operation. Thus, we can represent the data basis space as a tensor product space $\mathcal{X} \otimes \mathcal{Y}$, similar to a bipartite state space in the quantum mechanics framework.

Methods

This section presents a general methodology that can easily be used to reuse medical data that contains either only categorical or categorical and quantitative variables.

Categorical data representation

We assume we have a series of n observed data points, from a given a population Ω, which are obtained from a survey containing q-questions. These questions are represented by the categorical variables $X_1, \ldots, X_q$. Each X_i has an assigned set of m_i-answers denoted by $\mathcal{O}_i = \{O_1^{(i)}, \ldots, O_{m_i}^{(i)}\}$ where $m_i \geq 2$ and $1 \leq i \leq q$. Moreover, each individual ω in the population generates an output from the categorical variable X_i given by

$$X_i(\omega) = \sum_{k=1}^{m_i} \alpha_k^{(i)}(\omega) O_k^{(i)} = \begin{bmatrix} \alpha_1^{(i)}(\omega) \\ \alpha_2^{(i)}(\omega) \\ \vdots \\ \alpha_{m_i}^{(i)}(\omega) \end{bmatrix},$$

where $\alpha_k^{(i)}(\omega) = 1$ if ω has the modality $O_k^{(i)}$ and $\alpha_k^{(i)}(\omega) = 0$, otherwise $(1 \leq i \leq d)$.

Under this formalism, a Bernoulli categorical variable X_i takes values over two mutually incompatible modalities $\{O_1^{(i)}, O_2^{(i)}\}$ and hence $X_i(\omega)$ belongs to

$$\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} \subset \{0,1\}^2.$$

Thus, in general for a categorical variable over m_i mutually incompatible modalities $\{O_1^{(i)}, \dots, O_{m_i}^{(i)}\}$, each $X_i(\omega)$ will take values in the set

$$\mathcal{X}_{m_i} := \{\mathbf{e}_1, \dots, \mathbf{e}_{m_i}\} \subset \{0,1\}^{m_i},$$

for some $m_i \geq 2$, and where $\mathbf{e}_k$ is a vector containing 1 in the i-th position and 0 in the rest.

We now define a survey $\mathbf{X}$ from categorical variables $X_1, \dots, X_q$ as

$$X = X_1 + \cdots + X_q.$$

Each individual ω in Ω, has an assigned observation represented by a vector with binary entries $\mathbf{X}(\omega)$, defined as

$$\mathbf{X}(\omega) = X_1(\omega) + \cdots + X_q(\omega) = \sum_{i=1}^{d} \sum_{k=1}^{m_i} \alpha_k^{(i)}(\omega) O_k^{(i)} = \begin{bmatrix} \alpha_1^{(1)}(\omega) \\ \vdots \\ \alpha_{m_1}^{(1)}(\omega) \\ \vdots \\ \alpha_1^{(q)}(\omega) \\ \vdots \\ \alpha_{m_q}^{(q)}(\omega) \end{bmatrix},$$

In consequence, the set $\{\mathbf{X}(\omega) : \omega \in \Omega\}$ can be seen as the data set of a particular clinical survey. Since each categorical variable takes values in $\mathcal{X}_{m_i}$, the survey $\mathbf{X}$ takes values in the set

$$\mathcal{X}_d = \mathcal{X}_{m_1} \times \cdots \times \mathcal{X}_{m_q},$$

Moreover,

$$\operatorname{Card} \mathcal{X}_d = m_1 m_2 \cdots m_q \geq 2^q.$$

Finally, the output data Y is a univariate categorical variable taking values over $\mathcal{Y}_k = \{\mathbf{e}_1, \dots, \mathbf{e}_k\} \subset \{0,1\}^k$, for some $k > 1$. . Throughout this paper we will assume that d is much larger that k.

Observe that $\mathbf{X} \in \{0,1\}^d \subset \mathbb{R}^d$ and $\mathbf{e}_y \subset \{0,1\}^k \subset \mathbb{R}^k$, then we can define a tensor product operation similar to that used in quantum mechanics as

$$\mathbf{X} \otimes \mathbf{e}_y = \mathbf{X}\, \mathbf{e}_y^T,$$

which is a $d \times k$ matrix where all columns entries are zero except the y-th column that is equals to $\mathbf{X}$. Consequently, our data basis will be described with the following tensor product space

$$\mathcal{X}_d \otimes \mathcal{Y}_k = \left\{ \mathbf{X}\,\mathbf{e}_y^T : \mathbf{X} \in \mathcal{X}_d \text{ and } \mathbf{e}_y \in \mathcal{Y}_k \right\}.$$

Note that not only categorical variables fall within the above framework. Remember that indicator functions over measurable sets can approximate – under a convenient norm – any measurable function. As a result, any quantitative variable can be easily codified following the above ideas.

To describe the proposed methodology, we will consider that we are working with a generic training dataset

$$\mathcal{D} := \{\mathbf{X}_i\, Y_i^T \in \mathcal{X}_d \otimes \mathcal{Y}_k : 1 \leq i \leq n\}.$$

Now, the goal is to use $\mathcal{D}$ to give the classification model by constructing an empirical classification map, namely $\ell_{\mathcal{D}} : \mathcal{X}_d \longrightarrow \mathcal{Y}_k$, as an approximation of an "ideal" classification map ℓ.

The construction of the empirical classification map $\ell_{\mathcal{D}}$ will be given in three steps. In the first one we will use a density matrix from non-classical probability theory to construct a probability representative basis. We will then use this basis to parametrise the covariates $\mathbf{X}_i$ and construct a surrogate training dataset. Finally, we will use this surrogate training set to obtain a conditional empirical density function for each class in the independent set using a kernel method. This will be used to propose our empirical classification map.

Constructing non-classical probabilities from the training set

To construct the density matrix associated to $\mathcal{D}$ we will do the following. For each output data $\mathbf{e}_y \in \mathcal{Y}_k$, where $1 \leq y \leq k$, we compute the vector $\mathbf{f}_y$ of non-negative integers, which is defined by the sum of all elements in $\mathcal{D}$ sharing the output label $\mathbf{e}_y$, i.e.,

$$\sum_{\mathbf{X}\,\mathbf{e}_y^T \in \mathcal{D}} \mathbf{X}\mathbf{e}_y^T = \begin{pmatrix} f_{y,1} \\ f_{y,2} \\ \vdots \\ f_{y,d} \end{pmatrix} \mathbf{e}_y^T = \mathbf{f}_y\,\mathbf{e}_y^T.$$

Observe that vector $\mathbf{f}_y$ contains the frequencies for all categories related to the output label $\mathbf{e}_y$. Next, we construct a $d \times k$ matrix

$$D := [\mathbf{f}_1\,\mathbf{f}_1\,\cdots\,\mathbf{f}_k] = \begin{pmatrix} f_{1,1} & f_{2,1} & \cdots & f_{k,1} \\ f_{1,2} & f_{2,2} & \cdots & f_{k,2} \\ \vdots & \vdots & \ddots & \vdots \\ f_{1,d} & f_{2,d} & \cdots & f_{k,d} \end{pmatrix} = \sum_{y=1}^{k} \mathbf{f}_y\,\mathbf{e}_y^T.$$

The columns of matrix D are related to the values in $\mathcal{Y}_k$, whereas the rows of D are related to the values in $\mathcal{X}_d$. Consequently, matrix D contains the frequencies of $\mathcal{D} \subset \mathcal{X}_d \otimes \mathcal{Y}_k$ like a bivariate distribution. Now, we would want to extract the information about the input variable $\mathcal{X}_d$ contained in matrix D to asses a value in

$\mathcal{Y}_k$. For this, we transform D by using element-wise the function $\sqrt{\cdot}$ and obtaining a matrix

$$X = \sqrt{D} = \begin{pmatrix} \sqrt{f_{1,1}} & \sqrt{f_{2,1}} & \cdots & \sqrt{f_{k,1}} \\ \sqrt{f_{1,2}} & \sqrt{f_{2,2}} & \cdots & \sqrt{f_{k,2}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{f_{1,d}} & \sqrt{f_{2,d}} & \cdots & \sqrt{f_{k,d}} \end{pmatrix} = \sum_{y=1}^{k} \sqrt{\mathbf{f}_y}\, \mathbf{e}_y^T.$$

