Universidad Autónoma de Baja California

Facultad de Ingeniería, Arquitectura y Diseño

Maestría y Doctorado en Ciencias e Ingeniería

Dimensionality Reduction in Medical Image Matrices

THESIS

presented to partially fulfill the needed requirements to obtain the degree of

MASTERS IN SCIENCE

By:

Gener José Avilés-Rodríguez

Ensenada, Baja California, July of 2018.

Universidad Autónoma de Baja California

Facultad de Ingeniería, Arquitectura y Diseño

Dimensionality Reduction in Medical Image Matrices

THESIS

presented to partially fulfill the needed requirements to obtain the degree of

MASTERS IN SCIENCE

Presented by

Gener José Avilés Rodrííguez

Approved by:

Dra. María de los Ángeles Cosío León
Principal Advisor

Dr. Miguel Enrique Martínez Rosas
Committee Member

Dr. Oscar Barroso Huertas
Committee Member

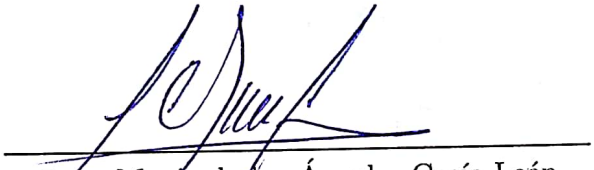
Dr. Rubén César Villarreal Sánchez
Committee Member

Ensenada, Baja California, July of 2018.

Summary of the thesis of **Gener José Avilés-Rodríguez**, presented to partially fulfill the needed requirements to obtain the degree of MASTERS IN SCIENCE from the Maestría y Doctorado en Ciencias e Ingeniería (MYDCI) from UABC. Ensenada Baja California, México, July of 2018.

Dimensionality Reduction in Medical Image Matrices

Summary Approved by:



Dra. María de los Ángeles Cosío León
Principal Thesis Advisor

In this work the process of performing dimensionality reduction of data in medical image matrices is presented, as well as a method validation adapted from the field of chemometrics to assure reproducibility of the process. A lineal approximation is proposed in order to keep clinically relevant results. The approach is generated as a transdisciplinary work between the areas of computational statistics, digital image processing and medicine.
The process and the results fulfill the definition of translational research and set the basis to further explore medical images from a translational approach.

Keywords: *Dimensionality Reduction, Medical Images, Algorithms, Computational Statistics.*

Dedication

To my wife Aghavni, for her endless patience and love. To my daughter Mané for becoming the tangible expression of the sublime. To my mother in law Hasmik for supporting us beyong expectations.

To my mother Beatriz for supporting and loving me even when everybody else did not. To my father Gener for planting the seed of an academic life in me.

To my principal advisor Dra Mary for teaching by example what it is to have a vision and be consecuent in deeds to it.

Acknowledgments

To the members of my thesis committee Dr. Miguel Enrique Martínez-Rosas, Dr. Med. Zeno L. Charles-Marcel, Dr. Oscar Barroso-Huertas, and Dr. Rubén César Villarreal-Sánchez for their guidance, feedback and support througout the realization of this project and thesis.

To my classmates, for their supportive friendship and the good times in "the battles".

To the MyDCI program in Facultad de Ingeniería, Arquitectura y Diseño from Universidad Autónoma de Baja California, for allowing a clinician to pursue trainning in the areas where his convictions took him.

To Dr. Félix David Ramón-Concepción for generously providing X-ray images of lateral projections from the cervical spine out of his private traumatology practice.

To the *Consejo Nacional de Ciencia y Tecnología* for the tuition and stipened support provided during my studies.

Contents

List of Figures

List of Tables

Listings

Chapter 1

Introduction

The coming of the era of big data, where information and communication technologies facilitate the generation of a significantly greater number of data available across many different disciplines, has also impacted the medical field. There is not a proportional amount of information derived from the growing pool of data available in the medical data [10]. The need for different approximations to data analytics in this field is very important.

One of the areas in which the phenomenon described in the previous paragraph is very tangible is with medical images. Digital images are processed as matrices of data where the 2D structure confines them particular characteristics for their analysis and interpretation.

Dimensionality reduction is one of the steps of the analytical process where areas of opportunity to facilitate further processes down the analytical pipeline can be explored.

1.1 Motivation

The exponential growth of data in the medical environment has been noticeable since the 1950's. Back then the volume of data available would duplicate every 50 years, by 1980 this trend had changed and the amount of data would double in around 7 years. By 2010, this phenomenon happened every 3.5 years and, by 2020 it is expected to happen every 0.2 years [10].

This exponential growth in medical data has a direct impact on the day to day clinical practice of health related professions [11]. The trend is not particular to health, it can be observed in every line of work where technology has facilitated data generation capabilities [12].

Paradoxically, health related professions have shown to be slower in the adoption of tools to facilitate the analysis and process of large amounts of data to generate significant information [13].

One of the areas in which technology has had a strong involvement within health is digital imaging, they are an integral part of the clinical process from prevention to

diagnosis, treatment and rehabilitation.

In the context of Electronic Medical Records, digital images are handled within the Digital Imaging and Communications in Medicine (DICOM) standard. This allows for the image matrix to be stored with a header of meta data with varied information such as name of the patient, physician, hospital, sociodemographic indicators, and technical information of the machine used to capture the image [14].

Images are usually stored using a 16-bit color depth in the DICOM standard, data contained in image matrices is of a multidimensional nature given the size of the array in which is organized and the amount of luminescence values each cell can take. Therefore the use of dimensionality reduction algorithms to preprocess medical images within the DICOM standard takes a significant role to optimize computational processing time, feature identification and extraction, and automation, while maintaining accuracy and clinical significance.

1.2 Problem Statement

1.2.1 Research Question

What is the magnitud of the effect on information pertinence generated from medical data when using dimensionality reduction algorithms, during the process of feature extraction?

1.2.2 Hypothesis

To use dimensionality reduction algorithms during the process of feature extraction in medical data raises information pertinence and allows for the increment on detection of areas of interest.

1.2.3 General Objective

To measure the effect of the most relevant dimensionality reduction algorithms in the state of the art, over the pertinence of information generated in the medical and biomedical context.

1.2.4 Specific Objectives

1. To determine the most commonly used dimensionality reduction algorithms for data analytics in the clinical and biomedical context.

2. To analyze dimensionality reduction algorithms selected from the literature and metrics to quantify the effect over pertinence of information.

3. To apply the algorithms selected to data from the medical and biomedical contexts.

4. Quantify the effect on pertinence of information in clinical and biomedical data.

1.3 Structure of thesis

This work is divided into the following chapters:

Chapter 1 A brief introduction to the thesis project is given in this chapter, a justification for the approach taken is presented, as well as the problem statement and the sequence in content for the following chapters.

Chapter 2 Chapter 2 starts with a review of dimensionality reduction techniques in medical image matrices, covering key concepts and definitions. Then it goes on to a review of the literature from 2007 to 2017 where the most published algorithms used in medical and biomedical data are shown. This with the goal of narrowing down the amount of algorithms to study in depth for the time of this project.

Chapter 3 In this chapter the method for dimensionality reduction of Principal Components Analysis is covered in depth, an explanation of the process of dimensionality reduction achieved in lateral projections of the cervical spine in X-ray images is detailed and the model developed is presented.

Chapter 4 This chapter shows and discusses the results obtained by applying the model presented in Chapter 3, it covers the technical details of implementing dimensionality reduction with a lineal approximation, advantages and disadvantages of the process.

Chapter 5 Finally, conclusions are presented in this chapter, as well as future work and lines of research to be implemented as follow up.

Chapter 2

Review of the Literature

This chapter presents the results of a literature review in dimensionality reduction techniques used in medical images. Then it elaborates on algorithms used to reduce dimensionality in medical and biomedical data. Finally it explores the use of Principal Components Analysis (PCA) given that this was the method selected in the experimental stages of this project.

2.1 Image Matrices in Medicine

Medical images are central to clinical day to day practice, from prevention [15], diagnostics [16], treatment [17] and rehabilitation [18] of patients. Nevertheless, with the coming of the era of big data to health care, the amount of medical images generated and stored in electronic health records has considerably grown in the last years, and this tendency is expected to stay into the foreseeable future [19].

Clinicians and medical administrators are being faced with the challenge of having large amounts of data from medical images, and not having enough human resources to process them into significant information for the benefit of their patients [20, 10].

A way to ameliorate this situation is to make use of computational statistics and digital imaging processing techniques, to pursue the implementation of Computer-aided Diagnosis (CAD) approaches [21]. This facilitates automation and standardization to the analysis of medical images, but brings it's own challenges with it.

In the context of image matrices, the effect of applying dimensionality reduction can be appreciated in the reduction of the cardinality of grayscale values, referred to as depth of luminescence in the area of digital image processing [8]. This means that a vector with the possible values that a given pixel could take can be reduced in length [22].

2.2 Dimensionality Reduction in Medical Images

2.2.1 Approximations to Dimensionality reduction

Dimensionality reduction is understood as the transformation of high-dimensional data into a meaningful representative subset that, ideally, keeps the minimum number of parameters needed to account for the observed properties of the data [23].

To a computer, images are data arranged in a matrix format of $n \times m$ dimensions, therefore numerical and algorithmic methods can be applied since the essence of the image is still raw data. Dimensionality reduction is a component of the analytical process, independent of the domain to which the data belongs to.

Linear approximations to reduce dimensionality are those in which the methods used comply with the superposition property, which means that the total result is equal to the sum of the individual contributions in the system or database [24]. Nonlinear approximations use methods in which the superposition property is not met.

Linear Approximations

Given a set of data graphed in a two dimensional space, we can represent the values with a line determined by a first degree polynomial of the form $y = ax + b$, also called a linear polynomial [25].

Techniques that seek to represent the dynamics of a given set of data with linear polynomials are called linear methods. There are algorithms designed with this concept as a key premise in their approach [25], in the next paragraphs we discuss a relevant subset mentioned in the literature.

Principal Components Analysis (PCA).

PCA is considered to be the most commonly used dimensionality reduction algorithm in image processing [1]. It has the quality of retaining most of the variation in the data set [26].

The previous is accomplished through the generation of new variables, called principal components, which represent the planes in which the variation in the data is maximal [27].

These components are found by determining orthogonal planes where maximal variance of the dataset is found, through linear algebra approximations. Smith an collaborators give a clear and comprehensive approximation to the mathematics of PCA [28].

Figure 2.1 shows a typical approach to a database with high multidimensionality, where the first two principal components are organized as a 2D array and visualized in a graph.

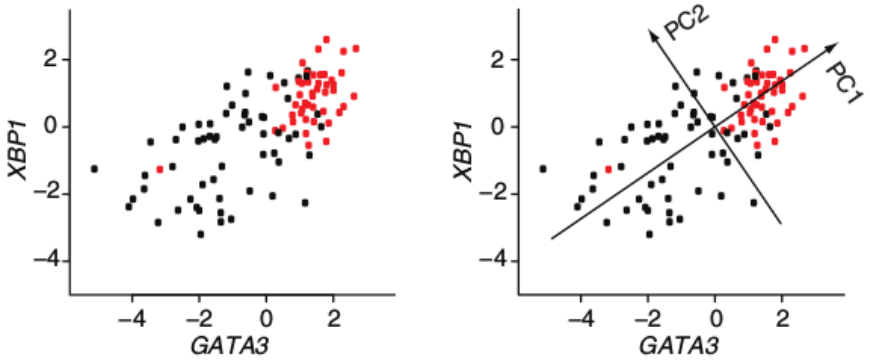


Figure 2.1: Graph of the 2 first principal components of a genome-wide expression database analysis.These studies are typical for their multidimensionality. Dots in the graph represent breast cancer tissue samples against its expression levels of two genes, by [26].

Principal components can not be correlated to the individual variables in the original database, they are the product of a feature-generation process which maintains the majority of the variance in the original database.

Even though this facilitates the use of analysis techniques in a lower dimension database to be extrapolated to the original multidimensional set, the spatial relationship in the original image matrix is lost in the process.

Figure 2.2 shows Computed Tomography (CT) images of scans of the thorax reduced with PCA, published by [1].

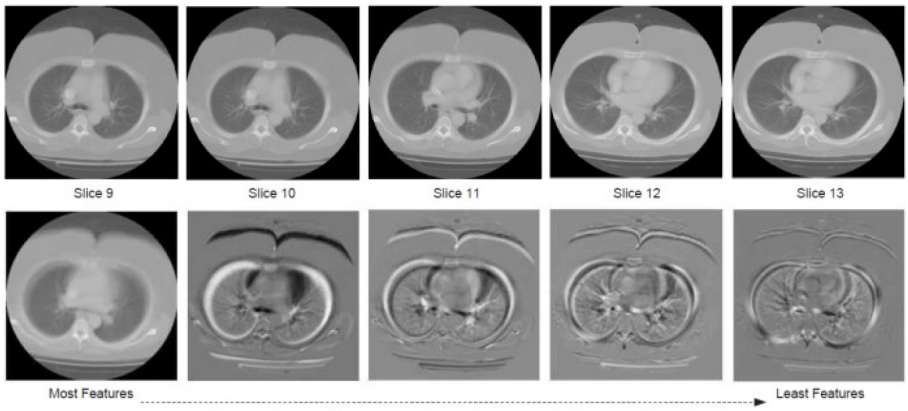


Figure 2.2: Images from Computed Tomography reduced by PCA, [1].

Canonical Contextual Correlation Projection (CCCP)

When we approach pixel based image segmentation in terms of the given characteristics of each pixel, an organization of possible classes can be generated around these pixel features. An example could be the luminescence value in a grayscale image, or the neighboring data.

CCCP is an algorithm that performs well given the circumstances above, it was developed from Linear Discriminant Analysis, it is important to understand this concept if the situation requires the researcher or clinician to pursue a deeper understanding of the functioning of the algorithm. Algorithm 1 shows the steps CCCP takes to accomplish dimensionality reduction.

Linear Discriminant Analysis (LDA) has been proven to have good performances segmenting image data without considering spatial labels. An example can be found in the work by Caan and collaborators [29]. As mentioned above, CCCP is considered an extension from LDA [2], it adds to it the capability of keeping spatial relations, context and labeling in the reduced data. This feature makes CCCP attractive for image analysis. Algorithm 1 shows the approach to CCCP used by [2].

Algorithm 1: CCCP Algorithm

Require: Image matrix X from n x d dimensions.

Output: Reduced image matrix X' with d dimensions.

1 Initialization;
2 Define what image feature information to use and which neighboring pixels to take for the class label content.;
3 Determine from the train images and associated segmentations the gray level feature vectors x_i;
4 Determine from the same data the class label feature vectors y_i;
5 Estimate the covariance matrices S_{XX}, S_{XY} and S_{YY};
6 Eigenvalue decomposition of matrix S_X where: $S_X = S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{XY}^t$;
7 Take the d rows of the d x n linear dimension reducing Transformation matrix L equal to the d eigenvectors associated to the d largest eigenvalues;
8 transform all x_i using L to Lx_i;

Output: Reduced image matrix X' with d dimensions.

Figure 2.3 shows the result of applying CCCP to a chest radiography, the algorithm is able to find clusters of pixels grouped by characteristics shared in the input and output, thus keeping spatial and contextual relations.

Population Value Decomposition (PVD)

This algorithm finds common features in a set of matrices of images, which become relevant when the amount of images becomes relatively large. Medical images are usually managed within electronic health records, thus a good example of the challenge of a high amount of images in one repository.

A mathematical representation was published by [30], from which some generalities are mention in the following paragraph:

$$Y_i \simeq PV_iD, \tag{2.1}$$

Where P and D are population specific matrices of size $F \times A$ and $B \times T$ respectively.

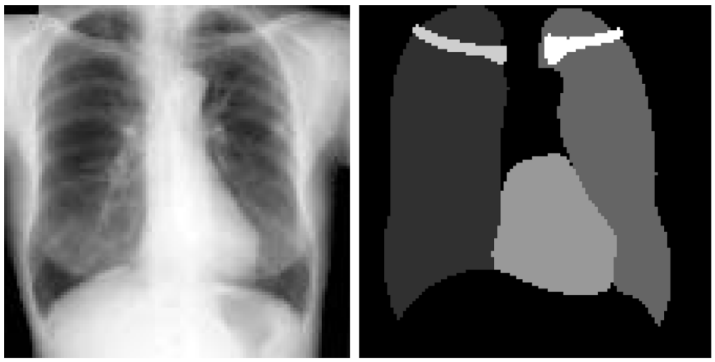


Figure 2.3: Images from Computed Tomography analyzed by CCCP, [2].

If A or B are much significantly smaller than F and T then Equation 2.1 generates a useful representation of the set of images. V_i is the reduced matrix and P and D provide a reference frame for the analysis.

2.2.2 Non-Linear approximations

Discrete Fourier Transform (DFT)

This approach can be used when the border of a structure in an image has a cyclic and closed pattern.

