UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
Instituto Universitario de Microelectrónica Aplicada
Sistemas de información y Comunicaciones

Master of Science in Telecommunication Technologies

Master Thesis

Semi-Supervised Classification of Hyperspectral Images for Brain Tumours detection

Author:	Mrs. Patricia Beltrán Alonso
Supervisor(s):	Dr. Gustavo Marrero Callicó
	Dr. Samuel Ortega Sarmiento
	Mrs. Beatriz Martínez Vega
Date:	September 2021

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
Instituto Universitario de Microelectrónica Aplicada
Sistemas de información y Comunicaciones

Master of Science in Telecommunication Technologies

IUMA

Master Thesis

Semi-Supervised Classification of Hyperspectral Images for Brain Tumours detection

Signatures

Author:	Mrs. Patricia Beltrán Alonso	Signature:
Supervisor(s):	Dr. Gustavo Marrero Callicó	Signature:
	Dr. Samuel Ortega Sarmiento	Signature:
	Mrs. Beatriz Martínez Vega	Signature:

Date: **September 2021**

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
Instituto Universitario de Microelectrónica Aplicada
Sistemas de información y Comunicaciones

Master of Science in Telecommunication Technologies

IUMA

Master Thesis

Semi-Supervised Classification of Hyperspectral Images for Brain Tumours detection

Evaluation

Grade: ....

President: Signature:

Secretary: Signature:

Member: Signature:

Date: **September 2021**

Abstract

This Master Thesis is related to the classification of medical hyperspectral (HS) data. The objective of this Master Thesis is to design a semi-supervised algorithm to carry out the labelling of the new acquired hyperspectral (HS) images, with the goal to incorporate these data in a supervised classification scheme. To develope this Master Thesis, a database obtained at the University Hospital Doctor Negrín was employed. This HS database is composed by 26 HS cubes belonging to a total of 16 different patients diagnosed with Glioblastoma primary brain tumour, where the test set consisted of 6 captures corresponding to 4 patients. The images were labelled with 4 different classes: normal tissue, tumour tissue, hypervascularized tissue, and background.

The main idea is to solve the problem that arises in these operating rooms, where there is a previously labelled database and the new data acquired from the patient who is undergoing surgery. The objective is to include this data from the current patient to the database with which the classification model is trained and generated. With this proposal it is possible to generate a learning model using the labelled data obtained in previous surgical interventions and the unlabelled data of the current patient. The main goal is to be able to improve the classification results by including data from the new patient.

To carry out the automatic generation of the current patient labels, it was decided to use the k-means algorithm. The chosen method uses the Euclidean distance by default, but a preliminary study was carried out to select the distance metric that better fits our database. According to this study, the cosine distance was chosen. Subsequently, to optimize the algorithm performance, a study was made to select the value of the parameter k.

Once these parameters have been selected, the current patient data are automatically labelled. Labelling was done in two ways, first looking what is the majority class for each cluster and then, with the proviso that only those clusters containing more than 60% of the same class will be taken.

These data are merged together with the database of previous patients (which are annotated by skilled neurosurgeons) in a Support Vector Machines (SVM) classifier to generate the model and subsequently evaluate its performance. Due to the high computation times of SVM training, the same procedure was developed with the Random Forest (RF) algorithm, where a study was carried out to evaluate the number of trees to be used and the parameter k was redefined. With a k equal to 15 and a number of trees of 100, the data were evaluated.

Since most clusters were identified as being of the background class, it is proposed to achieve the same procedure, but using only the 3 clusters that best represent the normal tissue, hypervascularized tissue and the background class in the generation of the current patient labels. All results were evaluated with various evaluation metrics, including the kappa coefficient, which is useful both for multiclass cases and when classes are unbalanced.

Resumen

Este Trabajo Fin de Máster está relacionado con la clasificación de datos médicos hiperespectrales (HS). El objetivo principal es desarrollo de un algoritmo semi-supervisado para poder realizar el etiquetado de las nuevas imágenes hiperespectrales (HS) adquiridas, con el objetivo de incorporar estos datos al esquema de clasificación supervisada. Para la realización se utilizó una base de datos obtenida en el Hospital Universitario Doctor Negrín. Esta base de datos de imágenes HS está compuesta por 26 cubos HS pertenecientes a un total de 16 pacientes diferentes con un tumor cerebral primario de glioblastoma, donde el conjunto de prueba consta de 6 capturas correspondientes a 4 pacientes. Para realizar el etiquetado de cada una de las imágenes, se definieron 4 clases: tejido normal, tejido tumoral, tejido hipervascularizado y la clase background.

La idea principal es la de poder solventar el problema que surge en los quirófanos, donde existe una base de datos previamente etiquetada y los nuevos datos adquiridos del paciente que está siendo intervenido. El objetivo es el de lograr con este estudio incluir estos datos actuales del paciente que se encuentra en la sala de operaciones a la base de datos con la que se entrena y se genera el modelo de clasificación. Con esta propuesta se consigue generar un modelo de aprendizaje utilizando tanto los datos etiquetados obtenidos en intervenciones quirúrgicas previas como los no etiquetados del paciente en cuestión. El objetivo principal es poder mejorar los resultados de la clasificación al incluir datos del nuevo paciente.

Para realizar la generación automática de las etiquetas del paciente actual se decide utilizar el algoritmo k-means. El método elegido utiliza la distancia euclidiana por defecto, por lo que se realiza un estudio preliminar para seleccionar la distancia que mejor se adapta a nuestra base de datos. Se escogió la distancia coseno. Posteriormente, para optimizar el rendimiento del algoritmo, se realizó un estudio para seleccionar el valor del parámetro k.

Una vez seleccionados estos parámetros, los datos del paciente actual se etiquetaron automáticamente. El etiquetado se realizó de dos maneras, primero teniendo en cuenta la clase mayoritaria que conformaba cada uno de los clústeres y luego, con la condición de que sólo se utilizaran para la generación de etiquetas aquellos clústeres que contuvieran al menos un 60% de algunas de las clases.

Estos datos etiquetados junto con la base de datos de los pacientes previos (que son etiquetados por neurocirujanos expertos) son introducidos en el clasificador *Support Vector Machine* (SVM) para generar el modelo y posteriormente testearlo. Debido a los altos tiempos de cómputo, se elaboró el mismo procedimiento con el algoritmo Random Forest (RF), donde se realizó un estudio para evaluar el número de árboles a utilizar y se redefinió el parámetro k. Con una k igual a 15 y un número de árboles de 100 se evaluaron los datos.

Debido a que la mayoría de los clústeres se identificaron como de la clase background, se propuso realizar el mismo procedimiento, pero utilizando en la generación de las etiquetes del paciente actual solo los 3 clúster que mejor representen las clases tejido normal, tejido hipervascularizado y la clase background. Todos los resultados fueron evaluados con varias métricas de evaluación, incluido el coeficiente kappa, que es útil tanto para los casos multiclase como para cuando las clases están desbalanceadas.

Acknowledgements

What we are today is due to our successes and achievements but also to our mistakes. Thank you so much to those who have always been there with support and encouragement, the road has been much easier with you.

Contents

List of Figures

List of Tables

List of Acronyms

Acronym	Meaning
CAS	Celiac Artery Stenosis
CF	Computed Tomography
DSI	Integrated Systems Division
FET	Future and Emerging Technologies
GDA	Gastroduodenal Artery
GMM	Gaussian Mixture Models
HS	Hyperspectral
HSI	Hyperspectral Imaging
IGS	Stereotactic Imaging Guide
IUMA	Institute of Applied Microelectronics
MRI	Magnetic Resonance Imaging
NIR	Near-Infrared
PD	Pancreatoduodenectomy
PLS-DA	Partial Least Squares
QTH	Quartz Tungsten Halogen
REA	Executive Research Agency
RF	Random Forest
SIMCA	Soft Independent Modeling of Class Analogy
SSL	Semi-supervised Learning
SVM	Support Vector Machine
ULPGC	University of Las Palmas de Gran Canaria
VNIR	Visual and Near-Infrared

Chapter 1: Introduction

This chapter presents the context, objectives and methodology to be followed in this Master Thesis. In addition, a detailed description of the structure of the document is provided.

1.1 Context

This Master Thesis is developed within the research lines for the acquisition and processing of hyperspectral (HS) images (HSI) that is currently being carried out by the Integrated Systems Division (DSI) of the Institute of Applied Microelectronics (IUMA) of the University of Las Palmas de Gran Canaria (ULPGC), specifically in the field of medical diagnosis.

Furthermore, the IUMA has been involved in several projects funded by public and private entities, in the field of HSI processing and its application. Among these projects are these two, being the first one financed by the European Commission:

- **HELICoiD Project (CNET-IST-618080)**

HELICoiD is a European collaboration project between a total of four universities (*The University of Las Palmas de Gran Canaria, Technology and Medicine of London, Imperial College of Science, Association pour la Recherche et le Développment des Methodes et Processus Industriels de Paris and Polytechnic University of Madrid*), three industrial partners and two hospitals. This project is financed by the Executive Research Agency (REA) of the European Union and was coordinated by the IUMA.

Its main objective was to apply HSI techniques to differentiate between healthy and tumour tissue during surgical procedures. This project developed an intraoperative experimental configuration based on non-invasive HS cameras connected to a platform running a set of algorithms capable of discriminating between healthy and pathological tissues in the brain. The database provided by this project will be used in this Master Thesis.

- **ITHaCA Project (ProID2017010164)**

The ITHaCA project (IndenTificación Hiperespectral de tumores CerebrAles), is a multidisciplinary project consisted of engineers, neurosurgeons and pathologist. It has been funded by the Canarian Agency for Research, Innovation and the Information Society (ACIISI) of the Canary Islands Government and was promoted by the IUMA and

FUNCANIS (Canarian Foundation for Health Research), under the coordination of IUMA. The aim of this project was to perform a real-time classification of the brain tumour area using HSI.

1.2 Objectives

The main objective of this Master Thesis is to design a semi-supervised classification system for the detection of brain tumours using HSI. Likewise, the main objective is broken down into the following partial objectives, that must be achieved during the development of the project.

- Study of the state of the art.
- Analysis of patient samples.
- Propose and implement semi-supervised algorithms for HS data.
- Study the evaluation metrics to evaluate the classification performance.
- Determine the technique applied to the semi-supervised classification that offers the best precision of the results.

1.3 Methodology

The methodology that will be carried out in this Master Thesis to achieve the objectives consist of the following steps:

- A thorough investigation of the scenario and the tools that will be used during the development of the work to understand its operation. To this end, emphasis is placed on the basic concepts of HS and its applications.
- To investigate the different semi-supervised classification techniques to be used for his in this context.
- Research the different evaluation metrics to use. We select the ones that best adapt to the comparison we want to make.
- To analyze and evaluate the results obtained.

1.4 Document organization

The document of this Master Thesis is structured in the following chapters:

Chapter 1: **Introduction.** The objectives of the project are exposed, and a general introduction is made about the scenario and the tools used in this Master Thesis.

Chapter 2: **State-of-art.** The basic concepts necessary for the development of the project are explained. Likewise, a study of the different applications of HSI is made. Finally, a brief study of the different types of classifiers is carried out.

Chapter 3: **Hyperspectral Image Database.** This chapter exposes the procedure applied to the captures of HS images of brain tumours. As well as the database that will be used during the development of this project.

Chapter 4: **Methodology.** This chapter explains the steps implemented to carry out the design of the semi-supervised algorithm.

Chapter 5: **Experimental results.** This chapter analyses the results obtained with the semi-supervised algorithm designed in this Master Thesis.

Chapter 6: **Conclusions and future lines.** The conclusions are presented from the in-depth analysis of the results. In addition, possible future directions for this final work are considered.

Chapter 2: State-of-the-art

2.1 Introduction

In this Master Thesis, the classification of brain tumours using semi-supervised algorithm is proposed. In this section, fundamental concepts will be introduced, which are essential to understand the development of this thesis. First, it is necessary to understand what hyperspectral (HS) imaging (HSI) are and their different application fields since this type of images will be used during the development of this thesis. Secondly, different methods of classification and distances will be explained. Finally, several studies where semi-supervised algorithms are applied will be analysed.

2.2 Hyperspectral images

Hyperspectral Imaging (HSI) is also known as imaging spectroscopy. The word "imaging” stand for the representation of the appearance or morphology of the object, and the term “spectroscopy” indicates the study of the interaction of electromagnetic radiation with different materials. Therefore, HSI contains both the spatial (x, y) and the spectral (λ) information of a given object.

This technology can acquire hundreds of contiguous spectral bands, obtaining the spectral signature of any material, as shown in the Figure 2-1 where the spectral signature extracted from a brain tumour is shown. The spectral signature identifies different types of materials[1] by measuring the radiation reflected by each material at each wavelength. This thesis tries to identify between four tissues: normal tissue, tumour, blood vessels and background.

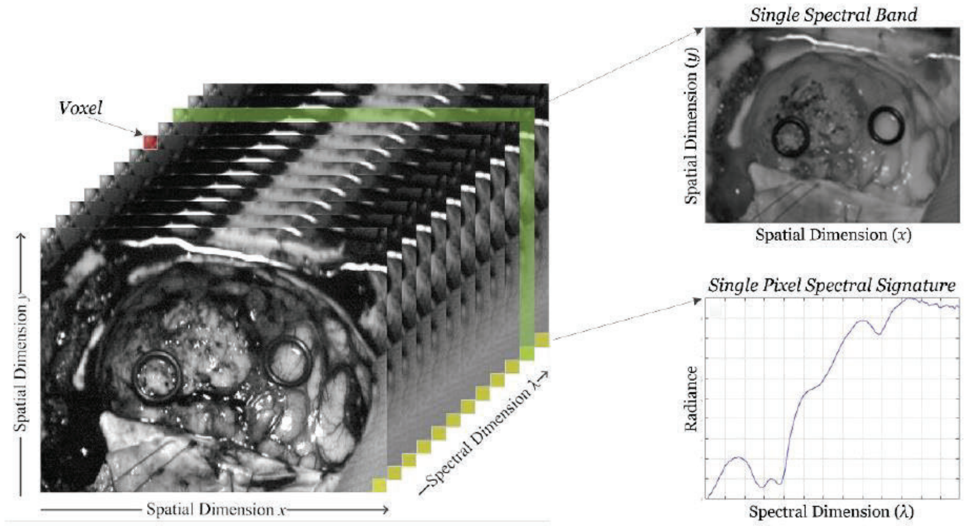


Figure 2-1. Spectral signature and hypercube of brain tumour.

All the information sampled with HSI is stored in a three-dimensional (3D) data structure called HS cube. The axis X and Y correspond to the spatial information, which indicate the position of the pixels, while λ shows the different wavelength that compose the spectral information.

HSI allows to obtain information about the observed object beyond what is possible to perceive to the naked eye. The spectrum of visible light ranges from 400 nm to 750 nm, and the eye perceives each of these wavelengths as a different colour (red, green and blue). On the contrary, HS images are able to sample hundreds of spectral bands, from the ultraviolet (UV, 200 nm) to the infrared (IR, 2500 nm) [2][3][4].

This type of technology was created for remote sensing applications. However, HSI is currently used in several applications, such as remote sensing in the field of the meteorology [5], the agriculture [6][7] or the environmental pollution[8]. It also supposes a solution for the food industry [9][10][11][12] or medical diagnosis[13][14]. In this thesis, this technique will be applied in the medical field, specifically as an aid tool for the detection of human brain tumours.

2.2.1 Medical applications

In the medical field, HS images have represented a technological breakthrough due to their non-invasive nature and because they provide useful information for the diagnosis of diseases. For example, for the assessment of oxygenation, perfusion and haemoglobin in various tissues during abdominal surgery and the identification of malignant breast tissues [15].

The work by Moulay Y. [15] presents a system that provide colorized images for surgeons during pancreatoduodenectomy (PD). These images indicate the different tissue characteristics (tissue oxygenation, organ haemoglobin index and lactate). The aim of this study is to contribute to the decision of the best surgical approach during the PD.

Table 2-1 shows liver and stomach HSI measurements in patients with celiac artery stenosis (CAS) before and after gastroduodenal artery (GDA) clamping.

Table 2-1. HSI measurements of liver and stomach.

Patient No.		Before GDA Clamping			30 min after GDA Clamping			
Nr.		StO_2 in %	OHI (0–100)	Lac in mmol/L	StO_2 in %	OHI (0–100)	Lac. in mmol/L	Additional Surgical Procedure
Pat. No. 1 Liver Stomach	Type A	63 89	18 33	0.6	79 83	42 58	0.7	none
Pat. No. 2 Liver Stomach	Type A	70 91	39 48	0.8	75 94	42 44	0.8	none
Pat. No. 3 * Liver Stomach	Type B	78 98	82 35	1.1	61 92	85 79	2.3	dissection of MAL
Pat. No. 4 Liver Stomach	Type C	67 91	74 60	1.2	59 91	59 70	1.2	dissection of MAL

Figure 2-2 shows the comparison of a normal and a color-coded image of liver oxygenation which can help the surgeon avoid ischemic complications.

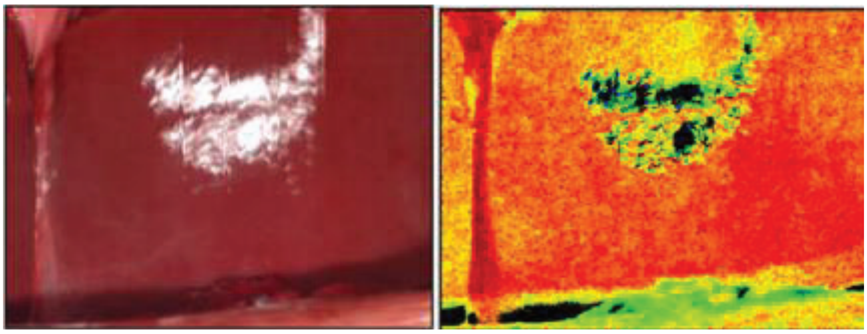


Figure 2-2. Color-coded images of the tissue oxygenation in % (Right) and color image (left).

Another study by Aref, Mohamed H. [16], presents a system capable of differentiating normal and malignant breast tissues. The HSI is used to measure the diffuse reflection (R_d) of breast samples (Figure 2-3).

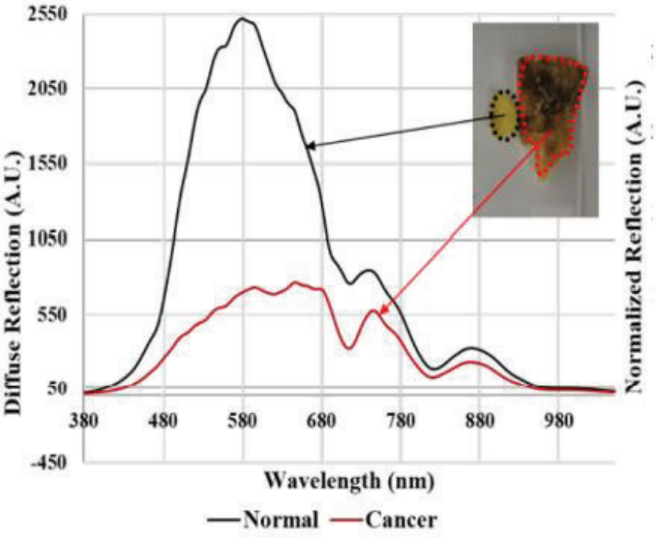


Figure 2-3. Raw data of the Rd of normal (black line) and cancer (red line) sample.

After the selection of the spectral image, the custom algorithm is applied to increase the image contrast and delineation of tumour regions as shown in Figure 2-4.

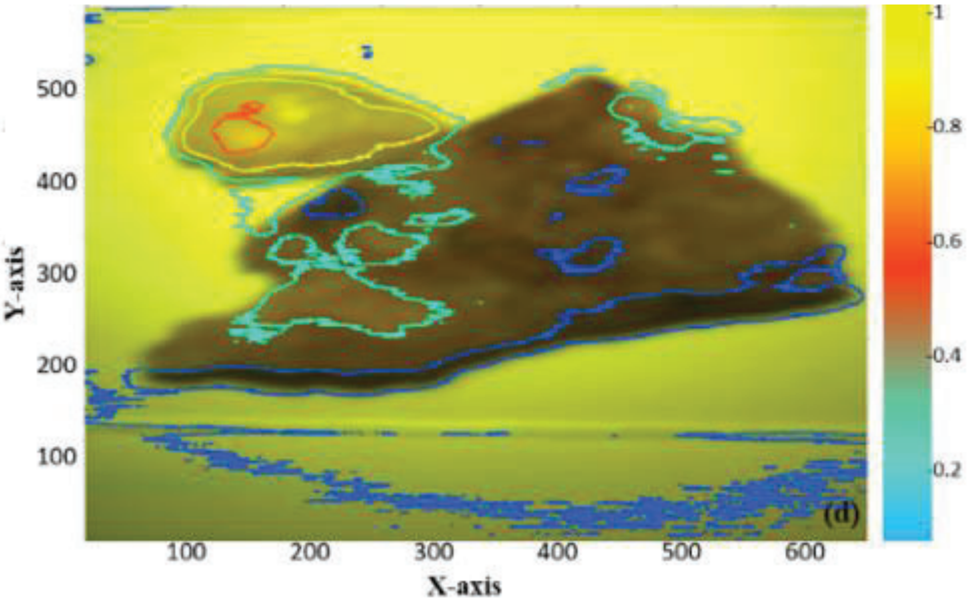


Figure 2-4. Contour delineation of tumour and normal tissue in breast sample.

The results of the samples were validated by comparing them with the pathological reports. The application of this technology in the medical field makes it possible to avoid invasive techniques such as biopsies. As a result, a more accurate diagnosis of many diseases can be achieved. The real challenge is to design algorithms capable of extracting the data in real time.

2.3 Machine Learning Algorithms

Machine Learning algorithms are used to extract information from data. Depending on the type of learning that the algorithm uses to perform the data analysis, there are three classes: supervised, semi-supervised and unsupervised algorithms.