By using $X = \sqrt{D}$, we construct a square matrix related to the frequencies of each category of the input space $\mathcal{X}_d$ as follows. Let

$$\begin{aligned} \rho_{\mathcal{D}} &:= \frac{1}{\operatorname{tr}(XX^T)} XX^T \\ &= \frac{1}{\operatorname{tr}(XX^T)} \sum_{y=1}^{k} \sum_{y'=1}^{k} \sqrt{\mathbf{f}_y}\, \mathbf{e}_y^T\, \mathbf{e}_{y'}\, (\sqrt{\mathbf{f}_{y'}})^T \\ &= \frac{1}{\operatorname{tr}(XX^T)} \sum_{y=1}^{k} \sqrt{\mathbf{f}_y}\, (\sqrt{\mathbf{f}_{y'}})^T \text{ (remember that } \mathbf{e}_y^T\, \mathbf{e}_{y'} = \delta_{y,y'}) \\ &= \frac{1}{\operatorname{tr}(XX^T)} \begin{pmatrix} \sum_{i=1}^k f_{i,1} & \sum_{i=1}^k \sqrt{f_{i,1} f_{i,2}} & \cdots & \sum_{i=1}^k \sqrt{f_{i,1} f_{i,d}} \\ \sum_{i=1}^k \sqrt{f_{i,2} f_{i,1}} & \sum_{i=1}^k f_{i,2} & \cdots & \sum_{i=1}^k \sqrt{f_{i,2} f_{i,d}} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^k \sqrt{f_{i,d} f_{i,1}} & \sum_{i=1}^k \sqrt{f_{i,d} f_{i,2}} & \cdots & \sum_{i=1}^k f_{i,d} \end{pmatrix}, \end{aligned}$$

where tr denotes the trace function (which is defined over the set of square matrices and returns the sum of its diagonal elements). Observe that $\operatorname{tr}(XX^T) = \sum_{i=1}^k \sum_{j=1}^d f_{i,j} = N$, is the total sum of all frequencies. Moreover, the matrix $\rho_{\mathcal{D}}$ has the following properties

1 $\rho_{\mathcal{D}}$ is a symmetric matrix, i.e., $\rho_{\mathcal{D}} = \rho_{\mathcal{D}}^T$ where the superscript ^{T} denotes the transpose matrix operation.

2 $\rho_{\mathcal{D}}$ is a semi-definite positive, i.e., it can be decompose as $\rho_{\mathcal{D}} = B^T B$ where $B = (\operatorname{tr}(XX^T))^{-1/2} X^T$.

3 $\operatorname{tr} \rho_{\mathcal{D}} = 1$.

In quantum mechanics a matrix satisfying 1. 2. and 3. is called a *density matrix* and it is a measure of quantum probability (also called non-classical probability). It represents a set of *mixed states* (in our setting it is produced by the interaction of the input and output outcomes) and where a mixed state refers to any case in which we subdivide a microscopic or macroscopic system into an ensemble. These mixed states are generated by a convex combination of *pure states* or rank-one tensors, i.e., density matrices $\rho = \mathbf{U}\mathbf{U}^T$ for an unitary vector $\mathbf{U}$.

Also as a consequence of 1. and 2. we obtain non-negative singular values of matrix $\rho_{\mathcal{D}}$. Another key feature is given by the Fundamental Theorem of the ranks, which asserts

$$\operatorname{rank} XX^T = \operatorname{rank} X^T X = \operatorname{rank} X = \operatorname{rank} X^T \leq \operatorname{Card} \mathcal{Y}_k,$$

hence

$$\operatorname{rank} \rho_{\mathcal{D}} \leq k. \tag{1}$$

Thus, the rank of the density matrix $\rho_{\mathcal{D}}$ and the matrix X are bounded by the number of classes in $\mathcal{Y}_k$. Property (1) will be essential to reduce the degree of freedoms (DoF) to represent the vectors in the input space data. From now one, we will assume

$$\operatorname{rank} \rho_{\mathcal{D}} = \mathfrak{r} \leq k.$$

To write each input data by means a set of $\mathfrak{r}$-coordinates or DoF, we will use the Singular Value Decomposition (SVD) of the density matrix $\rho_{\mathcal{D}}$. The SVD of $\rho_{\mathcal{D}}$ allows the explicit construction of an orthonormal basis

$$\mathfrak{B}(\rho_{\mathcal{D}}) = \{\mathbf{U}_i \in \mathbb{R}^d : 1 \leq i \leq d\}$$

of $\mathbb{R}^d$, where $\mathbf{U}_i^T \mathbf{U}_j = \delta_{i,j}$ holds for all $1 \leq i, j \leq d$, together with a set of non-negative values

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$$

such that

$$\rho_{\mathcal{D}} = \sum_{i=1}^{d} \sigma_i \mathbf{U}_i \, \mathbf{U}_i^T.$$

We known that only some of the non-negative values $\sigma_1, \ldots, \sigma_{\mathfrak{r}}$, called the singular values, are different from zero. More precisely, it is known that $\sigma_i = 0$ if and only if $i > \operatorname{rank} \rho_{\mathcal{D}}$. Thus

$$\rho_{\mathcal{D}} = \sum_{i=1}^{\mathfrak{r}} \sigma_i \mathbf{U}_i \, \mathbf{U}_i^T, \tag{2}$$

and $\mathrm{tr}\rho_{\mathcal{D}} = \sum_{i=1}^{\mathfrak{r}} \sigma_i = 1$. The quantum mechanics interpretation of (2) is that the density matrix $\rho_{\mathcal{D}}$ is generated by a convex combination of the pure states $\{\mathbf{U}_i \, \mathbf{U}_i^T : 1 \leq i \leq \mathfrak{r}\}$.

Constructing a surrogate training set from non-classical probabilities

Any input data $\mathbf{X} \in \mathcal{X}_d$, can be normalised under the Euclidean norm as

$$\widetilde{\mathbf{X}} := \frac{\mathbf{X}}{\sqrt{\mathbf{X}^T \mathbf{X}}} = \frac{\mathbf{X}}{\sqrt{q}} \in \mathbb{R}^d,$$

and then we write $\widetilde{\mathbf{X}}$ by using the basis $\mathfrak{B}(\rho_{\mathcal{D}})$ as $\widetilde{\mathbf{X}} = \sum_{i=1}^{d} \lambda_i \mathbf{U}_i$, where $\lambda_i = \mathbf{U}_i^T \widetilde{\mathbf{X}}$ for $1 \leq i \leq d$. On the other hand, we can identify any normalized input data $\widetilde{\mathbf{X}}$ with the symmetric matrix $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ (a random variable in quantum probability), i.e.,

$$\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T = \sum_{i=1}^{d} \sum_{j=1}^{d} \lambda_i \lambda_j \mathbf{U}_i \mathbf{U}_j^T.$$

Since $\|\widetilde{\mathbf{X}}\|^2 = \widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}} = 1$ then $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ it is a rank-one projection (and a pure state) over the linear subspace generated by the vector $\mathbf{X}$.

The $\rho_{\mathcal{D}}$-expected value of the rank-one projection $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ is given by

$$\mathbb{E}_{\rho_{\mathcal{D}}}[\mathbf{X}\mathbf{X}^T] = \mathrm{tr}(\rho_{\mathcal{D}}\mathbf{X}\mathbf{X}^T).$$

Since

$$\begin{aligned}
\rho_{\mathcal{D}}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T &= \sum_{k=1}^{\mathfrak{r}}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_k\mathbf{U}_k\,\mathbf{U}_k^T\lambda_i\lambda_j\mathbf{U}_i\mathbf{U}_j^T \\
&= \sum_{k=1}^{\mathfrak{r}}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_k\lambda_i\lambda_j\delta_{k,i}\mathbf{U}_k\mathbf{U}_j^T \\
&= \sum_{k=1}^{\mathfrak{r}}\sum_{j=1}^{n}\sigma_k\lambda_k\lambda_j\mathbf{U}_k\mathbf{U}_j^T,
\end{aligned}$$

the diagonal of $\rho_{\mathcal{D}}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ is the matrix $\sum_{k=1}^{\mathfrak{r}}\sigma_k\lambda_k^2\mathbf{U}_k\mathbf{U}_k^T$. Thus, we conclude that $\mathbb{E}_{\rho_{\mathcal{D}}}[\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T] = \sum_{k=1}^{\mathfrak{r}}\sigma_k\lambda_k^2$. Moreover, we deduce

$$\rho_{\mathcal{D}}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T = \rho_{\mathcal{D}}\widetilde{\mathbf{X}}_{\mathfrak{r}}\widetilde{\mathbf{X}}_{\mathfrak{r}}^T \text{ where } \widetilde{\mathbf{X}}_{\mathfrak{r}} := \sum_{i=1}^{\mathfrak{r}}\lambda_i\mathbf{U}_i. \tag{3}$$