Considering the outline to be a discrete-time complex periodical signal $z = \langle z_0, z_1, ..., z_{N-1} \rangle$ define $z_1 = x_l + iy_l$ $(i = \sqrt{-1})$, where x_l and y_l are the coordinate values of the lth $(l = 0, ..., N-1)$ sampled points. Consequently, z can be mapped to the frequency domain by the Discrete Fourier Transform [31], as shown in equation 2.2.

$$\begin{gathered} Z_m = \sum_{l=0}^{N-1} z_l e^{-i(2\pi lm/N)} = R_m e^{i\theta_m}, \\ \left(m = -\frac{N}{2}, ..., -1, 0, 1, ..., \frac{N}{2} - 1 \right). \end{gathered} \tag{2.2}$$

Where R_m and θ_m are the module and argument of the mth DFT coefficient. Then we obtain dimensionality reduction by selecting only the low frequency DFT coefficients. Techniques used to perform selection of these coefficients are explored in [32].

Figure 2.4 shows a simple application of this technique to the sketch of a heart.

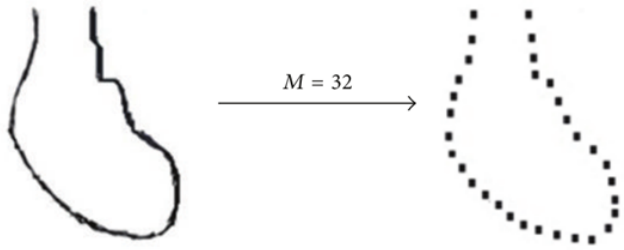


Figure 2.4: Reduced outline of a heart to 32 points using Discrete Fourier Transform, [32].

Sparse Eigenvectors

As we mentioned at the beginning of this section, dimensionality reduction by PCA has setbacks when dealing with data from image matrices because it causes pixels to lose their spatial relations. Thus, neighborhood information is not available with these approximations.

Sparse eigenanatomy deals with these problems by enforcing a sparsity constrain on the matrix decomposition.

The eigenanatomy function can be mathematically expressed as shown in equation 2.3:

$$\begin{aligned} \underset{U,V}{\arg\min} \; |||X - UV^T||_2^2 + smooth(G^*(V)) + cluster(V) \\ subject\ to\ ||V||_0 < \gamma, V \succeq 0 \end{aligned} \tag{2.3}$$

Where γ is the desired level of sparseness, the ℓ_0 "norm" operator $||*||_0$ the number of nonzeros entries in the argument, and G^* is the manifold in image space representing the geometric object [3]. Figure 2.5 shows the results of implementing PCA, Independent Component Analysis and eigenanatomy to a Computerized Axial Tomography (CAT) series of the brain.

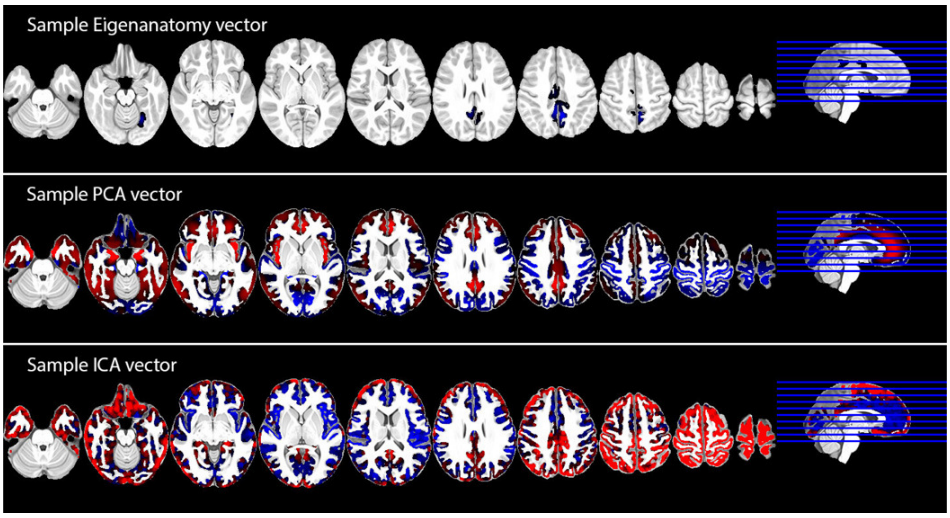


Figure 2.5: PCA and Independent Component Analysis (blue for positives, red for negatives) compared to eigenanatomy returning unsigned results and allowing for spatial based analysis, [3].

Isometric Feature Mapping (Isomap)

When high-dimensional data lies on a curved manifold, euclidean distance does not deliver an accurate measuring of neighboring data points; this can cause the method to consider two points as close to each other when their position on the curved lattice could be greater. This is overcome by Isomaps by measuring by geodesic distances (GD) instead. In graph theory, the number of edges in the shortest path connecting two specific vertices is known as geodesic distance [33].

The shortest-path algorithm by Dijkstra is usually the one implemented to calculate GD [34].

Locally Linear Embedding (LLE)

This method contrasts with Isomaps in which the first preserves the local properties of data. LLE is used for data embedded in non-convex manifolds. This method constructs a linear mapping of the hyperplane to a space of lower dimensionality by elaborating on the rationale that the global manifold can be reconstructed by small connecting regions that overlap with each other.

This causes the cost function shown in equation 2.4 to be minimized for x_i finding a reduced dimensional representation of x:

$$\varepsilon(W) = \sum_{i=1}^{n} \|X_i - \sum_{j=1}^{k} W_{ij} X_{ij}\|^2, \tag{2.4}$$

where X is the input data, n is the number of points, and k, neighborhood size. This equation is subject to two constraints:

$\sum_{j=1}^{k} w_{ij} = 1$ and $w_{ij} = 0$ when $x_j \notin R^{D(images)}$ [35].

Diffusion Maps (DfM)

This algorithm defines a Markov random walk on a Laplacian graph generated from the data, as a result finding the subspace that best preserves diffusion interpoint distances.

Weights (K) of the edges in the graphs are calculated with a Gaussian kernel function then normalized and organized in a matrix K. Finally diffusion distance is calculated on K [36].

Diffusion distance performs better with noise than geodesic distance because it uses more than one path throughout the graph to obtain the reduced image.

2.3 Dimensionality Reduction in Segmentation of Medical Images

According to Oge Marques [8], image segmentation is defined as "the process of partitioning an image into a set of nonoverlapping regions whose union is the entire image". The components of the set should, in an ideal scenario, correspond with meaningful regions as interpreted by a clinician.

The following paragraphs discuss how dimensionality reduction is used for 2D and 3D segmentation of medical images.

2.3.1 2D Images

Images in two dimensions are the core of medical praxis, higher dimension images are reconstructions from 2D image matrices.

Medical images are stored and shared using the Digital Imaging and Communications in Medicine (DICOM) international standard [14]. Usually carrying a 16-bit (2^{16}) luminescence depth. These characteristics tend to translate into high dimensionality in their image matrices. This is why dimensionality reduction in medical images is a key step in the analytical process.

Principal Componets Analysis

Irem Ersöz Kaya and collaborators have used PCA and variations of it as a preprocess (Probabilistic Principal Components Analysis– PPCA, Expectation Maximization Based Principal Components Analysis– EM-PCA, Generalized Hebbian Algorithm– GHA, Adaptive Principal Component Extractor– APEX) to reduce dimensionality in 2D images to ready them for a clustering stage [4].

Figure 2.6 shows some of their results with each of the 5 approximations.

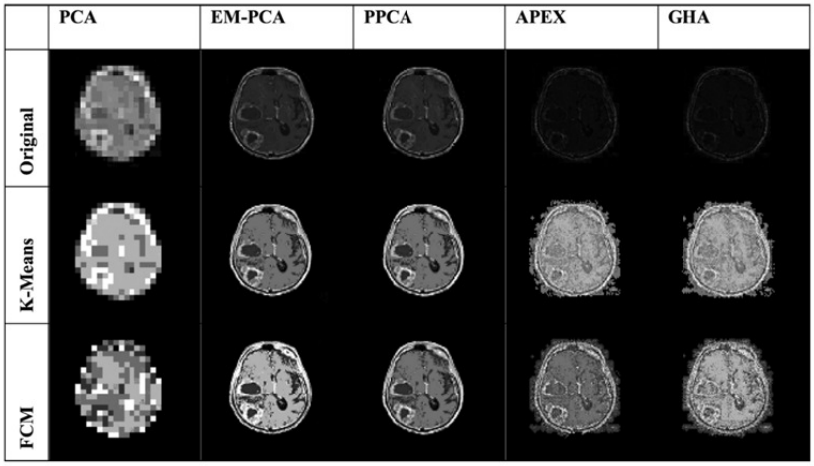


Figure 2.6: PCA algorithms used to reduce dimensionality as a preprocess for image clustering in tomographies with brain tumors, [4]

Artificial Neural Networks (ANN)

This bioinspired information processing paradigm encompasses algorithms that share a set of basic structures: perceptrons, inputs, outputs, weights and a mechanism to reduce error (which gives them the ability to produce results closer to the ideal) [5].

ANNs were proposed as an "unsupervised adaptive resolution clustering" algorithm in 1996 by A. Schenone and collaborators [37]. Eventhough computational power was very limited, they proposed this algorithm as a potentially useful tool in the dimensionality reduction steps of medical images.

Nowadays, with substantially greater computational processing power, ANNs have taken a key role not only in the preprocess but also the analytical steps of medical image studies.

Amin and collaborators [5] proposed a neural network where PCA is the main component and is used as a feature extraction system. In their approach eigenvalues of the Magnetic Resonance Images (MRI) analyzed were obtained and fed to a Multi Layer Perceptron (MLP) network to be classified into similar cases. They called their system a PCA-MLP network and it follows the structure of a hybrid ANN, as shown in Figure 2.7.

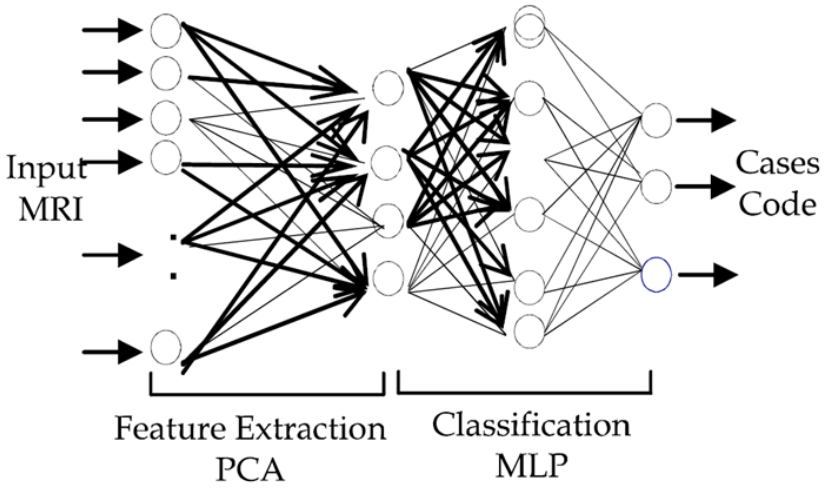


Figure 2.7: Structure of a PCA-MLP network as proposed by [5]

2.3.2 3D Images

Medical images in three dimensions are reconstructions from 2D projections [6]. Image matrices from CAT studies and Magnetic Resonance Imaging (MRIs) are the most commonly used approaches to generate 3D projections [38]. A basic process of 3D reconstruction is shown in Figure 2.8.

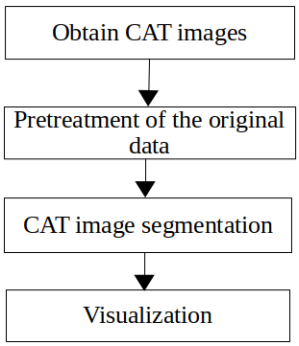


Figure 2.8: The process for 3D reconstruction. As mentioned by [6]

CATs and MRIs sample slices of variable thickness from the patient representing the result in a 2D image, in a subsequent step all the images obtained are stacked to form a 3D volume (voxels), an example of the result is shown in Figure 2.9.

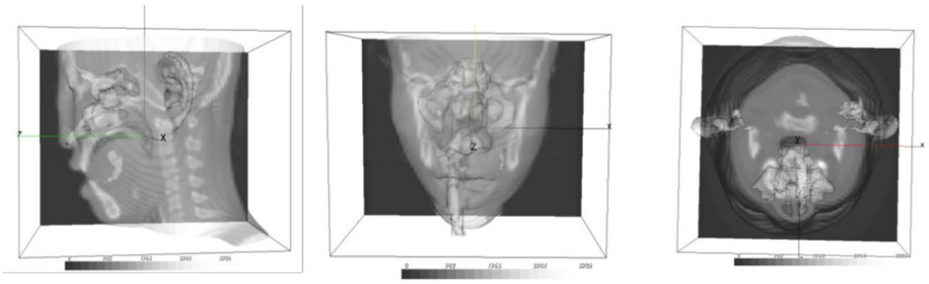


Figure 2.9: Three different perspectives in 2D from an MRI with matching x,y and z coordinates to form a hypercube. As shown by [6]

Once a 3D model is obtained the following analysis will be focused on segmentation techniques which will allow to identify areas of interest with clinical meaning for the specific case.

Methods of dimensionality reduction employed in 2D images are used to facilitate the analytical process and lighten the burden of computational cost of the process [39].

2.4 Dimensionality Reduction and Content Retrieval in Medical Images

As mentioned in section 1.1, we are able of generating and collecting a significant amount of data, including images, through Electronic Health Records (EHR) and digitalized medical imaging systems [40].

When a set of medical images is stored in an EHR. One of the most challenging tasks is to search images or their characteristics rapidly, accurately, and efficiently, this process is called content retrieval [41].

Techniques based on feature detection have been used and proven extensively in the area of digital image processing, lately implemented for content retrieval in medical image systems. Despite their robustness when it comes to the identification of geometric forms and elements, they lack the ability to interact with the clinician in a smoother manner. This capability is known as semantic content based retrieval.

Feature Based

When a machine reads a matrix image, it analyzes it in function of the values of each pixel available [40]. After performing a preprocess, feature selection or extraction is the natural next step [42], this sets the basis for higher processes as segmentation, shown in Figure 2.8.

The fundamental approach in feature based content retrieval (FBCR) is the identification of structures by a reduction to geometrical forms. In as much as the latter can be expressed in mathematical notation a computer can be programed to identify them.

Once the fundamental geometric elements of interest are identified in the image, a process of correlating these findings with clinical areas of interest can be performed [43]. This analysis prepares the approximation for higher level content retrieval where semantics play key role.

Semantic Content Based

Computers and humans interpret images very differently. A machine analyzes an image matrix and gives low level numerical interpretations of the pixels, whereas a human can naturally describe the overall topology, regions of interest (ROI) and associations within the same image and with others. This notable difference between the analytical capabilities of a computer and a human being is called the semantic gap [41].

Most of the search engines available today rely on metadata to perform image retrieval. This approach matches keywords from the queries in the input to a subset of tags in the image database, which were previously added by humans. This approach is known as Textual-based Image Retrieval (TBIR) [41].

A significant challenge with TBIR is that it depends on the metadata of the image, assigning these tags is a job done by humans and how it is accomplished varies from individual to individual, hence the results in a TBIR-based system are not always within the semantic context of the query.

In order to achieve better results in image and content retrieval it is important to go beyond analyzing the numerical description of the pixels of an image matrix.

As Smeulders and collaborators express it: "pictures have to bee seen and searched as pictures: by objects, by style, by purpose" [7]. This approach is known as Content Based Image Retrieval (CBIR) and it uses semantics as a key indicator to execute performance. Figure 2.10 shows an example of a CBIR approximation to an image.

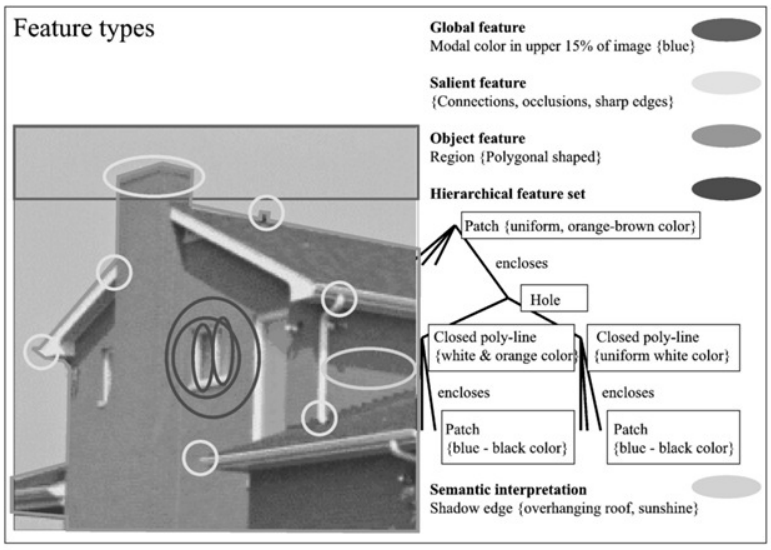


Figure 2.10: Example of a Content Based Image Retrieval by [7].