Supervised learning algorithms employ a set of data with known labels to generate a model in order to classify new data. Respect to the unsupervised learning, the technique cluster the data in different groups applying a similarity criterium. In the case of semi-supervised learning, to generate the predictions, both labelled and unlabelled data [17] are employed together.

2.3.1 Supervised learning algorithms

Supervised learning algorithms take as input data whose labels are known to realize predictions. Thus, from the data in the training set, the algorithm can generate a model which performs predictions for a new data sample.

In a conceptual way, the supervised classification can be formulated as an optimization problem. Consider $X \epsilon \mathbb{R}_n$as the domain of the attributes and the labelled set

$G = \{ G1, G2, \dots , Gk\}$. The aim is to find a mathematic function known as a classifier that allows the mapping $h(\cdot): X \rightarrow G$, which optimizes a suitable measure of predictive capacity when applied to entities with unknown groups data [18].

Some examples of supervised learning algorithms are: Support Vector Machine (SVM) and Random Forest (RF) among others [19][20][21].

2.3.1.1 Support Vector Machines algorithm

SVM is a supervised binary classification algorithm based on the principle of structural risk minimization. A simple way to develop a binary classifier is to generate an optimal hyperplane as a decision surface, which divides the data according to each feature with a maximum margin of separation [22].

It proceeds from a separate training data set $C = \{(x_1, y_1) \dots, (x_n, y_n)\}$ which consist of an ordered sequence of data (x_i) and labels (y_i), where $x_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$ that are defined as a separating hyperplane whose lineal function is able to separate the data of the set given by the expression (2.1), where $w_i \in \mathbb{R}$ $\forall_i = 1, \dots, dyb \in \mathbb{R}$

$$D(x) = (w_1 x_1 + \cdots + w_d x_d) + b = \langle w, x \rangle + b_1 \tag{2.1}$$

The hyperplane must fulfil the inequalities (2.2) and (2.3)

$$\langle w, x_i \rangle + \quad b \geq 0 \; if \; yi{=}{+}1 \; \forall i{=}1, \dots, n \tag{2.2}$$

$$\langle w, x_i \rangle + \quad b \geq 0 \; if \; yi{=}{-}1 \;\; \forall i = 1, \dots, n \tag{2.3}$$

Where if $y_i = +1$ the class is positive and, otherwise, when $y_i = -1$ the class is negative.

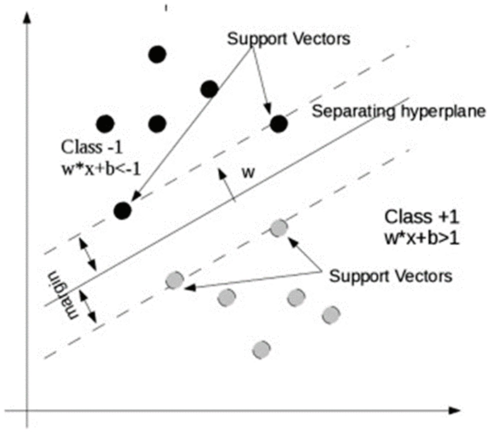


Figure 2-5. Separating Hyperplane (linear separable case).

Figure 2-5 shows an example of the SVM algorithm, which employs the support vectors of the training data to develop the decision surface. From a practical point of view, the maximum margin has proven to have a good generalization capability, avoiding the problem of overfitting to the training examples [23].

2.3.1.2 Random Forest

Within the decision trees that recursively segment the feature space and assign a label to each resulting partition there are the RF. This decision allows to classify the new data points [24].

This type of supervised algorithm generates several decision trees, each tree gives a prediction, and it is the forest that chooses the most repeated prediction.

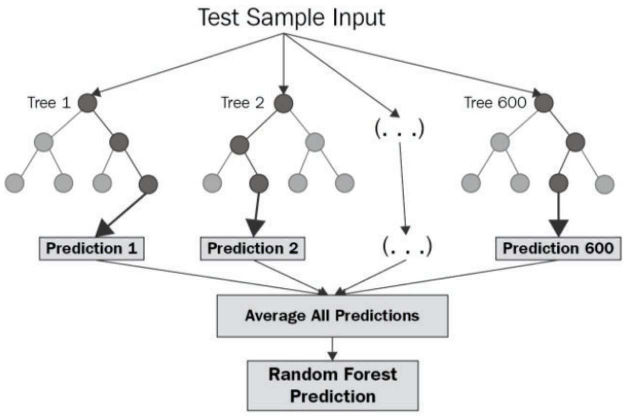


Figure 2-6. General structure of the Random Forest algorithm.

As it can be seen in Figure 2-6, the forest is the one which selects the average of the results of all the trees. RF has been proven to offer high generalization performance and high computational efficiency in HSI [25].

2.3.2 Unsupervised algorithms

Unsupervised learning is a kind of automatic learning algorithms, which the dataset is mined without the requirement of human intervention. In other words, information can be extracted from the input data even when the labels are unknown.

The most common algorithms are clustering methods, which are based on exploring the data and finding patterns or clusters. Some examples of clustering algorithms are spectral clustering, k-means and k-medoids clustering, or Gaussian mixtures models [26].

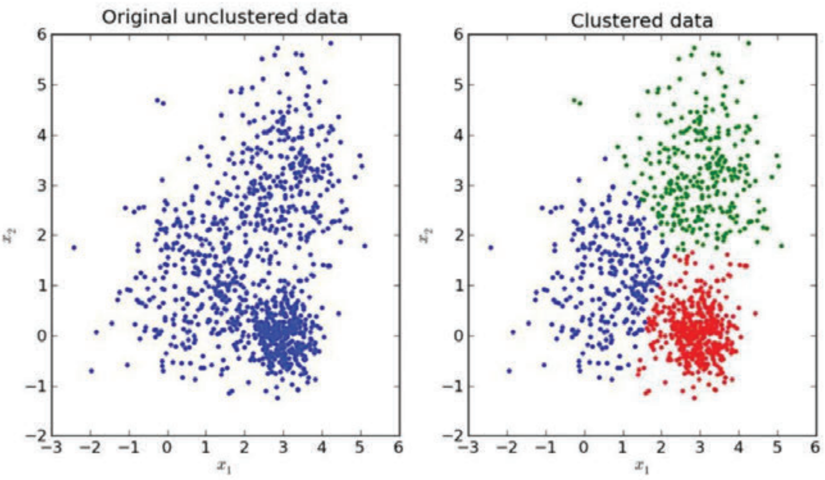


Figure 2-7. Visual example of data grouping by the clustering method

In Figure 2-7, an example of the clustering algorithm is shown. In the first image (Original unclustered data, Figure 2-7) at least two groups of points can be identified with the naked eye. The first, in the lower right quadrant and the second made up of the rest of the points. The second image (clustered data, Figure 2-7) shows the result of using a clustering algorithm, where three different data sets are identified.

2.3.2.1 K-means algorithm

The k-means method assembles the existing data into groups. It looks for the similarity levels between the different groups to have a low value. This similarity is calculated from the mean value of the group or centroid [27].

This algorithm has an input parameter k, which represents the number of clusters. K-means partitions a set of n objects into k clusters looking for high intra-cluster similarity and low inter-cluster similarity. From a given data set, it associates each point to the nearest centroid.

The K-means algorithm consists of the following steps:

1. The k centroids are initialized with random data samples from the data set.
2. A similarity metric between each sample of the dataset and the k centroids is calculated.
3. Each data sample is assigned to the cluster whose centroid is the nearest.
4. The value of the centroid is updated to the mean value of the data composing the cluster. This last stage is performed until the centroids do not move or change, elsewhere repeat from 2 [28].

2.3.3 Semi-supervised algorithms

In supervised learning, the Machine Learning model is generated from a large, labelled training set. However, in many practical classification applications the number of available unlabelled samples is larger, since the collection of labelled samples is complicated. The assignment of labels to these unlabelled data is a process that requires human effort and experience. For this reason, it is interesting to develop algorithms that can use both labelled and unlabelled samples in the classification process to obtain high-performance classifiers [29].

In this case, semi-supervised learning (SSL) algorithms are interesting to be applied when the number of labelled data is limited, and there are available unlabelled samples. SSL is a powerful tool to generate learning models when the number of labelled samples is low [30]. Most SSL approaches rely on the design of specialized learning algorithms to effectively use the data. Generally, supervised learning algorithms are the ones selected to generate the models, so SSL aims to improve the performance of the selected supervised algorithm using the available unlabelled data [31][32] [33]. In the literature, the most important types of SSL are:

- Generative models, which involve the estimation of conditional density $p(x|y)$ [34].
- Low density separation algorithms, which seek the maximum margin of the labelled and unlabelled samples simultaneously, such as inductive and transductive SVMs [35].

- Graph-based methods, where each sample scatters its labelled information to its neighbours. This is repeated until a steady state is achieved across the entire dataset [36][37].

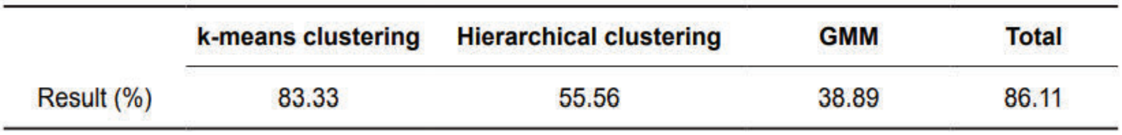
The use of HSI have increased in many application fields such remote sensing [38][39][40], military, agriculture or medical field [41][16]. However, assigning a label to each pixel is a difficult task due to the shortage of labelled samples. For this reason, it is becoming necessary to design new solutions such as the implementation of semi supervised algorithms.

2.3.3.1 Applications/Examples

The following paper focuses on a semi-supervised classification and aims to carry out a study of tomatoes using the segmentation: vine tomatoes, background, stalk and flesh. The first step in the design is the unsupervised classification of the hypercube with three different algorithms: k-means clustering, the agglomerative hierarchical and the multivariate Gaussian Mixture Models (GMM). The second step is a classification of new data in a supervised way with three different techniques: SVM, Partial Least Squares Discriminant Analysis (PLS-DA) and Soft Independent Modelling of Class Analogy (SIMCA). Finally, the last step is the evaluation phase where a supervisor should decide if the segmentation is acceptable [42].

From the results obtained from the unsupervised learning approaches, the 5 spectra of pixels with the most information were selected then, they are used to build the supervised model.

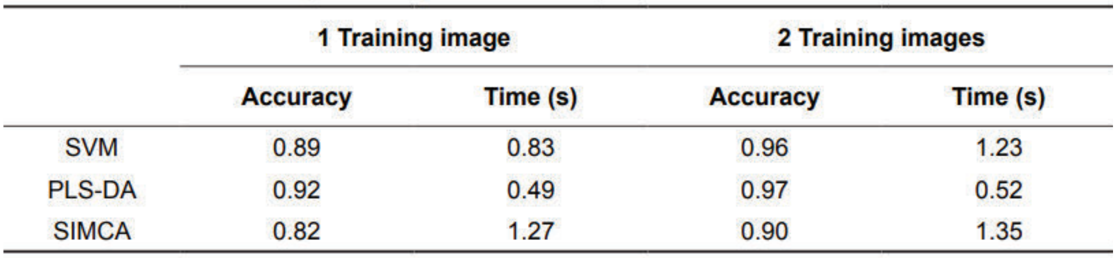
Table 2-2. Results of the three unsupervised techniques.



	k-means clustering	Hierarchical clustering	GMM	Total
Result (%)	83.33	55.56	38.89	86.11

On the one hand, the first table (Table 2-2) shows the results of the three unsupervised techniques. The k-means with the square Euclidean distance achieved the best result. On the other hand, the best supervised method was the PLS-DA with an accuracy of 97%, what is shown in the table Table 2-3.

Table 2-3. Results of the three supervised techniques.



	1 Training image		2 Training images	
	Accuracy	Time (s)	Accuracy	Time (s)
SVM	0.89	0.83	0.96	1.23
PLS-DA	0.92	0.49	0.97	0.52
SIMCA	0.82	1.27	0.90	1.35

For the former example, the semi-supervised segmentation algorithm that has been elaborated with the combination of the supervised and unsupervised techniques have achieved for this database a correct classification rate of 97%. In this case, the semi supervised approach improved the results of the supervised classification.

In the study performed by Tatyana V. Bandos, a semi-supervised graph-based method was proposed. This design is intended to produce smoother classifications. To do this, the authors exploit both the spatial and contextual information of the images through composite kernels (Spatial, spectral, stacked, summation and cross-information) [37]. They used a reduced training with 3, 5, 10 samples per class through 3-fold cross validation. The design was tested with whole image where better integration of the spatial information is achieved by the graph-based (Figure 2-8).

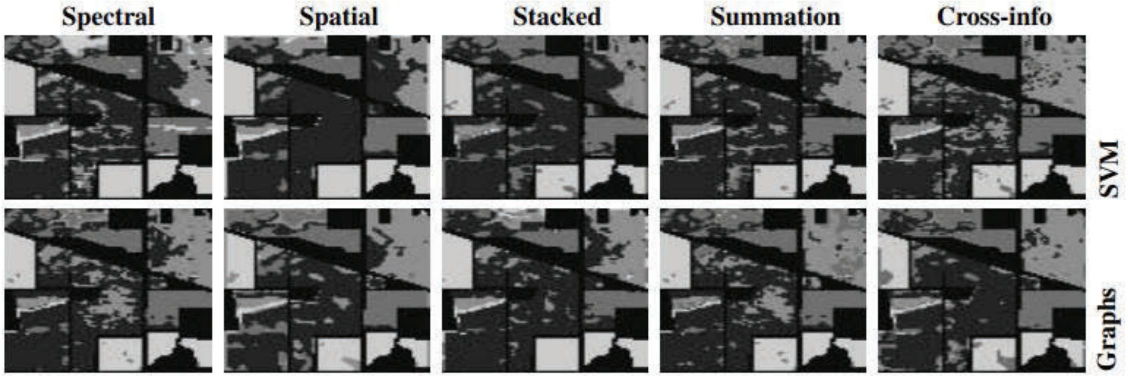
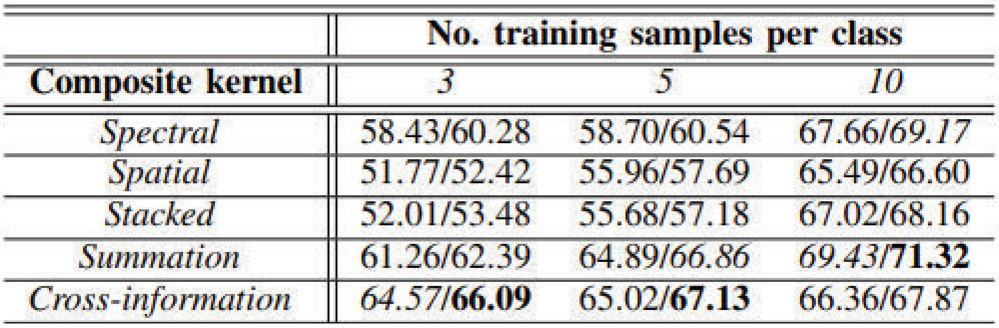


Figure 2-8. Classified images with the SVM-based and the graph-based with 5 training pixels by class.

The evaluation metric used was the hit rate. The Table 2-4, shows the overall accuracy (OA) obtained for each kernel as well as the number of labelled samples per class. The results are shown as SVM/GRAPH. As it can be observed for the graph method, a better OA value is obtained respect to the SVM method.

Table 2-4. Overall Accuracy (OA%) obtained with SVM and Graph methods.



	No. training samples per class		
Composite kernel	*3*	*5*	*10*
Spectral	58.43/60.28	58.70/60.54	67.66/*69.17*
Spatial	51.77/52.42	55.96/57.69	65.49/66.60
Stacked	52.01/53.48	55.68/57.18	67.02/68.16
Summation	61.26/62.39	64.89/*66.86*	*69.43*/**71.32**
Cross-information	*64.57*/**66.09**	65.02/**67.13**	66.36/67.87

2.4 Similarity Metrics

In this Master Thesis, different similarity metrics have been used. For this reason, this section explains different types of mathematical distance metrics that allow to calculate the similarity between data samples. The most commonly used mathematical models are: Minkowski, Euclidean, Cosine, City-block and Chebyshev distances.

2.4.1 Minkowski distance

This distance is calculated using the following formula (2.8), where d_{st} is the Minkowski distance between the vectors x_s and y_t,:

$$dst = \sqrt[p]{\sum_{j=1}^{n} |x_{sj} - y_{tj}|^p} \tag{2.8}$$

The Table 2-5 shows the special cases of the Minskowski distance [43].

Table 2-5. Special cases of Minskowski distance.

Minskowski Distance	p=1	Manhattan or City-block distance
	p=2	Euclidean distance
	p=∞	Supremum or Chebyshev distance

This distance can be considered a generalization of the Euclidean and Manhattan distances.

2.4.2 Euclidean distance

The Euclidean distance is used to calculate the distance between two points in a two-dimensional space, a smaller value indicates more similar points [44].

The Euclidean distance is defined (2.4) as follows:

$$d^2{}_{st} = (x_s - y_t)(x_s - y_t)' \tag{2.4}$$

Where d_{st} is the Euclidian distance between the vectors x_s and y_t, which is given in a data matrix. It is a special case of the Minkowski distance, but with p=2 (see section **¡Error! No se encuentra el origen de la referencia.**). Its use is recommended when the variables are homogeneous and are measured in similar units.

2.4.3 Cosine distance

The Cosine distance (2.5) calculates the angle between two vectors projected in a multidimensional space. This measure deals with the magnitude and the result is confined by the interval (-1, 1) [45].

This distance is calculated as follow:

$$d_{st} = 1 - \frac{x_s\, y_t'}{\sqrt{(x_s\, x_s')(y_t\, y_t')}} \tag{2.5}$$

It is recommended because if the vectors are similar and separated by Euclidean distance, due to the large size of the data, they may still be oriented closer together. The Cosine distance is related to the Spectral Angle similarity metric, which is commonly used in HSI applications [46][47].

2.4.4 City-block distance

The City-block or Manhattan distance calculates the exact distance between two points and not the estimate of the shortest distance[48].

This measure is represented by the following equation (2.6):

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - y_{tj}| \tag{2.6}$$

It is the Minkowski distance with p=1. In most cases this distance produces similar results to the Euclidean distance.

2.4.5 Chebyshev distance

The Chebyshev distance gives the maximum difference and is a special case of the Minkowski distance with p=∞ [49]. It is determined by the next formula (2.7):

$$dst = \ max_j\{|x_{sj} - y_{tj}|\} \tag{2.7}$$

2.5 Summary

In the medicine field, collecting labelled samples is often costly, since label assignment is a process that requires human effort and expertise, in this case, from medical experts. For this reason, it is an interesting challenge to develop algorithms that can use both labelled and unlabelled data for classification.

This chapter has presented the fundamental concepts that must be taken into account to understand the development of this thesis. It was considered necessary to understand both the composition and specifications of HSI systems. For this purpose, a brief introduction of the different fields of application was made to understand the scope of this technique. In addition, different types of algorithms and classifiers that will be used in this Master Thesis to classify the content of a HS image using SSL are presented. SSL is a promising technique in cases where the proportion of labelled data instances is small compared to the unlabelled instances.

Chapter 3: Hyperspectral image database

3.1 Introduction

This chapter describes the database used to evaluate the classification of HS with the SSL algorithm designed in this Master Thesis. Likewise, the procedure carried out to obtain the images in the HELICoiD research project is described.

HELICoiD was a European project of the Future and Emerging Technologies program (FET-Open), within the framework of the seventh Framework Program of the European Union [50]. This project applied advanced hyperspectral image classification techniques for the detection of brain tumours. The aim was to generate a demonstrator capable of discriminating between healthy and tumour tissue in real time during neurosurgery interventions. Thus, an intraoperative experimental system was developed which allows neurosurgeons to confirm the complete resection of the tumour tissue in real time from maps indicating the area of the tumour.

3.2 HSI procedure

For HSI of the surface of the human brain during neurosurgical operations, the hyperspectral pushbroom cameras selected were the Hyperspec® VNIR (Visual and Near-Infrared) Series A model and the Hyperspec® NIR (Near-Infrared)100/U model.

Figure 3-1 shows the platform installed in the preoperative area of University Hospital Doctor Negrín that was used in the HELICoiD project to acquire the images with the selected pushbroom cameras. The illumination system is a 150 W Quartz Tungsten lamp (QTH), with a broadband emission between 400 and 2200 nm, due to the great homogeneity of its spectrum that it offers throughout the spectral range.

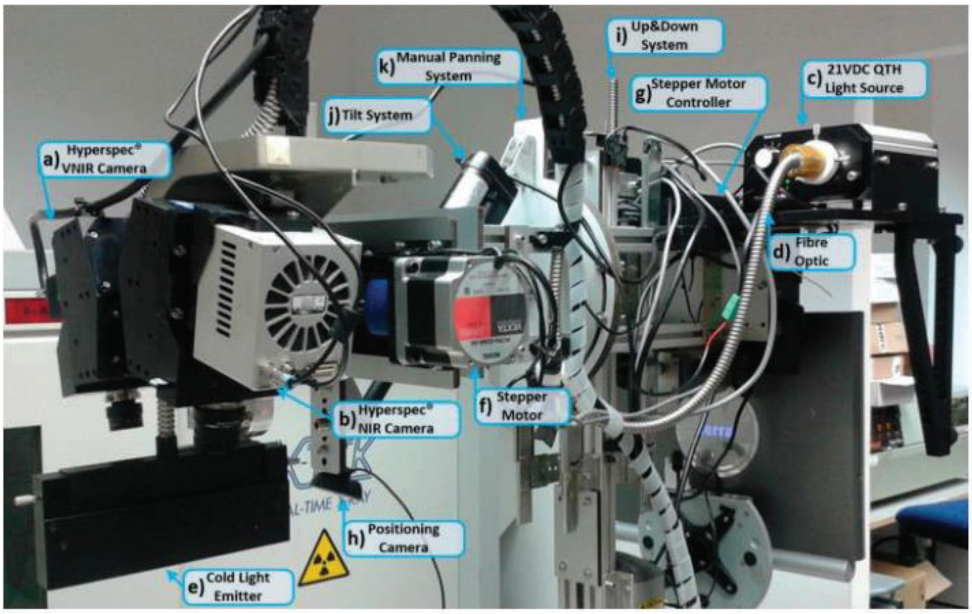


Figure 3-1. HELICoiD demonstrator acquisition platform. (a, b) VNIR and NIR HS cameras mounted on the scanning platform; (c-e) Light source QTH connected to the fiber optic system for the transmission of light to obtain a light emission on the scanning platform; (f, g) stepper motor coupled to the shaft and connected to the stepper motor controller to perform the linear movement of the cameras; (h) Positioning of the camera used to identify the position of the field of vision of the cameras (FOV); (i) The Up & Down system used to focus the HS cameras; (j) and (k) Manual pan and tilt systems used to correctly orient the scanning platform [51].