Thus, (3) implies that

$$\mathbb{E}_{\rho_{\mathcal{D}}}[\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T] = \mathbb{E}_{\rho_{\mathcal{D}}}[\widetilde{\mathbf{X}}_{\mathfrak{r}}\widetilde{\mathbf{X}}_{\mathfrak{r}}^T], \tag{4}$$

holds for every $\mathbf{X} \in \mathcal{X}_d$. From (4), the rank-one projection $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ and the symmetric matrix $\widetilde{\mathbf{X}}_{\mathfrak{r}}\widetilde{\mathbf{X}}_{\mathfrak{r}}^T$ have the same $\rho_{\mathcal{D}}$-expected value. This allows us to identify the elementary events representing the one dimensional linear subspaces generated by $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{X}}_{\mathfrak{r}}$, respectively. Indeed, the vector $\widetilde{\mathbf{X}}_{\mathfrak{r}}$ written in the $\mathfrak{B}(\rho_{\mathcal{D}})$-basis is given by

$$\widetilde{\mathbf{X}}_{\mathfrak{r}} = \sum_{i=1}^{\mathfrak{r}}\lambda_i\mathbf{U}_i = \sum_{i=1}^{\mathfrak{r}}\frac{\mathbf{U}_i^T\mathbf{X}}{\sqrt{q}}\,\mathbf{U}_i = \begin{bmatrix}\frac{\mathbf{U}_1^T\mathbf{X}}{\sqrt{q}} \\ \frac{\mathbf{U}_2^T\mathbf{X}}{\sqrt{q}} \\ \vdots \\ \frac{\mathbf{U}_{\mathfrak{r}}^T\mathbf{X}}{\sqrt{q}}\end{bmatrix}.$$

In consequence, we will consider the following reduced input space

$$\widetilde{\mathcal{X}}_{\mathfrak{r}} := \left\{\widetilde{\mathbf{X}}_{\mathfrak{r}} \in \mathbb{R}^{\mathfrak{r}} : \mathbf{X}\,\mathbf{e}_y^T \in \mathcal{D} \text{ for some } \mathbf{e}_y \in \mathcal{Y}_k\right\},$$

which is contained in

$$B_1^{\mathfrak{r}}(\mathbf{0}) = \left\{\sum_{i=1}^{\mathfrak{r}}\lambda_i\,\mathbf{U}_i \in \mathbb{R}^{\mathfrak{r}} : \sum_{i=1}^{\mathfrak{r}}\lambda_i^2 \leq 1\right\},$$

the closed unit ball of $\mathbb{R}^{\mathfrak{r}}$. So, to construct the empirical classification map $\ell_{\mathcal{D}}$, we will use the following surrogate training set

$$\widetilde{\mathcal{D}} = \left\{\widetilde{\mathbf{X}}_{\mathfrak{r}}\,\mathbf{e}_y^T \in \mathbb{R}^{\mathfrak{r}} \otimes \mathcal{Y}_k : \mathbf{X}\,\mathbf{e}_y^T \in \mathcal{D}\right\}.$$

Since the vectors $\widetilde{\mathbf{X}}_{\mathfrak{r}} = \sum_{i=1}^{\mathfrak{r}} \lambda_i \, \mathbf{U}_i \in \widetilde{\mathcal{X}}_{\mathfrak{r}}$ satisfy $\sum_{i=1}^{\mathfrak{r}} \lambda_i^2 \leq 1$ we can also use the $\mathfrak{r}$-dimensional spherical coordinate system to represent $\widetilde{\mathbf{X}}_{\mathfrak{r}}$, i.e.,

$$\begin{aligned}
\lambda_1 &= r\cos(\varphi_1)\\
\lambda_2 &= r\sin(\varphi_1)\cos(\varphi_2)\\
\lambda_3 &= r\sin(\varphi_1)\sin(\varphi_2)\cos(\varphi_3)\\
&\vdots\\
\lambda_{\mathfrak{r}-1} &= r\sin(\varphi_1)\sin(\varphi_2)\cdots\sin(\varphi_{\mathfrak{r}-2})\cos(\varphi_{\mathfrak{r}-1})\\
\lambda_{\mathfrak{r}} &= r\sin(\varphi_1)\sin(\varphi_2)\cdots\sin(\varphi_{\mathfrak{r}-2})\sin(\varphi_{\mathfrak{r}-1})
\end{aligned}$$

where $\varphi_i \in [0,\pi]$ for $1 \leq i \leq \mathfrak{r}-1$, $\varphi_{\mathfrak{r}-1} \in [0,2\pi)$ and $r \in [0,1]$. Hence we can characterize each $\widetilde{\mathbf{X}}_{\mathfrak{r}}$ as

$$\widetilde{\mathbf{X}}_{\mathfrak{r}} := \begin{bmatrix} r(\widetilde{\mathbf{X}}_{\mathfrak{r}}) \\ \varphi_1(\widetilde{\mathbf{X}}_{\mathfrak{r}}) \\ \vdots \\ \varphi_{\mathfrak{r}-1}(\widetilde{\mathbf{X}}_{\mathfrak{r}}) \end{bmatrix} \in [0,1] \times [0,\pi]^{\mathfrak{r}-2} \times [0,2\pi).$$

A surrogate training set in polar coordinates to be used in a binary classification model

Next, we will concentrate on the binary classification case, i.e., when $k=2$ and $\mathcal{Y}_2 = \{\mathbf{e}_1, \mathbf{e}_2\}$. Under this condition $\operatorname{rank}\rho_{\mathcal{D}} \leq 2$ holds. Without loss of generality assume that $\operatorname{rank}\rho_{\mathcal{D}} = \mathfrak{r} = 2$ and hence each $\widetilde{\mathbf{X}}_2 \in \widetilde{\mathcal{X}}_2$ obtained from a $\mathbf{X} \in \mathcal{X}_d$ can be written as

$$\widetilde{\mathbf{X}}_2 = \frac{\mathbf{U}_1^T\mathbf{X}}{\sqrt{q}}\,\mathbf{U}_1 + \frac{\mathbf{U}_2^T\mathbf{X}}{\sqrt{q}}\,\mathbf{U}_2 \text{ where } 0 < \frac{(\mathbf{U}_1^T\mathbf{X})^2}{q} + \frac{(\mathbf{U}_2^T\mathbf{X})^2}{q} \leq 1.$$

Using polar coordinates we have

$$\begin{aligned}
\frac{\mathbf{U}_1^T\mathbf{X}}{\sqrt{q}} &= r\cos\theta,\\
\frac{\mathbf{U}_2^T\mathbf{X}}{\sqrt{q}} &= r\sin\theta,
\end{aligned}$$

then

$$\begin{aligned}
r(\mathbf{X}) &= \sqrt{\frac{(\mathbf{U}_1^T\mathbf{X})^2}{q} + \frac{(\mathbf{U}_2^T\mathbf{X})^2}{q}},\\
\theta(\mathbf{X}) &= \arctan\frac{\mathbf{U}_2^T\mathbf{X}}{\mathbf{U}_1^T\mathbf{X}},
\end{aligned}$$

and hence

$$\widetilde{\mathbf{X}}_2 = r(\mathbf{X})(\sin\theta(\mathbf{X})\mathbf{U}_1 + \cos\theta(\mathbf{X})\mathbf{U}_2).$$

Thus, we obtain an equivalent parametrization of $\widetilde{\mathbf{X}}_2$ by the parameters $(\theta(\mathbf{X}), r(\mathbf{X})) \in [-\pi/2, \pi/2] \times [0, 1]$, which allows us to define the next surrogate training data set

$$\widetilde{\mathcal{D}} = \left\{ \begin{bmatrix} \theta(\mathbf{X}) \\ r(\mathbf{X})) \end{bmatrix} \mathbf{e}_y^T \in ([-\pi/2, \pi/2] \times [0, 1]) \otimes \mathcal{Y}_k : \mathbf{X}\, \mathbf{e}_y^T \in \mathcal{D} \right\}.$$

On the empirical conditional density functions for the dependent variables and the classification map

In this third step, for each fixed $\mathbf{e}_y \in \mathcal{Y}_k$ put

$$\widetilde{\mathcal{X}}_{\mathfrak{r},\mathbf{e}_y} := \left\{ \widetilde{\mathbf{X}}_{\mathfrak{r}} : \widetilde{\mathbf{X}}_{\mathfrak{r}}\, \mathbf{e}_y^T \in \widetilde{\mathcal{D}} \right\},$$

and we check -for example with the help of some multivariate Kolmogorov-Smirnov test [6]- that the probability distributions of the k-cloud sets:

$$\{\widetilde{\mathcal{X}}_{\mathfrak{r},\mathbf{e}_y} : \mathbf{e}_y \in \mathcal{Y}_k\} \subset B_1^{\mathfrak{r}}(\mathbf{0})$$

are different.