Table 2.1: Algorithms and Their Use in Dimensionality Reduction-related Tasks

ALGORITHM	TYPE	USE
PCA	Linear	Preprocess, reduction of depth of luminosity [1], [44], [45].
CCCP	Linear	Contextual value with input and output data, considers pixel neighborhood, keeps idea of spatial pixel relations [2].
PVD	Non Linear	Finds common features in a set of images [30].
DFT	Non Linear	For images with cyclic and closed patterns[32].
Sparse Eigenvectors	Non Linear	It respects the neighborhood through a sparsity constrain [46].
Isomap	Non Linear	When the data is distributed in a curved manifold, it uses geodesic distance [3].
LLE	Non Linear	Does not preserve local properties of data, preferred for data in non-convex manifolds [46], [22].

The area of CBIR remains an active field of study by the Digital Image Processing community. Table 2.1 summarizes the algorithms mentioned in this text, their mathematical approach and when they are more suitable for implementation.

2.5 Dimensionality Reduction in Medical and Biomedical Data

Even though image matrices are data organized in 2D arrays, a tendency was noted in the literature where specific methods of analysis are used for images in comparison to other kind of data structures. This prompted an interest in exploring the algorithms used to analyze medical and biomedical data not arranged in image matrices.

Figure 2.11 shows the most cited algorithms found in a time interval from 2007 to 2017 in the scientific literature.

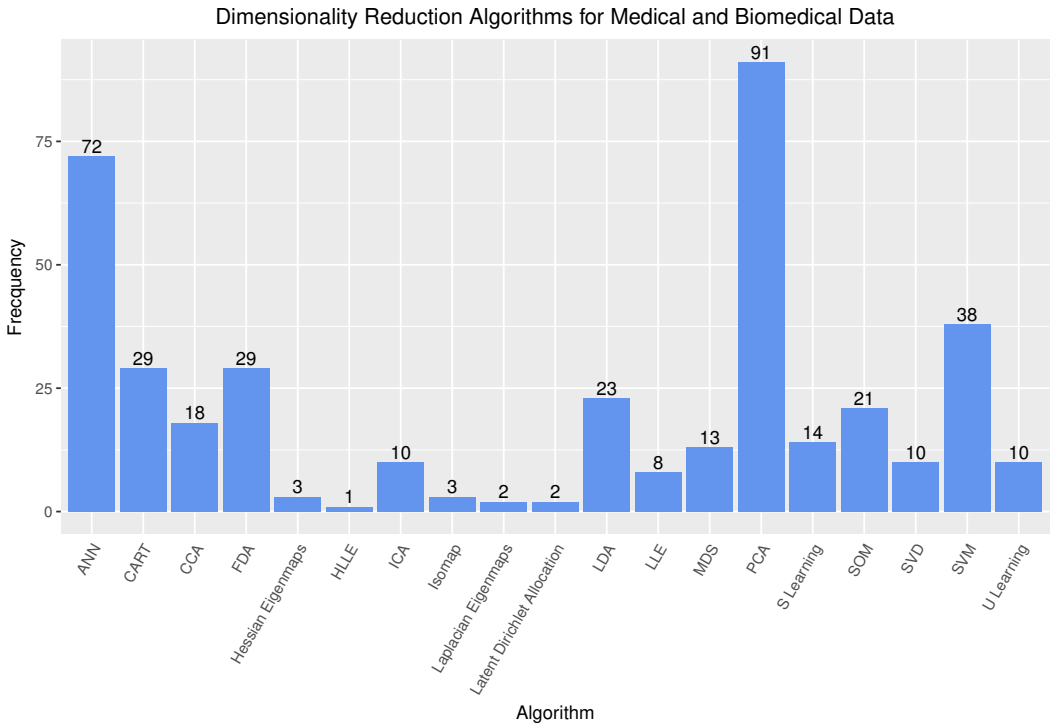


Figure 2.11: Frequency of appearance in literature review of algorithms used for dimensionality reduction in the context of medical and biomedical data.

As it can be seen, the most published algorithms found in the review stage were:

1. Principal Components Analysis,
2. Artificial Neural Networks and,
3. Support Vector Machines.

Once these three algorithms were selected it became important to find out if the overall prevalence in publications was similar to the tendency throughout the period of time selected. Figure 2.14 shows the amount of publications per year of the five most published algorithms from Figure 2.11.

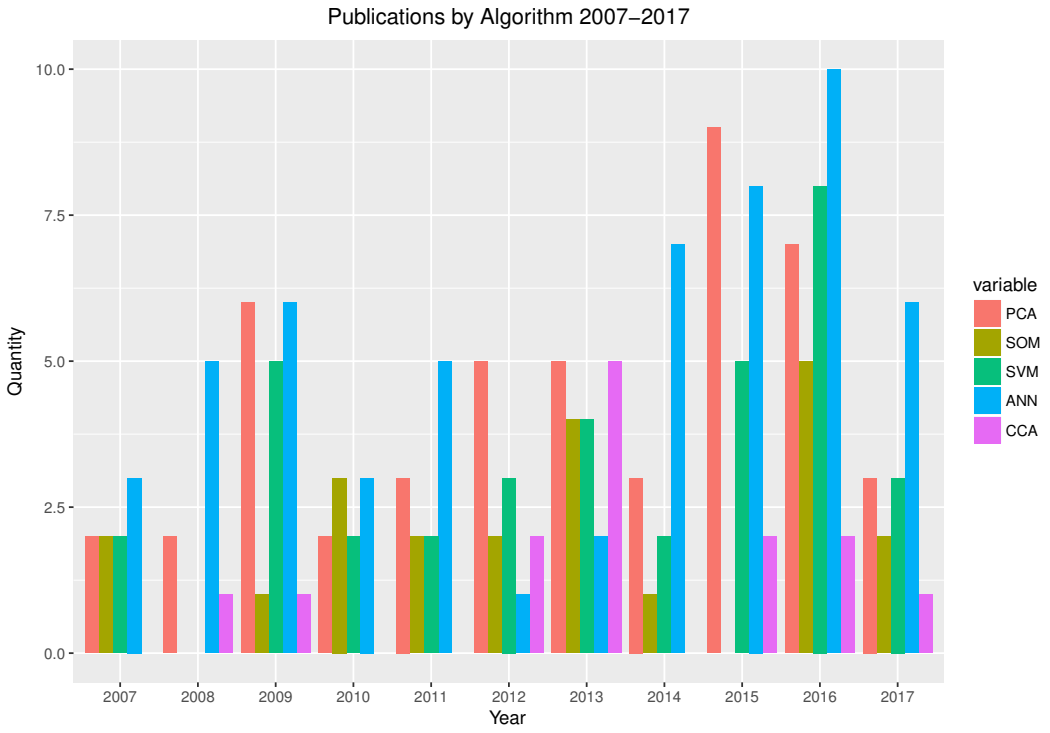


Figure 2.12: Amount of Publications by Year per Algorithm.

Looking at Figure 2.14 a clear tendency from years 2013 to 2016 can be identified where Artificial Neural Networks are the most published algorithm in this context, followed by Principal Components Analysis and Support Vector Machines.

Since this project was carried out in the context of medical data, an area where a clear explanation of how results are obtained in order to make them reproducible to more patients or populations, black box algorithms [47] were not prioritized. Once this criterion was considered, Principal Components Analysis was selected to be used in the phase of mathematical explorations for this project.

2.6 Digital Image Processing

2.6.1 Generalities of Digital Images

Understanding how an image is processed in a digital environment is fundamental to be able to use the classic approaches used to perform analysis in the area of digital image processing.

Image Formation

Any image (natural or digital) requires a source of illumination to be produced, this is mathematically expressed with the greek letter λ (lambda) and is interpreted as the

wavelength of the ray of light involved in the generation of the image [48]. Therefore an incident ray of light on a specific point generating the elements needed for an image is expressed in equation 2.5:

$$E(x, y, z, \lambda) \tag{2.5}$$

Where E represents the light falling on a specific point, x, y, and z the coordinates in space. Furthermore, once light incises, the object tends to reflect some of it. This phenomenon is expressed by the reflectivity function shown in the following equation:

$$r(x, y, z, \lambda) \tag{2.6}$$

Where r represents the reflectivity of an object. This light reflected on a point can be captured by some kind of image device, represented in the following equation by c:

$$c(x, y, z, \lambda) = E(x, y, z, \lambda) \times r(x, y, z, \lambda) \tag{2.7}$$

While the device receives data from the processes represented in equations 2.5 and 2.6, the image generated by it has a significant difference from the previous steps in which it is a 2D representation of a 3D phenomenon as shown on Figure 2.13. This process is, intrinsically a dimensionality reduction strategy to represent an 3D environment in a 2D expression, capable of being expressed in a numerical matrix.

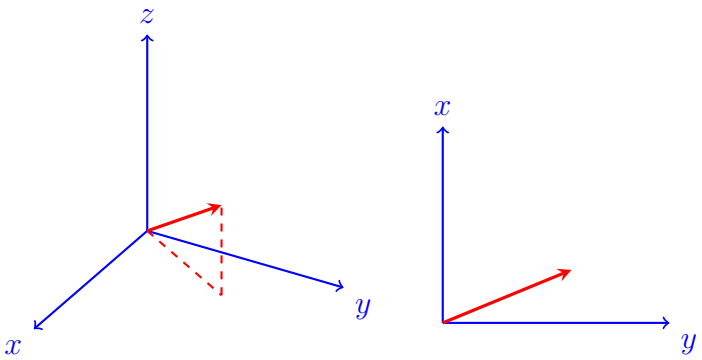


Figure 2.13: Example of a 3D (left) to 2D (right) projection

The transformation performed by the device allows for a new representation in 2D which is mathematically expressed as:

$$c_p(x\prime, y\prime, \lambda) \tag{2.8}$$

Where $x\prime$ and $y\prime$ are the new coordinates in 2D for the date in the image, and c_p represents the projection by the camera. Once an image has been processed and reinterpreted digitally following Equation 2.8, the λ value is not needed any further and the data, then, can be expressed by the image function as an image matrix:

$$f_c(x\prime, y\prime) = \int c_p(x\prime, y\prime, \lambda) V_c(\lambda) d\lambda \tag{2.9}$$

Where $c_p(x\prime, y\prime, \lambda)$ is the two coordinate dimensional system obtained from the three coordinate dimensional system of the projection of the light reflected on the sensor of the device, $V_c(\lambda)$ is the function of sensibility of the device.

Sampling

Once the image function is obtained it is important to have the values discretized. Even though a 2D projection has been generated, the available values for it are still expressed in a continuum interval where $x\prime \in [x\prime_{min}, x\prime_{max}]$ and $y\prime \in [y\prime_{min}, y\prime_{max}]$.

Also, values the image function could take belong to the set of real numbers and vary in a continuum $f_c(x\prime, y\prime) \in [f_{min}, f_{min}]$. Computers cannot handle data that varies in a continuum, therefore these values have to be discretized.

Sampling is the process through which a discrete subset of values is selected to represent the image in a matrix such that it can be processed by a digital computation device. It can also be understood as the process in which coordinate values are obtained.

It implements the function `comb()` which takes an image function and samples on regular intervals along the x and y axis using the following formula:

$$\texttt{comb}(x\prime, y\prime) = \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \delta(x\prime - i\Delta_x, y\prime - j\Delta_y) \tag{2.10}$$

To perform a sampling the following operation is done:

$$f_c(x\prime, y\prime) \times \texttt{comb}(x\prime, y\prime) \tag{2.11}$$

This process gives as a result an image matrix of dimensions $n \times m$ expressed as:

$$A_{i,j}(i = 0, ..., N-1, j = 0, ..., M-1) \tag{2.12}$$

Figure 2.14: Illustration of the sampling process as a product of subsequent impulses sampling the continuous phenomenon and producing a discrete signal representing it, as presented by [8].

Quantization

For each point acquired during the sampling process, a voltage reading is obtained on the image sensor. The voltage read relates to the light wavelength, which is projected through the device lens to the point in the image plane.

These analogue signals can be discretized into a number of bins representing levels of intensity which will become values to be recorded in A_{ij}, [49]. In other words, through the quantization process amplitude values are digitized.

The values of $f_c(i,j)$ are discretized to P as follows:

$$\text{Be } \Delta_Q = \frac{f_{max} - f_{min}}{P}. \tag{2.13}$$

$$\widehat{f}_c(i,j) = Q(f_c(i,j)) \tag{2.14}$$

where

$$\begin{aligned} Q(f_c(i,j)) &= (k + \frac{1}{2})\Delta_Q + f_{min} \\ \text{if and only if } f_c(i,j) &\in [f_min + k\Delta_Q, f_{min} + (k+a)\Delta_Q] \\ \text{if and only if } f_{min} + k\Delta_Q &\leq f_c(i,j) < f_{min} + (k+1)\Delta_Q \\ \text{for } k &= 0, ..., P-1 \end{aligned} \tag{2.15}$$

This process allows for a digital computation device to handle values of luminescence in grayscale or color spaces. Usually $P = 2^8 = 256$ levels of luminescence, when a image matrix has this depth of luminescence we say that it has been quantized to 8 bits [8]. Medical images used in clinical settings through electronic medical records follow the DICOM standard [14], which uses a 16-bit quantization for image matrices, in these images the depth of luminescence is expressed as:

$$A_{i,j} \in \{0, ..., 65535\} \tag{2.16}$$

2.6.2 Images as Matrices

Once an image has been formed through the processes discussed in section 2.6.1, and given that in this thesis we will be dealing specifically with grayscale images, the image function can be expressed as $\widehat{f}_{gray}(i,j)$ which is a matrix of $n \times m$ cells with possible values between $0, \ldots, 255$ and $0, \ldots, 65535$ for 8 and 16-bit images respectively. For images in the DICOM format their matrix and elements can be expressed as shown in Equation 2.16, with the characteristics of containing integer values from the natural numbers. Image matrices can also have continuous numbers as long as the values are adequately scaled and rounded [8].

Understanding an image as the representation of a matrix of $n \times m$ dimensions allows methods of data analytics to process it and, potentially produce significant results.

2.6.3 Point-based Operations

Point transforms are the most basic type of image-processing operations, in these processes values in the points of the input image are mapped to corresponding points in the output image [49]. From a mathematical perspective these operations are one-to-one functional mappings from input to output. Arithmetic operations can be performed as simple examples of this concept. Equation 2.17 shows this idea and Figure 2.15 shows the process applied to a grayscale picture.

$$\begin{aligned} I_{output}(i,j) &= I_A(i,j) + I_B(i,j) \\ I_{output}(i,j) &= I_A(i,j) + C \end{aligned} \tag{2.17}$$

Figure 2.15: Original 8-bit picture on the left, on the right the result of adding 100 units to each value to the matrix of the original picture. $I_{output} = I_A + 100$.

There are many other operations that can be applied on matrices as point based operations, such as thresholding, logical operations, logarithmic transforms, exponential transforms, and power-law (gamma) transforms, among others [8] [49]. They all follow the same principle, which is to apply a specific mathematical process to each value contained in a cell of the matrix or pixel of the image.

Image histograms

An image histogram is a visualization of the cardinality of the values of luminescence in a given image matrix. It can be expressed as relative or total frequency [8]. In grayscale images the histogram can be generated by counting the number of times each

value occurs within the image matrix. Equation 2.18 shows the basic mathematical process to obtain a histogram from an 8-bit grayscale image.

$$h(k) = n_k = card\{(x, y)|f(x, y) = k\} \tag{2.18}$$

Where:

$k = \{0, 1, ..., L - 1\}$,

L**:** is the depth of luminescence in a given image matrix,

*card...***:** refers to the cardinality of a set, that is, the number of elements with different value in that set (n_k).

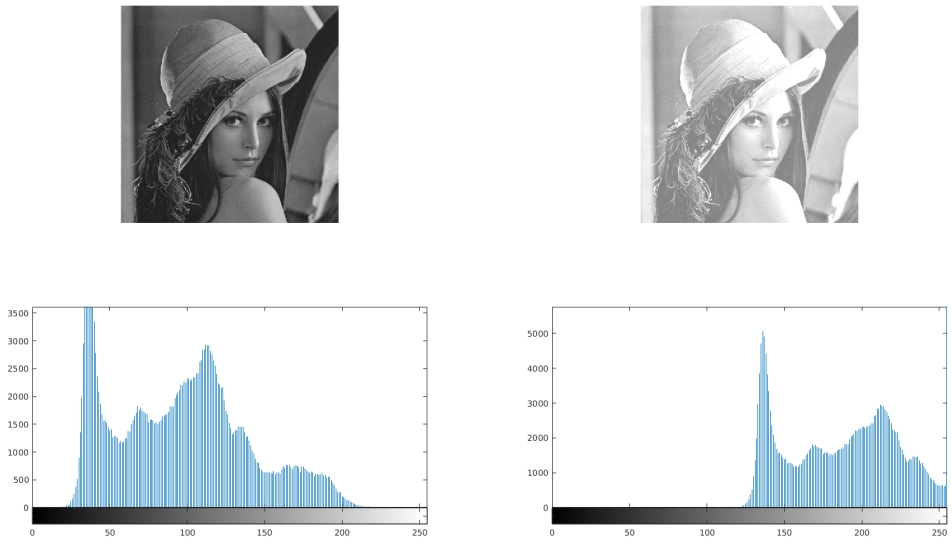


Figure 2.16: Grayscale images with their corresponding histograms.