On the one hand, in the Figure 3-1.a, Hyperspec® VNIR Series A model is presented covering a spectral range from 400 to 1000 nm and is capable of capturing 826 spectral bands and 1004 spatial pixels. On the other hand, the Hyperspec® NIR 100 / U model ranges from 900 to 1700 nm, with 172 spectral channels and 320 spatial pixels is shown in the Figure 3-1 b. Both are based on the linear scan technique (line scan), a method used to obtain the hyperspectral cube. This method covers a spectral range of 400 and 1700nm (VNIR and NIR), where the most relevant spectral regions for the application of this thesis are shown.

Finally, the sensor is a two-dimensional array of detectors, a spatial dimension and a complete spectral one where the scene is captured in a single shot or frame. The technique used by these cameras offers a compromise between spectral and spatial resolution, as well as acquisition time [51].

3.3 Acquisition of HSI

The HELICoiD project has developed a demonstrator capable of simultaneously obtaining two hyperspectral cubes. Figure 3-2 shows the flowchart of the project.

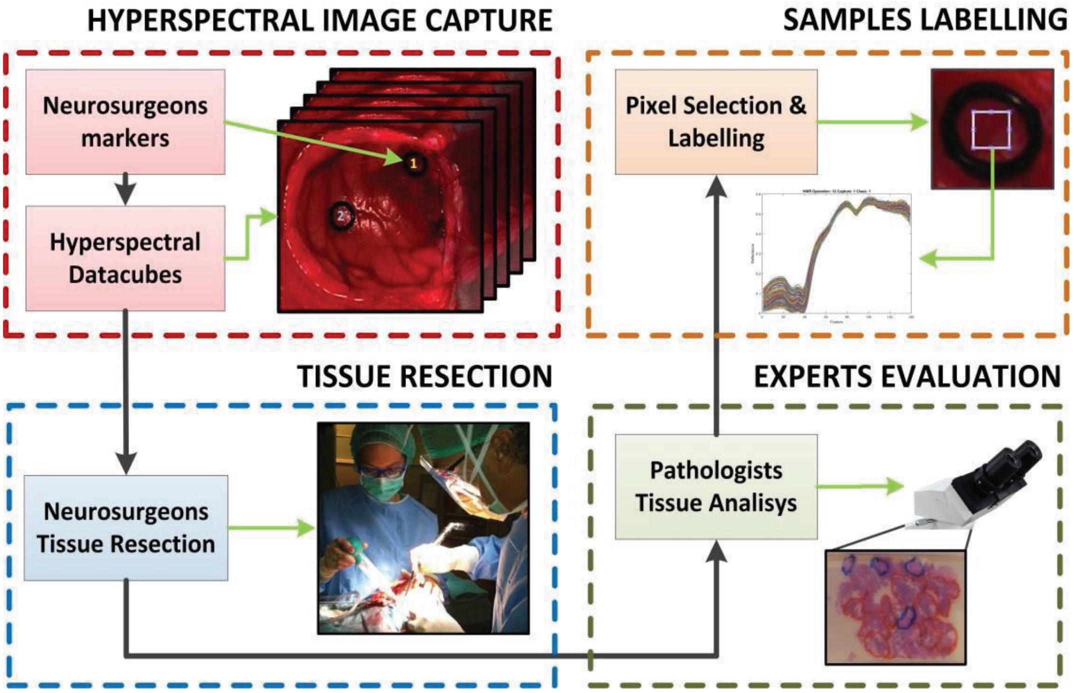

HYPERSPECTRAL IMAGE CAPTURE
Neurosurgeons markers
Hyperspectral Datacubes
1
2
SAMPLES LABELLING
Pixel Selection & Labelling
TISSUE RESECTION
Neurosurgeons Tissue Resection
EXPERTS EVALUATION
Pathologists Tissue Analisys

Figure 3-2. Data acquisition and labelling procedure of HELICoiD project.

First, prior to the operation, the patient is submitted to a stereotactic imaging guide (IGS) with compatible Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans that are loaded into the IGS system. Once the necessary tests have been performed, general anaesthesia is applied to the patient, then an incision is made in the scalp.

At that time, a craniotome is inserted and a craniotomy is performed, a part of the skull bone known as the bone flap is removed. Images are captured after durotomy and before the arachnoids and pia mater have ruptured. In case the tumour can be seen on the surface, by visual appearance, two sterilized markers are placed in the shape of a rubber ring, as seen in Figure 3-3. Thus, at the surgeon's judgment, both the tumour position and the healthy part of the brain tissue are identified [52].

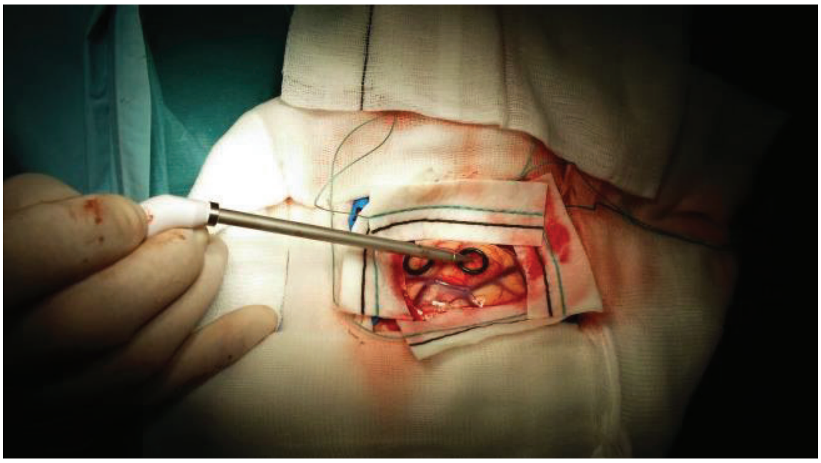


Figure 3-3. Pointer of the IGS system on the HELICoiD tumour marker located on the exposed surface of the brain.

Lastly, after using the demonstrator, several hyperspectral images were captured with and without markers. These markers provide an area of the image where the pixels can be labelled with the surgeon's prior assessment, and this is finally contrasted with the pathology results [53]. Following the use of the demonstrator, a database of hyperspectral images of the human brain in vivo is created in hyperspectral cubes.

3.3.1 Tissue resection

After the first hyperspectral image capture, the HELICoiD demonstrator is removed from the surgical site. Subsequently, the neurosurgeons begin resection of the tumour and take a sample of the tissue within the tumour marker. Tissue samples obtained from the marker position are sent to the pathology laboratory for the final tissue diagnosis. These samples will later be used as a reference for algorithm development.

3.3.2 Expert evaluation

Samples are sent to the Pathology Laboratory where they are histologically processed and subjected to standard H&E staining and any other if it is required to establish a definitive histopathological diagnosis. The only ones who can determine whether a tissue within the marker is a tumour are neuropathologists. This is performed by analysing the biopsies taken during surgery. The samples are diagnosed as tumour (subdivided into type and grade) or normal brain (white or grey matter) [54].

3.3.3 Samples labelling

From the information provided by the pathologists, and using the MATLAB tool, some pixels of each hyperspectral image were labelled (see Figure 3.4). In this way, the ground truth for training the algorithm is generated. The pixels were labelled in four classes and assigned a colour to each one: healthy tissue (class 1) represented by green colour, tumour tissue (class 2) drawn in red, hypervascularized tissue (class 3) figured in blue and the background (class 4) represented with the black color.

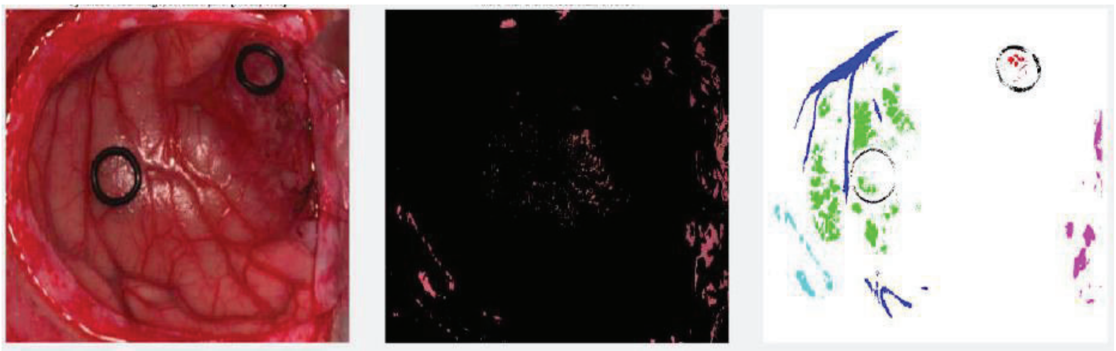
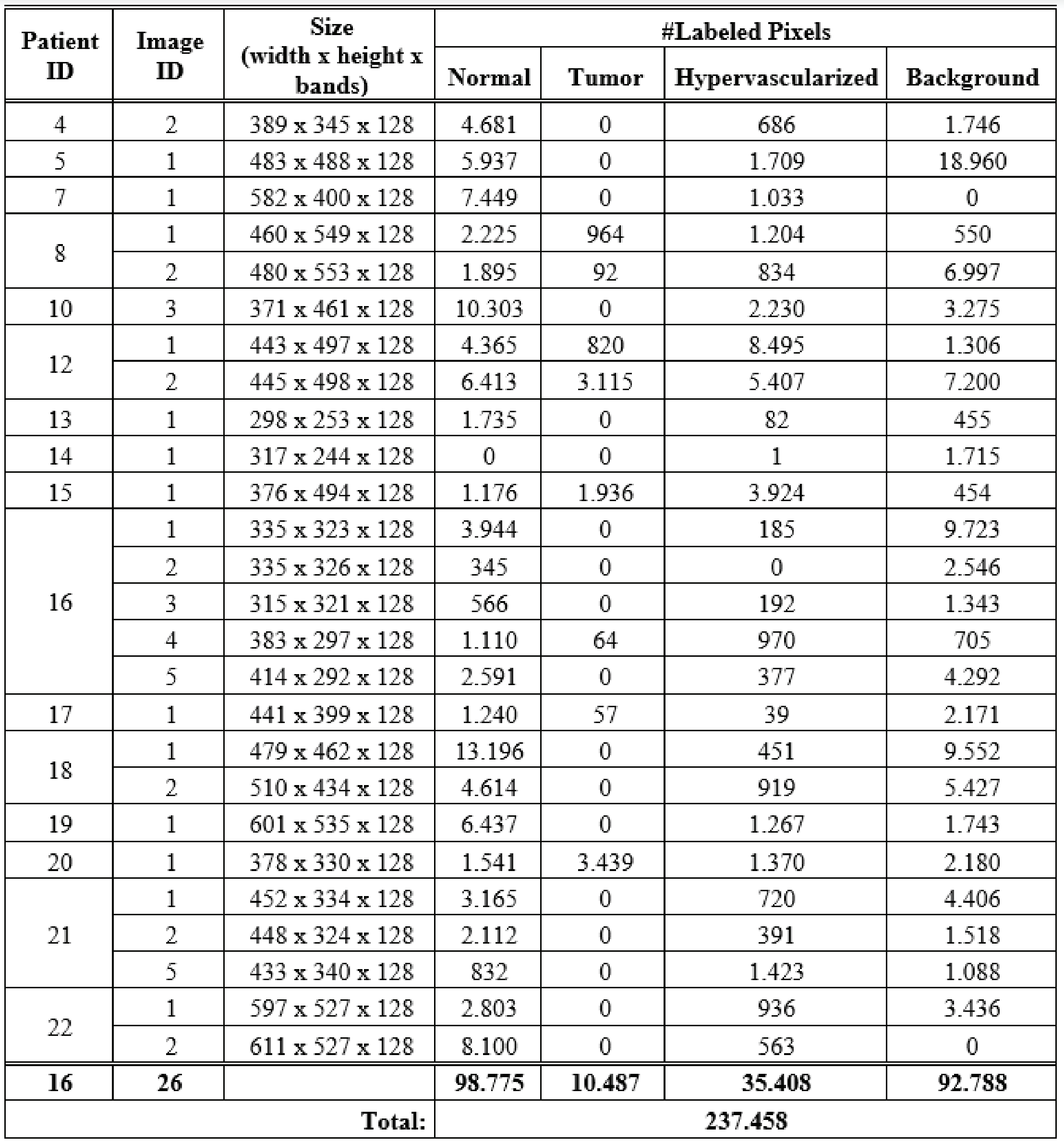


Figure 3-4. Screenshot of the HELICoiD labeling tool.

To carry out this Master Thesis a database subset of images from the HELICOiD and ITHACA projects was employed. This HS database is composed by 26 HS cubes belonging to a total of 16 different patients with Glioblastoma primary brain tumour.

In order to evaluate the design of semi-supervised algorithms for this thesis, the database is divided into two sets. One set for the training (Table 3-1) and the other for testing (Table 3-2). The selection of images for the test set was based on the images that contained the four classes with the aim of trying to predict both classes.

Table 3-1. HELICoiD labelled pixel train dataset.



Patient ID	**Image ID**	**Size (width x height x bands)**	**#Labeled Pixels**			
			Normal	**Tumor**	**Hypervascularized**	**Background**
4	2	389 x 345 x 128	4.681	0	686	1.746
5	1	483 x 488 x 128	5.937	0	1.709	18.960
7	1	582 x 400 x 128	7.449	0	1.033	0
8	1	460 x 549 x 128	2.225	964	1.204	550
	2	480 x 553 x 128	1.895	92	834	6.997
10	3	371 x 461 x 128	10.303	0	2.230	3.275
12	1	443 x 497 x 128	4.365	820	8.495	1.306
	2	445 x 498 x 128	6.413	3.115	5.407	7.200
13	1	298 x 253 x 128	1.735	0	82	455
14	1	317 x 244 x 128	0	0	1	1.715
15	1	376 x 494 x 128	1.176	1.936	3.924	454
16	1	335 x 323 x 128	3.944	0	185	9.723
	2	335 x 326 x 128	345	0	0	2.546
	3	315 x 321 x 128	566	0	192	1.343
	4	383 x 297 x 128	1.110	64	970	705
	5	414 x 292 x 128	2.591	0	377	4.292
17	1	441 x 399 x 128	1.240	57	39	2.171
18	1	479 x 462 x 128	13.196	0	451	9.552
	2	510 x 434 x 128	4.614	0	919	5.427
19	1	601 x 535 x 128	6.437	0	1.267	1.743
20	1	378 x 330 x 128	1.541	3.439	1.370	2.180
21	1	452 x 334 x 128	3.165	0	720	4.406
	2	448 x 324 x 128	2.112	0	391	1.518
	5	433 x 340 x 128	832	0	1.423	1.088
22	1	597 x 527 x 128	2.803	0	936	3.436
	2	611 x 527 x 128	8.100	0	563	0
16	**26**		**98.775**	**10.487**	**35.408**	**92.788**
		Total:	**237.458**			

Table 3-2. HELICoiD labelled pixel test dataset.

Patient ID	Image ID	Size (width x height x bands)	#Labeled Pixels			
			Normal	Tumor	Hypervascularized	Background
8	1	460 x 549 x 128	2.225	964	1.204	550
	2	480 x 553 x 128	1.895	92	834	6.997
12	1	443 x 497 x 128	4.365	820	8.495	1.306
	2	445 x 498 x 128	6.413	3.115	5.407	7.200
15	1	376 x 494 x 128	1.176	1.936	3.924	454
20	1	378 x 330 x 128	1.541	3.439	1.370	2.180
4	**6**		**17.615**	**10.366**	**21.234**	**18.687**
Total:			**67.902**			

Since the leave-one-out methodology has been used during the development of this thesis, the remaining HS test images are also included in the training database (see Chapter 4: 4.2.1 Leave-one-out Cross-Validation technique). Table 3-3 and Table 3-4 illustrate the RGB images and the gold reference of the patients who were included in the test set.

Table 3-3. RGB of hyperspectral test set.

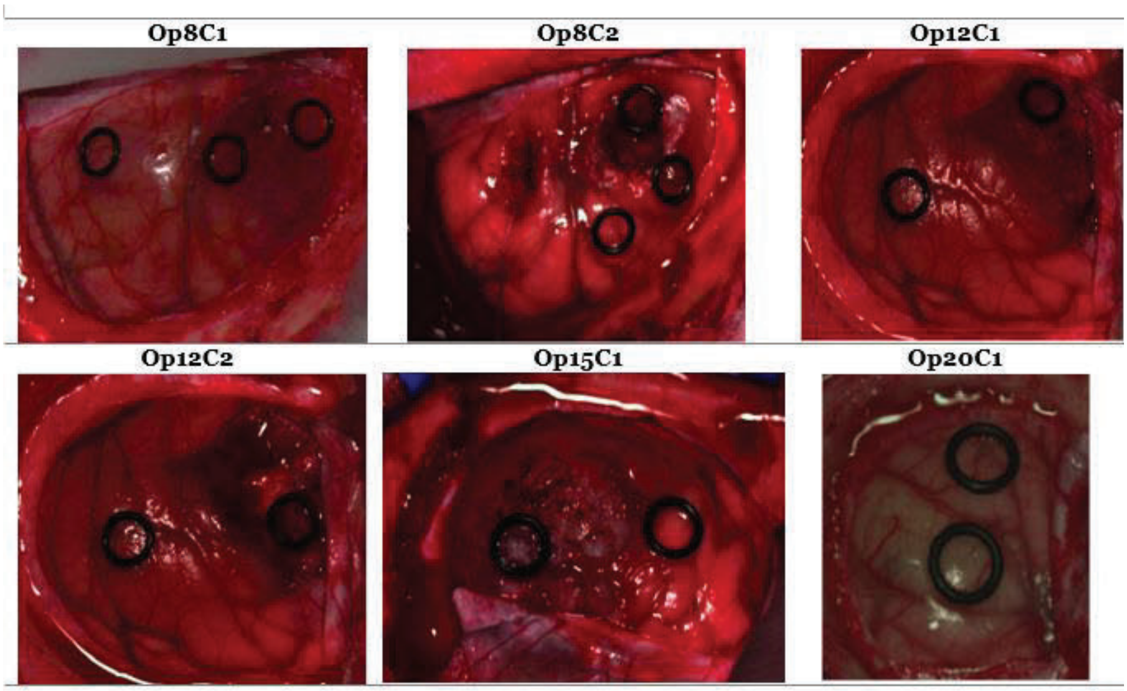
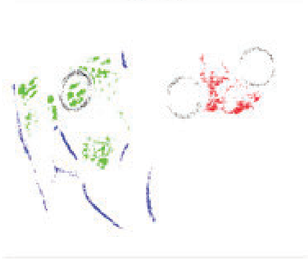
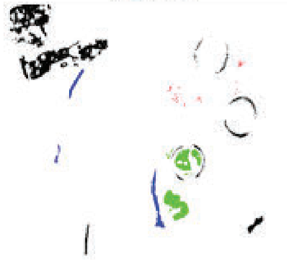
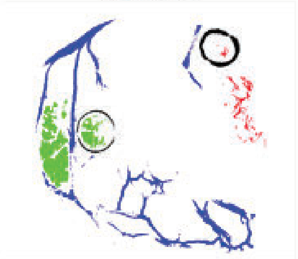
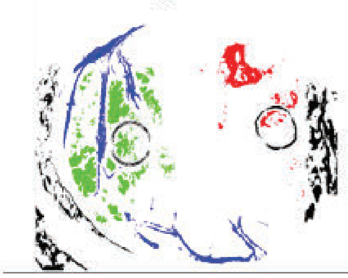
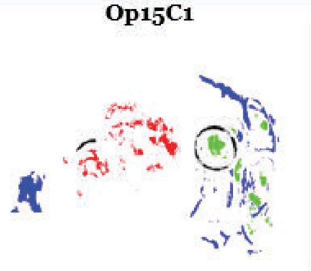
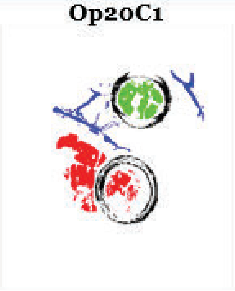


Table 3-4. Ground truth maps of hyperspectral test set.

Op8C1	Op8C2	Op12C1
		
Op12C2	**Op15C1**	**Op20C1**
		

3.4 Summary

This chapter has presented the procedure for the capture of hyperspectral images of brain tissues. It is noted that, in order to obtain these images, the neurosurgeon must follow a strict procedure to extract the images that are part of the database of this thesis. For this reason, it is necessary to perform a craniotomy and extract some tissues (fibrous covers) to reach the brain tissue that will later be indicated with the markers. Finally, the samples were labelled obtaining a total of 26 HS cubes from 16 patients. The dataset was divided into two sets: the training set composed by 16 patients and 26 HS cubes (containing a total of 237.458 labelled pixels) and the test set (consisting of 6 captures from 4 patients). This test set have a total of 67.902 labelled pixels of which 10.366 correspond to tumour samples.

Chapter 4: Methodology

4.1 Introduction

In this chapter, the methodology followed to perform a semi-supervised classification of our database is explained.