Under the assumption of a positive answer to the above test, we then construct an empirical probability distribution function for each individual set $\widetilde{\mathcal{X}}_{\mathfrak{r},\mathbf{e}_y}$ $(\mathbf{e}_y \in \mathcal{Y}_k)$, using a multivariable kernel method. For this we will consider a function

$$K : \mathbb{R}^{\mathfrak{r}} \longrightarrow [0, \infty)$$

with a compact support in $B_1\mathfrak{r}(\mathbf{0})$ and satisfying

$$\int_{\mathbb{R}^{\mathfrak{r}}} K(\mathbf{x}) d\mathbf{x} = \int_{B_1^{\mathfrak{r}}(\mathbf{0})} K(\mathbf{x}) d\mathbf{x} = 1$$

Usually K will be a radially symmetric unimodal probability density function, for example is the multivariate Epanechnikov kernel defined by

$$K(\mathbf{x}) = \begin{cases} \frac{1}{2} c_{\mathfrak{r}}^{-1} (\mathfrak{r} + 2)(1 - \mathbf{x}^T \mathbf{x}) & \text{if } \mathbf{x}^T \mathbf{x} < 1 \\ 0 & \text{otherwise,} \end{cases},$$

where $c_{\mathfrak{r}}$ is the volume of $B_1^{\mathfrak{r}}(\mathbf{0})$. We recall that a multivariate kernel density estimator with kernel K and window width h in $\widetilde{\mathcal{X}}_{\mathfrak{r},\mathbf{e}_y}$ is defined by

$$\widehat{f}_h(\mathbf{x}|\mathbf{e}_y) = \sum_{\widetilde{\mathbf{X}}_{\mathfrak{r}} \in \widetilde{\mathcal{X}}_{\mathfrak{r},\mathbf{e}_y}} \frac{1}{n h^{\mathfrak{r}}} K\left(\frac{\mathbf{x} - \widetilde{\mathbf{X}}_{\mathfrak{r}}}{h} \right). \tag{5}$$

Since the family of maps $\{\widehat{f}_h(\mathbf{x}|\mathbf{e}_y) : \mathbf{e}_y \in \mathcal{Y}_k\}$ are continuous with compact support, we can introduce the family on non-empty sets

$$\mathcal{M}_y(\widetilde{\mathcal{D}}) := \arg \max_{\mathbf{x} \in B_1(\mathbf{0})} \widehat{f}_h(\mathbf{x}|\mathbf{e}_y) \text{ for each } \mathbf{e}_y \in \mathcal{Y}_k, \text{ for } 1 \leq y \leq k.$$

The multivariate Epanechnikov kernel is unimodal, thus the map $\widehat{f}_h(\mathbf{x}|\mathbf{e}_y)$ has an absolute maximum and hence $\mathcal{M}_y(\widetilde{\mathcal{D}}) = \{\mathbf{x}_y^{\max}\}$.

Now, we have everything to define

$$\ell_{\widetilde{\mathcal{D}}} : B_1^{\mathfrak{r}}(\mathbf{0}) \subset \mathbb{R}^{\mathfrak{r}} \longrightarrow \mathcal{Y}_k,$$

by

$$\ell_{\widetilde{\mathcal{D}}}(\widetilde{\mathbf{X}}_{\mathfrak{r}}) = \mathbf{e}_y \text{ if and only if } \|\widetilde{\mathbf{X}}_{\mathfrak{r}} - \mathbf{x}_y^{\max}\| < \|\widetilde{\mathbf{X}}_{\mathfrak{r}} - \mathbf{x}_{y'}^{\max}\| \text{ holds for all } y' \neq y.$$

Evaluation of the classification map

Evaluating a classification model consists of determining how often labels are correctly or incorrectly classified for the testing samples. In other words, it is counting how many times a sample is correctly or incorrectly labelled into a particular class. We distinguish four qualities:

- TP (True Positive): the correct classification of a sample into a class;
- TN (True Negative): the correct classification of a sample out of a class;
- FP (False Positive): the incorrect classification of a sample into a class;
- FN (False Negative): the incorrect classification of a sample out of class.

To assess the quality of the classification map, confusion matrix, sensitivity, specificity, the predictive values, accuracy, F1 score, Cohen's kappa [7] and Matthews Correlation Coefficient [8] will be computed by using the above qualities. Recall that

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \ \text{(worst value = 0; best value = 1)},$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \ \text{(worst value = 0; best value = 1)},$$

$$\text{Positive predictive value} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \ \text{(worst value = 0; best value = 1)},$$

$$\text{Negative predictive value} = \frac{\text{TN}}{\text{TN} + \text{FN}}, \ \text{(worst value = 0; best value = 1)},$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \ \text{(worst value = 0; best value = 1)},$$

$$\text{F1 Score} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}, \ \text{(worst value = 0; best value = 1)}.$$

The definition of the Cohen's kappa of a binary classification confusion matrix is given by

$$\kappa = \frac{2(\text{TP} \times \text{TP} - \text{FN} \times \text{FP})}{(\text{TP} \times \text{FP}) + (\text{FP} \times \text{TN}) + (\text{TP} \times \text{FN}) + (\text{FN} \times \text{TN})}$$
$$\text{(worst value} = -1\text{; best value} = 1\text{; agreement expected by chance} = 0).$$

and the Matthews Correlation Coefficient (MCC) which is defined by

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}},$$

$$(\text{worst value} = -1;\ \text{best value} = 1;\ \text{agreement expected by chance} = 0).$$

To measure the quality of our performance we will use the *receiver operating characteristic* (ROC) curve. This curve is computed using Sensitivity on the vertical axis and 1-Specificity on the horizontal axis. Evaluating the performance is given by the so-called *area under the curve* (AUC): the greater the AUC is, the better model is performing. ROC curves are insensitive to changes in class distribution. If the proportion of positive to negative instances changes in a test set, the ROC curves will not change [9]. However, we often have a data set with many negative instances and few positives instances. For these kind of imbalanced sets, according to Chicco [10], the *Precision-Recall curve* is a more reliable and informative indicator of the statistical performance of the proposal method.

Results

We proved the usefulness and effectiveness of the proposed method using routine clinical data. Our aim was to establish a classification model to distribute patients – with a prior diagnosis of a Solitary Pulmonary Nodule (SPN) – into two classes; class 1 patients diagnosed with lung cancer, or class 0 free of lung cancer after five-years of follow-up. For that purpose, we reused routine clinical data from patients of two public university general hospitals. Briefly, SPN, defined as a pulmonary opacity up to 30 mm in diameter, is a common finding in routine clinical practice when performing chest imaging tests such as X-rays or computed tomography (CT) for any reason [11, 12]. The vast majority of these nodules are benign, and only a small proportion (around 10–20%) are malignant [13, 14]. A Spanish cohort found, after five years of follow-up, a prevalence of malignant SPN, detected by chest radiography or CT, of 12.1% and 18.2% respectively [15]. In routine clinical practice, the risk of malignant SPN can be clinician-assigned based on clinician judgement, which can lead to the use of inadequate clinical tests and treatments with potential side effects such as excess radiation, or calculated using a validated risk prediction model. The applicability of a predictive model, rather than just using clinical judgement based on intuition/experience, offers a standardised and reproducible approach to nodule risk assessment [16].

Results from a recent systematic review which evaluated 15 SPN malignancy predictive models applicable to routine clinical practice, indicated that most models were derived from multivariable logistic regression models [17]. All models constructed in these studies were classified as having a high risk of bias compromising their clinical applicability. The quality of these models was assessed using the Prediction model Risk Of Bias Assessment Tool (PROBAST) [18], to construct a structured judgement of the applicability and transferability of predictive models to clinical practice.

For the purpose of this work, we used data from a retrospective cohort study of patients aged $\geq$35 years referred for thoracic imaging for non-screening reasons (such as preoperative evaluation) in two hospitals from the Valencian Autonomous Community (south-east Spain) from within the hospital and from primary health care centres during 2010 and 2011. Out of 25,422 patients with an imaging test performed during that period at both hospitals, 893 patients were found to have an SPN. Patients previously diagnosed with lung cancer or with intrapulmonary lymph nodes and pseudo-lesions were excluded from the study. The detailed methodology has been described elsewhere [19, 20]. All patients with SPN (893) were followed up for five years or up to a diagnosis of lung cancer from nodule detection. Selected clinical and demographic variables were collected from medical records.

We limited our study to 404 patients for which complete information on the eight chosen predictor variables were available: sex, previous malignancy, smoking habit (non-smokers, current or former smokers), Chronic Obstructive Pulmonary Disease (COPD), more than one SPN, SPN diameter (mm), SPN location (lung lobe) and SPN border. These variables are usually considered by most medical guidelines to the most significant in managing the diagnoses procedures when an SPN is found in a routine clinical context. The eight chosen predictors were found to be associated with a risk of lung cancer in two previous analyses with the same population [15, 20]. For missing values, it was decided not to impute missing data and rely only on data collected.

The modalities of the eight variables and their distribution according to the two classes are shown in Table 1. Sex, Previous Malignancy, Smoker, COPD, More than one SPN have two modalities, however, SPN diameter and SPN location (lobe) use three modalities and SPN border uses four. The dependent variable collected (which functions as a class label) was the development or not of lung cancer during the five years following the detection of the SPN. In this period, 22% of patients were diagnosed with lung cancer. The percentage is higher than normal because the diagnosed patients were slightly better documented than the undiagnosed. Of the initial 893, the percentage was 14.9%. These predictor variables can be considered as a survey of $q = 8$ questions with $d = 20$ answers.

Although patient characteristics, such as advanced age have been associated with a higher risk of malignant SPN [11, 21, 22], age was not considered because its inclusion in the training dataset did not change the results and therefore did not contribute to the construction of the classification map, as explain below. Diameter, on the other hand, followed a very specific pattern and its introduction was considered significant.