Figure 2.16 shows the histograms from the images presented in Figure 2.15. The process implemented with Equation 2.17 becomes obvious with the help of the histogram.

Image histograms facilitate the initial exploration of values in the matrix, they simplify the identification of regions of interest (ROI) where relevant information might be located, like background or clusters of similar highly repeated values of luminescence.

2.6.4 Histogram equalization

This technique changes the grayscale value distribution in an image so that the resulting histogram shows a uniform distribution or the closest possible to it, as it can be

appreciated in Figure 2.17. In order to do this an auxiliary function must be used, generally called the *transformation function*, (Tr), any specific function used here has to satisfy two criteria shown in Equation 2.19, [50].

1. $T(r)$ must be a monotonically increasing function in the interval
$$0 \leq r \leq L-1 \tag{2.19}$$
2. $0 \leq T(r) \leq L-1$ for $0 \leq r \leq L-1$

The cumulative distribution function (cdf) of the original probability mass function is the most commonly used transformation [49], and it is given by

$$sk = T(r_k) = \sum_{j=0}^{k} \frac{n_j}{n} = \sum_{j=0}^{k} p(r_j) \tag{2.20}$$

where:

sk**:** is the new (mapped) gray level for all cells whose original gray level used to be r_k,

n_j**:** the jth gray level,

n**:** is the total number of cells in an image matrix,

$p(r_j)$**:** is the percentage of cells in the image matrix whose gray level is r_j.

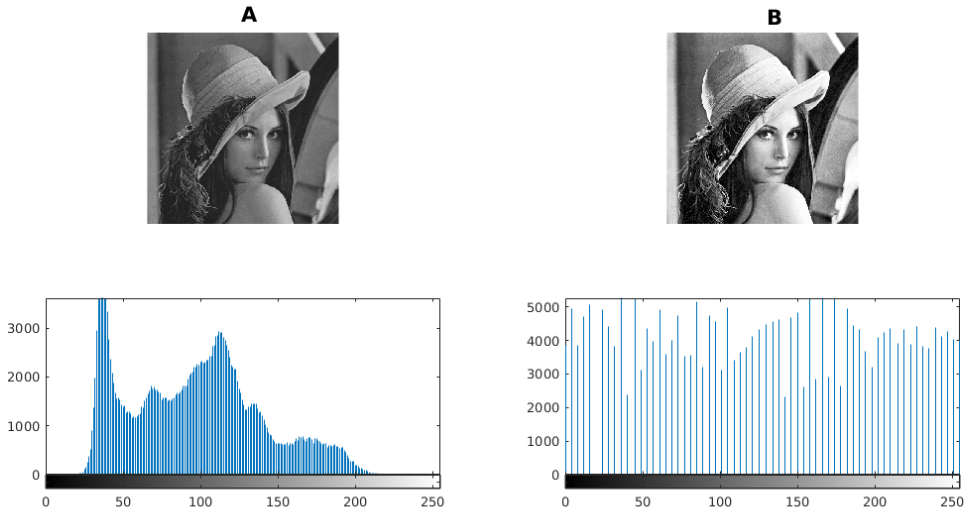


Figure 2.17: Image B results from performing a histogram equalization on image A, comparing their corresponding image histograms shows the tendency to a uniform distribution in histogram B.

The visual result of performing a histogram equalization to an image matrix will be a contrast enhancement of the elements in the image. This facilitates further analysis like border detection, identification of ROI, or even visualizations by humans.

Chapter 3

Development of Research Strategy

3.1 Methodology for Validation

In order to maintain a standardized approach of the research process an adaptation of the work by Kruve and collaborators for validation of liquid chromatography-mass spectrometry (LC-MS) methods [9] was made. This allows to build on a published and used approach to method validation. Figure 3.1 shows their original proposal.

3.1.1 Stability

In the context of LC-MS stability refers to the analyte ability to avoid decomposition, therefore maintaining trueness and precision of the procedure [9]. Parameters like long-term stability, short-term stability and stock solution stability become very relevant when evaluating this in fluids.

For the purpose of image matrices, a set of characteristics can be determined as the minimum requirements for stability of the image, these are proposed to be ordered under three main categories:

1. Format related.
    - Same protocol or format for all images.
    - Depth of luminescence.
    - Color space.
2. Security related.
    - It is important to make sure that the data is accessible through decryption protocols provided by the source of the clinical images [51].
    - If data is provided without restriction, a preprocess for eliminating sensitive information about the stakeholders involved in the clinical scenario has to take place before moving forward on the analytics pipeline.

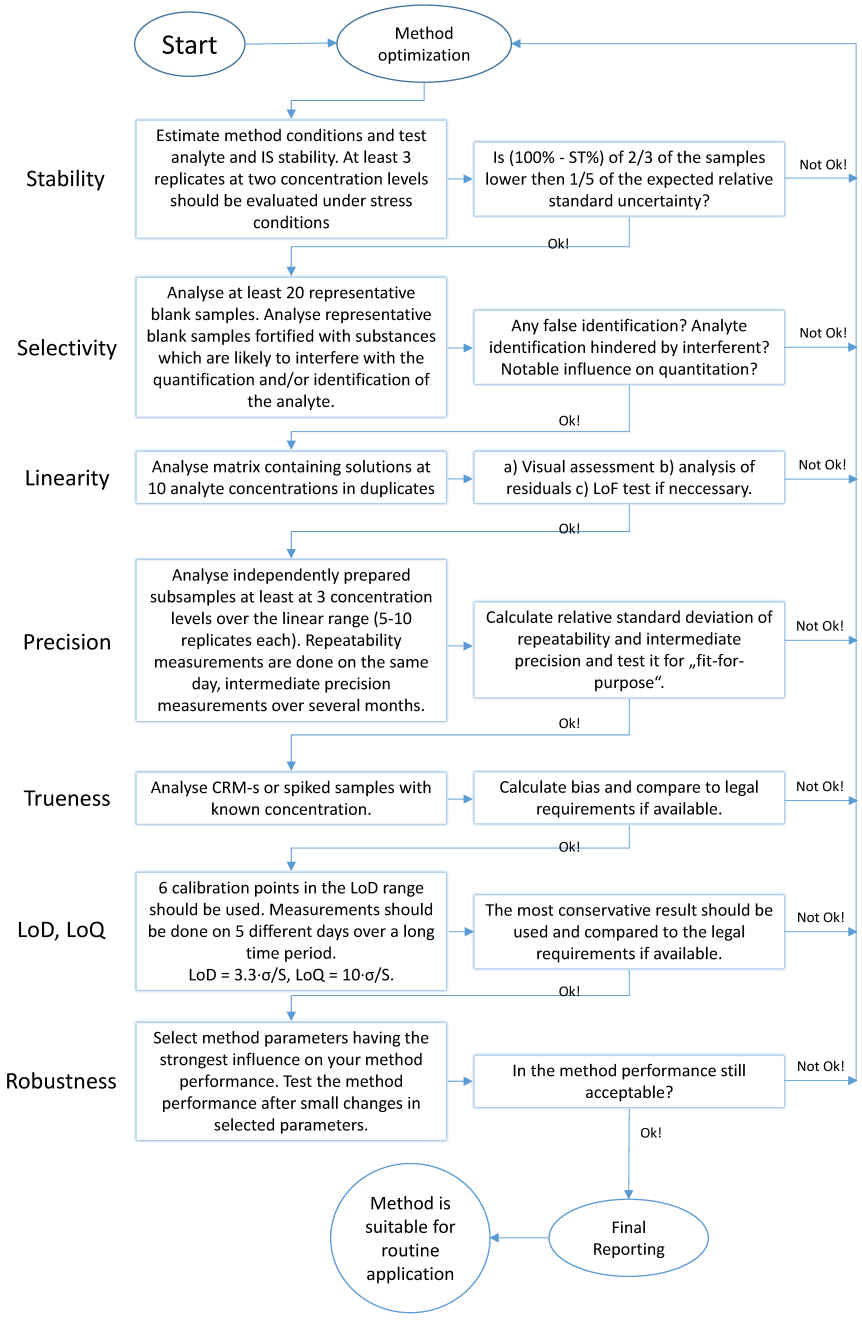


Figure 3.1: Proposal by [9] for method validation in liquid chromatography mass spectrometry.

3. Related to the clinical practice.
    - Radiologic technique used to perform the study.
    - Angle of projection in which the image was captured.
    - Anatomic region visualized.
    - Structures of interest in the anatomic region.

Figure 3.2 shows an adaptation from Kruve and collaborators [9] to evaluate stability in medical images and Table 3.1 show the specific values used for the images of this project.

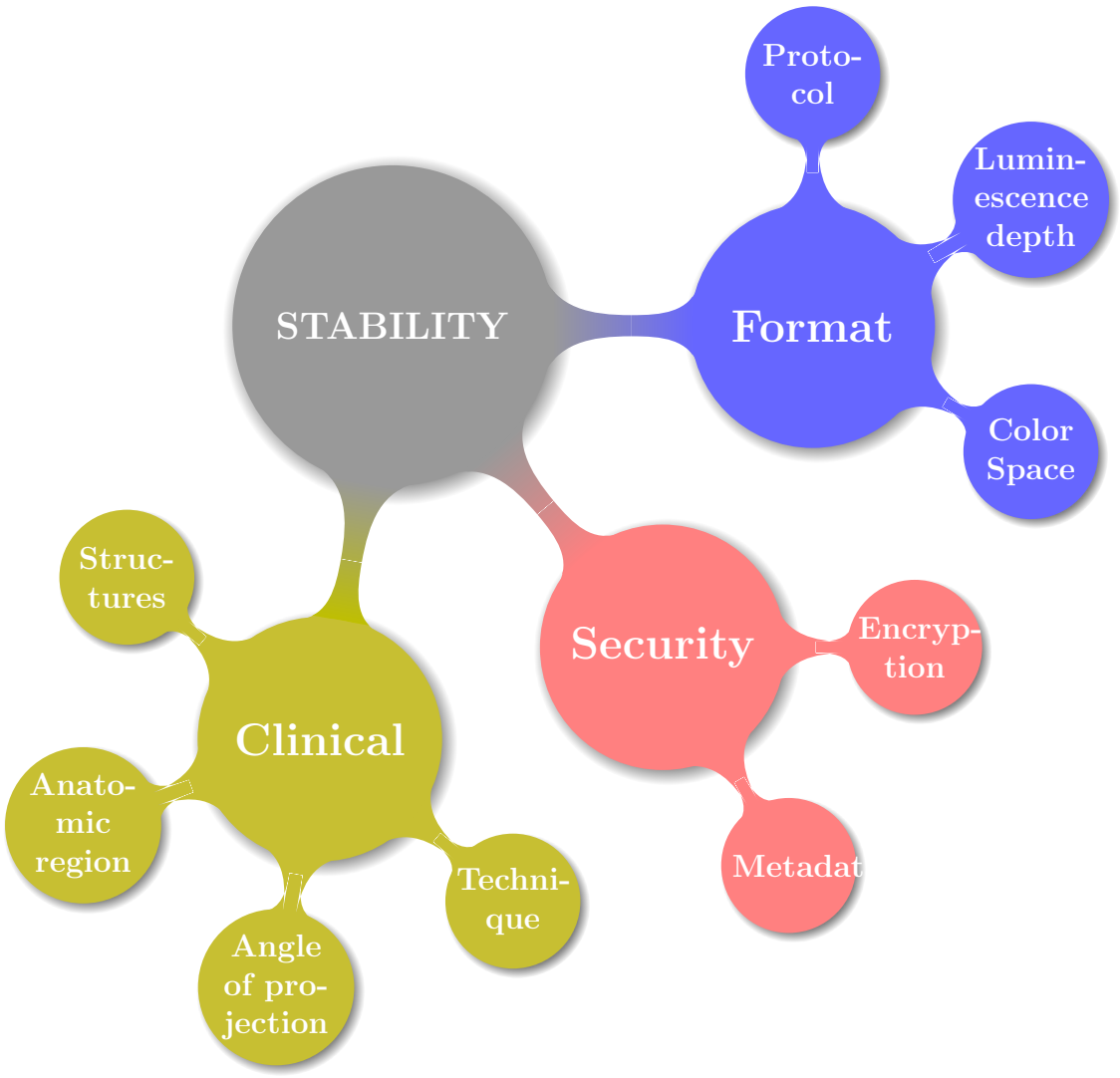


Figure 3.2: Indicators considered for **stability** in the medical images analyzed.

Table 3.1: Specific **stability** indicators used for this project.

INDICATOR	VALUES
Format	
Protocol	DICOM
Luminescence	2^{16}
Color Space	Grayscale
Security	
Encryption	Non-encrypted images
Header	Without metadata
Clinical	
Technique	X-rays
Angle of projection	Lateral
Anatomic Region	Cervical spine
Structures	Vertebrae

3.1.2 Selectivity

In the context of LC-MS selectivity is closely related to specificity, they both refer to the ability of a method to measure the amount of the analyte that is claimed to be measured [52]. Selectivity, in this context, is used to refer to the extent to which other substances interfere with the determination of an anlayte [52].

For medical images the concept of selectivity can be adapted to where the information by cell or pixel allows for an adequate identification of regions of interest (ROI).

The minimum requirements for selectivity in this process are proposed in the following list, under three categories:

1. Image histogram related.
    - Continuity in bin values.
    - Probability distribution in histograms. **ADD FORMULA AND CODE TO IMPLEMENT THIS**
    - Regions of interest (ROIs) in the histogram.
2. Image matrix related.
    - Absolute values.
    - Size of image ($n \times m$).
3. Clinical practice related [53].
    - Technical evaluation of the image.

Figure 3.3 shows the adaptation used to the concept of selectivity in medical images from Kruve and collaborators [52]. Table 3.2 shows the specific values used for the evaluation of selectivity and, therefore methodical implementation of the process. It is important to highlight that concepts of digital image processing, data analytics and clinical practice were required to integrate this approach.

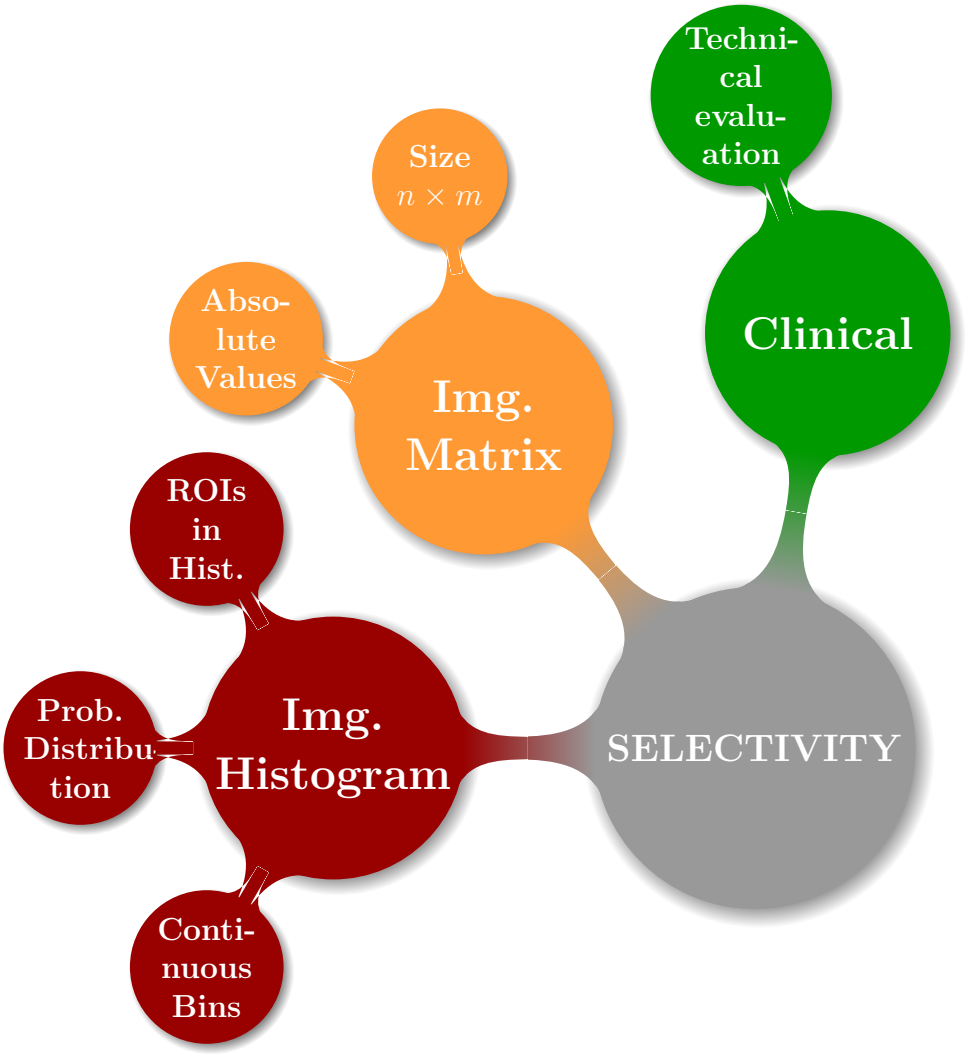


Figure 3.3: Indicators considered for **selectivity** in medical images analyzed.