The objective of this Master Thesis is to design a semi-supervised algorithm. It is decided to make a pre-labelling using the data set of previous patients to label the current patient, with the goal of using the data for the current patient for the classification. The idea is to simulate real cases in the operating room, where there is a previously labelled database and the new acquired data of the patient who is going to receive the intervention. The objective is to include this current patient data in the database with which to train and generate the SVM and RF models.

First, the k-means method will be used to generate the labels of the current patient. The chosen method uses the default Euclidean distance, so a preliminary study is carried out to select the distance that better suits our database. The study has been performed for the following distances: Euclidean, Cosine, City-block, Minkowski and Chebyshev. The sample label generation decision is made in two ways. 1) By voting, this method counts the label value given by each distance. For example, if three distances identified that pixel belonged to the normal tissue class and two distances identified it as background, the pixel will be labelled as normal tissue, label with value 1. 2) By the best result, which selects the distance that best fits the database. In this case, the label is assigned to the pixel based on the distance that obtained the smallest value. After evaluating the results, the best selection method is chosen to determine the distance that best suits our database.

Once the most suitable distance has been selected, based on the results of the k-means method, the data are processed to find out to which class each cluster belongs. First, it is decided which clusters are assigned to a class based on the majority. This decision is initially made by the percentage of pixels which belong to the same class (the cluster is assigned the class that has the highest number of labelled pixels for a certain class). However, in this approach it may happen that there are clusters in which there is a high number of pixels of various classes. This could lead to errors in the final label assignment. In order to avoid this issue, a second labelling generation approach is proposed. In this second approach, only those clusters which contain more than 60% of the same class will be taken (the cluster is assigned a label if one class has a hit greater than 60%). From there, the current patient data is labelled taking into account the minimum distance between the pixels and the closest centroid.

Finally, these semi-automatic labels (unsupervised assignment of labels to the current patient data) are merged with the training dataset (corresponding to previous patient data) directly to the classifier to generate the model. The results were assessed with several evaluation metrics, including the kappa metric which is a useful coefficient for multiclass cases and when classes are unbalanced.

4.2 Proposed methodology

The methodology proposed in this Master Thesis to develop the semi-supervised classification of HS images of brain tumours is as follows: It starts from a database that consist of pre-processed and previously labelled hyperspectral images. With this database, the labelling of a new patient is performed by using the distance of each pixel with respect to the mean of the complete database without such patient. Once the distance metric selection is done, an evaluation is made of which distance metric best fits our database and then use this parameter in the k-means method. The database of the previous patients without the new patients used in the k-means to get the different clusters. Once it is known which cluster belongs to which class, the labels of the new patient are generated (Figure 4-1).

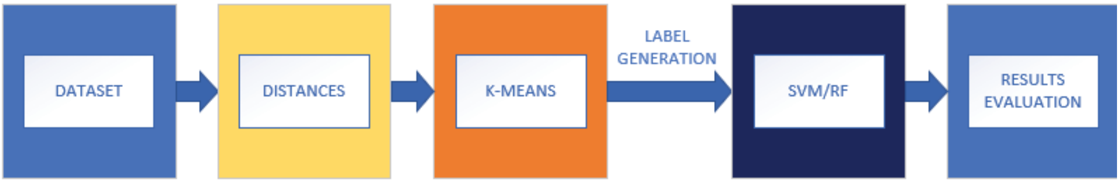


Figure 4-1. Block diagram corresponding to the proposed procedure.

The new patient labelled with this methodology and the dataset of the previous patients are fed into the classifiers, in order to train it, generate a model and finally evaluate its performance. For this purpose, the following evaluation metrics have been employed: accuracy, which is the success rate or precision, the confusion matrix, specificity, sensitivity and, finally, the kappa coefficient. To obtain these metrics, a data partition based on the leave-one-out cross-validation technique (explained in the next section) are used.

In the classification stage, it was decided to use the SVM algorithm. After conducting an exhaustive analysis on different studies from the literature. Being this algorithm the one that has obtained the best results in the cases in which high-dimensional data classification is used and the training samples are limited [55][56]. This same procedure is subsequently proposed using the RF algorithm, with the aim of reducing computation times and being able to compare the results of the proposed procedure (Figure 4-1).

The procedure followed consists of using the spectral signatures to train the classifier. Likewise, to separate the training and test data and to provide validity to the evaluation metrics, the leave-one-out cross-validation technique is used.

4.2.1 Leave-one-out Cross-Validation technique

The cross-validation technique allows estimating the precision of the generated model. For this, a partition of the data is performed where, on the one hand, there will be a training set and, on the other hand, a group of test data to assess the model performance [57].

In the cross-validation process there are different methodologies: k-fold and leave-one-out.

The k-fold cross-validation is an improvement to the hold-out method. In the hold-out data partition strategy, the data is divided into two subsets, a training part and a test part. While in the k-fold method the data are divided into k subsets. So that, this process repeats a set of k iterations [58]. Both methods use the same principle of operation, testing with data that have not been used during the training so in each subset some samples are part of the training and others of the test. Finally, all samples are part of the test set.

In this study, a leave-one-out data partition is implemented. The leave-one-out is a cross-validation on n, where n is the number of instances in the data set. Each instance of n is left out and the classifier is trained on all other instances (see Figure 4-2) [59].

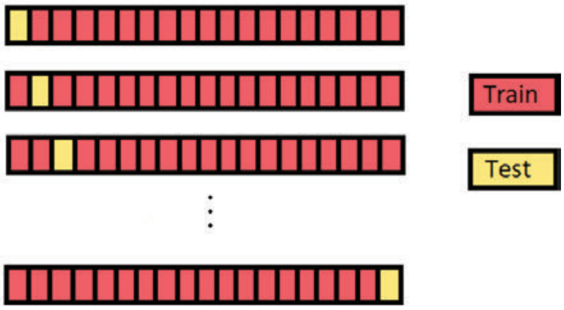


Figure 4-2. Leave-one-out (Cross-validation method).

The n results are averaged, with this average being the representation of the final error estimate. In addition, the randomly generated subsets for each of the methods are made to contain approximately the same proportion of labels as the original data set. In our case, each patient is used as test data (emulating a new patient arriving to the surgical room), and the data from the remaining patients is used to train the supervised classifier.

4.2.2 First part of the proposed SSL method

The first part of the proposed processing framework aims to evaluate which distance is best suited to the database employed to optimize the k-means method. For this, the following steps were followed: first it is necessary to calculate the class means of the previous set of patients without the current patient to evaluate (performing a leave one out cross-validation). Then, once the mean signatures have been calculated, pairwise distances are calculated between each of the mean classes and the mean signatures of each class. Finally, a comparison is made between the dataset corresponding to the current patient and the mean signatures obtained for the current patient pixels with the highest similarity to the mean signatures corresponding to each class. The objective is to identify which mean signature (4 classes, 4 means) has the greatest similarity with the pixel to be labelled. This is how the labels are assigned.

To automatically perform the new labelling of the surgical samples of the current patient, two methods are initially proposed, by voting, i.e. how many distances claim that this pixel belongs to the same class? and by the best result obtained (Show Figure 4-3).

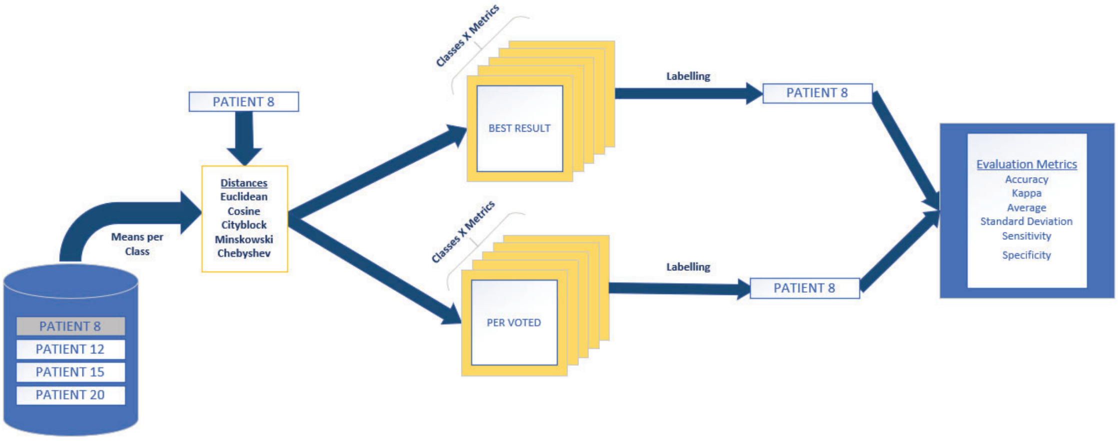


Figure 4-3. Flow diagram of the study of distances.

A total of five distances were evaluated: Euclidean, Cosine, Cityblock, Minkowski and Chebyshev distance. Once the results were obtained, the voting method was discarded due to the autocorrelation between distances, so for this study the best result method was used to generate the new labels of the current patient and evaluate the different distances. The evaluation metrics are applied when comparing the result with the reference image, where the value of the labels was already known.

4.2.3 Second part of the proposed method

Once the distance is selected, an attempt is made to optimize the k-means algorithm a little more. A study is made to select the best value of k and the index of the MaxIter parameter, which is the maximum number of iterations. This last parameter has a default value of 100. Three different values 100, 1000 and 3000 are tested and the value with the best result is selected.

The next step is to use the k-means algorithm in this process to find and cluster the data in k groups by evaluating a similarity metric between samples.

The k-means algorithm needs the input parameter k, with which it divides the ensemble samples into k clusters. This method tries to find that the similarity level between the members of a cluster is high and with the samples of other clusters very low. The similarity of the cluster to the members is measured by the proximity of the object to the mean value of the cluster or centroid [27].

For this case, the similarity seeks to measure the distance between the hyperspectral signatures of each pixel. The goal is to find the signatures that are most similar to each other and group them together. The k-means uses the default Euclidean distance to perform this calculation.

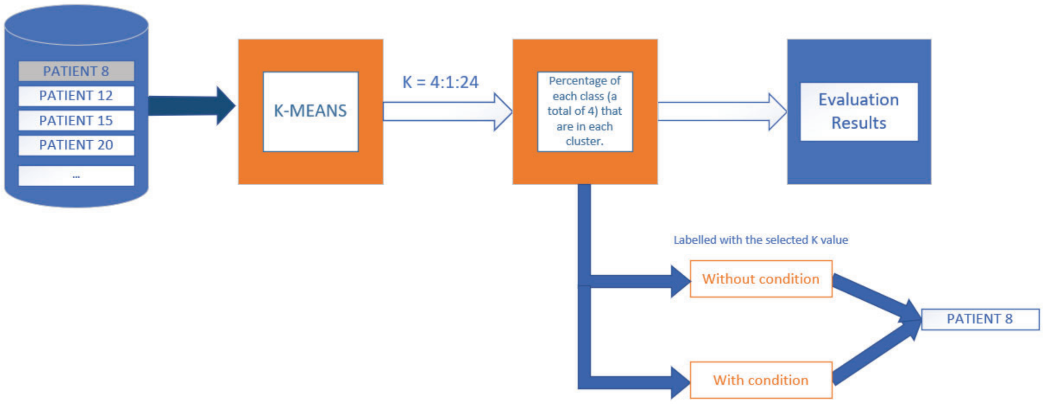


Figure 4-4. Flow diagram of k-means method.

Once the distance that best suits our database is selected, (see Section 4.2.2), the following steps are illustrated in Figure 4-4. In order to identify the best value for the parameter *k* a sweep is performed with a range value from 4 to 24, which represents the number of groups into which the set of observations will be divided [60].

After studying the data, it was analysed how the k-means grouped the data. As the k value is increased, it was possible to separate the classes more efficiently, however, it is possible that one cluster has pixels from different classes, for that reason, it is decided to develop two conditions for the final classification.

The next step is to identify to which class each cluster belongs. As it is commented before, two mechanisms are proposed for this decision. The first mechanism is the unconditional one. It consists of assign the cluster to the majority class, i.e., if in cluster 1 (C1) 61% of the pixels belong to class 1 (Normal), 1% to class 2 (Tumour), 7% to 3 (Blood vessel) and 32% to 4 (Background), then that cluster C1 belongs to class 1. The second proposed mechanism consist in using a condition, i.e., if the class with the highest percentage of that cluster does not exceed 60%, its means that this cluster has a lot of data variability, so it is discarded for the final labelling.

4.2.4 Third part of the proposed method

To evaluate the obtained results, the SVM and RF classifiers were used. The final block of the general process is shown in more detail in Figure 4-5. As it is a semi-supervised algorithm, what is sought is to increase the database with which the model is generated.

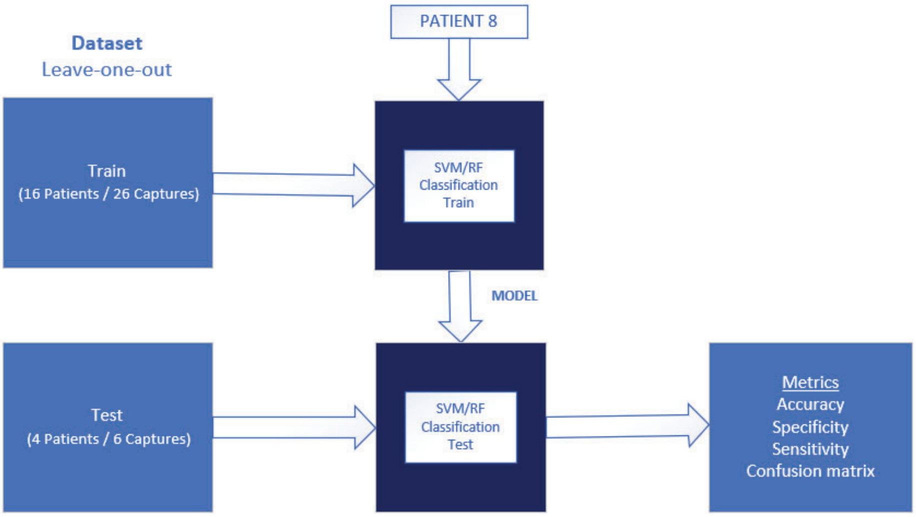


Figure 4-5. Final flow diagram of the evaluation metrics computation with the implementation of SVM and RF algorithms.

The Train set together with the current patient labelled are entered into the SVM and RF algorithm. With these data, the SVM is trained, and the RF algorithm generates the decision trees. Once the generated models are obtained, the Test set is used to evaluate its performance with the different evaluation metrics.

4.3 Evaluation metrics

To validate the results, it is advisable to contrast with at least more than one metric, since using only one metric could not be enough to describe the goodness of the results of a classifier and could lead to wrong conclusions. This section describes the evaluation metrics used to assess classification accuracy.

4.3.1 Overall Accuracy

Overall Accuracy (OA) refers to the hit rate and determines the accuracy with which the classifier is able to correctly predict the classes of the pixels samples. It indicates how many pixels could be correctly identified by the classification algorithm. This metric is calculated as the success rate of the predictions of the classifier and is defined by the equation (4.1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Where TP (True Positives), corresponds to the correctly detected conditions. In other words, the sample label is positive, and the result of the classification is positive. False Positives (FP) are the incorrectly detected conditions. The sample label is negative, but the result of the classification is positive. True Negatives (TN) correctly rejected conditions. The sample label is negative, and the result of the classification is negative and the False Negatives (FN) the incorrectly rejected conditions. The sample label is positive, but the result of the classification is negative [61].

4.3.2 Specificity

Specificity is the proportion of true negatives which the classifier identifies as such. It is calculated by the expression (4.2):

$$Specificity = \frac{TN}{TN + FP} \tag{4.2}$$

4.3.3 Sensitivity

The sensitivity corresponds to the proportion of true positives that have been correctly identified such as positives. The equation to calculate this percentage is shown in (4.3).

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.1}$$

4.3.4 Kappa coefficient

The Kappa Coefficient (k) determines the interobserver agreement, it can be calculated on tables of any dimension. This coefficient is constructed from a quotient shown in the following equation (4.4) [62].

$$k = \frac{[(\sum observer\ concordances) - (\sum random\ concordances)]}{[(Total\ observations)\ - (\sum random\ concordances)]} \tag{4.4}$$

The Kappa coefficient is, in other words, the ratio between observed and non-random agreement divided by the total possible concordance not produced by chance.

The range of values that the kappa coefficient is between +1 and -1. A positive kappa indicates that the observers agree more frequently than would be randomly expected, meaning that there is a strongest degree of interobserver concordance. If the value is k=1 it indicates a completed agreement. Conversely, if k=0, it denotes that the concordance is as expected by chance. A negative kappa indicates that disagreement between observers is more frequent than expected by chance. Finally, if k=-1, it indicates a totally disagreement [63].

4.4 Summary

This chapter has described the methodology followed throughout this Master Thesis to design a semi supervised algorithm and evaluate the results.

First, to optimize the k-means method, a study of the distances best suited to our database was carried out. Then, a study was conducted to select the parameter k for the k-means clustering. Once the parameters of the k-means algorithm were defined, the train database was introduced into the k-means method with the objective of dividing the samples in k clusters. The class presence in each cluster was calculated to generate the current patient labels. For this purpose, two methods were proposed. The first method is the unconditional one, where the cluster gets the label value of the class with the highest percentage. The second method, consist in using a condition, i.e., if the class with the highest percentage of that cluster does not exceed 60% its means that this cluster has a lot of data variability, so it is discarded for the final labelling. Once the clusters have been identified, the algorithm was trained, and the models were generated for both approaches to label the current patient data. Finally, both ways of automatic label generation were evaluated.

Chapter 5: Experimental results

5.1 Introduction

This section shows the results obtained for the proposed semi supervised algorithms, with the aim of evaluating and defining which proposed method performs best for the classification of brain HS data. This chapter will be divided as follows:

1. Selection and evaluation of the results obtained with each of the distances.
2. Choice of the value of the parameter *k* to optimize the unsupervised k means algorithm.
3. Use of the unsupervised algorithm for the generation of new patient labels and evaluation of the supervised classification models generated with the SVM and RF algorithms.

First, the different distances are analysed to observe which one best fits the HS brain database and use it for the proposed semi-supervised classification. Once the distance metric is chosen, the value of the parameter *k* is selected to optimize the k-means algorithm. After selecting the k-means parameters, the labels are generated for the current patient with and without conditions. These data and annotations from the current patient are included to the database of previous patients and are introduced into the SVM supervised classifier, where the model is trained and generated for further evaluation. To be able to make a broader comparison, the study is performed again by changing the SVM algorithm for the RF allowing to analyse different values of *k*.

For the analysis and selection of the best distance to use, the following evaluation metrics were used: OA (Overall accuracy), kappa coefficient, standard deviation, sensitivity, specificity. While for the evaluation of the final results the selected metrics were: accuracy, specificity, sensitivity and confusion matrix.

5.2 Distance selection and evaluation.

In this section the results obtained for Euclidean, Cosine, Cityblock, Minkowski and Chebyshev distances are presented. To perform this evaluation, the current patient labels were generated in two ways: Voting and Best Result methods. Once the labels have been generating according to these similarity measurements, the different evaluation metrics are calculated.

5.2.1 Voting method results

The voting selection process consists of assigning to each pixel of the HS image the class that has appeared most often among the distances. That is, if three distances identify a pixel as the background class and two distances identify the same pixel as the healthy class, then, being the majority, the label assignment for that pixel will be the background class.

Table 5-1 shows the overall accuracy, sensitivity, specificity, and kappa coefficient obtained after classifying the pixels by the voting method. The results show that OA has a value of 52.88%. As explained in Chapter 4, there is an autocorrelation. For example, the City-block distance is a special case of the Minkowski but with p=1 (see 2.4.1 Minkowski distance). For this reason, the results obtained by this labelling method are discarded.

Table 5-1. Results obtained with the voting method.

		Sensitivity				Specificity				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
P8C1	28.10%	0.06	0.24	0.36	1.00	0.90	0.42	0.95	0.32	0.12
P8C2	67.69%	1.00	0.31	0.94	0.55	0.61	0.97	1.00	1.00	0.48
P12C1	73.66%	0.47	0.03	0.89	1.00	0.91	0.85	0.81	0.98	0.58
P12C2	43,.41%	0.59	0.01	0.92	0.13	0.46	0.81	0.63	0.96	0.24
P15C1	72.14%	1.00	0.81	0.55	0.98	0.69	1.00	0.99	0.98	0.62
P20C1	32.29%	0.96	0.00	0.61	0.16	0.24	0.58	0.98	0.98	0.11
AVG	**52.88%**	**68.03%**	**23.27%**	**71.26%**	**63.73%**	**63.47%**	**77.01%**	**89.23%**	**86.81%**	**0.36**
Std	**±0.18**	**±0.32**	**±0.26**	**±0.20**	**±0.35**	**±0.22**	**±0.19**	**±0.12**	**±0.23**	**±0.19**

5.2.2 Best result method results

In this section, the samples are labelled, making the decision based on the best result obtained with each of the distances. For this method, like the previous one, several evaluation metrics are used, highlighting the kappa coefficient widely used for multiclass and unbalanced cases.

The OA obtained for the Euclidean distance (Table 5-2) is 51.91% and a kappa of 0.35. For the Cosine distance (Table 5-3), the OA is 65.98% and the kappa has a moderate value of 0.51. The Chebyshev distance (Table 5-4) obtained an OA of 50.28% and a kappa coefficient of 0.34 while the City-block distance (

Table 5-5) manages to improve a bit on this value with a success rate of 53.84% and a 0.37 kappa. Finally, the Minkowski distance (Table 5-6) obtained an OA of 51.91%, however, the kappa coefficient decreased with respect to the City-block distance, with a value of 0.35. As can be observed, the Minkowski distance provides the same results as the Euclidean distance do. The reason may be due to the fact that the Euclidean distance is a special case of the Minkowski with p = 2 as discussed in Section 2.4.1. It can be observed that the Cosine distance is the one with the highest hit rate and the best kappa with a moderate value of 0.51, being the selected distance method. This indicates that the disagreement between the observations is less frequent than is expected.