We constructed the classification map with 404 participants with SPN after five years of follow-up with a mean age of 65 ± 12 years. We consider each of these 20 different answers as a measurement for an individual patient. For each question, every individual could only have one modality marked as 1 with the other modalities being 0. Moreover, each individual was classified in class 0 : non SPN diagnosis, or otherwise in class 1. Table 2 gives an example of two data vectors from two different individuals according to the two different classes.

After we separated patients into two groups: one consisting of two thirds of the patients to construct the model (the training dataset), and the remaining third (the test dataset) was used to validate the model. With the individual data vectors of the first group, matrix D was computed (Table 3). It contains the frequency of each answer in each class in our database. Calculating the root of all its elements, the X matrix is obtained (Table 4). The density matrix ($\rho_{\mathcal{D}}$) is then calculated (Table 5). In our model, the range of $\rho_{\mathcal{D}}$ is equal to the number of classes, so $\mathfrak{r}$ is equal to 2.

Subsequently, the SVD of $\rho_{\mathcal{D}}$ was performed obtaining the orthonormal basis $\mathfrak{B}(\rho_{\mathcal{D}})$. With it, each input data $\widetilde{\mathbf{X}}$ can be then considered by means of its truncated representation $\widetilde{\mathbf{X}}_{\mathfrak{r}}$.

These $\widetilde{\mathbf{X}}_{\mathfrak{r}}$ are obtained from the individual data vectors of the training dataset, obtaining a surrogate training dataset. Likewise, the $\widetilde{\mathbf{X}}_{\mathfrak{r}}$ of the test dataset are calculated using the same basis vectors $\mathfrak{B}(\rho_{\mathcal{D}})$ previously computed with the training set. The true classification for each of them is known.

The polar coordinates $(\theta(\mathbf{X}), r(\mathbf{X}))$ of both the training and test datasets were then calculated. With this new surrogate training dataset, having used the multivariate Kolmogorov-Smirnov test to check that the probability distribution of the coordinates in the training dataset is different between the two classes, an empirical probability density function of both coordinates of each class $(f(\mathbf{X}|\mathbf{e}_y))$ was constructed using the multivariate Gaussian kernel and the variance matrix as the bandwidth matrix. An example of the density functions are showed in Figure 1. They were constructed using the R package `ks` [23]. Each individual in the test dataset is assigned to the class whose maximum probability has the smallest distance to that individual's coordinate values.

Nevertheless, as shown in Figure 2, it was observed that the classes were discriminated primarily by the θ coordinate. It was therefore decided to carry out the classification following the same process as above but using only this coordinate. In this case, the multivariate kernel used was the Epanechnikov kernel. The bandwidth was chosen using Silverman's rule of thumb. An example of the density functions is shown in Figure 3.

In total, 1000 analyses were performed by randomising the training dataset and the test dataset. With the difference between the predicted and the real classes, both with the two coordinates and with θ only, the confusion matrices were calculated. The average of the matrices of the 1000 analyses using the two coordinates is given in Table 6, using only θ in Table 7. Furthermore, to quantify the model's goodness of fit, the predictive values, the accuracy, the sensitivity, the specificity, the F1 score and the Cohen's kappa coefficient were computed. The averages of these are shown in Table 8.

Subsequently, to assess the goodness of the model, an ROC curve was constructed. This curve is normally used with analyses that yield a probability, but can be constructed using another statistic or score, i.e. a numeric value that represents the degree to which an individual is a member of a class [9]. In our case, we have used the difference between the distances of the patient's coordinate to the maximum of each of the two classes.

To summarise the 1000 curves in one, we use the vertical averaging method [9]. It allows the mean and confidence interval to be plotted at different points on the curve. The averaged curve of the analysis using the two coordinates with 95% confidence intervals is shown in Figure 4. It was constructed using the R package `ROCR` [24]. The result using the θ coordinate was very similar (data not shown). Moreover, the averages of the area under the ROC curve of the 1000 analyses, both using the two coordinates and only using θ, are shown in Table 8.

We also present the graph reflecting the relationship between the positive predictive value and the sensitivity (usually called the Precision/Reception graph) of the analysis employing the two coordinates, using the same output score, the same averaging method and the same package to construct it (Figure 5).

Discussion

In our model, the classification using the two coordinates (r, θ) and only θ garners very similar results. It can therefore be concluded, as seen in Figure 2, that discrimination is mainly produced by the θ coordinate.

The results obtained (Tables 6 and 7) are acceptable, with better discrimination of patients who were actually diagnosed with lung cancer. This is explained by the higher number of undiagnosed patients in the area where the density functions of the diagnosed and undiagnosed overlap (Figure 3).

All parameters in Table 8, except the positive predictive value (PPV), also result in acceptable values. The low value of the PPV is partly explained by a higher proportion of undiagnosed patients. It is worth noting the good value of the MCC. The MCC, as described by Chicco *et al.* [25], is a more comprehensive coefficient than Cohen's kappa coefficient in checking the performance of a classification map. The good assessment obtained for our proposed methodology is reinforced by the results of the ROC curve. However, the information provided by the Precision/Recall graph is not as strong due to the worse positive predictive value results, due in part to using an unbalanced dataset, i.e., the number of individuals with a diagnosis of lung cancer being much lower (22%) than those without a diagnosis (78%).

Conclusions

This classification model is based on concepts borrowed from non-classical probability theory arising in quantum mechanics and provides an acceptable performance that could be a used with routine clinical data..

Funding
This research was funded by the grant number COMCUANTICA/007 from the Generalitat Valenciana, Spain, by the grant number INDI22/15 from the Universidad CEU Cardenal Herrera, Spain, and by the Fondo de Investigaciones Sanitarias, Instituto de Salud Carlos III (Plan Nacional de I+D+I) PI18/00258 and cofunded by the Fondo Europeo de Desarrollo Regional (FEDER).

Availability of data and materials
The data and the code used to generate the tables and the figures in this paper is freely available at https://github.com/afalco/quantum_based_learning.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Conceptualization, R. B-R., J. L., A. F., M. P-V.; obtaining funding, J. L., A. F., M. P-V.; data collection, J.M. S-S.; data curation, R. B-R.; methodology R. B-R., J. L., A. F., M. P-V.; mathematical modelling and analysis, R. B-R., M.C. S-A., J. F-M., B. S-G., J. C., A. F.; validation, R. B-R., A. F.; software, R. B-R., J. F-M., A. F.; supervision, J. L., A. F., M. P-V.; interpretation of the results, R. B-R., J.L., M. S-V., A. F., M. P-V. All authors cowrote the manuscript. All authors critically revised the paper for important intellectual content and approved the final version.

Author details
[1]Departamento de Matemáticas, Física y Ciencias Tecnológicas, Universidad Cardenal Herrera-CEU, CEU Universities, Alfara del Patriarca, Spain. [2]ESI International Chair@CEU-UCH, Universidad Cardenal Herrera-CEU, CEU Universities, Alfara del Patriarca, Spain. [3]NAVARRABIOMED, Centro de Investigación Biomédica, Pamplona, Spain. [4]Salud Pública, Historia de la Ciencia y Ginecología, Universidad Miguel Hernández, Alicante, Spain. [5]Departamento de Farmacia, Universidad Cardenal Herrera-CEU, CEU Universities, Alfara del Patriarca, Spain. [6]Hospital Universitario Sant Joan d'Alacant, Alicante, Spain. [7]Departamento de Ciencias Biomédicas, Universidad

Cardenal Herrera-CEU, CEU Universities, Alfara del Patriarca, Spain. [8]Red de Investigación en Servicios de Salud en Enfermedades Crónicas (REDISSEC), Valencia, Spain. [9]CIBER en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [10]Servicio de Dermatología, Hospital General Universitario de Alicante, Alicante, Spain.