Table 3.2: Specific **selectivity** indicators used for this project.

INDICATOR	VALUES
Image Histogram	
Continuous bins	Histogram bins should have values all along the interval between the minimum and maximum value.
Probability distribution	Images without a strong tendency towards a uniform distribution are preferred.
ROIs	There should be a subset of bins in the histogram with a high frequency of specific values.
Image Matrix	
Absolute values	Luminescence values of absolute black and absolute white should be no more than 30% of the total values.
Image size	An image matrix of size 1700×400 as minimum.
Clinical	
Technical evaluation [53]	All seven cervical vertebrae should be distinguishable. C4 should be in the center of the radiography. Bone and soft tissue details should be visible as well as the air column. Spinous processes should show in profile. Rami of the mandible should not superimpose C1 to C2. No rotation of the projection, corroborated by: no superimposition of both rami of mandible, both sides apophyseal joints, and posterior borders of the vertebrae bodies.

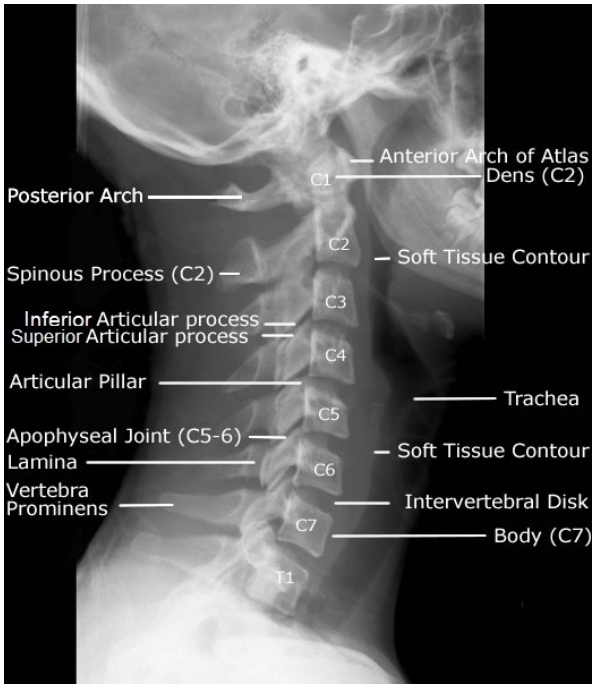


Figure 3.4: Points of reference for a technical evaluation of X-ray lateral projections of the cervical spine. Image taken from wikiRadiography.

To add clarity and elaborate in more detail about the technical evaluation done within the clinical context, as mentioned in Figure 3.3, Figure 3.4 shows a lateral projection of the cervical spine taken with X-rays and points of reference to understand the technical evaluation proposed for the clinical component.

3.1.3 Linearity

In the context of LC-MS to asses the linearity of a signal refers to the exploration of linear relationships between analyte signals and analyte concentrations in calibration samples or matrix components [9]. Therefore the ion source in the fluid can be associated to concentrations of substances given their linear behavior.

For medical images, this concept can be applied to luminescence values. Taking the example of X-rays, given the context in which the image is generated (rays bouncing on a sensitive plaque and generating an image) each cell will contain a luminescence value linearly related to tissue density. Figure 3.5 shows in more detail two different processes in which x-ray images are obtained.

This concept of linearity is used in clinical practice in the process of interpreting the image to obtain a medical conclusion that benefits a given patient.

The minimum requirements for linearity are proposed in the following list:

1. Machine performance.

- Technical calibration and/or certification of the equipment used to capture the image.

2. Image structure.

   - The amount of luminescence values in the image matrix, should be linearly relatable to different tissue densities evaluated by the clinician.

3. Interpretation of image.

   - Most features should be detectable by a linear method measuring luminescence.

Figure 3.6 shows the adaptation used to the concept of linearity in medical images from Kruve and collaborators [52]. Table 3.3 shows the specific values used for the evaluation of linearity and, therefore methodical implementation of the process.

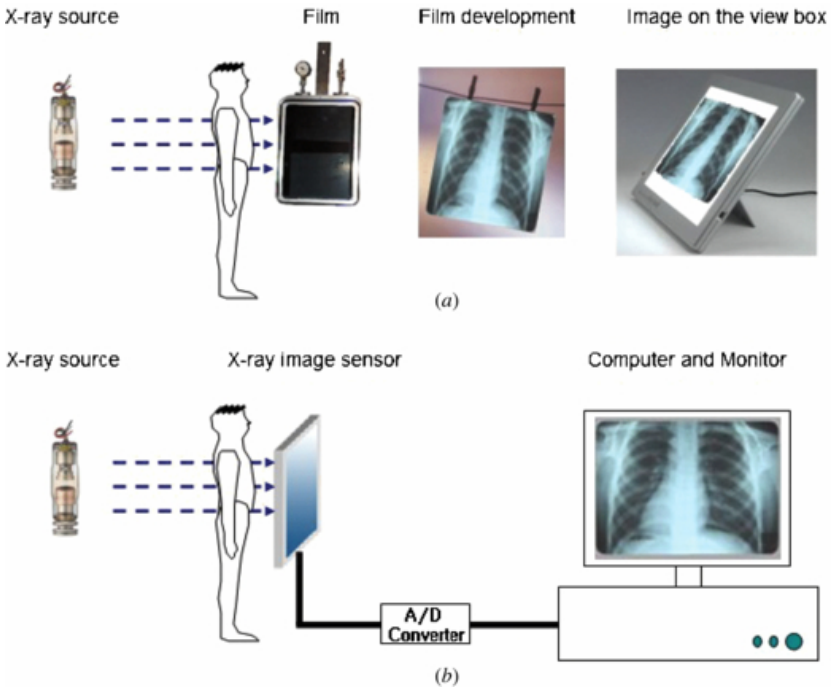


Figure 3.5: Diagram of two x-ray detector systems as published by [54]: *(a)* analog-type detector, *(b)* a digital version of the detector.

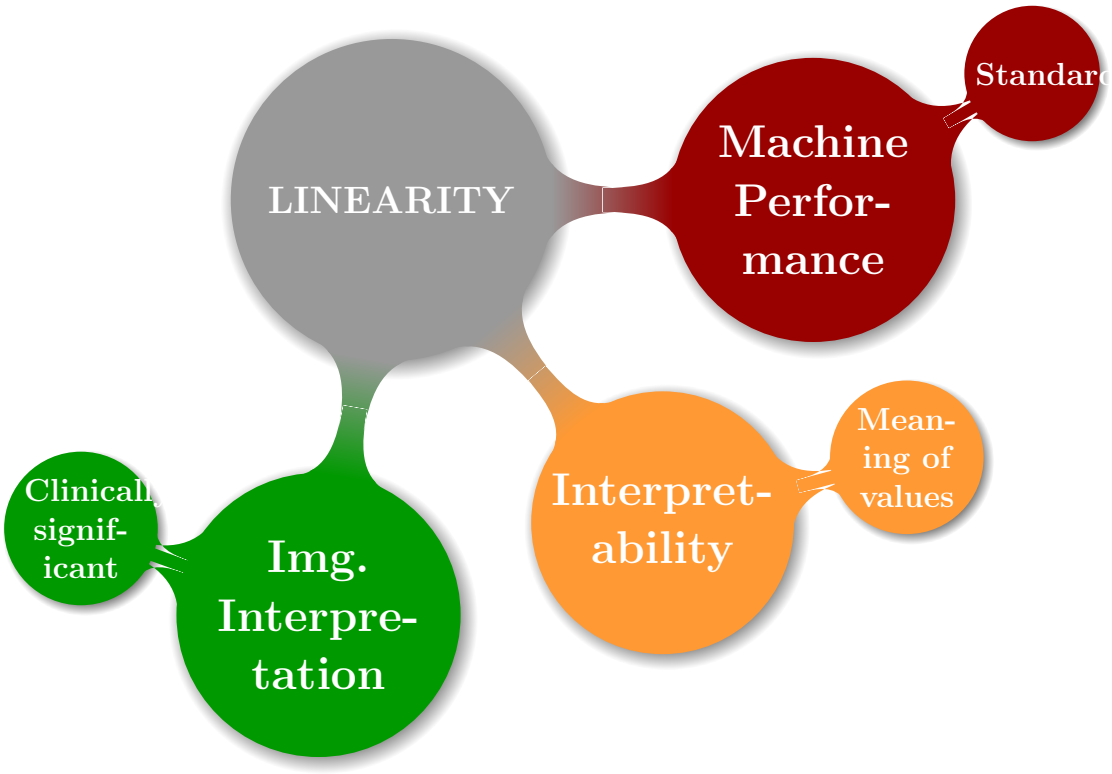


Figure 3.6: Indicators to be considered for **linearity** in medical images.

Table 3.3: Specific **linearity** indicators used for this project.

INDICATOR	VALUES
Machine Performance	
Standards of Performance	Previously calibrated machine
Interpretability	
Meaning of values	Luminescence = tissue density
Image Interpretation	
Clinically significant	Detectable features

3.1.4 Dimensionality Reduction Model

Step by step integration of the linear model

The following flowchart summarizes an approximation used to face the challenge of dimensionality reduction in the image matrices selected for this project. Each of the steps mentioned in the flowchart will be discussed to detail in the following subsections

of this chapter.

Components in the flowchart are to be understood as follows:

A, B, C**:** Image matrices of $n \times m$ dimensions.

$histEq()$**:** Histogram equalization function.

hac**:** Commulative histogram.

PCA**:** Principal Components Analysis.

$PC1$**:** First Principal Component.

$himhist()$**:** Image histogram function.

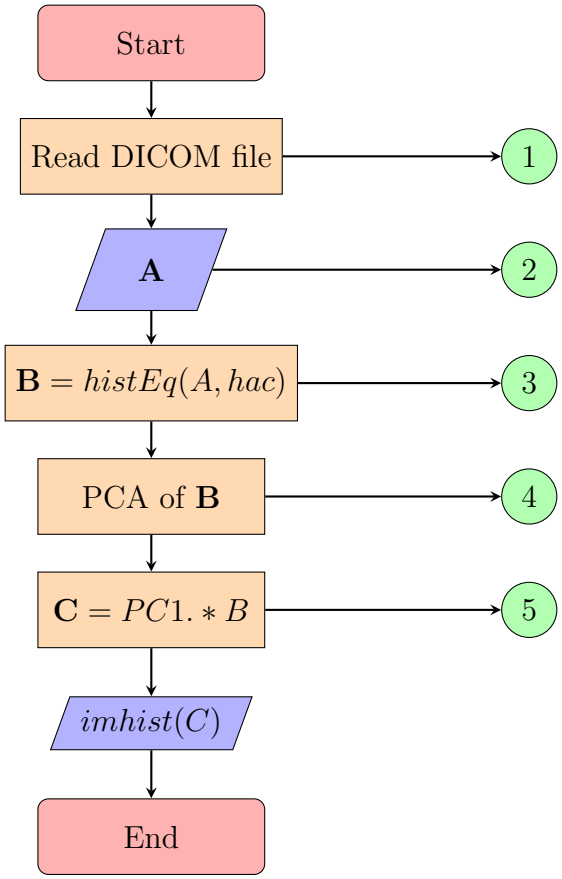


Figure 3.7: Flowchart of the general steps in the linear model used in this project.

About the input image

We begin with an image matrix as an input with the following characteristics:

1. A DICOM [1] file, which is a tuple containing 2 elements: a header with metadata and the image matrix: $A = (header, imageMatrix)$.
2. $A_{i,j}$
    - Where $A_{i,j}$ is of $n \times m$ and the depth of luminescence is expressed as $L = 2^{16} = (0, 1, 2, ..., 65536)$.

Histogram equalization

As mentioned in section 2.6.4, a histogram equalization is performed to enhance contrast in an image making changes in luminescence more obvious.

The previous is achieved with the following formula:

$$B_{i,j} = s_k = T(r_k) = \sum_{i=0}^{k} \frac{n_j}{n} = \sum_{j=0}^{k} p(rj) \tag{3.1}$$

where:

B**:** is the new image matrix of $n \times m$ dimensions with equalized values of luminescence.

s_k**:** represents the new (mapped) values of luminescence for all pixels whose original gray level used to be r_k.

$T(r_k)$**:** is the transformation function satisfying the following criteria:

- $T(r_k)$ is a monotonically increasing function in the interval $0 \leq r \leq L - 1$.
- $0 \leq T(r) \leq L - 1$ for $0 \leq r \leq L - 1$

Principal Components Analysis of equalized image

Once an equalized image matrix is generated (B) a Principal Components Anlaysis can be performed on it ($PCA(B)$), for this process the following steps are followed:

1. Obtaining the mean of the data in the matrix:

$$\overrightarrow{\mu} = \tfrac{1}{n}(\overrightarrow{x_1} + ... + \overrightarrow{x_m})$$

2. Re-center data so that $\mu = 0$.

[1] Digital Imaging and Communications in Medicine.

$$B_{i,j} = [\overrightarrow{x_1} - \overrightarrow{\mu}|...|\overrightarrow{x_n} - \overrightarrow{\mu}]$$

3. Define covariance matrix $S_{i,j}$, which will be of size $m \times m$.

$$S_{i,j} = \frac{1}{n-1} BB^T$$

4. Calculate eigenvalues and eigenvectors:

   - Considering the spectral theorem [55]:
     If S is symmetric (meaning $S^T = S$), then S is orthogonally diagonalizable and has only real eigenvalues. In other words, there exist real numbers $\lambda_1, ..., \lambda_n$ (or eigenvalues) and orthogonal, non-zero real vectors (or eigenvectors) such that for each $i = 1, 2, ..., n$: $A\overrightarrow{v_i} = \lambda\overrightarrow{v_i}$.
   - Since $S_{i,j}$ is a symmetric matrix it can be orthogonally diagonalized by applying the spectral theorem and let $\lambda_1 \geq \lambda_2 \geq ...\lambda_m \geq 0$ be the decreasing eigenvalues of $S_{i,j}$ with corresponding orthonormal eigenvectors (or principal components) $\overrightarrow{u_1}, ..., \overrightarrow{u_m}$.

Using the first principal component

The approach in this project was to use the first principal component and perform a point multiplication with each of the columns of the equalized image matrix, the decision to use only the first principal component was made because the variability in the values was explained majorly by the first component as shown in Figure 3.8. The resultant matrix C is of the same dimensions of the original and equalized matrices but has a significantly reduced number of values. The operation is mathematically expressed in equation 3.2.

$$C = \overrightarrow{u_1}. * B \tag{3.2}$$

- Where C is the image matrix with a reduced dimension of luminescence values.

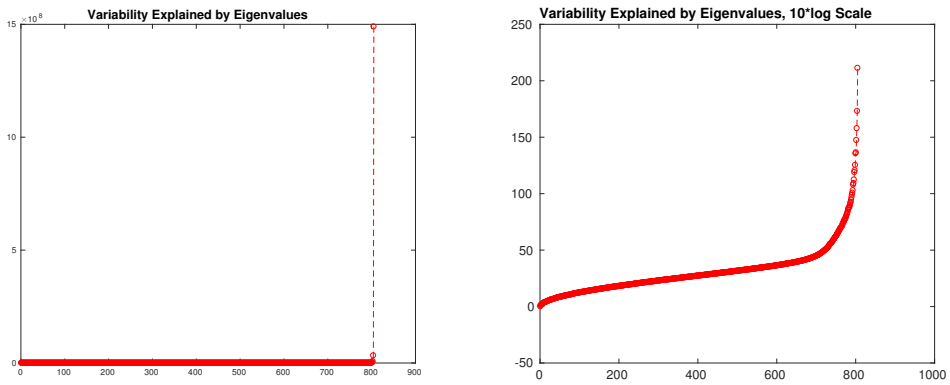


Figure 3.8: Variability explained by eigenvalues related to eigenvectors (principal components) from an image of a lateral projection of the cervical spine. Normal and $log10$ scales.

3.1.5 Precision

In the context of LC-MS, precision is related to the agreement between replicate measure results or between the measure value and a reference value [9]. This step in the validation of LC-MS methods takes a significant amount of theory and calculations, because each measurement will most likely not deliver the same result.

This is not the case when the method is applied to an image matrix, given that the acquisition of the image has already taken place and, almost always, the analyst is not involved in this process. Nevertheless, there are some indicators that can still be of benefit for the validation of this step.

For image matrices this indicator is evaluated considering all the matrices in the set of images being analyzed in the process. Each measurement will be an individual result given the values of the rest of the matrices in the set.

The determinants of precision are mentioned in the following list:

1. Repeatability.
    - Reproducibility, which is the delivering of the smallest variation in results and is obtained by repeatedly analyzing independently prepared subsamples from a homogeneous sample [56].
    - Intermediate precision, which is the precision obtained within a single laboratory over a longer period of time [56].
2. Error source [57].
    - Systematic error.
        - An error that occurs every time the measurement is taken, in this case an image.
    - Random error.

– An error that appears occasionally with no obvious pattern of occurrence. It does not have a definite set of causes.