Table 5-2. Euclidean distance results.

		Sensitivity				**Specificity**				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
P8C1	27.15%	0.06	0.24	0.32	1.00	0.90	0.40	0.95	0.30	0.11
P8C2	67.29%	1.00	0.31	0.93	0.55	0.61	0.96	1.00	1.00	0.48
P12C1	71.65%	0.41	0.03	0.89	1.00	0.91	0.82	0.80	0.98	0.55
P12C2	42.47%	0.56	0.00	0.92	0.13	0.46	0.80	0.62	0.95	0.23
P15C1	70.64%	1.00	0.80	0.52	0.98	0.67	1.00	0.98	0.97	0.60
P20C1	32.27%	0.96	0.00	0.61	0.15	0.24	0.58	0.99	0.98	0.11
AVG	**51.91%**	**66.35%**	**23.18%**	**69.87%**	**63.67%**	**63.01%**	**76.08%**	**88.97%**	**86.39%**	**0.35**
Std	**±0.19**	**±0.36**	**±0.28**	**±0.23**	**±0.38**	**±0.23**	**±0.21**	**±0.14**	**±0.25**	**±0.20**

Table 5-3. Cosine distance results.

		Sensitivity				**Specificity**				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
P8C1	54.87%	0.48	0.52	0.52	0.91	0.90	0.61	0.80	0.93	0.39
P8C2	87.58%	0.97	0.25	0.99	0.84	0.97	0.91	1.00	0.97	0.77
P12C1	77.64%	0.85	0.02	0.93	0.16	0.92	0.86	0.85	1.00	0.63
P12C2	58.44%	0.93	0.01	0.95	0.28	0.60	0.90	0.74	1.00	0.44
P15C1	78.97%	0.86	0.92	0.75	0.51	0.83	0.93	0.98	0.96	0.69
P20C1	38.40%	0.97	0.00	0.75	0.30	0.83	0.67	0.99	0.44	0.17
AVG	**65.98%**	**84.43%**	**28.67%**	**81.45%**	**50.07%**	**84.18%**	**81.25%**	**89.30%**	**88.31%**	**0.51**
Std	**±0.17**	**±0.17**	**±0.34**	**±0.16**	**±0.29**	**±0.12**	**±0.12**	**±0.10**	**±0.20**	**±0.20**

Table 5-4. Chebyshev distance results.

		Sensitivity				**Specificity**				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
P8C1	23.46%	0.03	0.01	0.43	1.00	0.96	0.57	0.48	0.21	0.08
P8C2	66.52%	0.98	0.41	0.91	0.55	0.60	0.97	1.00	0.99	0.47
P12C1	66.04%	0.73	0.27	0.60	1.00	0.69	0.89	0.89	0.97	0.48
P12C2	42.17%	0.77	0.10	0.62	0.13	0.37	0.82	0.71	0.98	0.23
P15C1	50.83%	0.99	0.53	0.27	0.99	0.57	0.93	0.91	0.73	0.37
P20C1	52.67%	0.91	0.00	0.55	0.98	0.45	0.98	0.97	0.91	0.40
AVG	**50.28%**	**73.65%**	**21.99%**	**56.26%**	**77.38%**	**60.71%**	**86.05%**	**82.56%**	**79.94%**	**0.34**
Std	**±0.15**	**±0.33**	**±0.20**	**±0.20**	**±0.33**	**±0.19**	**±0.14**	**±0.18**	**±0.28**	**±0.14**

Table 5-5. Cityblock distance results.

		Sensitivity				Specificity				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
P8C1	28.99%	0.06	0.24	0.40	1.00	0.89	0.45	0.95	0.31	0.13
P8C2	68.26%	1.00	0.09	0.96	0.56	0.62	0.97	0.99	1.00	0.49
P12C1	75.08%	0.45	0.03	0.93	1.00	0.94	0.86	0.75	0.98	0.59
P12C2	43.12%	0.56	0.01	0.95	0.13	0.47	0.82	0.59	0.95	0.24
P15C1	76.55%	1.00	0.78	0.65	0.99	0.74	1.00	0.98	0.98	0.67
P20C1	31.03%	0.96	0.00	0.66	0.09	0.23	0.55	0.98	0.98	0.10
AVG	**53.84%**	**66.91%**	**19.07%**	**75.72%**	**62.81%**	**64.94%**	**77.72%**	**87.26%**	**86.79%**	**0.37**
Std	**±0.20**	**±0.35**	**±0.28**	**±0.21**	**±0.40**	**±0.25**	**±0.20**	**±0.15**	**±0.25**	**±0.22**

Table 5-6. Minkowski distance results.

		Sensitivity				Specificity				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
P8C1	27.15%	0.06	0.24	0.32	1.00	0.90	0.40	0.95	0.30	0.11
P8C2	67.29%	1.00	0.31	0.93	0.55	0.61	0.96	1.00	1.00	0.48
P12C1	71.65%	0.41	0.03	0.89	1.00	0.91	0.82	0.80	0.98	0.55
P12C2	42.47%	0.56	0.00	0.92	0.13	0.46	0.80	0.62	0.95	0.23
P15C1	70.64%	1.00	0.80	0.52	0.98	0.67	1.00	0.98	0.97	0.60
P20C1	32.27%	0.96	0.00	0.61	0.15	0.24	0.58	0.99	0.98	0.11
AVG	**51.91%**	**66.35%**	**23.18%**	**69.87%**	**63.67%**	**63.01%**	**76.08%**	**88.97%**	**86.39%**	**0.35**
Std	**±0.19**	**±0.36**	**±0.28**	**±0.23**	**±0.38**	**±0.23**	**±0.21**	**±0.14**	**±0.25**	**±0.20**

With this labelling method, Best Result, it is observed that the average OA obtained is 65.98% (see Table 5-7), being slightly higher than the discard method shown in Section 5.2.1. In addition, a higher kappa coefficient is also obtained, with a value of 0.51 compared to 0.36 of the voting method. There are some images in which the method works well (P8C2, P12C1), others in which it works more or less (P8C1 and P12C2), and a patient in which it works poorly (P20C1). Although labelling of the current patient is not 100% effective, there is a possibility that adding some labelled samples may improve the classification. The sensitivity is especially low in the tumour class, so adding data from this class could worsen the classification.

Table 5-7. Results obtained with the Best Result mehotd.

		Sensitivity				Specificity				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
P8C1	54.87%	0.48	0.52	0.52	0.91	0.90	0.61	0.80	0.93	0.39
P8C2	87.58%	0.97	0.25	0.99	0.84	0.97	0.91	1.00	0.97	0.77
P12C1	77.64%	0.85	0.02	0.93	0.16	0.92	0.86	0.85	1.00	0.63
P12C2	58.44%	0.93	0.01	0.95	0.28	0.60	0.90	0.74	1.00	0.44
P15C1	78.97%	0.86	0.92	0.75	0.51	0.83	0.93	0.98	0.96	0.69
P20C1	38.40%	0.97	0.00	0.75	0.30	0.83	0.67	0.99	0.44	0.17
AVG	**65.98%**	**84.43%**	**28.67%**	**81.45%**	**50.07%**	**84.18%**	**81.25%**	**89.30%**	**88.31%**	**0.51**
Std	**±0.17**	**±0.17**	**±0.34**	**±0.16**	**±0.29**	**±0.12**	**±0.12**	**±0.10**	**±0.20**	**±0.20**

After comparing all the results obtained with each of the distances, it is found that the best result was always obtained with the cosine distance (Figure 5-1). Therefore, all the values shown in the table above (Table 5-7) correspond to those of the cosine distance. With this type of distance, all classes, except background, obtain the best specificity and sensitivity results.

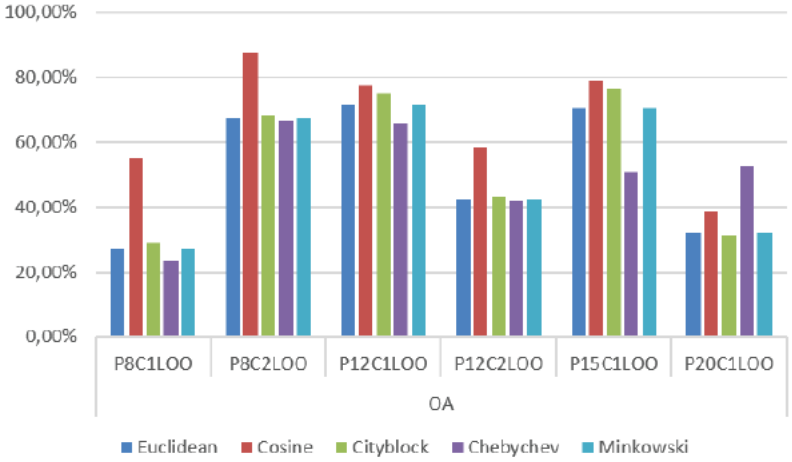


Figure 5-1. Graph of the OA obtained with each distance.

5.3 Optimization k for the k-means clustering.

Once the most suitable distance metric is selected, the value of the input parameter *k* is calculated, with which the k-means method divides the samples into *k* groups. The range evaluated was from 4 to 24. To make this decision, the k-means algorithm is executed with that range of values, the decision was based on observing what the percentage of each class was grouped in each cluster. This evaluation has been performed using the dataset belonging to the previous patients.

Evaluating the results shown in the following graphs (Figure 5-2, **¡Error! No se encuentra el origen de la referencia.**, Figure 5-3, Figure 5-4 and Figure 5-5) it is found that as the value of *k* increases, more of the background class is present. This was to be expected due to the high variability of this class, but the rest of the classes are well identified with this method, except for the tumour class. No value of *k* can group in such a way that in at least one cluster its majority class is the tumour tissue. The data are evaluated and *k*=10 is enough to be able to identify the other three classes in at least one cluster, a higher *k* would increase the computation time unnecessarily.

As we can see in the Figure 5-2, for patient P8C1LOO with k=5, cluster 1 contains most of the pixels of the healthy class. However, for this value of k, no cluster is formed mostly by the blood vessel class. Most of the rest of the clusters belong to the background class. With a k = 10, it is possible to identify some cluster as of the blood vessel class, being identified in some of the clusters three of the four classes.

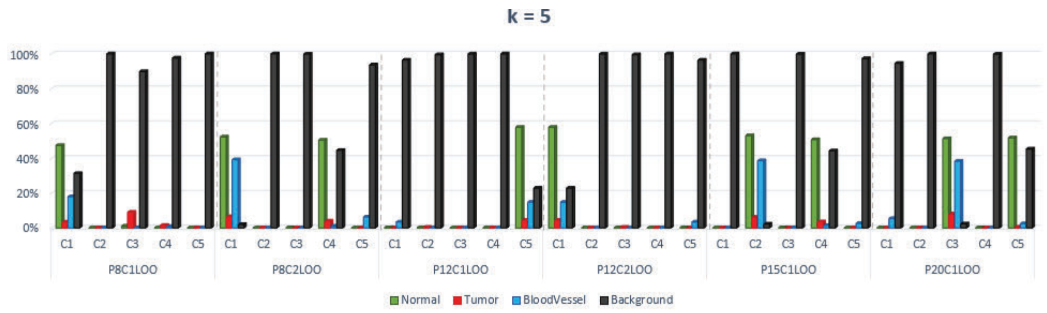


Figure 5-2. Percentage graph of the of the classes contained in each each cluster (k = 5).

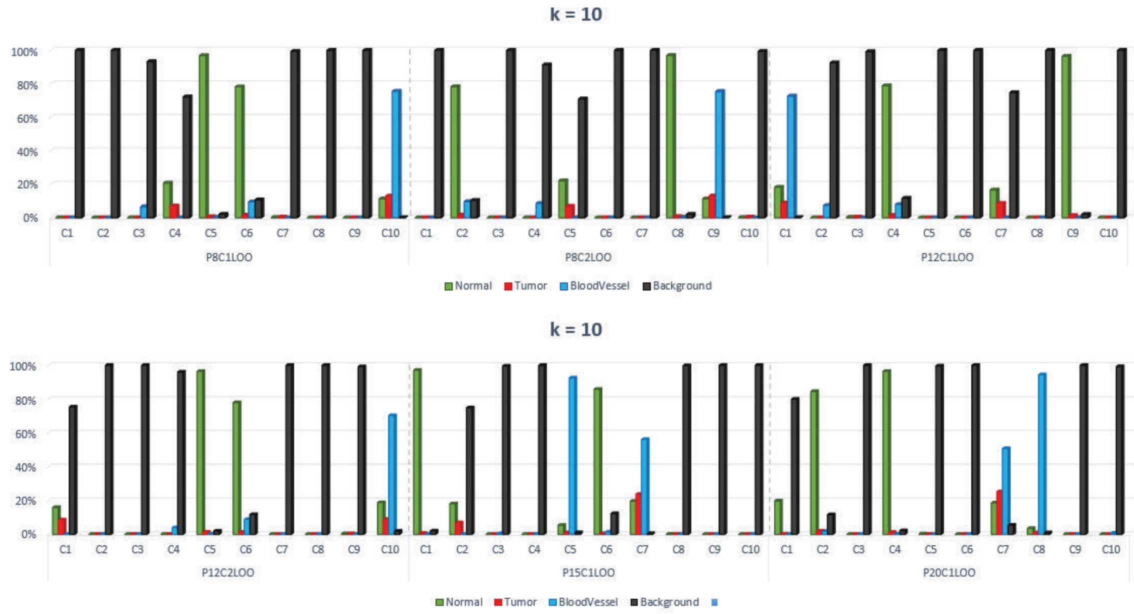


Figure 5-3. Percentage graph of the of the classes contained in each each cluster (k = 10).

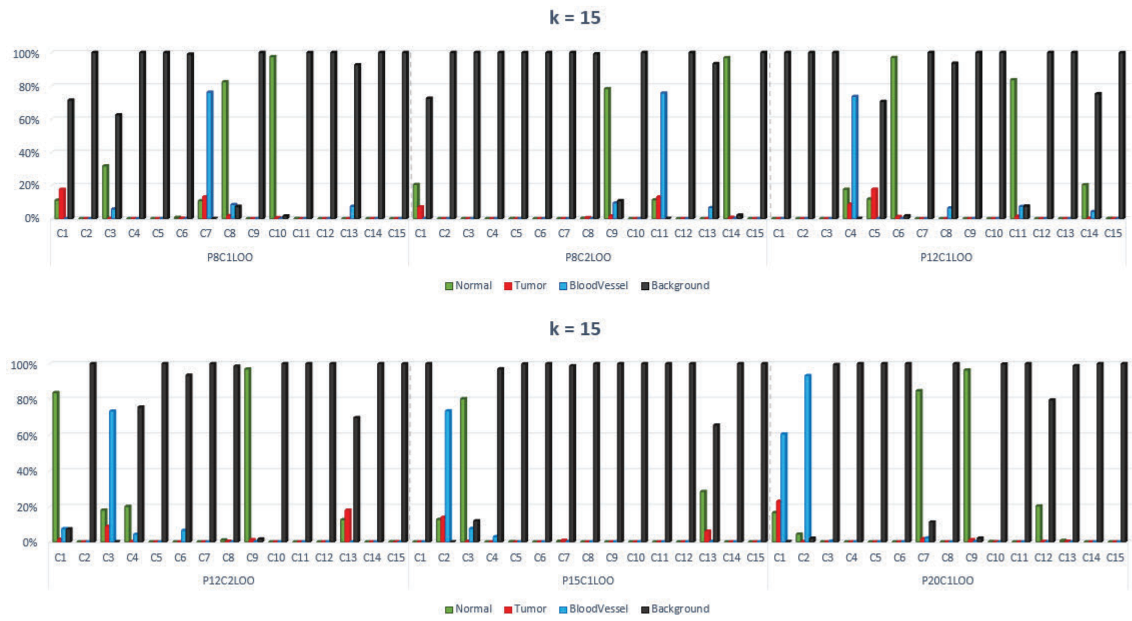


Figure 5-4. Percentage graph of the of the classes contained in each each cluster (k = 15).

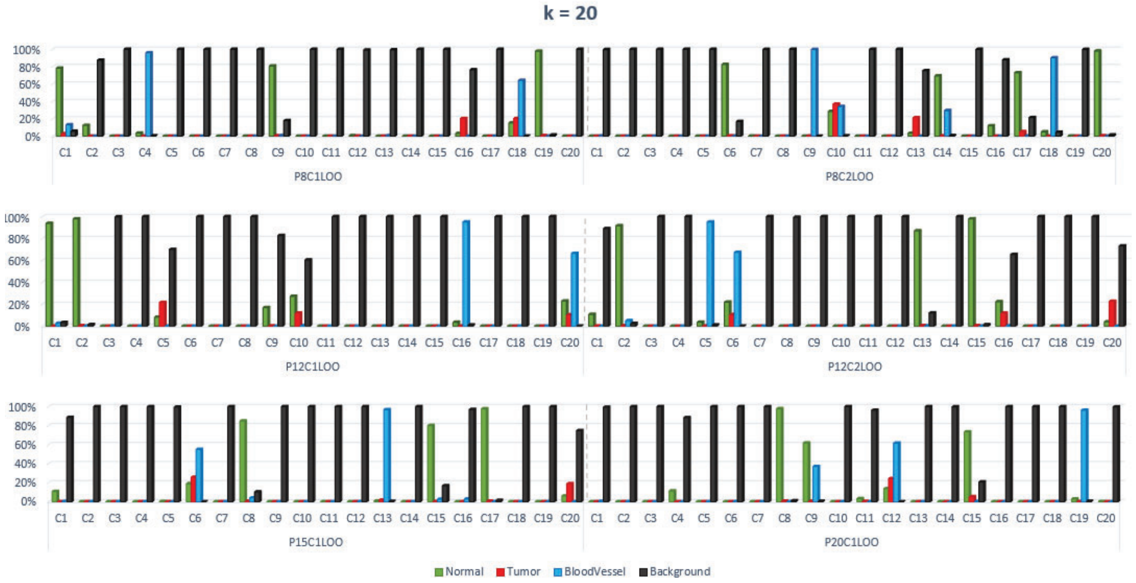


Figure 5-5. Percentage graph of the of the classes contained in each each cluster (k = 20).

The rest of the graphs can be found in Annex I where it is possible that, as the k-value increases, the results remain constant.

5.4 Evaluation of semi-supervised algorithm using the SVM classifier.

Once it is known to which class each cluster belongs, the current patient is labelled by calculating the distance between the current patient pixels and the centroids of the resulting clusters. First without any conditions (all clusters are used) and then with a condition (only the clusters that presents high presence of classes, more than 60%, are used) for automatic label generation. Finally, the semi-automatic labelled data are introduced together with the database of previous patients in the supervised classification algorithm to train and generate the model that will subsequently be evaluated.

The code is firstly executed without the semi-supervised part to have a reference result to compare with. The Table 5-8 shown these reference results with the SVM algorithm.

Table 5-8. Results obtained with the SVM algorithm.

		Sensitivity				Specificity			
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background
P8C1	58.23%	0.68	0.15	0.61	1.00	0.80	0.90	0.80	0.73
P8C2	95.12%	0.97	0.34	0.99	0.95	0.96	0.99	1.00	0.97
P12C1	93.32%	0.99	0.47	0.94	0.99	0.92	1.00	0.98	1.00
P12C2	79.27%	0.97	0.04	0.98	0.80	0.87	1.00	0.83	0.99
P15C1	88.31%	1.00	0.67	0.94	0.98	0.87	1.00	1.00	0.99
P20C1	58.37%	0.97	0.00	0.80	1.00	0.50	1.00	1.00	0.96
AVG	**78.77%**	**93.00%**	**28.03%**	**87.44%**	**95.39%**	**81.96%**	**98.09%**	**93.32%**	**93.94%**
Std	**±0.14**	**±0.10**	**±0.22**	**±0.12**	**±0.07**	**±0.14**	**±0.04**	**±0.08**	**±0.09**

5.4.1 Results of SVM without condition

Table 5-9 shows the results obtained with the semi-supervised processing framework developed, in which the automatic generation of labels from the current patient samples is performed without discarding any of the 10 clusters. Remember that, as mentioned in Section 5.3, it was decided to work with k = 10.

For this proposed methodology, an average OA of 45,57% with a standard deviation (Std) of ±23% is achieved. This value shows the dispersion of the data with respect to the mean. A high Std value indicates a greater dispersion of the data and therefore a lower precision.

Table 5-9. Results obtained in the semi-supervised process with the SVM algorithm with the generated label method without condition.

		Sensitivity				Specificity				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	Kappa
P8C1	39.51%	0.81	0.01	0.22	0.00	0.15	0.98	0.99	0.59	.
P8C2	92.34%	0.98	0.01	0.62	0.97	0.97	1.00	1.00	0.83	0.83
P12C1	52.42%	0.36	0.08	0.56	1.00	0.61	1.00	0.55	0.92	0.22
P12C2	44.61%	0.56	0.02	0.62	0.39	0.46	1.00	0.64	0.89	0.23
P15C1	11.04%	0.00	0.00	0.10	0.68	0.79	1.00	1.00	0.06	0.00
P20C1	33.48%	0.00	0.00	0.65	0.85	0.89	1.00	1.00	0.14	0.10
AVG	**45.57%**	**45.18%**	**1.90%**	**46.34%**	**64.69%**	**64.53%**	**99.62%**	**86.34%**	**57.10%**	**0.25**
Std	**±0.23**	**±0.35**	**±0.02**	**±0.20**	**±0.33**	**±0.26**	**±0.01**	**±0.18**	**±0.33**	**±0.25**

Focusing on the sensitivity, it is observed that it is not possible to identify the tumour class, obtaining an average value of 1.90%. This result was to be expected, since for automatic labelling it was not possible to obtain any cluster that consisted mainly of this class. This value improves for the rest of the classes, although it is only exceeded 50% with the background class.