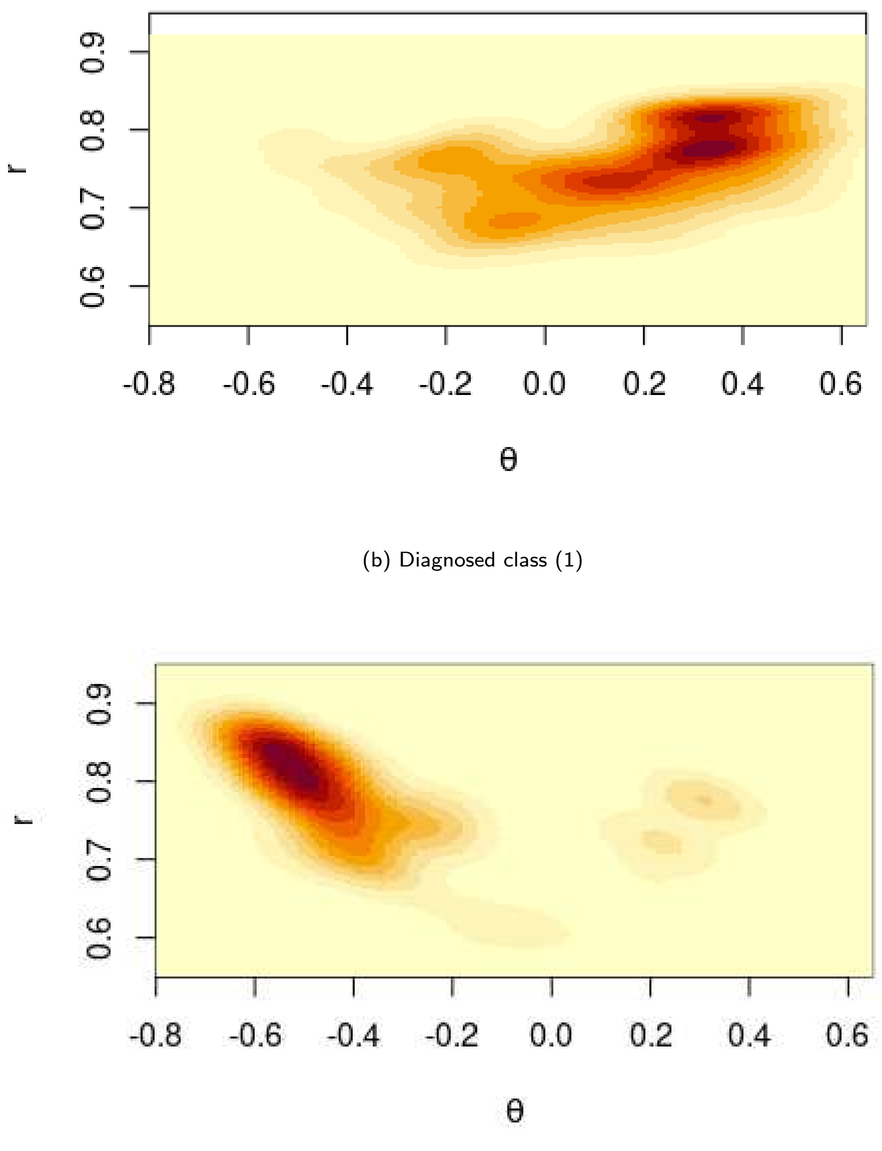
References

1. Chang, P.C., Afifi, A.A.: Classification based on dichotomous and continuous variables. Journal of the American Statistical Association **69**(346), 336–339 (1974). doi:10.1080/01621459.1974.10482949
2. Krzanowski, W.J.: Discrimination and classification using both binary and continuous variables. Journal of the American Statistical Association **70**(352), 782–790 (1975). doi:10.1080/01621459.1975.10480303
3. Olkin, I., Tate, R.F.: Multivariate correlation models with mixed discrete and continuous variables. The Annals of Mathematical Statistics **32**(2), 448–465 (1961). doi:10.1214/aoms/1177705052
4. Murphy, K.P.: Probabilistic Machine Learning: An Introduction. MIT Press, Cambridge, MA (2012). probml.ai
5. Abdi, H.: Discriminant Correspondence Analysis. Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA (2007)
6. Fasano, G., Franceschini, A.: A multidimensional version of the kolmogorov–smirnov test. Monthly Notices of the Royal Astronomical Society **225**(1), 155–170 (1987). doi:10.1093/mnras/225.1.155
7. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
8. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta **405**(2), 442–451 (1975)
9. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters **27**(8), 861–874 (2006). doi:10.1016/j.patrec.2005.10.010
10. Chicco, D.: Ten quick tips for machine learning in computational biology. BioData Mining **10**(1) (2017). doi:10.1186/s13040-017-0155-3
11. Gould, M.K., Donington, J., Lynch, W.R., Mazzone, P.J., Midthun, D.E., Naidich, D.P., Wiener, R.S.: Evaluation of individuals with pulmonary nodules: When is it lung cancer? Chest **143**(5), 93–120 (2013). doi:10.1378/chest.12-2351
12. Lumbreras, B., Vilar, J., González-Álvarez, I., Gómez-Sáez, N., Domingo, M.L., Lorente, M.F., Pastor-Valero, M., Hernández-Aguado, I.: The fate of patients with solitary pulmonary nodules: Clinical management and radiation exposure associated. PLOS ONE **11**(7), 0158458 (2016). doi:10.1371/journal.pone.0158458
13. Alzahouri, K., Velten, M., Arveux, P., Woronoff-Lemsi, M.-C., Jolly, D., Guillemin, F.: Management of SPN in france. pathways for definitive diagnosis of solitary pulmonary nodule: a multicentre study in 18 french districts. BMC Cancer **8**(1) (2008). doi:10.1186/1471-2407-8-93
14. Wiener, R.S., Gould, M.K., Slatore, C.G., Fincke, B.G., Schwartz, L.M., Woloshin, S.: Resource use and guideline concordance in evaluation of pulmonary nodules for cancer. JAMA Internal Medicine **174**(6), 871 (2014). doi:10.1001/jamainternmed.2014.561
15. Chilet-Rosell, E., Parker, L.A., Hernández-Aguado, I., Pastor-Valero, M., Vilar, J., González-Álvarez, I., Salinas-Serrano, J.M., Lorente-Fernández, F., Domingo, M.L., Lumbreras, B.: The determinants of lung cancer after detecting a solitary pulmonary nodule are different in men and women, for both chest radiograph and CT. PLOS ONE **14**(9), 0221134 (2019). doi:10.1371/journal.pone.0221134
16. Fox, A.H., Tanner, N.T.: Approaches to lung nodule risk assessment: clinician intuition versus prediction models. Journal of Thoracic Disease **12**(6), 3296–3302 (2020). doi:10.21037/jtd.2020.03.68
17. Senent-Valero, M., Librero, J., Pastor-Valero, M.: Solitary pulmonary nodule malignancy predictive models applicable to routine clinical practice: a systematic review. Systematic Reviews **10**(1) (2021). doi:10.1186/s13643-021-01856-6
18. Moons, K.G.M., Wolff, R.F., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J., Mallett, S.: PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. Annals of Internal Medicine **170**(1), 1 (2019). doi:10.7326/m18-1377
19. Gómez-Sáez, N., González-Álvarez, I., Vilar, J., Hernández-Aguado, I., Domingo, M.L., Lorente, M.F., Pastor-Valero, M., Parker, L.A., Picazo, N., Calbo, J., Lumbreras, B.: Prevalence and variables associated with solitary pulmonary nodules in a routine clinic-based population: a cross-sectional study. Eur. Radiol. **24**(9), 2174–2182 (2014)
20. Gómez-Sáez, N., Hernández-Aguado, I., Vilar, J., González-Alvarez, I., Lorente, M.F., Domingo, M.L., Valero, M.P., Parker, L.A., Lumbreras, B.: Lung cancer risk and cancer-specific mortality in subjects undergoing routine imaging test when stratified with and without identified lung nodule on imaging study. Eur. Radiol. **25**(12), 3518–3527 (2015)
21. Chung, K., Mets, O.M., Gerke, P.K., Jacobs, C., den Harder, A.M., Scholten, E.T., Prokop, M., de Jong, P.A., van Ginneken, B., Schaefer-Prokop, C.M.: Brock malignancy risk calculator for pulmonary nodules: validation outside a lung cancer screening population. Thorax **73**(9), 857–863 (2018)
22. MacMahon, H., Naidich, D.P., Goo, J.M., Lee, K.S., Leung, A.N.C., Mayo, J.R., Mehta, A.C., Ohno, Y., Powell, C.A., Prokop, M., Rubin, G.D., Schaefer-Prokop, C.M., Travis, W.D., Schil, P.E.V., Bankier, A.A.: Guidelines for management of incidental pulmonary nodules detected on CT images: From the fleischner society 2017. Radiology **284**(1), 228–243 (2017). doi:10.1148/radiol.2017161659
23. Duong, T.: Ks: Kernel Smoothing. (2022). R package version 1.13.5. https://CRAN.R-project.org/package=ks
24. Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T.: Rocr: visualizing classifier performance in r. Bioinformatics **21**(20), 7881 (2005)
25. Chicco, D., Warrens, M.J., Jurman, G.: The matthews correlation coefficient (MCC) is more informative than cohen's kappa and brier score in binary classification assessment. IEEE Access **9**, 78368–78381 (2021). doi:10.1109/access.2021.3084050

Figures

Tables

Figure 1: An example (among the 1000 analyses performed) of the density functions $f(\mathbf{X}|\mathbf{e}_y)$ for $y \in \{0, 1\}$ of the test dataset.