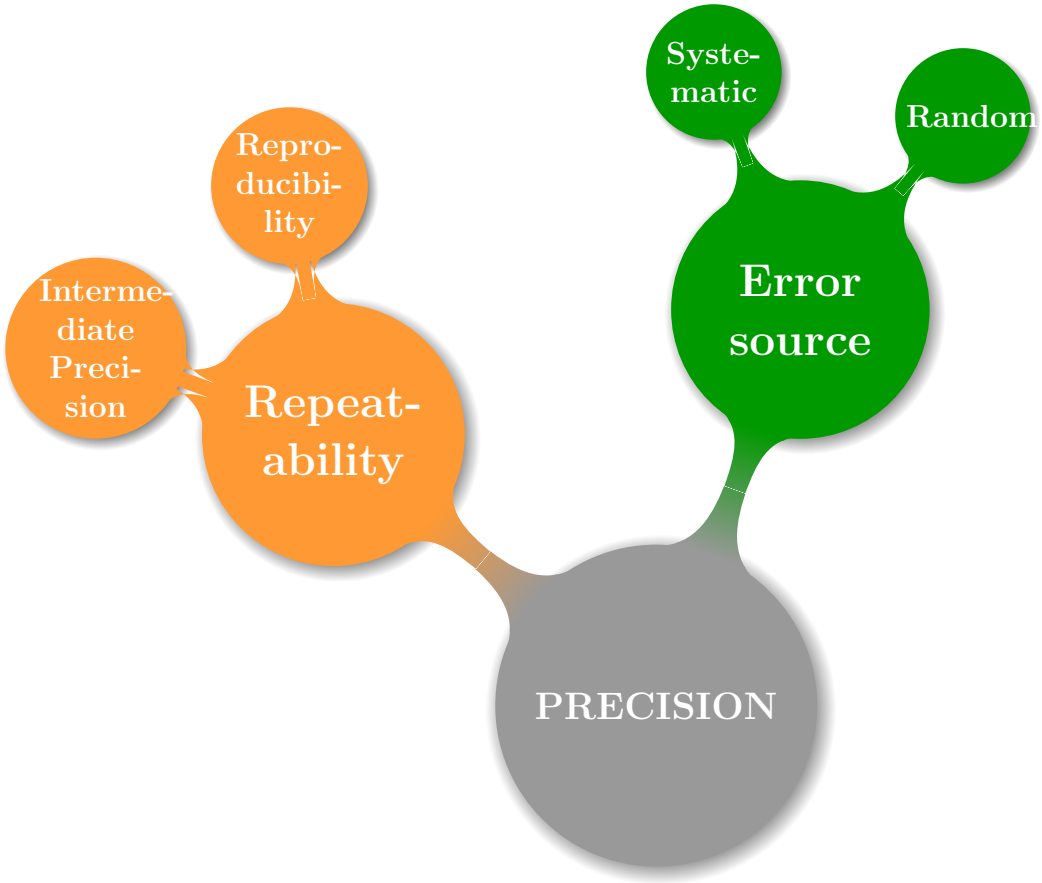


Figure 3.9: Indicators considered for **precision** for the analysis of medical images.

Table 3.4: Specific **precision** indicators used for this project.

INDICATOR	VALUES
Reproducibility	
Intermediate precision	Same results with same code, machine and software version.
Repeatability	Same results by other machine or different operative system, with or without same code but with same technique and different software version.
Error source	
Systematic error	Not possible to evaluate image in a clinical context given the error.
Random error	Script needing debugging for delivering expected results.

3.1.6 Limit of Detection and Limit of Quantitation

Limit of detection (LoD) is referred to by Kruve and collaborators [52] as the lowest amount or lowest concentration of the analyte in a sample which can be reliably detected and identified with the method. This definition works in general terms to introduce the concept, still it does not take into consideration the possibility for false positives and false negatives [58], still it lays a foundation to propose an adaptation for medical images.

Limit of quantitation (LoQ) is defined as the lowest concentration of analyte that can be determined with an acceptable repeatability and trueness. Meaning that LoD will be the lowest possible value detected by the method and LoQ will be the smallest change possibly detected from the lowest value onward.

These two concepts can be implemented to medical images by analyzing their depth of luminescence and corresponding image histogram as shwon in section 2.6.3 of this document and Figure 3.10. Additionally a normalization of the values can be done to obtain zero as the lowest value and adjust to a normalized rate of change in the luminescence vector (ΔL), Table 3.5.

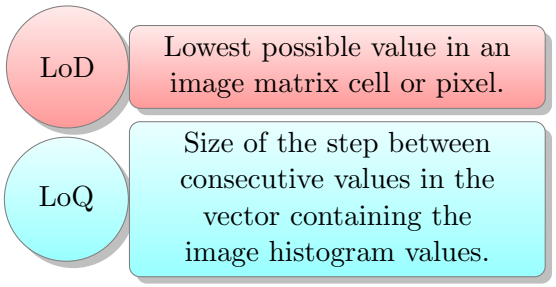


Figure 3.10: Indicators considered for **LoD** and **LoQ** for the analysis of medical images.

Table 3.5: Specific **LoD** and **LoQ** indicators used for this project.

INDICATOR	VALUES
LoD	Minimum value in a 16-bit image matrix, a normalization of values can also by implemented to use 0 as the LoD.
LoQ	Image histogram bin values compared before and after linear transformations.

3.1.7 Robustness

This concept refers to the ability of an analytical method to remain unaffected by small variations in method parameters and environmental factors, allowing for the crit-

ical performance characteristics to remain unchanged by the variation of the method parameters [59].

For medical images, an evaluation of the parameters that add to the method validation strategy (Figure 3.11) is what will give a clearer idea of the robustness of the method and the bottle necks within it.

This step has to be customize to each method, given that the specific indicators will vary depending on the model.

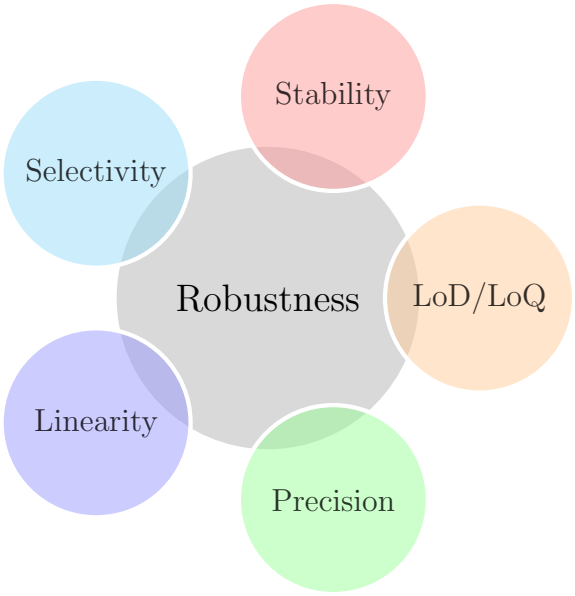


Figure 3.11: Indicators used to assess **robustness**.

3.2 Integration of the Model

Figure 3.12 integrates the steps used in this project for method validation and ensuring reproducibility of the approximation, the process was originally published by Kruve and collaborators [9] as a method validation strategy in fluid analysis, specifically LC-MS. Figure 3.1 shows the original visualization presented in their published work.

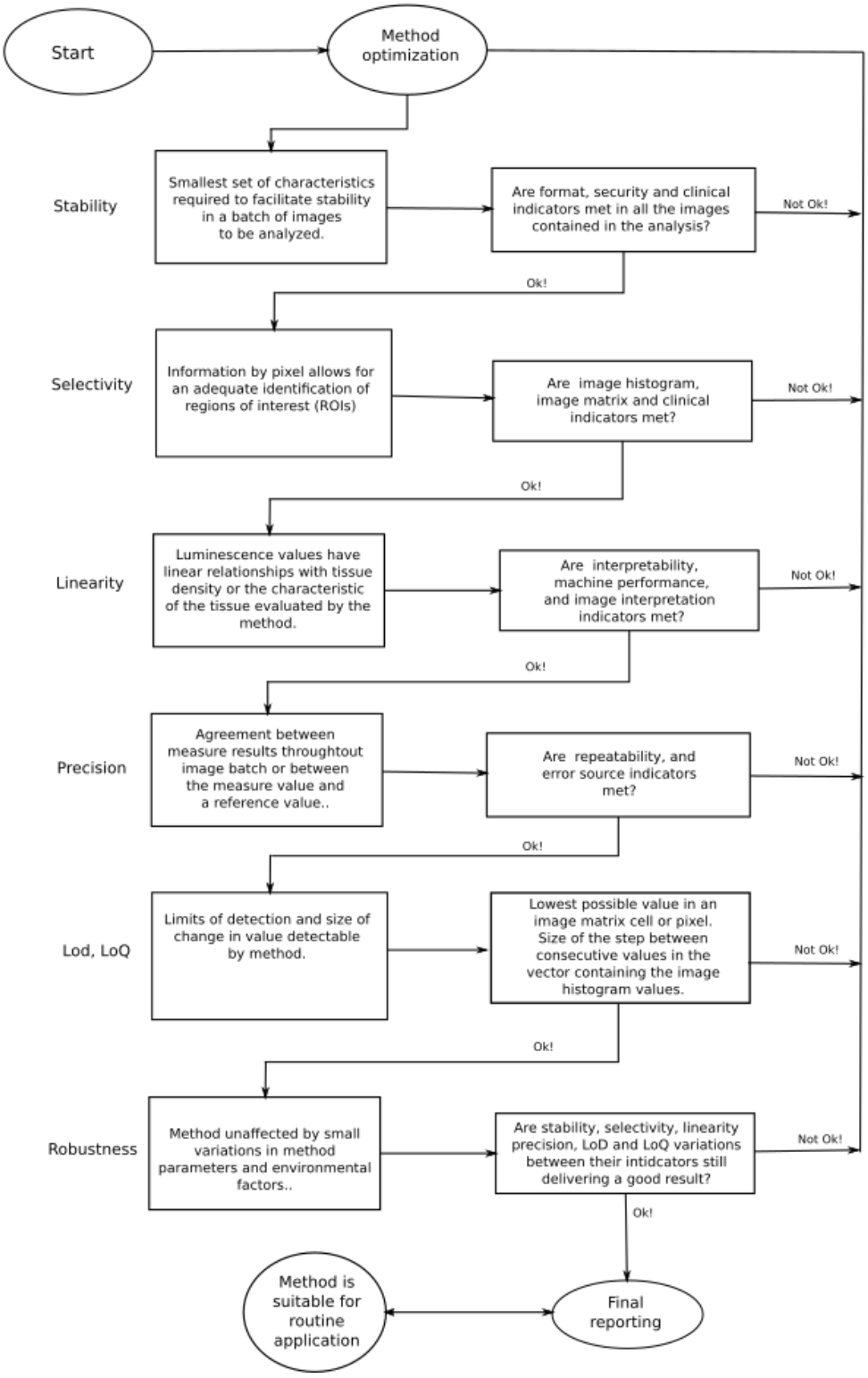


Figure 3.12: Integration of the model adapted from [9] for medical images.

Chapter 4

Results

This chapter presents the tests and results obtained by following the method validation proposed in Chapter 3. The results of the linear approximation utilized are also presented and discussed at length. Clinical implications of the results are also evaluated and discussed.

4.1 Generalities of the images used in this project

For this project we selected lateral projections of the cervical spine, obtained using X-rays. Figure 4.1 shows the image of a rather healthy patient in the projection used in this work. All X-rays used in this project were kindly provided by Dr. Félix David Ramón-Concepción, from his private traumatology practice in the state of Nuevo León, México.

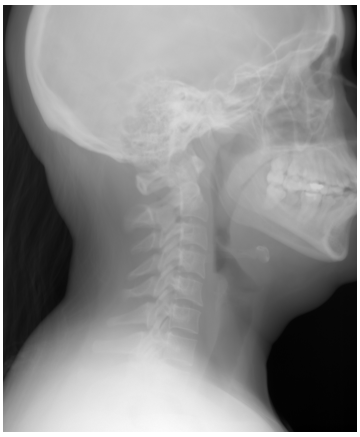


Figure 4.1: Lateral proyection of cervical spine in a healthy patient, obtained using X-rays.

X-ray images have the characteristic of being grayscale images, therefore they are represented with one matrix of $n \times m$ dimensions with available values of 2^{16}, $L = \{0, \ldots, 65535\}$, of luminescence. A particular trademark in medicine is that the negative

of the image is the one used in the clinical practice. Figure 4.2 shows the same image shown in Figure 4.1 with it's accompanying image complement which would be the one used in clinical settings.

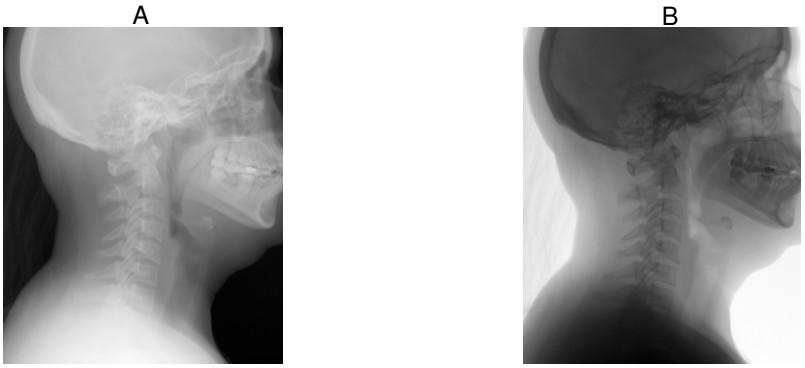


Figure 4.2: Image **A** would be used in a clinical setting, which is the complement of image **B**: $A = B^C$.

4.2 Implementation of Stability

As mentioned in section 3.1.1, three indicators are considered to assure stability of images in a batch for analytical purposes: format, security and clinical guidelines. Listing 4.1 shows an implementation in Python [1] of the format and clinical guidelines, the results generated by the computation can be appreciated in Table 4.1. Figure 4.3 show the resulting subset of images selected after implementing the aforementioned indicators.

```
from medpy.io import load
import cv2
import numpy as np
import funciones as fn
from matplotlib import pyplot as plt
import matplotlib.image as mpimg
import sys
import timepy as tp
nombre=sys.argv[1]

image_data, image_header = load(nombre)
image_data = np.transpose(image_data)

#===============STABILITY=========================
#============Format.protocol======================
format_ = str('image_header' in globals())
```

[1] Python 2.7.14 —Anaconda, Inc.— and IPython 5.4.1 were used for the implementations, unless indicated differently.

```
17
18 if format_ == "True":
19     format_ = "DICOM"
20 else:
21     format_ = "Not DICOM"
22
23 #Format.luminescence
24 lumDepth = image_data.dtype
25
26 #Format.ColorSpace
27 color = image_data.ndim
28 if color == 2:
29     color = "grayscale"
30 else:
31     color = "not grayscale"
32
33 #=======Clinical=====================================
34 #Clinical.technique
35 technique = image_header.SOPClassUID
36 if technique == "1.2.840.10008.5.1.4.1.1.1.1":
37     technique = "X-rays"
38 else:
39     technique = "Not X-rays"
40
41 #Clinical.AnatomicRegion
42 anatomicRegion = image_header.BodyPartExamined
43
44 #Clinical.AngleOfProjection
45 angle = image_header.ProtocolName
46
47 #=======Creating a .txt table with the results=======================
48 F = open("values_stability.txt",'a')
49 F.write(str(nombre)+","+str(format_)+","+str(lumDepth)+","+str(color)+","
       +str(technique)+","+str(anatomicRegion)+","+str(angle)+","+str(vMax)+"
       ,"+str(vMin)+"\n")
50 F.close()
```

Listing 4.1: Implementation of stability indicators.

Non of the images facilitated for this project were encrypted and, the metadata contained in the header section of the DICOM files was used only when necessary to evaluate compliance to the method validation process, sensitive information such as name or addresses were not accessed by any process of the analytic pipeline.

Table 4.1: Indicators for stability in a subset of 15 radiographies.