The specificity values obtained for this case study are mostly good, indicating a high probability that a pixel that is not of that class will be identified as not being of that class. This would explain the value obtained for the tumour class of 99.62%. Finally, a kappa value of 0.25 was obtained where it can be seen that the strength of agreement between the samples is fair.

5.4.2 Results of SVM with condition

Table 5-10 shows the results obtained when the decision to generate labels is made based on a condition. The clusters used must contain at least 60% of one of the classes, otherwise the cluster is discarded, and the data are not included in the training set.

Table 5-10. Results obtained in the semi-supervised process with the SVM algorithm with the generated label method with condition.

		Sensitivity				Specificity				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
P8C1	27.46%	0.23	0.00	0.26	1.00	0.98	0.96	0.99	0.18	0.14
P8C2	90.25%	0.79	0.01	0.56	1.00	1.00	1.00	1.00	0.69	0.77
P12C1	51.17%	0.95	0.06	0.24	0.99	0.35	1.00	1.00	0.89	0.31
P12C2	54.78%	0.91	0.00	0.19	0.72	0.48	1.00	0.88	0.84	0.36
P15C1	11.04%	0.00	0.00	0.10	0.68	0.79	1.00	1.00	0.06	0.00
P20C1	33.48%	0.00	0.00	0.65	0.85	0.89	1.00	1.00	0.14	0.10
AVG	**44.70%**	**48.05%**	**1.43%**	**33.26%**	**87.28%**	**74.95%**	**99.30%**	**97.64%**	**46.76%**	**0.28**
Std	**±0.23**	**±0.38**	**±0.02**	**±0.19**	**±0.13**	**±0.23**	**±0.01**	**±0.04**	**±0.32**	**±0.23**

For this semi-supervised design, an AVG of 44.70% was obtained, with a Std of ±23%. The OA obtained for the samples from the patient 15 (P15C1 and P15C2) and patient 20 (P20C1) is maintained with respect to the previous case (see Table 5-10). However, it is lower with respect to the SVM without condition in patient 8 (P8C1 and P8C2) and in patient 12 (P12C1). Improving only this value in patient 12 (P12C2). For this sample, the sensitivity of the normal and background class improves considerably, but there are still problems in identifying tumour pixels, obtaining a sensitivity of 1.43% and a specificity of 99.30% for this case study. Regarding the Kappa value, it has been improved a little with respect to the generated labels without condition method and now the coefficient value is 0.28. With the proposed semi-supervised method, the SVM classification results are worse compared to the supervised method.

5.5 Evaluation of semi-supervised algorithm using the RF classifier

As the SVM computation times are too long, the same procedure is proposed, but in this case using the RF classification algorithm. For the implementation, the value of the parameter *k* was re-evaluated to try to define the number of clusters with which the k-means will work. This time it is decided to use a k = 15 (Table 5-11), since it is the only one that meets the same criteria as k = 10 but with the difference that the condition no longer needs to be applied. The procedure will only have to be performed once, since the criterion used was that all clusters should consist of at least 60% of a class.

On the other hand, P8C1 was used as a sweep to select the number of trees. This capture is left as a validation set. If we look at the results shown in Table 5-11, it can be observed that the best result is obtained for 100 trees. Where 62.2% of OA was achieved and where the highest sensitivity values are also obtained for most of the classes, highlighting 81.0% for the normal class and 25.2% of the tumour class. Furthermore, in all cases a sensitivity of 100% was achieved for the background class.

Table 5-11. Results obtained with different values of trees (from 50 to 300) with RF.

		Sensitivity				Specificity			
RF	**OA**	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background
50	55.7%	63.4%	20.6%	53.6%	100.0%	73.8%	86.6%	81.3%	73.1%
100	62.1%	81.0%	25.3%	45.2%	100.0%	65.0%	85.7%	89.5%	87.3%
150	58.6%	71.1%	18.1%	54.6%	100.0%	73.5%	86.6%	82.5%	78.4%
200	60.7%	78.4%	16.8%	51.8%	100.0%	70.3%	87.0%	84.5%	82.8%
250	59.9%	74.8%	16.5%	54.9%	100.0%	72.3%	87.1%	83.3%	80.7%
300	60.5%	77.0%	15.8%	54.3%	100.0%	72.5%	87.4%	83.2%	81.6%
avg	**59.57%**	**74.29%**	**18.84%**	**52.40%**	**100.00%**	**71.24%**	**86.73%**	**84.05%**	**80.65%**
std	**±0.02**	**±0.05**	**±0.03**	**±0.03**	**±0.00**	**±0.03**	**±0.01**	**±0.02**	**±0.04**

Once the parameter *k* of the k-means algorithm and the number of trees in the RF algorithm have been selected, the code is executed without the semi-supervised part in order to have a reference result to compare with. These reference results are shown in Table 5-12)

Table 5-12. Results obtained with the RF algorithm.

		Sensitivity				Specificity				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
P8C2	94.08%	0.92	0.31	0.96	0.96	0.96	1.00	1.00	0.93	0.88
P12C1	90.94%	1.00	0.14	0.92	1.00	0.88	1.00	0.99	1.00	0.85
P12C2	73.01%	0.99	0.01	0.94	0.65	0.75	1.00	0.81	1.00	0.62
P15C1	68.88%	0.98	0.03	0.88	0.99	0.65	1.00	0.99	0.97	0.55
P20C1	58.05%	0.96	0.00	0.79	1.00	0.60	1.00	0.99	0.66	0.46
AVG	**76.99%**	**97.04%**	**9.91%**	**89.79%**	**91.73%**	**76.92%**	**99.88%**	**95.67%**	**91.09%**	**0.67**
Std	**±0.12**	**±0.03**	**±0.11**	**±0.06**	**±0.12**	**±0.12**	**±0.00**	**±0.07**	**±0.11**	**±0.15**

For the RF algorithm, an average accuracy value of 76.99% and a considerable kappa value of 0.67 were obtained (see Table 5-12), being the best kappa value achieved so far. In the semi-supervised process proposed with the RF algorithm, an AVG of 46.56% and an acceptable kappa value of 0.27 were obtained (Table 5-13).

Table 5-13. Results obtained in the semi-supervised process with the RF algorithm.

		Sensitivity				Specificity				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
P8C2	79.49%	0.29	0.00	0.52	0.99	0.99	1.00	1.00	0.35	0.45
P12C1	22.83%	0.01	0.00	0.21	1.00	0.98	1.00	1.00	0.14	0.08
P12C2	57.95%	0.62	0.00	0.30	0.95	0.96	1.00	0.89	0.42	0.38
P15C1	24.05%	0.66	0.00	0.10	1.00	0.93	1.00	1.00	0.17	0.14
P20C1	48.49%	0.66	0.00	0.56	1.00	0.97	1.00	1.00	0.30	0.31
AVG	**46.56%**	**44.80%**	**0.07%**	**33.96%**	**98.77%**	**96.55%**	**99.92%**	**97.67%**	**27.48%**	**0.27**
Std	**±0.20**	**±0.24**	**±0.00**	**±0.16**	**±0.02**	**±0.02**	**±0.00**	**±0.04**	**±0.10**	**±0.13**

With this new approach, the same problem is still detected when identifying the tumour class pixels. In general, the results of the classification worsen in all cases, except the specificity of the normal class, which increases.

As there are many clusters where the background class predominates, it was decided to carry out the whole procedure again, but this time only using 3 clusters for the automatic generation of labels instead of 15. The choice was based on selecting the one that is composed of a predominant class (background, blood vessel and normal tissue). With the tumour class, this criterion is never met.

5.5.1 Results of RF evaluating with three clusters

The following Table 5-14 shows the results obtained using only 3 clusters in the generation of labels. One made up entirely of the background class, and the other two where the normal and blood vessel class stand out with more than 60%.

Table 5-14. Results obtained in the semi-supervised process with the RF algorithm (3 clusters).

		Sensitivity				Specificity				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
P8C2	46.70%	0.65	0.00	0.53	0.41	0.44	0.99	0.95	0.67	0.14
P12C1	25.92%	0.03	0.00	0.26	0.99	0.39	1.00	0.26	0.85	-0.21
P12C2	55.57%	0.91	0.00	1.00	0.18	0.79	1.00	0.46	1.00	0.40
P15C1	45.51%	0.88	0.00	0.46	0.98	0.52	1.00	0.49	0.93	0.21
P20C1	55.73%	0.73	0.00	0.94	0.99	0.98	1.00	0.91	0.42	0.42
AVG	**45.89%**	**64.20%**	**0.03%**	**63.87%**	**70,.99%**	**62.45%**	**100.00%**	**61.55%**	**77.32%**	**0.19**
Std	**±0.10**	**±0.29**	**±0.00**	**±0.26**	**±0.32**	**±0.20**	**±3.27**	**±0.24**	**±0.19**	**±0.21**

Although the AVG obtained is 45.89%, a little below that obtained using all 15 clusters (see Table 5-13), the sensitivity of the Normal and the Blood Vessel class is considerably improved, with a 64,20% and 63,87% respectively. Also, the balance of the sensitivity and specificity values of the background class improve with 70.99% sensitivity and 77.32% specificity, thus giving great validity of diagnostic for this class. The specificity of the healthy class has decreased compared to the previous method, and the results are worse than without using a semi-supervised mechanism.

5.6 Summary

This chapter has analyzed the results obtained with the proposed semi-supervised procedure. The decisions taken during the development of this thesis have also been justified.

First, the choice of *k* was made, where initially its value was 10, since it was a high enough value to be able to identify in any of the clusters at least three of the classes (normal, blood vessel and background). Once the *k* is selected, the whole process is carried out with the SVM algorithm. In this way the model is generated and trained for later testing. The process was repeated with two different ways of generating the labels automatically: generating these labels without any conditions and then applying the criterion that the clusters must be formed by at least 60% of the same class. If this was not fulfilled, the cluster was discarded for decision making.

Considering the high computation times, it is decided to perform the same procedure with the RF algorithm. In addition, the *k* parameter is re-evaluated, and it is found that with a value of k = 15 the 60% criterion is met in all clusters, so no conditions need to be applied. The choice of the number of trees is made using the P8C1 and it is decided that the number of trees should be 100.

Take note of that most of the clusters were made up of the background class. It was decided to choose from the 15 clusters the 3 that best represented background, blood vessels and normal tissue class. With these three clusters, the process was repeated, seeking to improve the sensitivity of those same classes.

The results with the averages obtained from all patients are shown in Table 5-15, where, the highest success rate was obtained for the semi-supervised approach for RF with 46.40%. However, the sensitivity obtained from the main classes is lower than for the rest of the processes, except for the background class, which is not given with it.

Table 5-15. Results obtained in all semi-supervised process.

		Sensitivity				**Specificity**				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
Supervised process										
SVM	**78.77%**	93.00%	28.03%	87.44%	95.39%	81.96%	98.09%	93.32%	93.94%	-
RF	**76.99%**	97.04%	9.91%	89.79%	91.73%	76.92%	99.88%	95.67%	91.09%	**0.67**
Semi-supervised process										
SVM without Condition	**45.57%**	45.18%	1.90%	46.34%	64.69%	64.53%	99.62%	86.34%	57.10%	**0.25**
SVM with condition	**44.70%**	48.05%	1.43%	33.26%	87.28%	74.95%	99.30%	97.64%	46,.6%	**0.28**
RF	**46.56%**	44.80%	0.07%	33.96%	98.77%	96.55%	99.92%	97.67%	27.48%	**0.27**
RF (evaluating with three clusters)	**45.89%**	64.20%	0.03%	63.87%	70.99%	62.45%	100.00%	61.55%	77.32%	**0.19**

Finally, the highest sensitivity value for tumour class was obtained with the SVM algorithm without condition at 1.90%, which is still a too low. Focusing on the rest of the data, perhaps the RF evaluating with three clusters approach gives the best results for all kinds of class except for tumour. The proposed processing method may not be adequate to improve the results. The semi-supervised algorithm proposal worsens the classification results compared to the non-semi-supervised.

Chapter 6: Conclusions & Future Lines

6.1 Conclusions

The main problem in this field is working with a limited database. According to the exposed problem, the objective of this Master Thesis was to design a semi-supervised classification system. This type of classification was intended to increase the existing database for the supervised classification. During this process, the HS images described in Chapter 3 have been used.

The proposed methodology consists in that when the patient is in the operating room, the images taken by the surgeon will be automatically labelled by a SSL algorithm and then, together with the existing database, the model can be generated. To perform the labelling of the current patient samples, we proposed a method which rely in the k-means algorithm.

First, to optimize the k-means algorithm, the different types of distances were evaluated. The distance selected for this type of study was the cosine distance. Then, a study was carried out to choose the parameter k (number of target clusters for k-means), choosing a k of 10. Once the k-means parameters have been optimized, this algorithm is used to automatically generate the current patient labels. Such labels are included together with the database of previous patients and used as training data in the SVM to generate the model and evaluate it. This labelling process is implemented in two ways. Without any conditions and with the condition that the clusters used must contain at least 60% of one of the classes, otherwise the cluster is discarded.

Due to the long computational times, it was decided to carry out the same procedure but this time using the RF algorithm. After conducting a study, it was decided that for this database the best number of trees was 100. The parameter k is re-evaluated and a k of 15 is established. With this value, it is not necessary to apply any conditions since all clusters are made up of at least 60% of some class.

When evaluating all the results, it is seen that most of the clusters belong to the background class. This is due to the great variability of this class. To avoid this, the last proposed procedure is performed again but using only 3 clusters, those that represent the background, blood vessel and normal tissue class. It is assumed with them that there is no cluster identified as being of the tumour class. When analysing these last results, it is seen how it is possible to improve the sensitivity of these three classes.

It is considered that the image used in the semi-supervised to automatically label it and thus increase the database with which the model is generated, must be an image that does not include any tumour pixels. In this way we ensure that when the automatic labelled is generated there are no mislabelled tumour pixels. If we improve the balance of specificity and sensitivity of the rest of the classes, we will also be able to improve it for the tumour class. Finally, although a method to improve the classification has been proposed, this goal has not been achieved. The proposed methods worsen the original supervised classification (without semi-supervised).

6.2 Future lines

It is known that the captures made at the beginning of this process in the semi-supervised algorithms play an important role. As a future line, it is proposed to increase the database with HS images of the same patient that contain tumour and captures those which do not contain it. Right now, the database of this Master Thesis only contains captures with tumour, negatively influencing the results obtained. The data set of previous patients could also be balanced when applying the k-means. From this, the data would be better grouped avoiding so many clusters belonging to the background class.

On the other hand, it is proposed to continue improving the techniques proposed for the automatic generation of labels, trying to improve the identification of the rest of the classes. An option to this end is to make use of dimensionality reduction algorithm (such as Principal Components Analysis) in order to establish a more robust similarity metric that allows a more accurate label assignment, and hence an improvement of the results. Besides, the use of commonly used SSL methods found in the literature (Section 2.3.3) may improve the results of the classification. However, the study and implementation of such complex approaches are out of the scope of this Master Thesis. Additionally, it could be also interesting to evaluate different classifiers such as the ANN, using the same proposed methodology. Finally, in the design of the semi-supervised design that used the SVM algorithm, the results could be improved by performing an optimization of the hyperparameters.

References

[1] G. ElMasry and D.-W. Sun, “Principles of Hyperspectral Imaging Technology,” in *Hyperspectral Imaging for Food Quality Analysis and Control*, Elsevier, 2010, pp. 3–43.

[2] G. Shaw and D. Manolakis, “Signal processing for hyperspectral image exploitation,” *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 12–16, 2002.

[3] G. A. Shaw and H. K. Burke, “Spectral Imaging for Remote Sensing,” *LINCOLN Lab. J.*, vol. 14, no. 1, 2003.

[4] J. Salmelin *et al.*, “Hyperspectral Imaging of Macroinvertebrates—a Pilot Study for Detecting Metal Contamination in Aquatic Ecosystems,” *Water, Air, Soil Pollut.*, vol. 229, no. 9, p. 308, Sep. 2018.

[5] J. Satapathy and B. P. Jangid, “Monitoring a local extreme weather event with the scope of hyperspectral sounding,” *Meteorol. Atmos. Phys.*, vol. 130, no. 3, pp. 371–381, 2018.

[6] G. Hong and H. T. Abd El-Hamid, “Hyperspectral imaging using multivariate analysis for simulation and prediction of agricultural crops in Ningxia, China,” *Comput. Electron. Agric.*, vol. 172, p. 105355, May 2020.

[7] H.-J. He and D.-W. Sun, “Hyperspectral imaging technology for rapid detection of various microbial contaminants in agricultural and food products,” *Trends Food Sci. Technol.*, vol. 46, no. 1, pp. 99–109, Nov. 2015.

[8] J. Shan, J. Zhao, Y. Zhang, L. Liu, F. Wu, and X. Wang, “Simple and rapid detection of microplastics in seawater using hyperspectral imaging technology,” *Anal. Chim. Acta*, vol. 1050, pp. 161–168, Mar. 2019.

[9] J. Xu *et al.*, “Applications of hyperspectral and optical scattering imaging technique in the detection of food microorganism,” *Int. J. Comput. Vis. Robot.*, vol. 8, no. 3, p. 267, 2018.

[10] J. Li, W. Luo, Z. Wang, and S. Fan, “Early detection of decay on apples using hyperspectral reflectance imaging combining both principal component analysis and improved watershed segmentation method,” *Postharvest Biol. Technol.*, vol. 149, pp. 235–246, Mar. 2019.

[11] X. Fu and J. Chen, “A Review of Hyperspectral Imaging for Chicken Meat Safety and Quality Evaluation: Application, Hardware, and Software,” *Compr. Rev. Food Sci. Food Saf.*, vol. 18, no. 2, pp. 535–547, Mar. 2019.

[12] B. Li, M. Cobo-Medina, J. Lecourt, N. B. Harrison, R. J. Harrison, and J. V. Cross, “Application of hyperspectral imaging for nondestructive measurement of plum quality attributes,” *Postharvest Biol. Technol.*, vol. 141, no. March, pp. 8–15, 2018.

[13] S. Guan, H. Asfour, N. Sarvazyan, and M. Loew, “Application of unsupervised learning to hyperspectral imaging of cardiac ablation lesions,” *J. Med. Imaging*,

vol. 5, no. 04, p. 1, Dec. 2018.

[14] S. M. Blair, M. Garcia, C. Konopka, L. Dobrucki, and V. Gruev, “A 27-band snapshot hyperspectral imaging system for label-free tumour detection during image-guided surgery,” in *Label-free Biomedical Imaging and Sensing (LBIS) 2019*, 2019, p. 16.

[15] Y. Moulla *et al.*, “Hyperspectral imaging (Hsi)—a new tool to estimate the perfusion of upper abdominal organs during pancreatoduodenectomy,” *Cancers (Basel).*, vol. 13, no. 11, pp. 1–13, 2021.

[16] M. H. Aref, A. B. M. Youssef, I. H. Aboughaleb, and Y. H. El-Sharkawy, “Characterization of normal and malignant breast tissues utilizing hyperspectral images and associated differential spectrum algorithm,” *J. Biomed. Photonics Eng.*, vol. 7, no. 2, pp. 1–12, 2021.

[17] D. R. C. G. Samuel Ortega Sarmiento, Dr. Gustavo Marrero Callicó, “Técnicas de Reconocimiento Automático de Patrones Aplicadas a Imágenes Hiperespectrales Médicas.” 2016.

[18] A. Pedro Duarte Silva, “Optimization approaches to Supervised Classification,” *Eur. J. Oper. Res.*, vol. 261, no. 2, pp. 772–788, Sep. 2017.

[19] L. Ge, B. Huang, W. Wei, and Z. Pan, “Semi-Supervised classification of hyperspectral images using discrete nonlocal variation Potts Model,” *Math. Found. Comput.*, vol. 4, no. 2, p. 73, 2021.

[20] D. K. Pathak, S. K. Kalita, and D. K. Bhattacharya, “Hyperspectral image classification using support vector machine: a spectral spatial feature based approach,” *Evol. Intell.*, pp. 430–435, 2021.

[21] G. Y. Chen, “Multiscale filter-based hyperspectral image classification with PCA and SVM,” *J. Electr. Eng.*, vol. 72, no. 1, pp. 40–45, 2021.

[22] M. P. S. Brown *et al.*, “Knowledge-based analysis of microarray gene expression data by using support vector machines,” *Proc. Natl. Acad. Sci.*, vol. 97, no. 1, pp. 262–267, Jan. 2000.

[23] E. J. Carmona Suárez, “Máquinas de Vectores Soporte (SVM),” *Dpto. Intel. Artif. ETS Ing. Inforática, Univ. Nac. Educ. a Distancia*, pp. 1–25, 2014.

[24] D. Bertsimas and J. Dunn, “Optimal classification trees,” *Mach. Learn.*, vol. 106, no. 7, pp. 1039–1082, Jul. 2017.

[25] Y. Zhang, G. Cao, X. Li, B. Wang, and P. Fu, “Active semi-supervised random forest for hyperspectral image classification,” *Remote Sens.*, vol. 11, no. 24, pp. 1–21, 2019.

[26] MathWorks, “Unsupervised Learning,” *Matlab, Simulink*. [Online]. Available: Unsupervised Learning - MATLAB & Simulink (mathworks.com).

[27] A. A. Aldino, D. Darwis, A. T. Prastowo, and C. Sujana, “Implementation of K-Means Algorithm for Clustering Corn Planting Feasibility Area in South Lampung Regency,” *J. Phys. Conf. Ser.*, vol. 1751, no. 1, 2021.

[28] M. D. J. Bora and D. A. K. Gupta, “Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab,” vol. 5, no. 2, pp. 2501–2506, 2014.

[29] Z.-H. Zhou and M. Li, “Semi-supervised learning by disagreement,” *Knowl Inf Syst*, vol. 24, pp. 415–439, 2010.

[30] A. Z. Olivier Chapelle, Bernhard Scho ̈lkopf, *A semi-supervised learning*, vol. 1, no. 2. 2009.

[31] Z. Feng, S. Yang, M. Wang, and L. Jiao, “Learning Dual Geometric Low-Rank Structure for Semisupervised Hyperspectral Image Classification,” *IEEE Trans. Cybern.*, pp. 1–13, 2019.