(a) Not diagnosed class (0)

(b) Diagnosed class (1)

Figure 2: An example (among the 1000 analyses performed) of the distribution of the polar coordinates of the datatest set

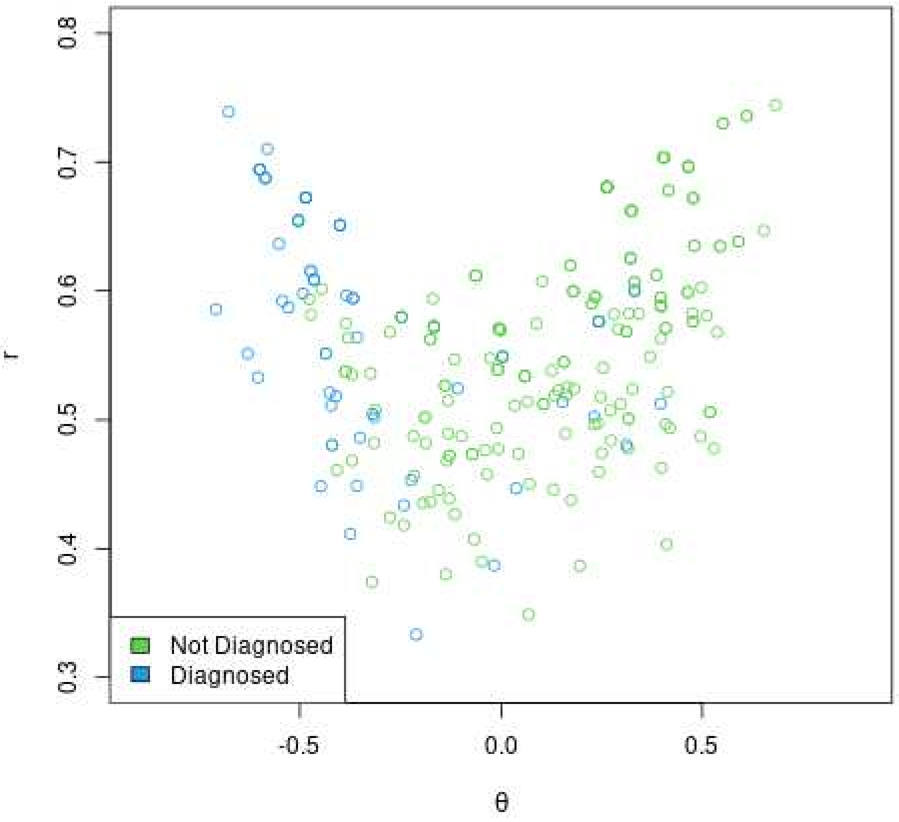


Figure 3: An example (among the 1000 analyses performed) of the density functions $f(\theta|y)$ for $y \in \{0,1\}$ of the test dataset.

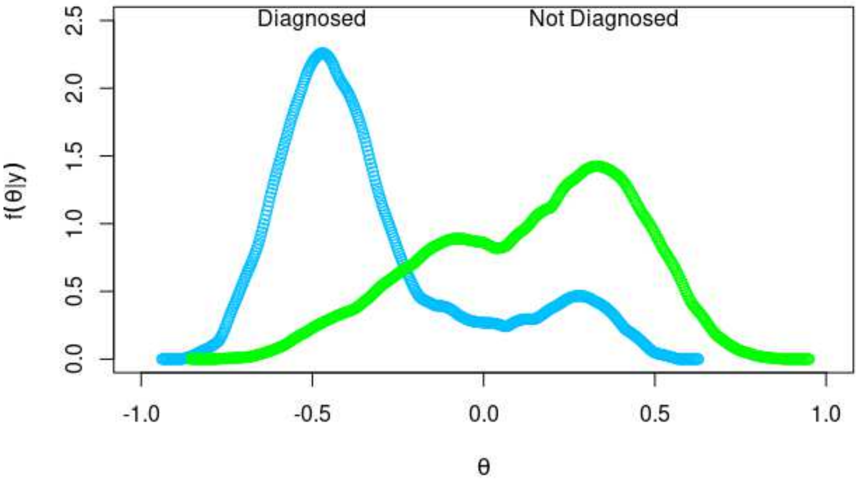


Figure 4: ROC curve averaged using the vertical averaging method. The ROC curve of 1000 analyses was averaged using the two coordinates as predictors. The bars show 95% confidence intervals.

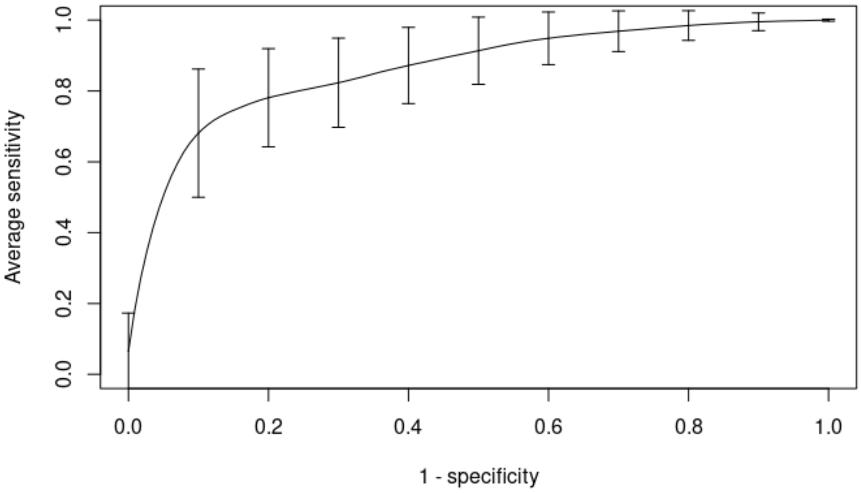


Figure 5: Positive predictive value-sensitivity curve averaged using the vertical averaging method. The positive predictive value-sensitivity curve of 1000 analyses was averaged using the two coordinates as predictors. The bars show 95% confidence intervals.

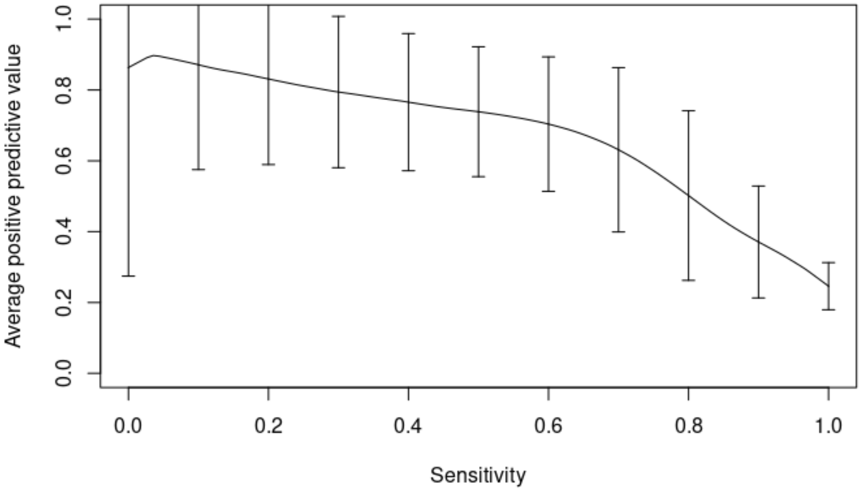


Table 1: The data used to construct the model.

N (%)	Not diagnosed	Diagnosed	Total	*p* value[1]
N	315 (78.0)	89 (22.0)	404	
Sex				
Male	197 (62.5)	68 (76.4)	265 (65.6)	0.021
Female	118 (37.5)	21 (23.6)	139 (34.4)	
Previous malignancy				
No	204 (64.8)	48 (53.9)	252 (62.4)	0.082
Yes	111 (35.2)	41 (46.1)	152 (37.6)	
Smoker				
Never	99 (31.4)	10 (11.2)	109 (27)	< 0.001
Former or current	216 (68.6)	79 (88.8)	295 (73)	
COPD				
No	228 (72.4)	52 (58.4)	280 (69.3)	0.017
Yes	87 (27.6)	37 (41.6)	124 (30.7)	
More than one SPN				
No	266 (84.4)	72 (80.9)	338 (83.7)	0.524
Yes	49 (15.6)	17 (19.1)	66 (16.3)	
SPN diameter (mm)				
< 11.3	232 (73.7)	15 (16.9)	247 (61.1)	< 0.001
11.3 - 20.7	63 (20)	34 (38.2)	97 (24)	
> 20.7	20 (6.3)	40 (44.9)	60 (14.9)	
SPN location (lobe)				
Lower	114 (36.2)	28 (31.5)	142 (35.1)	0.028
Middle	36 (11.4)	3 (3.4)	39 (9.7)	
Upper	165 (52.4)	58 (65.2)	223 (55.2)	
SPN border				
Smooth	155 (49.2)	6 (6.7)	161 (39.9)	< 0.001
Lobulation	47 (14.9)	21 (23.6)	68 (16.8)	
Spiculation	63 (20)	16 (18)	79 (19.6)	
Other irregular	50 (15.9)	46 (51.7)	96 (23.8)	

[1] *p* value of Pearson's chi-squared test

Table 2: An example of two individual data samples, one for class 0 and another for class 1.