Name	Protocol	Luminescence	Color Space	Technique	Anatomic Region	Angle
img/IMG2.2	DICOM	uint16	grayscale	X-rays	CSPINE	Cervical OM
img/IMG6.4	DICOM	uint16	grayscale	X-rays	LEG	Foot Bone
img/IMG8.2	DICOM	uint16	grayscale	X-rays	CSPINE	C-Spine LAT
img/IMG1	DICOM	uint16	grayscale	X-rays	CSPINE	C-Spine LAT
img/IMG3.3	DICOM	uint16	grayscale	X-rays	CSPINE	C-Spine LAT
img/IMG9.4	DICOM	uint16	grayscale	X-rays	CSPINE	C-Spine LAT
img/IMG5.2	DICOM	uint16	grayscale	X-rays	CSPINE	C-Spine LAT
img/IMG7.1	DICOM	uint16	grayscale	X-rays	CSPINE	C-Spine LAT
img/IMG4.6	DICOM	uint16	grayscale	X-rays	CSPINE	Neck LAT

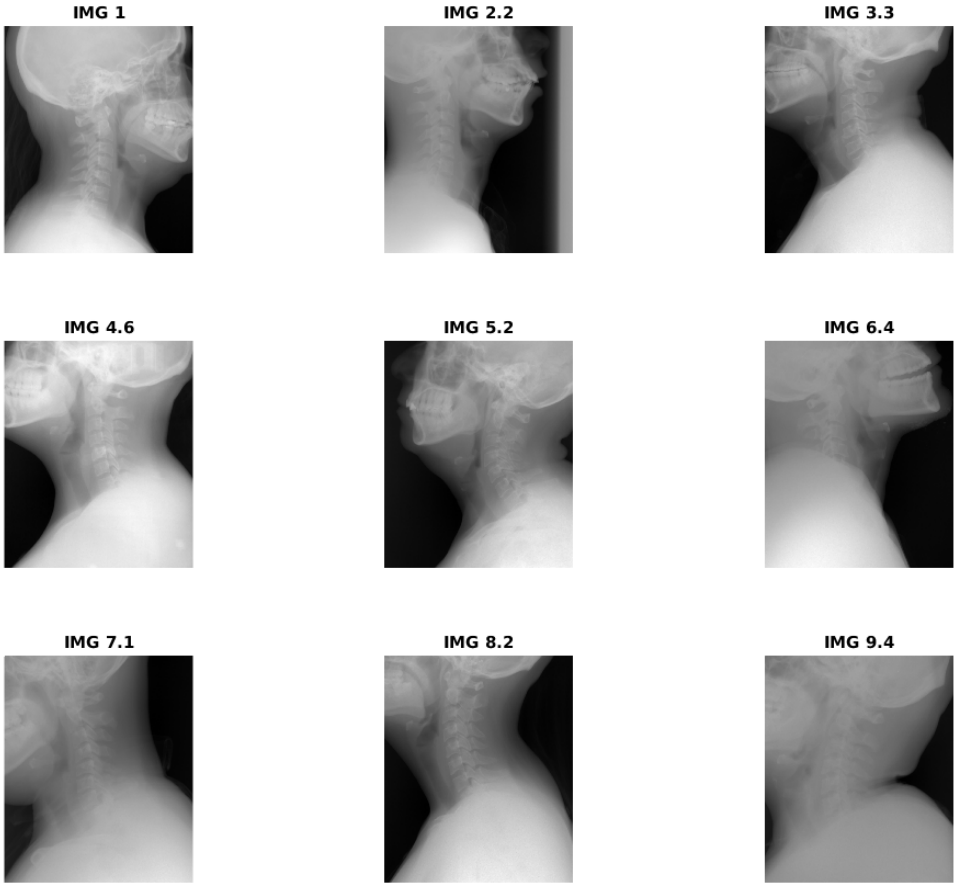


Figure 4.3: Subset of resulting images after applying the stability indicators of the method validation.

The implementation of the stability section of the method validation allowed us to obtain a subset of images ready for the next step of validation.

4.3 Selectivity

Section 4.3 explains to detail the concept behind selectivity. In summary, this step of the method validation is implemented to allow an adequate identification of regions of interest (ROIs) in the image or images analyzed.

4.3.1 Image histogram related

An image histogram is one of the initial analysis done to an image matrix, it shows the frequency of repetition of each luminescence value available in an image matrix. Section 2.6.3 explains the details of an image histogram. Image 4.4 shows a visualization of this concept.

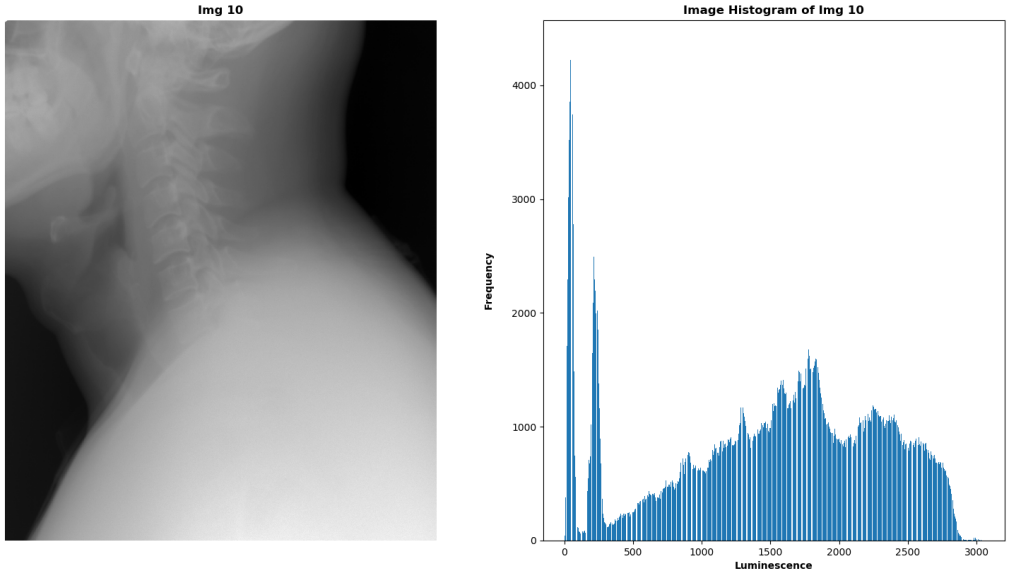


Figure 4.4: Image 10 from the set used for this project and it's corresponding histogram.

Continuity in bin values

This check point allows for the identification of previously manipulated images from raw images before they are submitted to the rest of the analytical pipeline. It is very important for this process to run the analysis on raw images without manipulation, to ensure access to all the data available. Image 4.5 shows the same image shown in Figure 4.4 but with a histogram equalization process performed to it. In this image histogram continuity in bin values is not met, since a previous process has affected the luminescence values in the image matrix.

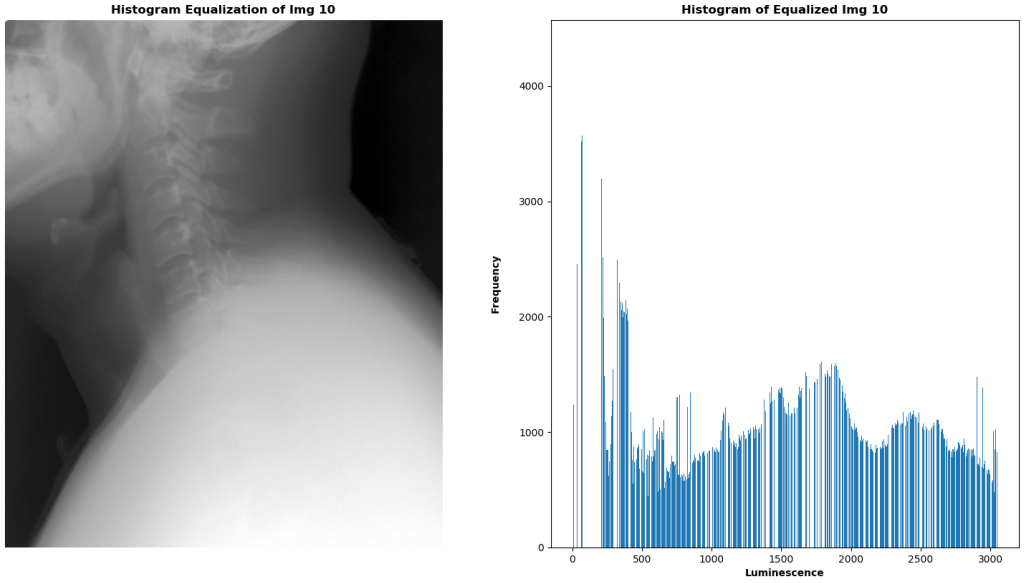


Figure 4.5: Image shown in Figure 4.4 with a histogram equalization process performed to it and the corresponding histogram.

Regions of interest identifiable in the histogram - ROIs

Table 3.2 refers to ROIs as a subset of bins in the histogram with a high frequency of specific values. These luminescence values translate into particular regions in the image that can, also, match anatomic regions of clinical interest.

Figure 4.4 shows an image histogram with many peaks in the graph, where the frequency of the specific values significantly surpasses those of the neighboring values, this is highlighted in Figure 4.6.

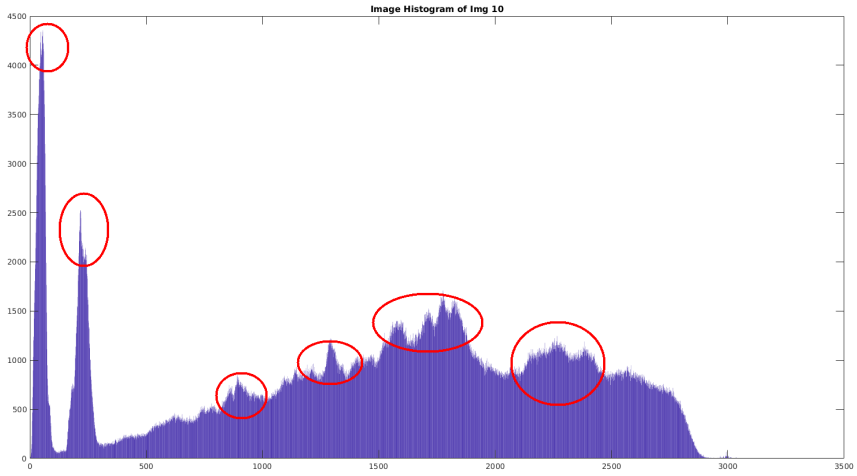


Figure 4.6: Image histogram with subsets of luminescence values with high frequency highlighted with red circles. These values could be suggesting ROIs.

To build on the concept of identifications of ROIs through image histograms a mask can be generated from the original image, leaving only values contained in high frequency areas and converting the rest to a standardized color (black or white). Figure 4.7 shows what is generated when a mask is applied in an interval from 1500 to 1800 values of luminescence from the minimum value, based on data showed on Figure 4.6. Listing 4.2 shows the implementation of this process in Matlab. [2]

[2]Matlab R2017a (9.2.0.538062) 64-bit(win64) with academic license number: 40553077.

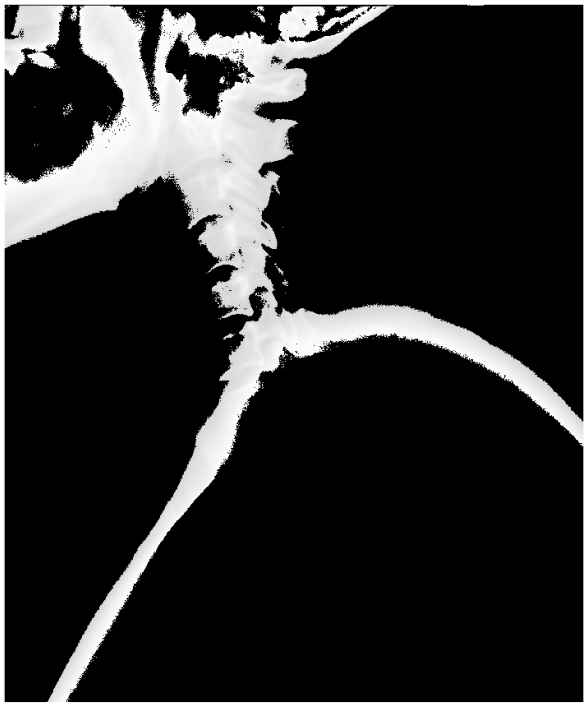


Figure 4.7: Mask generated from Figure 4.4 where values between 63 941 (minimo + 1500) and 64241 (minimo + 1800) are kept and the rest is assigned the minimum value, in this case 62441.

```
1
2  minimum = min(min(img10));
3  maximum = max(max(img10));
4
5  [n,m] = size(img10);
6  mask = zeros(n,m);
7  for i = 1:n
8      for j = 1:m
9          if img10(i,j) > (minimum + 1500) & img10(i,j) < (minimum+1800);
10             mask(i,j) = img10(i,j);
11         else
12             mask(i,j) = 62441;
13         end
14     end
15 end
16
17 mask = uint16(mask);
18
19 figure(),
20 imshow(mask,[]);
```

Listing 4.2: Implementation of a mask in Matlab.

The previous paragraphs and visualizations show the concept of using high frequency

values in an image histogram translated into potential ROIs in a medical image, where Figure 4.7 shows part of the cervical bodies plus skull and thorax structures.

Figure 4.8 shows the image histograms corresponding to the images shown in Figure 4.3. Es it can be appreciated, each one of the histograms contains areas of high frequency values where ROIs could be associated. This is how we know the indicators of the section related to image histograms of the Selectivity step are met.

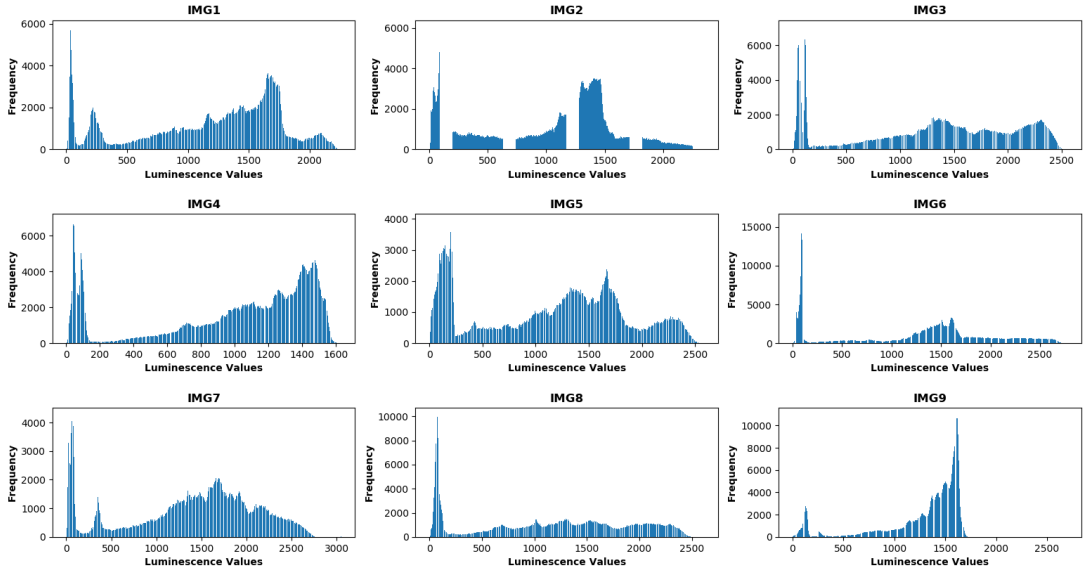


Figure 4.8: Image histograms of x-rays shown in Figure 4.3.

4.3.2 Image matrix related

Absolute values

Listing 4.3 shows the code used to obtain indicators in the minimum and maximum values for *Img10*. As stated in Table 3.2, for this project images selected should have no more than 30% of their values in their extreme ranges.

```
1  from medpy.io import load
2  import cv2
3  import numpy as np
4  import funciones as fn
5  #import timePython as tm
6  from matplotlib import pyplot as plt
7  import matplotlib.image as mpimg
8  import sys
9
10 #====Obtaining image matrix and metadata
11 nombre=sys.argv[1]
12
```

```
image_data, image_header = load(nombre)
image_data = np.transpose(image_data)

#=======Matrix Dimentions + calculations

N,M = fn.tamano(image_data)

vMax = np.max(image_data)
vMin = np.min(image_data)
totalPixels = N*M

minValue = np.count_nonzero(image_data <= (vMin+100))
maxValue = np.count_nonzero(image_data >= (vMax-100))

minInterval = maxValue/float(totalPixels)*100
maxInterval = minValue/float(totalPixels)*100

extremeRange = round(minInterval + maxInterval,2)
```

Listing 4.3: Obtaining a proportion of absolute values in Python.

We decided not to use a single value as a reference, rather use an interval of values within 100 units from the minimum and maximum. This resulted in:

- 8.08% of values in the interval considered as minimum values, and
- 0.034% of values in the interval considered as maximum values,
- Together they make up the 8.12 % of pixels in the absolute extreme values.

The results of applying the code found in Listing 4.3 recursively over the subset of images shown in section 4.2 generates the results shown in Table 4.2, where all images selected have less than 30% of the total amount absolute range values.

Table 4.2: Absolute values and related indicators from images in Figure 4.3.

Name	Min. Value	Max. Value	% in Min. Interval	% in Max. Interval	% in Extreme Ranges
img/IMG1	16204	171601	6.9687	0.6580	7.63
img/IMG2.2	9056	285149	11.5798	0.3678	11.95
img/IMG3.3	10265	251918	10.2303	0.4169	10.65
img/IMG4.6	65772	314991	12.7917	2.6710	15.46
img/IMG5.2	2413	153629	6.2388	0.0980	6.34
img/IMG6.4	7369	462175	18.7688	0.2993	19.07
img/IMG7.1	62	268329	10.8968	0.0025	10.90
img/IMG8.2	63	420359	17.0707	0.0026	17.07
img/IMG9.4	16	40614	1.6493	0.0006	1.65

4.3.3 Clinical practice related

Table 3.2 shows the indicators used by the clinician when evaluating an x-ray image, Figure 3.4 gives a visualization of the key areas and references used to be able to perform this in a clinical context.

Table 4.3 shows the results of having a clinician evaluate all nine images from Figure 4.3 following the indicators previously mentioned.

Table 4.3: Clinical evaluation of images in Figure 4.3 by standards proposed in Table 3.2.