[32] E. Adeli *et al.*, “Semi-Supervised Discriminative Classification Robust to Sample-Outliers and Feature-Noises,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 515–522, Feb. 2019.

[33] Y. Lei, L. Cen, X. Chen, and Y. Xie, “A Hybrid Regularization Semi-Supervised Extreme Learning Machine Method and Its Application,” *IEEE Access*, vol. 7, pp. 30102–30111, 2019.

[34] S. T. H. Shah, S. G. Javed, A. Majid, S. A. H. Shah, and S. A. Qureshi, “Novel Classification Technique for Hyperspectral Imaging using Multinomial Logistic Regression and Morphological Profiles with Composite Kernels,” *Proc. 2019 16th Int. Bhurban Conf. Appl. Sci. Technol. IBCAST 2019*, pp. 419–424, 2019.

[35] B. L, C. M, and M. M, “A Novel Transductive SVM for Semisupervised Classification of Remote-Sensing Images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, p. 3363, 2006.

[36] L. Gómez-Chova, G. Camps-Valls, J. Muñoz-Marí, and J. Calpe, “Semi-supervised image classification with Laplacian Support Vector Machines,” *ICET 2013 - 2013 IEEE 9th Int. Conf. Emerg. Technol.*, vol. XX, pp. 1–5, 2007.

[37] P. Sellars, A. Aviles-Rivero, N. Papadakis, D. Coomes, A. Faul, and C.-B. Schönlieb, “Semi-supervised Learning with Graphs: Covariance Based Superpixels For Hyperspectral Image Classification,” 2019.

[38] M. Chi and L. Bruzzone, “Semisupervised Classification of Hyperspectral Images by SVMs Optimized in the Primal,” *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1870–1880, Jun. 2007.

[39] K. Tan, E. Li, Q. Du, and P. Du, “An efficient semi-supervised classification approach for hyperspectral imagery,” *ISPRS J. Photogramm. Remote Sens.*, vol. 97, pp. 36–45, Nov. 2014.

[40] F. de Morsier, M. Borgeaud, V. Gass, J.-P. Thiran, and D. Tuia, “Kernel Low-Rank and Sparse Graph for Unsupervised and Semi-Supervised Classification of Hyperspectral Images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3410–3420, Jun. 2016.

[41] B. Manifold, S. Men, R. Hu, and D. Fu, “A versatile deep learning architecture for classification and label-free prediction of hyperspectral images,” *Nat. Mach. Intell.*, vol. 3, no. 4, pp. 306–315, Apr. 2021.

[42] J. van Roy, N. Wouters, B. De Ketelaere, and W. Saeys, “Intuitive, semi-supervised training of the segmentation of hyperspectral images,” *Near Infrared Spectrosc. Proc. Int. Conf.*, vol. I, no. October, 2015.

[43] P. Ghaderyan and S. M. Ghoreshi Beyrami, “Neurodegenerative diseases detection using distance metrics and sparse coding: A new perspective on gait symmetric features,” *Comput. Biol. Med.*, vol. 120, no. March, p. 103736, 2020.

[44] H. Wu, Y. Cao, H. Wei, and Z. Tian, “Face Recognition Based on Haar like and Euclidean Distance,” *J. Phys. Conf. Ser.*, vol. 1813, no. 1, 2021.

[45] L. Sahu and B. R. Mohan, “An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop,” *9th Int. Conf. Ind. Inf. Syst. ICIIS 2014*, 2015.

[46] C.-I. Chang, “Hyperspectral Target Detection: Hypothesis Testing, Signal-to-Noise Ratio, and Spectral Angle Theories,” *IEEE Trans. Geosci. Remote Sens.*, pp. 1–23, 2021.

[47] M. Wang, Z. Huang, X. Zhang, Y. Zhang, and M. Chen, “Altered mineral mapping based on ground-airborne hyperspectral data and wavelet spectral angle mapper tri-training model: Case studies from Dehua-Youxi-Yongtai Ore District, Central Fujian, China,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 102, p. 102409, Oct. 2021.

[48] A. Chakraborty, N. Faujdar, A. Punhani, and S. Saraswat, “Comparative study of K-means clustering using iris data set for various distances,” *Proc. Conflu. 2020 - 10th Int. Conf. Cloud Comput. Data Sci. Eng.*, pp. 332–335, 2020.

[49] X. Gao and G. Li, “A KNN Model Based on Manhattan Distance to Identify the SNARE Proteins,” *IEEE Access*, vol. 8, pp. 112922–112931, 2020.

[50] H. Fabelo *et al.*, “HELICoiD project: a new use of hyperspectral imaging for brain cancer detection in real-time during neurosurgical operations,” 2016, p. 986002.

[51] H. Fabelo *et al.*, “An intraoperative visualization system using hyperspectral imaging to aid in brain tumour delineation,” *Sensors (Switzerland)*, vol. 18, no. 2, 2018.

[52] P. Algorithms, C. Detection, and H. I. Deliverable, “Figure 4-1: Data capture and labelling process.,” vol. 2, pp. 33–52, 2016.

[53] H. Fabelo *et al.*, “Deep Learning-Based Framework for In Vivo Identification of Glioblastoma Tumour using Hyperspectral Images of Human Brain,” *Sensors (Basel).*, vol. 19, no. 4, 2019.

[54] D. N. Louis *et al.*, “The 2016 World Health Organization Classification of Tumours of the Central Nervous System: a summary,” *Acta Neuropathol.*, vol. 131, no. 6, pp. 803–820, Jun. 2016.

[55] K. Wang, L. Cheng, and B. Yong, “Spectral-Similarity-Based Kernel of SVM for Hyperspectral Image Classification,” *Remote Sens.*, vol. 12, no. 13, p. 2154, Jul. 2020.

[56] A. G. Vani and D. V Saravanan, “Hyperspectral Image Classification Using Svm Machine Learning Approach,” *Gedrag Organ. Rev.*, vol. 33, no. 02, pp. 68–75, 2020.

[57] T. Horvat, L. Havaš, and D. Srpak, “The Impact of Selecting a Validation Method in Machine Learning on Predicting Basketball Game Outcomes,” *Symmetry (Basel).*, vol. 12, no. 3, p. 431, Mar. 2020.

[58] K. Jung, D.-H. Bae, M.-J. Um, S. Kim, S. Jeon, and D. Park, “Evaluation of Nitrate Load Estimations Using Neural Networks and Canonical Correlation Analysis with K-Fold Cross-Validation,” *Sustainability*, vol. 12, no. 1, p. 400, Jan. 2020.

[59] M. Magnusson, M. R. Andersen, J. Jonasson, and A. Vehtari, “Leave-One-Out Cross-Validation for Bayesian Model Comparison in Large Data,” no. January, 2020.

[60] F. de J. Núñez Cárdenas, E. Aguirre Hernández, A. E. Guerrero Zenil, and A. M. Felipe Redondo, “Application Method of Data Mining Using the K means Algorithm for the Determination of Stress Level in High School Students Using the Beck Depression Inventory,” *Cienc. Huasteca Boletín Científico la Esc. Super. Huejutla*, vol. 8, no. 15, pp. 1–8, Jan. 2020.

[61] A. J. Alberg, J. W. Park, B. W. Hager, M. V. Brock, and M. Diener-West, “The use of ‘overall accuracy’ to evaluate the validity of screening or diagnostic tests,” *J. Gen. Intern. Med.*, vol. 19, no. 5 PART 1, pp. 460–465, 2004.

[62] J. Cerda Lorca and L. Villarroel Del P., “Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa,” *Rev. Chil. Pediatr.*, vol. 79, no. 1, pp. 54–58, 2008.

[63] J. A. Gras, M. T. A. Argilaga, and J. G. Benito, “Metodología observacional,” *Metodol. la Investig. en ciencias del Comport.*, vol. 1, pp. 125–236, 1990.

Annex I

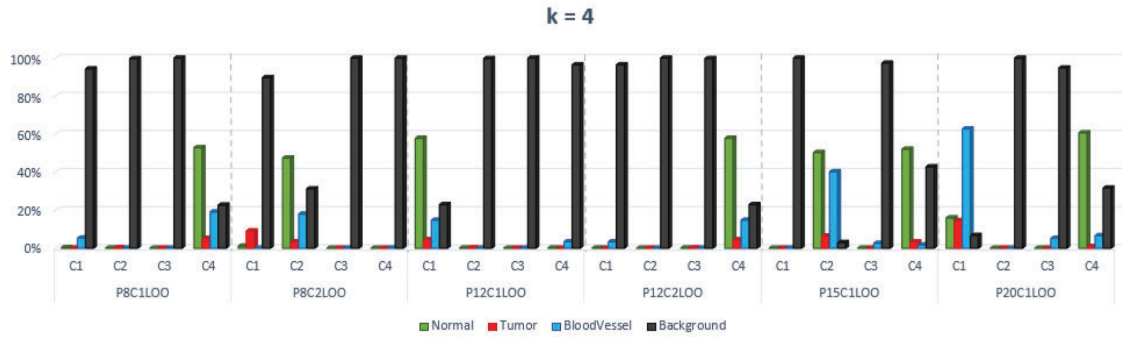


Figure 0-1. Graph of the percentages of the classes that each cluster contains (k = 4).

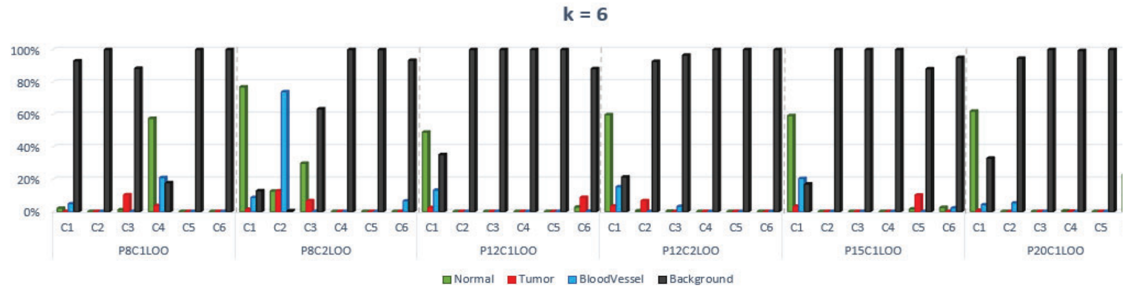


Figure 0-2. Graph of the percentages of the classes that each cluster contains (k = 6).

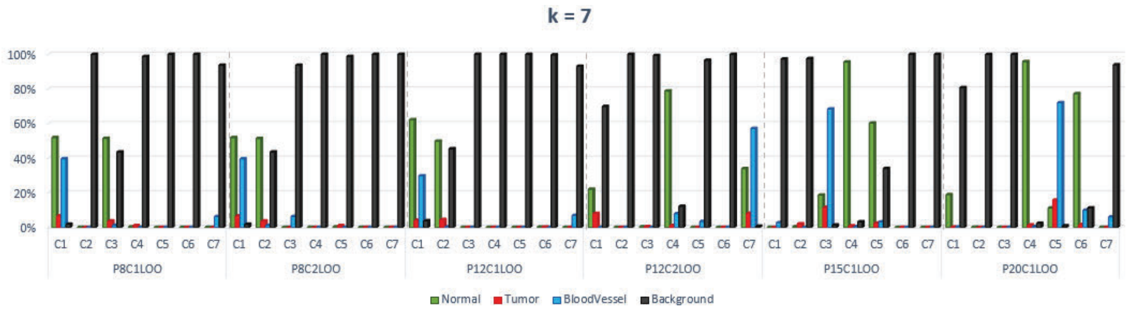


Figure 0-3. Graph of the percentages of the classes that each cluster contains (k = 7).

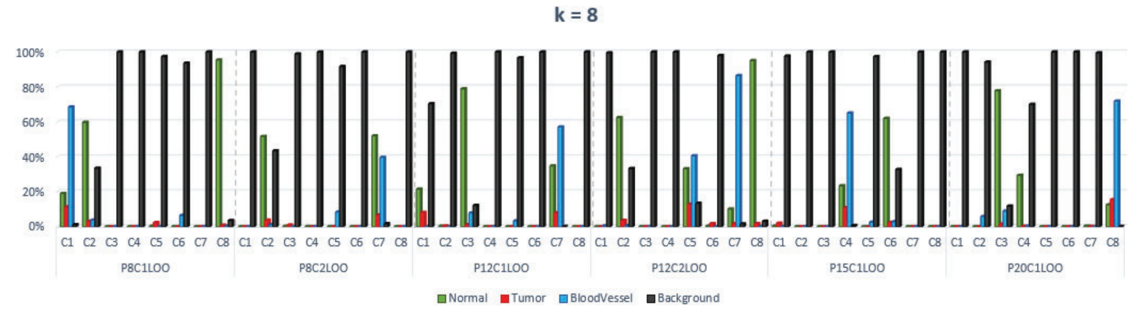


Figure 0-4. Graph of the percentages of the classes that each cluster contains (k = 8).

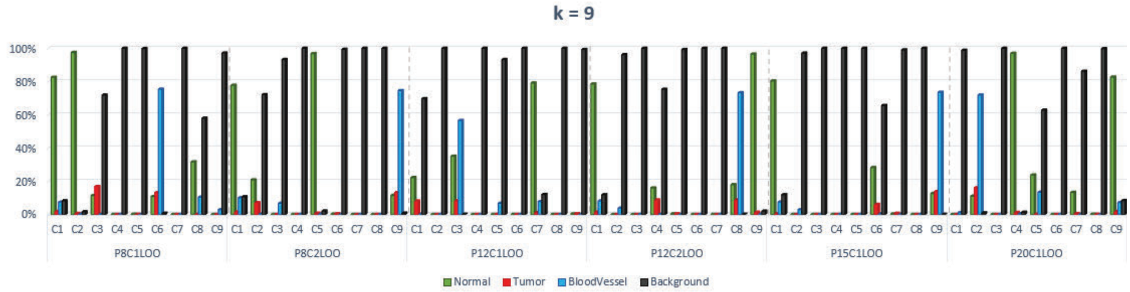


Figure 0-5. Graph of the percentages of the classes that each cluster contains (k = 9).

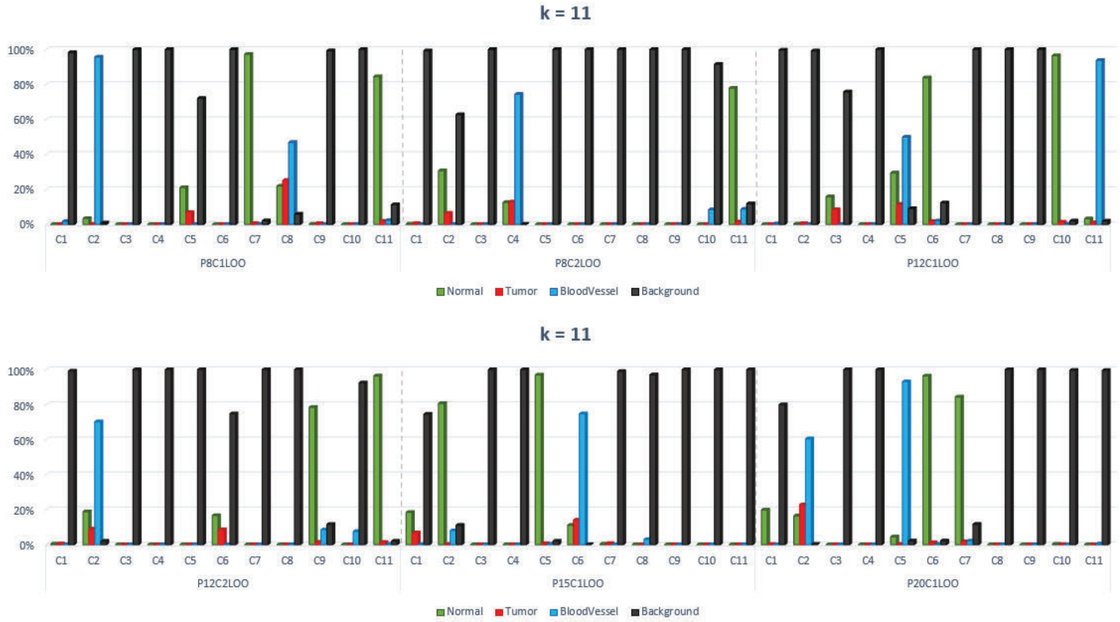


Figure 0-6. Graph of the percentages of the classes that each cluster contains (k = 11).

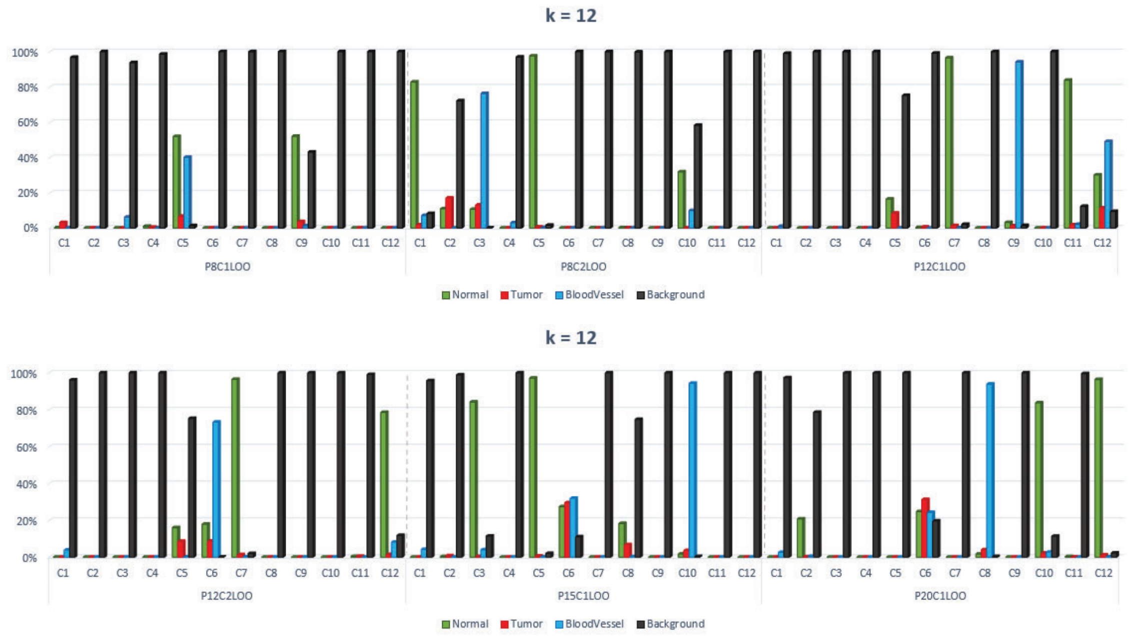


Figure 0-7. Graph of the percentages of the classes that each cluster contains (k = 12).

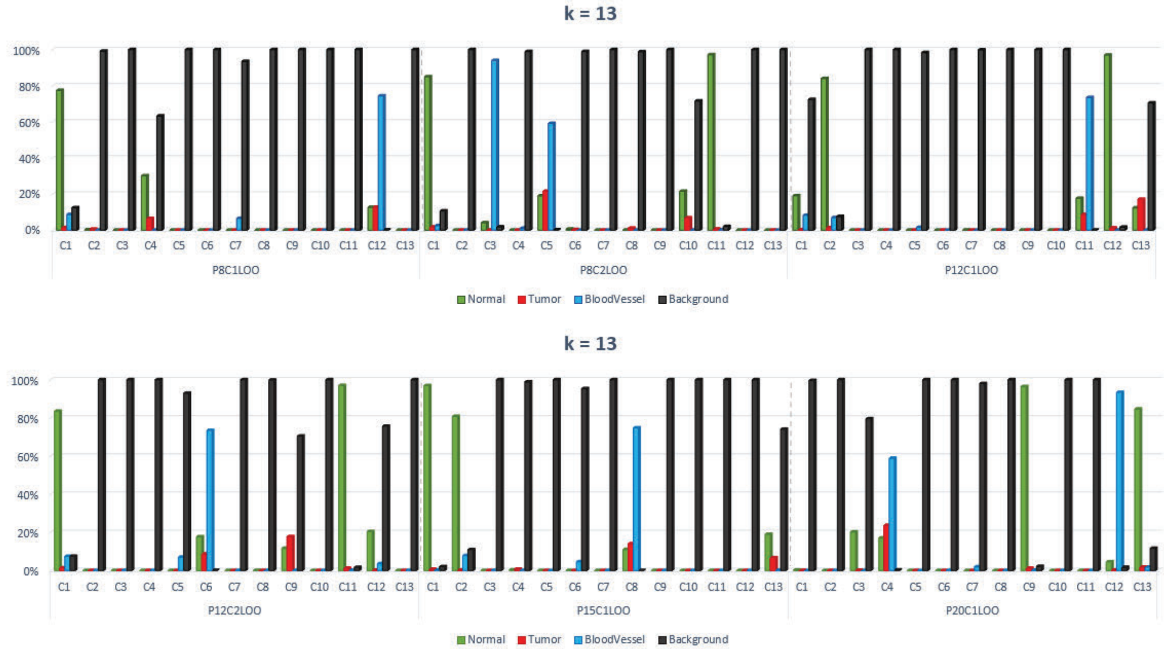


Figure 0-8. Graph of the percentages of the classes that each cluster contains (k = 13).