Class	0
Sex Male	1
Sex Female	0
More than one SPN Yes	1
More than one SPN No	0
SPN diameter (mm) $<$ 11.3	1
SPN diameter (mm) 11.3-20.7	0
SPN diameter (mm) $>$ 20.7	0
SPN location (lobe) Middle	0
SPN location (lobe) Upper	1
SPN location (lobe) Lower	0
SPN border Other	0
SPN border Spiculation	1
SPN border Lobulation	0
SPN border Smooth	0
Previous malignancy Yes	0
Previous malignancy No	1
Smoker Yes	1
Smoker Never	0
COPD Yes	1
COPD No	0

Class	1
Sex Male	0
Sex Female	1
More than one SPN Yes	1
More than one SPN No	0
SPN diameter (mm) $<$ 11.3	0
SPN diameter (mm) 11.3-20.7	0
SPN diameter (mm) $>$ 20.7	1
SPN location (lobe) Middle	0
SPN location (lobe) Upper	0
SPN location (lobe) Lower	1
SPN border Other	0
SPN border Spiculation	0
SPN border Lobulation	0
SPN border Smooth	1
Previous malignancy Yes	0
Previous malignancy No	1
Smoker Yes	0
Smoker Never	1
COPD Yes	0
COPD No	1

Table 3: An example of matrix D.

	0	1
Sex Male	124	49
Sex Female	84	13
More than one SPN No	174	51
More than one SPN Yes	34	11
SPN diameter (mm) $<$ 11.3	159	10
SPN diameter (mm) 11.3-20.7	37	26
SPN diameter (mm) $>$ 20.7	12	26
SPN location (lobe) Upper	116	42
SPN location (lobe) Middle	24	2
SPN location (lobe) Lower	68	18
SPN border Smooth	98	6
SPN border Spiculation	40	11
SPN border Lobulation	36	13
SPN border Other	34	32
Previous malignancy No	135	31
Previous malignancy Yes	73	31
Smoker Never	64	8
Smoker Yes	144	54
COPD No	150	36
COPD Yes	58	26

Table 4: An example of matrix X computed from matrix D.

	0	1
Sex Male	11.14	7.00
Sex Female	9.17	3.61
More than one SPN No	13.19	7.14
More than one SPN Yes	5.83	3.32
SPN diameter (mm) < 11.3	12.61	3.16
SPN diameter (mm) 11.3-20.7	6.08	5.10
SPN diameter (mm) > 20.7	3.46	5.10
SPN location (lobe) Upper	10.77	6.48
SPN location (lobe) Middle	4.90	1.41
SPN location (lobe) Lower	8.25	4.24
SPN border Smooth	9.90	2.45
SPN border Spiculation	6.32	3.32
SPN border Lobulation	6.00	3.61
SPN border Other	5.83	5.66
Previous malignancy No	11.62	5.57
Previous malignancy Yes	8.54	5.57
Smoker Never	8.00	2.83
Smoker Yes	12.00	7.35
COPD No	12.25	6.00
COPD Yes	7.62	5.10

Table 5: An example of matrix $\rho_{\mathcal{D}}$ computed from matrix X. $\rho_{\mathcal{D}}$

	Sex Male	Sex Female	More than one SPN No	More than one SPN Yes	SPN diameter (mm) $<$ 11.3	SPN diameter (mm) 11.3-20.7	SPN diameter (mm) $>$ 20.7	SPN location (lobe) Upper	SPN location (lobe) Middle	SPN location (lobe) Lower	SPN border Smooth	SPN border Spiculation	SPN border Lobulation	SPN border Other	Previous malignancy No	Previous malignancy Yes	Smoker Never	Smoker Yes	COPD No	COPD Yes
Sex Male	0.08	0.06	0.09	0.04	0.08	0.05	0.03	0.08	0.03	0.06	0.06	0.04	0.04	0.05	0.08	0.06	0.05	0.09	0.08	0.06
Sex Female	0.06	0.04	0.07	0.03	0.06	0.03	0.02	0.06	0.02	0.04	0.05	0.03	0.03	0.03	0.06	0.05	0.04	0.06	0.06	0.04
More than one SPN No	0.09	0.07	0.10	0.05	0.09	0.05	0.04	0.09	0.03	0.06	0.07	0.05	0.05	0.05	0.09	0.07	0.06	0.10	0.09	0.06
More than one SPN Yes	0.04	0.03	0.05	0.02	0.04	0.02	0.02	0.04	0.02	0.03	0.03	0.02	0.02	0.02	0.04	0.03	0.03	0.04	0.04	0.03
SPN diameter (mm) $<$ 11.3	0.08	0.06	0.09	0.04	0.08	0.04	0.03	0.07	0.03	0.05	0.06	0.04	0.04	0.04	0.08	0.06	0.05	0.08	0.08	0.05
SPN diameter (mm) 11.3-20.7	0.05	0.03	0.05	0.02	0.04	0.03	0.02	0.05	0.02	0.03	0.03	0.03	0.03	0.03	0.05	0.04	0.03	0.05	0.05	0.03
SPN diameter (mm) $>$ 20.7	0.03	0.02	0.04	0.02	0.03	0.02	0.02	0.03	0.01	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.02	0.04	0.03	0.02
SPN location (lobe) Upper	0.08	0.06	0.09	0.04	0.07	0.05	0.03	0.07	0.03	0.05	0.06	0.04	0.04	0.05	0.07	0.06	0.05	0.08	0.08	0.05
SPN location (lobe) Middle	0.03	0.02	0.03	0.02	0.03	0.02	0.01	0.03	0.01	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.03	0.03	0.02
SPN location (lobe) Lower	0.06	0.04	0.06	0.03	0.05	0.03	0.02	0.05	0.02	0.04	0.04	0.03	0.03	0.03	0.06	0.04	0.04	0.06	0.06	0.04
SPN border Smooth	0.06	0.05	0.07	0.03	0.06	0.03	0.02	0.06	0.02	0.04	0.05	0.03	0.03	0.03	0.06	0.05	0.04	0.06	0.06	0.04
SPN border Spiculation	0.04	0.03	0.05	0.02	0.04	0.03	0.02	0.04	0.02	0.03	0.03	0.02	0.02	0.03	0.04	0.03	0.03	0.05	0.05	0.03
SPN border Lobulation	0.04	0.03	0.05	0.02	0.04	0.03	0.02	0.04	0.02	0.03	0.03	0.02	0.02	0.03	0.04	0.03	0.03	0.05	0.04	0.03
SPN border Other	0.05	0.03	0.05	0.02	0.04	0.03	0.02	0.05	0.02	0.03	0.03	0.03	0.03	0.03	0.05	0.04	0.03	0.05	0.05	0.03
Previous malignancy No	0.08	0.06	0.09	0.04	0.08	0.05	0.03	0.07	0.03	0.06	0.06	0.04	0.04	0.05	0.08	0.06	0.05	0.08	0.08	0.05
Previous malignancy Yes	0.06	0.05	0.07	0.03	0.06	0.04	0.03	0.06	0.02	0.04	0.05	0.03	0.03	0.04	0.06	0.05	0.04	0.07	0.06	0.04
Smoker Never	0.05	0.04	0.06	0.03	0.05	0.03	0.02	0.05	0.02	0.04	0.04	0.03	0.03	0.03	0.05	0.04	0.03	0.05	0.05	0.03
Smoker Yes	0.09	0.06	0.10	0.04	0.08	0.05	0.04	0.08	0.03	0.06	0.06	0.05	0.05	0.05	0.08	0.07	0.05	0.09	0.09	0.06
COPD No	0.08	0.06	0.09	0.04	0.08	0.05	0.03	0.08	0.03	0.06	0.06	0.05	0.04	0.05	0.08	0.06	0.05	0.09	0.09	0.06
COPD Yes	0.06	0.04	0.06	0.03	0.05	0.03	0.02	0.05	0.02	0.04	0.04	0.03	0.03	0.03	0.05	0.04	0.03	0.06	0.06	0.04

Table 6: Average Confusion Matrix using the two coordinates

	Predicted classification	
	Not diagnosed	Diagnosed
Real classification		
Not diagnosed	0.7198	0.2802
Diagnosed	0.1862	0.8138

Table 7: Average Confusion Matrix using only the θ coordinate

	Predicted classification	
	Not diagnosed	Diagnosed
Real classification		
Not diagnosed	0.7147	0.2853
Diagnosed	0.1842	0.8158

Table 8: Classification performance parameters

	Two coordinates	θ coordinate only
Sensitivity	0.8138	0.8158
Specificity	0.7198	0.7147
Negative Predictive Value	0.9322	0.9323
Positive predictive value	0.4517	0.4469
Accuracy	0.7403	0.7368
F1 Score	0.5781	0.5751
Cohen's kappa coefficient	0.4131	0.4080
Matthews correlation coefficient	0.4521	0.4481
Area under the ROC curve	0.8617	0.8598