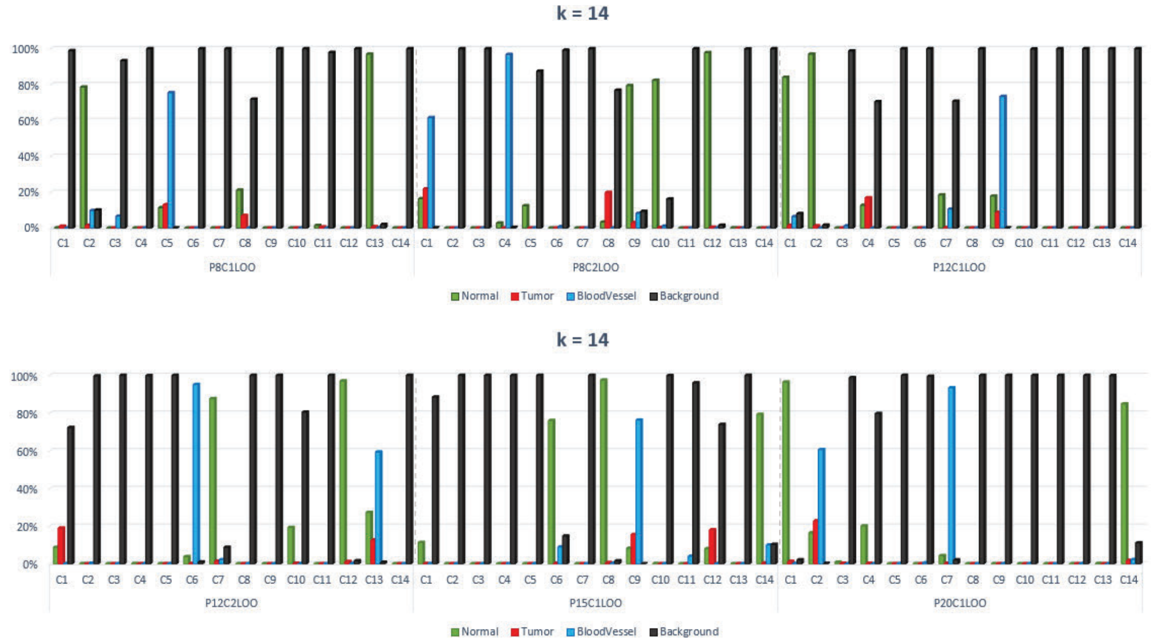


Figure 0-9. Graph of the percentages of the classes that each cluster contains (k = 14).

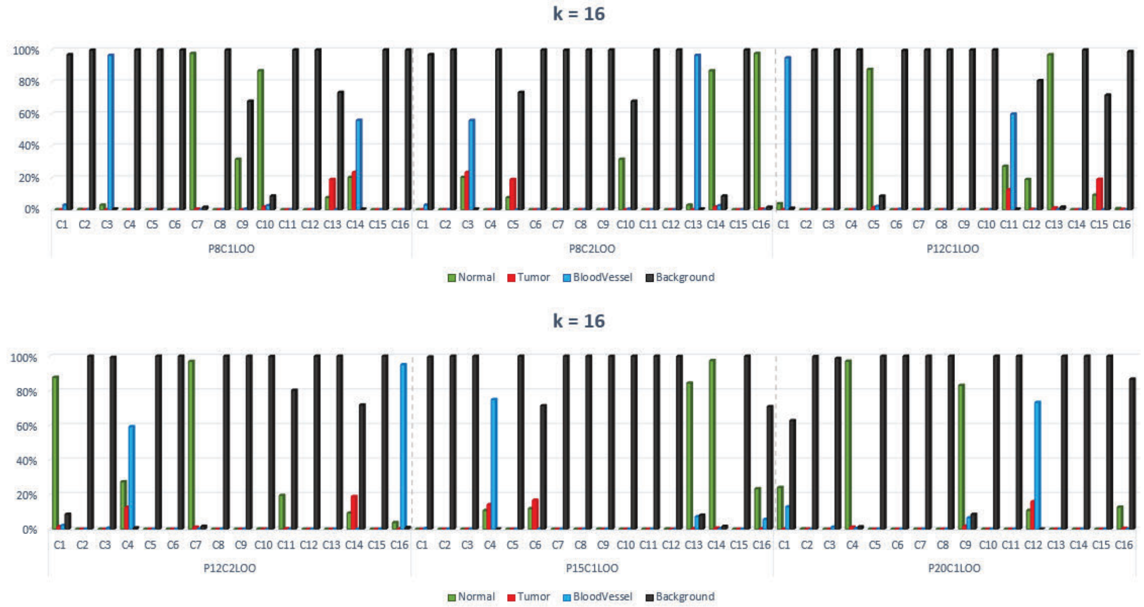


Figure 0-10. Graph of the percentages of the classes that each cluster contains (k = 16).

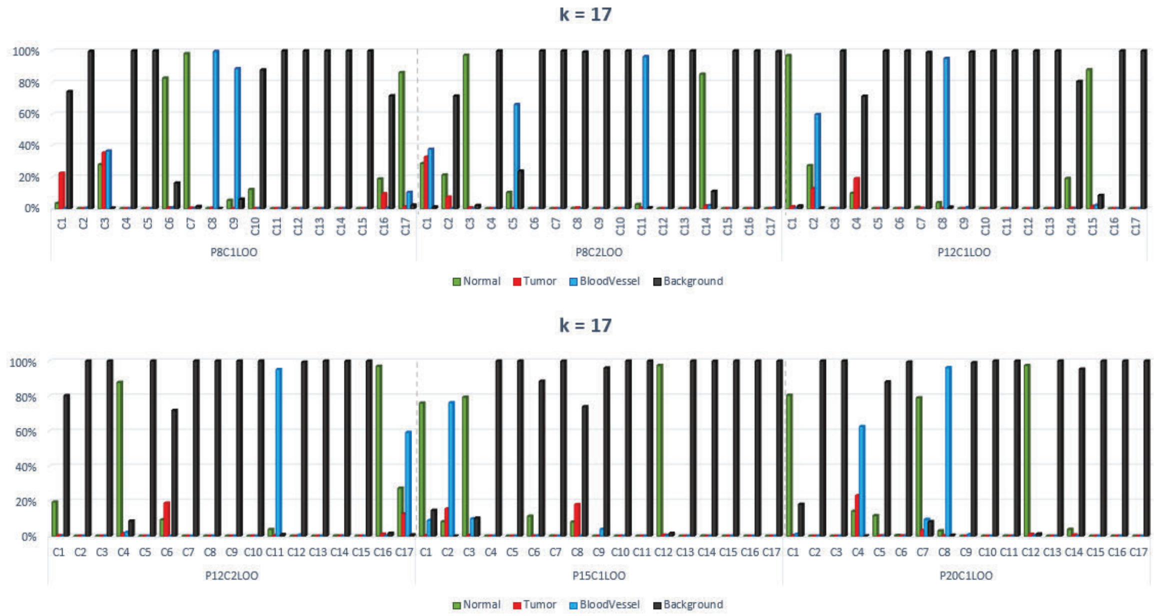


Figure 0-11. Graph of the percentages of the classes that each cluster contains (k = 17).

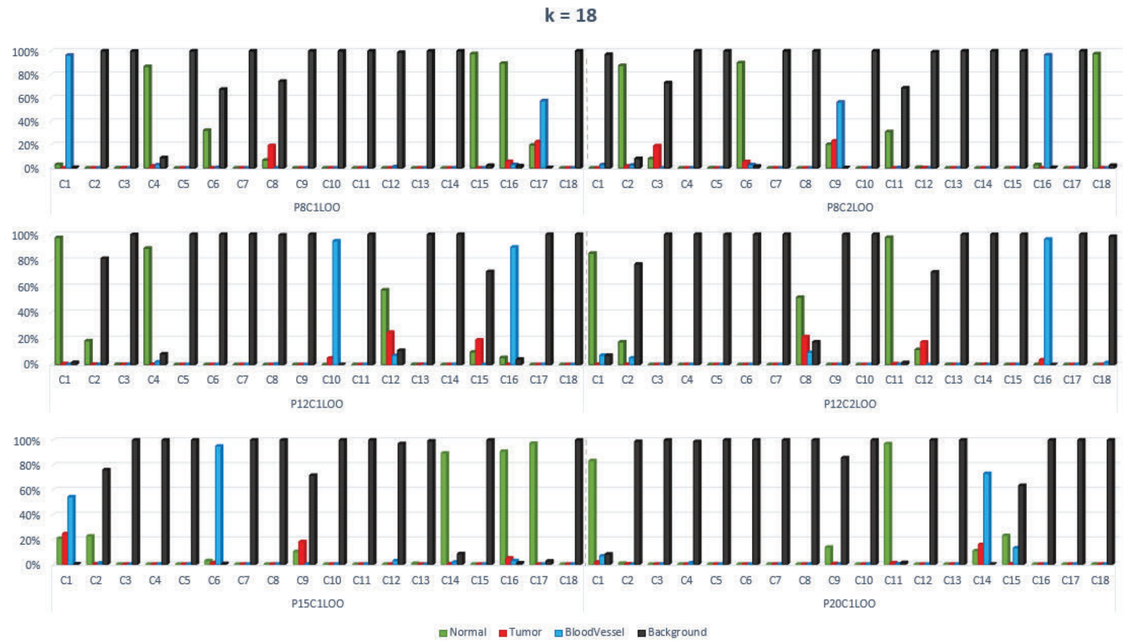


Figure 0-12. Graph of the percentages of the classes that each cluster contains (k = 18).

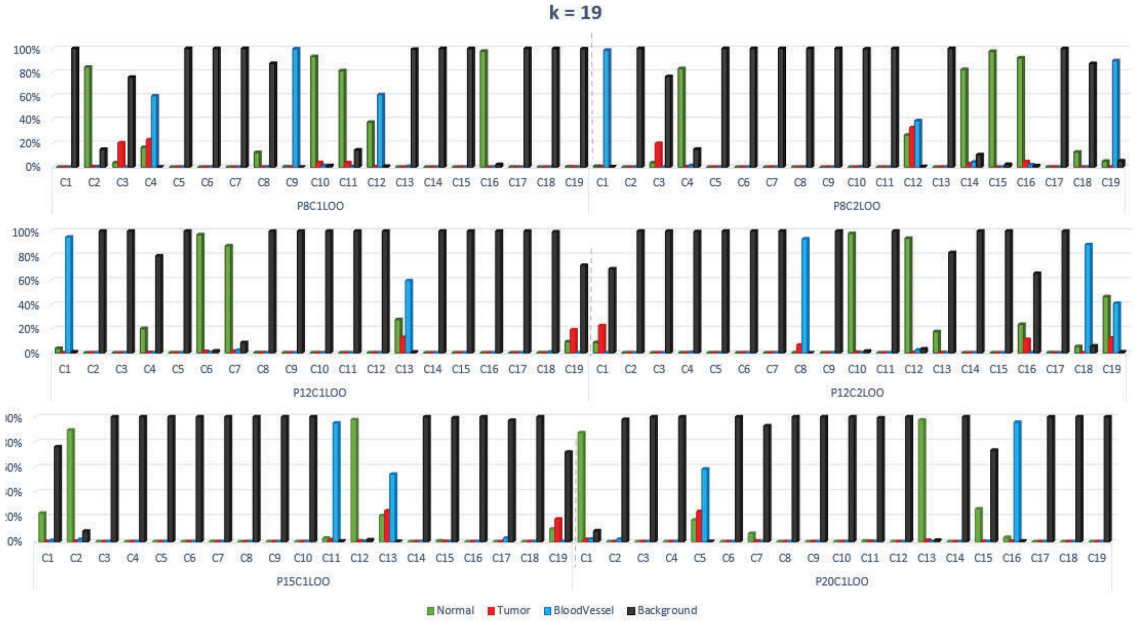


Figure 0-13. Graph of the percentages of the classes that each cluster contains (k = 19).

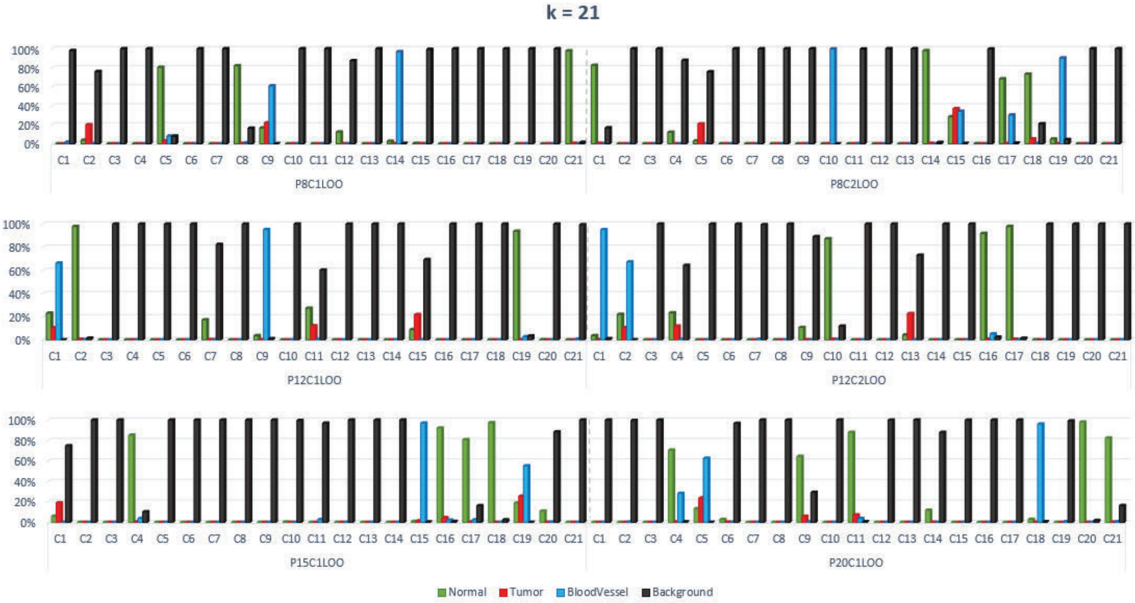


Figure 0-14. Graph of the percentages of the classes that each cluster contains (k = 21).

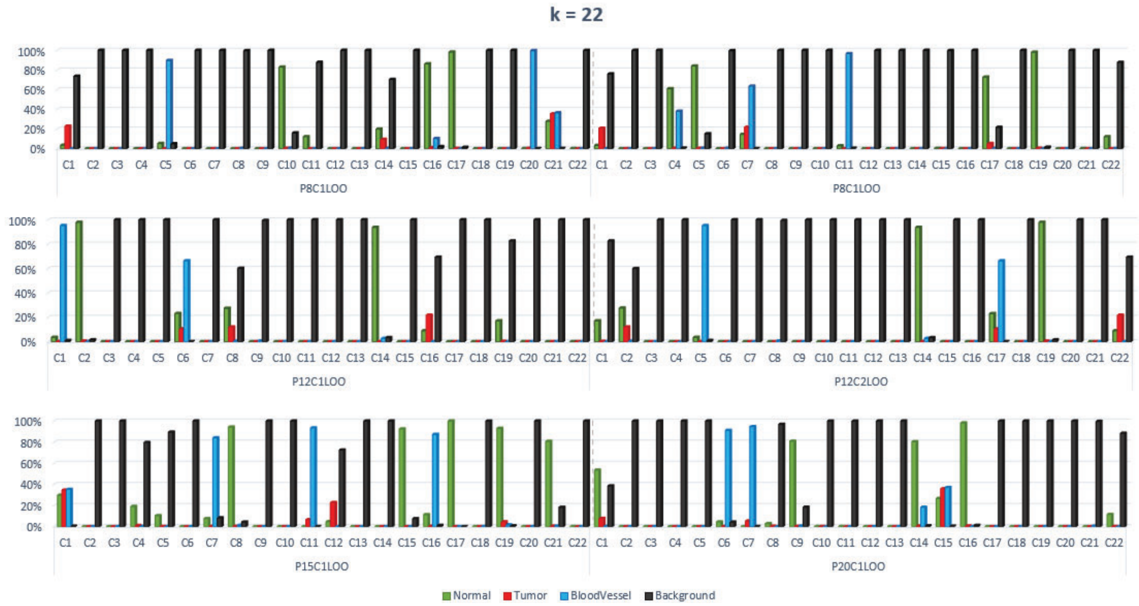


Figure 0-15. Graph of the percentages of the classes that each cluster contains (k = 22).

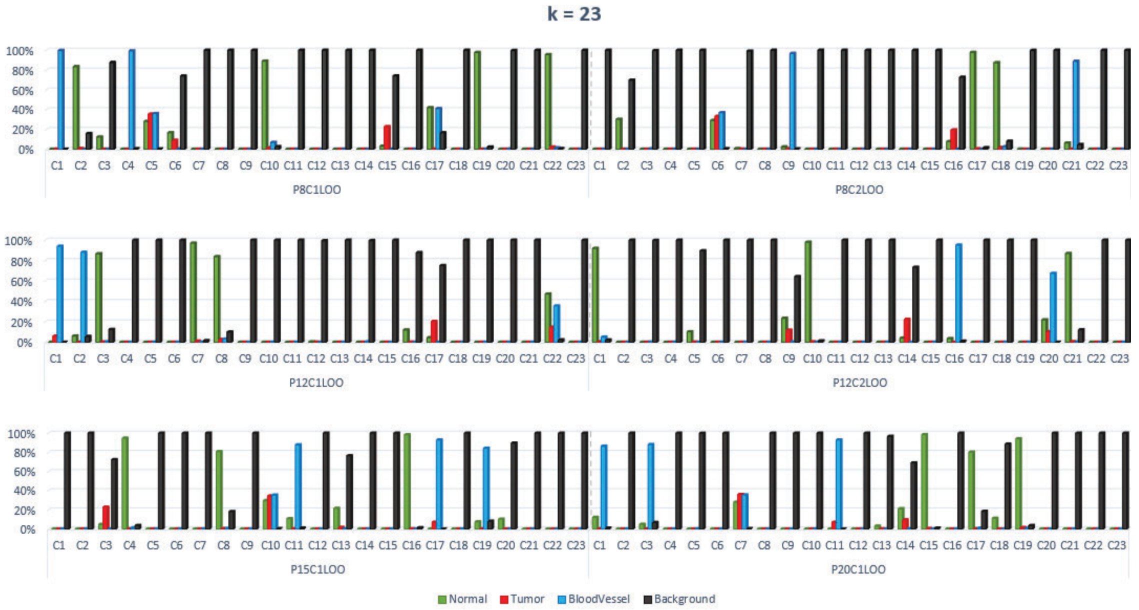


Figure 0-16. Graph of the percentages of the classes that each cluster contains (k = 23).

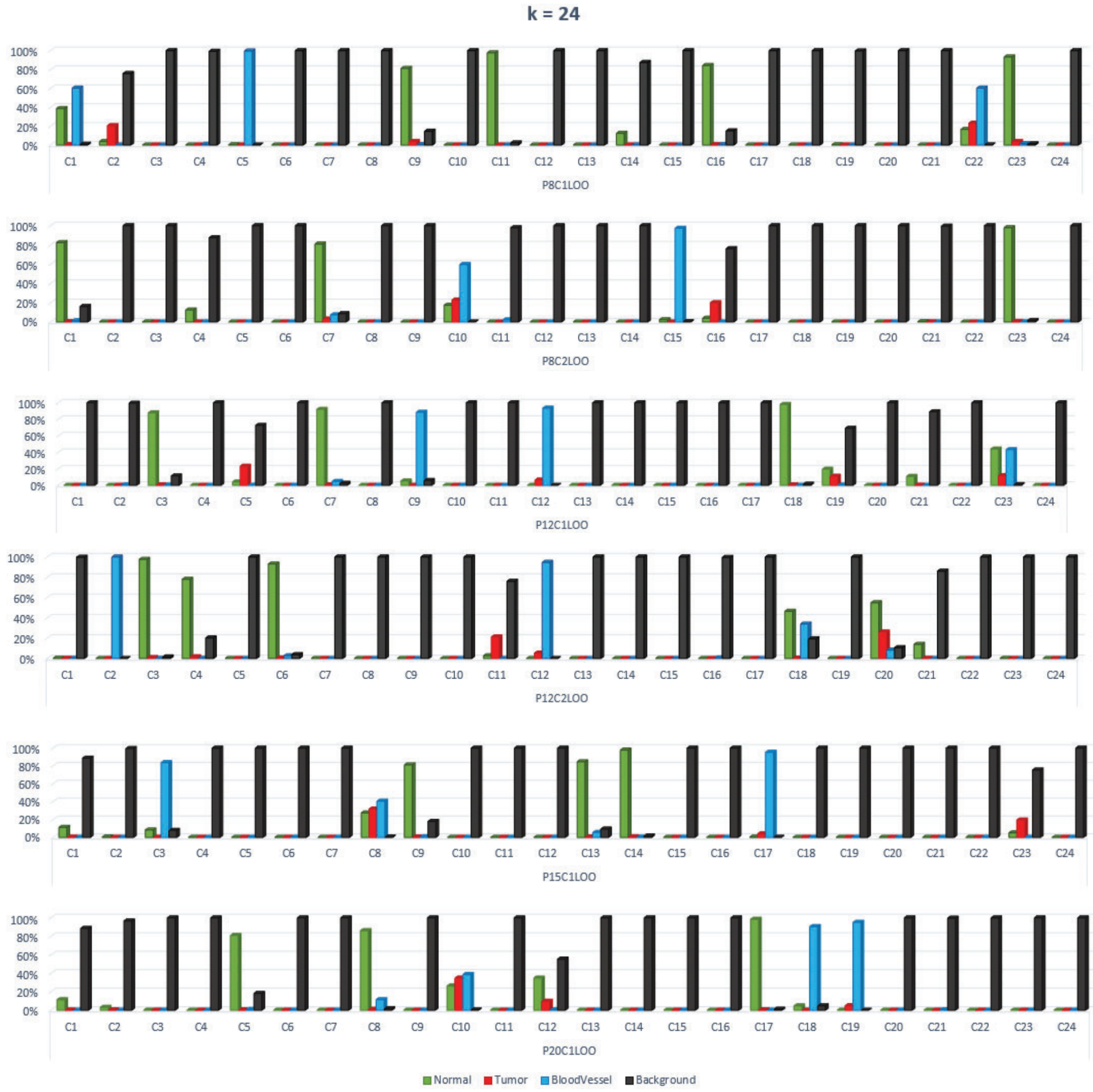


Figure 0-17. Graph of the percentages of the classes that each cluster contains (k = 24).