Name	No. of Cervicals Visible	C4 in Center of Img	Bone, Soft Tissue & Air Column Ok	Spinous Processes Visible	Rami of Mandibule Superimposed in C1 to C2	Rotation of Projection
img/IMG1	7	no	yes	7	no	no
img/IMG2.2	7	no	yes	7	no	no
img/IMG3.3	7	no	yes	6	yes	yes
img/IMG4.6	7	yes	yes	6	yes	yes
img/IMG5.2	7	no	yes	7	no	yes
img/IMG6.4	4	no	no	3	no	yes
img/IMG7.1	7	no	yes	6	no	yes
img/IMG8.2	7	no	yes	7	no	no
img/IMG9.4	6	no	no	5	yes	yes

Non of the images in Table 4.3 were reported to comply with all the indicators, it is important to highlight that this evaluation was performed by a medical doctor. Indicators with the least compliance were *C4 in Center of Img* and *Rotation of Projection*, both indicating the way a patient is set and how the x-ray penetrates the tissue, this being an issue with the technician in charge of performing the study.

A high variability in the results of the evaluation of a subset of images gives us an idea of the variability in encountered by practitioners in the every day clinical practice.

4.4 Linearity

Figure 3.6 and Table 3.2 in Section 3.1.3 provide the theoretical context for the results in this section. Three main indicators are proposed and they have to do with machine performance, image structure and the way an image is interpreted.

It is in this section that a method for dimensionality reduction is presented, based on Principal Components Analysis, as the main linear transformation.

4.4.1 Machine performance

The standards of performance for the machine used to acquire and digitize the image are usually out of the scope of the person in charge of the analytical process.

In this case, the radiology department is in charge of maintenance and technical adjustments of the machines used for clinical practice.

Still, the manufacturer can be extracted from the DICOM file as shown in Listing 4.4.

```
1 from medpy.io import load
```

```
2 import cv2
3 import numpy as np
4 import funciones as fn
5 #import timePython as tm
6 from matplotlib import pyplot as plt
7 import matplotlib.image as mpimg
8 import sys
9
10 #========== LINEARITY=============
11
12 #--Machine Performance-----
13 In[2]: image_header.Manufacturer
14 Out[2]: 'KONICA MINOLTA'
15
16 In[3]: image_header.ManufacturerModelName
17 Out[3]: '0817'
```

Listing 4.4: Obtaining x-ray machine manufacturer and model.

4.4.2 Image structure

As stated in Section 3.1.3: the amount of luminescence values in the image matrix, should be linearly relatable to different tissue densities evaluated by the clinician. To accomplish this, a linear method has to be proposed and implemented. The method used for this project is based on the algorithm of Principal Components Analysis and is described in detail in Section 3.1.4, this method allowed us to reduce the dimensionality of the luminescence vector shortening the computational time used to process the resulting images, also, special attention was put to the algorithm so that spatial relations within pixels were maintained, keeping the resulting image clinically relevant was also an objective.

Section 3.1.4 presents the flowchart for the linear model to achieve dimensionality reduction proposed for this section. It is a five step process taking as input a medical image of an x-ray in DICOM protocol at 16bit and delivering as an output an image matrix in the same protocol and characteristics but with a significantly smaller vector of luminescence values.

Figure 4.9 shows the implementation of the process shown in the flowchart in Figure 3.7.

All the explanations for the mathematics in Figure 4.9 are detailed in Section 3.1.4, for the purpose of facilitating the interpretation in this chapter some guidelines are provided in the following paragraph.
Components in the flowchart are to be understood as follows:

A, B, C**:** Image matrices of $n \times m$ dimensions.

$histEq()$**:** Histogram equalization function.

hac**:** Image cumulative histogram.

PCA**:** Principal Components Analysis.

$PC1$**:** First Principal Component.

$himhist()$**:** Image histogram function.

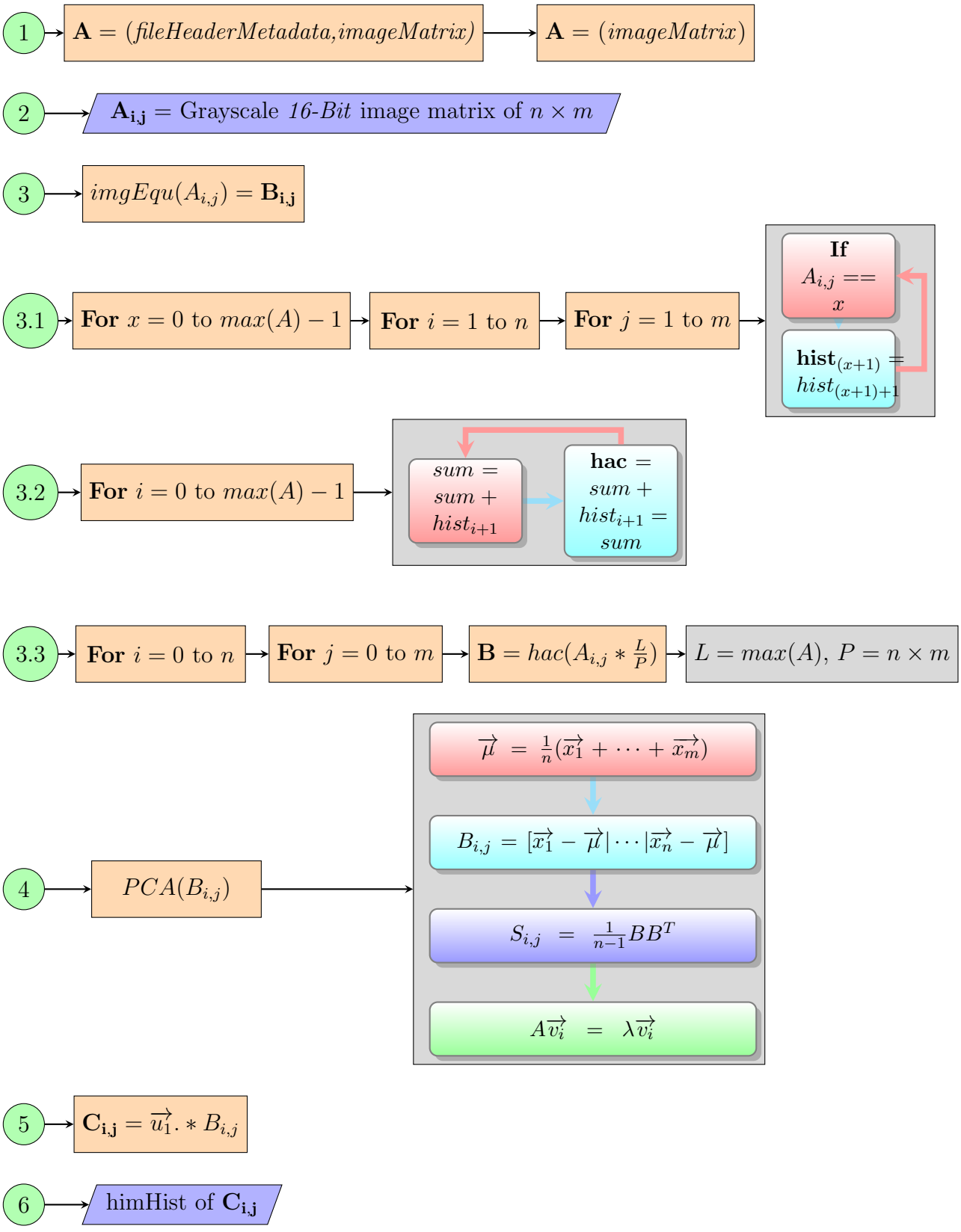


Figure 4.9: Flowchart for the implementation of the linear model used in this project.

The result of applying the method proposed to the image shown in Figure 4.4 is shown in Figures 4.10, 4.11 and 4.12.

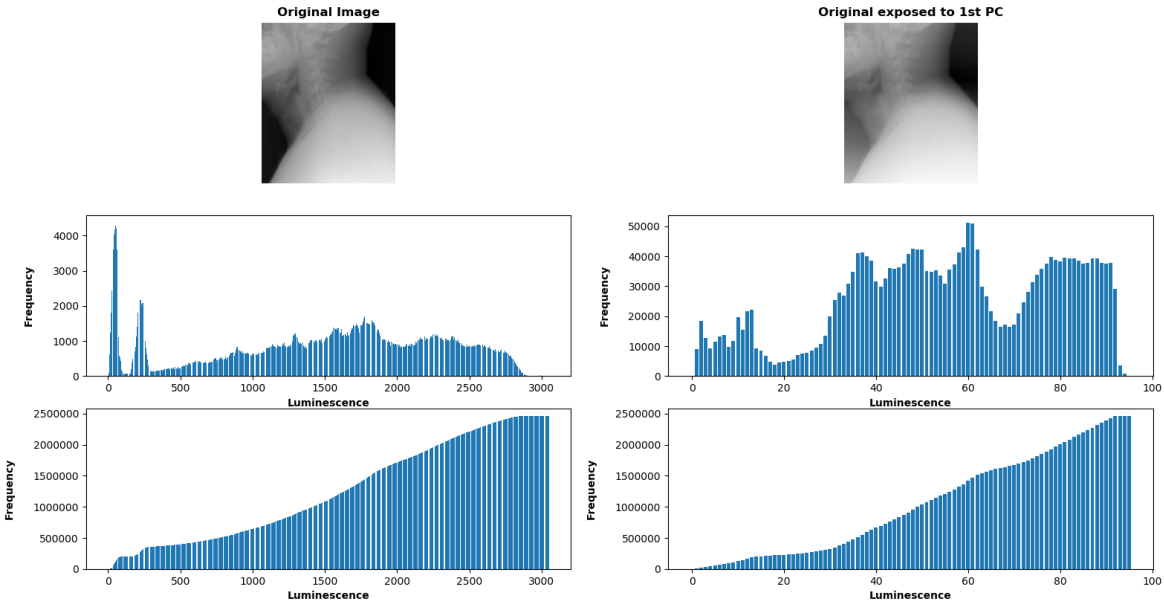


Figure 4.10: Result of applying dimensionality reduction to the original image as described in Figure 4.9.

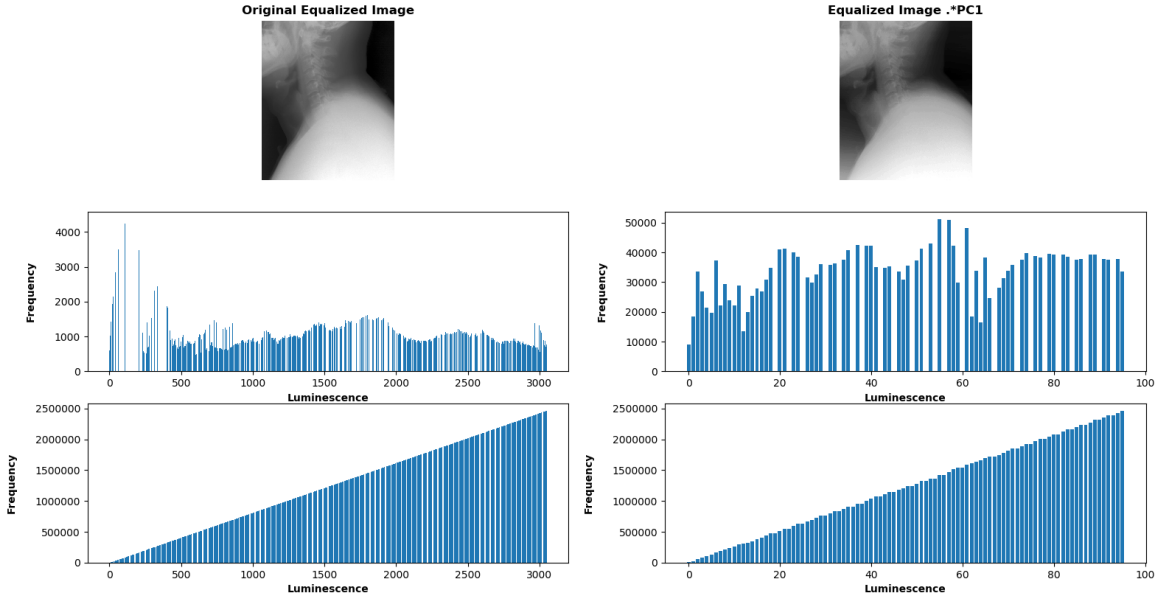


Figure 4.11: Result of applying dimensionality reduction to the equalization of the original image as described in Figure 4.9.

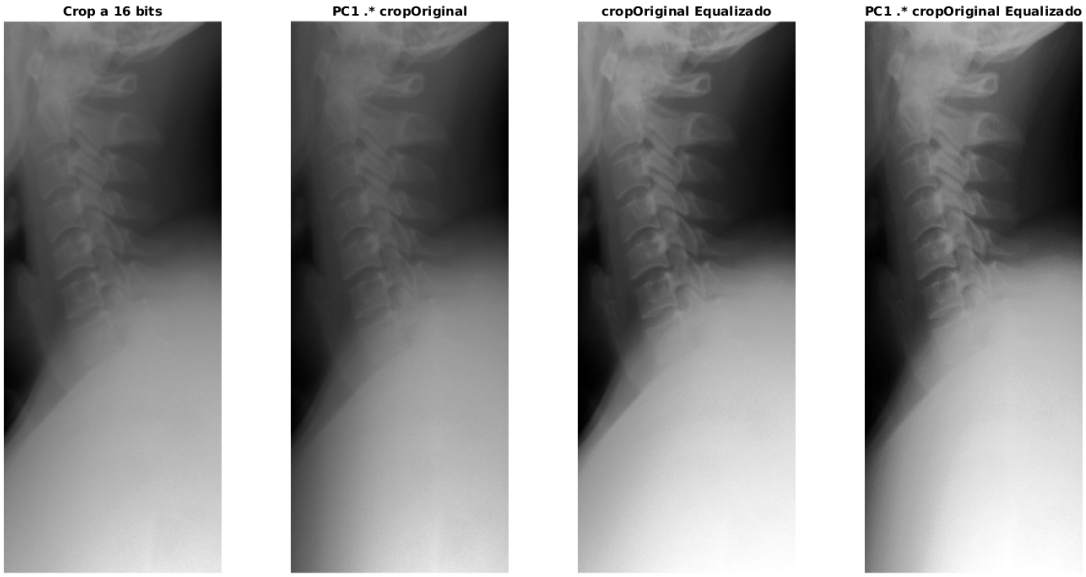


Figure 4.12: Resulting enlarged images from Figures 4.10 and 4.11.

The code implementation of this method is shown in the following listing (4.5).

```
1
2 clc, clear
3 cd '/home/...'
4
5 tic
6 % Getting and cropping DICOM file.
7
8 img = crop_cervicales4(uigetfile('*','Select MATLAB code file to generate
      crop'));
9
10 % Image dimensions:
11 [n,m] = size(img);
12
13 %---------------- COVARIANCE MATRIX ----------------------
14 imgDouble = double(img);
15 s = transpose(imgDouble)*imgDouble;
16 S = s/n;
17
18 %---------- CALCULATING THE RIGHT EIGENVALUES AND EIGENVECTORES.
19
20 eigenvalues = eig(S);
21 [eigenvectors,D] = eig(S);
22
23 %------------ MATRIX RECONSTRUCTION-------------------------------
24
25 imgReconstructDoubleT = diag(eigenvectors(:,m)) * transpose(imgDouble);
```

```
26
27 imgReconstructDouble = transpose(imgReconstructDoubleT);
28
29 imgReconstruct = uint16(imgReconstructDouble);
30
31 [x,hac] = hist16(img,3,1);
32
33
34 %-------------IMAGE MATRIX VISUALIZATION-------------------------------
35
36 figure()
37 subplot(3,4,1)
38 imshow(img,[]);
39 title('Original')
40
41 subplot(3,4,2)
42 imshow(imgReconstruct,[]);
43 title('Original exposed to 1st PC')
44
45 subplot(3,4,3)
46 imgEq = img16Eq(img,x,hac,4);
47 imshow(imgEq,[]);
48 title('Original + Equalization')
49
50 subplot(3,4,4)
51 imgPCAEqComp = imgPCAEq;
52 imshow(imcomplement(imgPCAEqComp),[]);
53 title('PC1 .* Original + Equalization')
54 %--------
55 subplot(3,4,5)
56 hist16(img,4,1);
57
58 subplot(3,4,6)
59 hist16(imgReconstruct,4,1);
60
61 subplot(3,4,7)
62 hist16(imgEq,4,1);
63
64 subplot(3,4,8)
65 hist16(imgPCAEqComp,4,1);
66 %---------
67 subplot(3,4,9)
68 hist16(img,4,2);
69
70 subplot(3,4,10)
71 hist16(imgReconstruct,4,2);
72
73 subplot(3,4,11)
74 hist16(imgEq,4,2);
75
76 subplot(3,4,12)
```

```
77 hist16(imgPCAEqComp,4,2);
```

Listing 4.5: Implementation of the process shown in Figure 4.9 in Matlab.

Figure 4.13 shows the resulting image histograms of the raw images compared to the ones corresponding to the images resulting after the analysis shown in Figure 4.9.

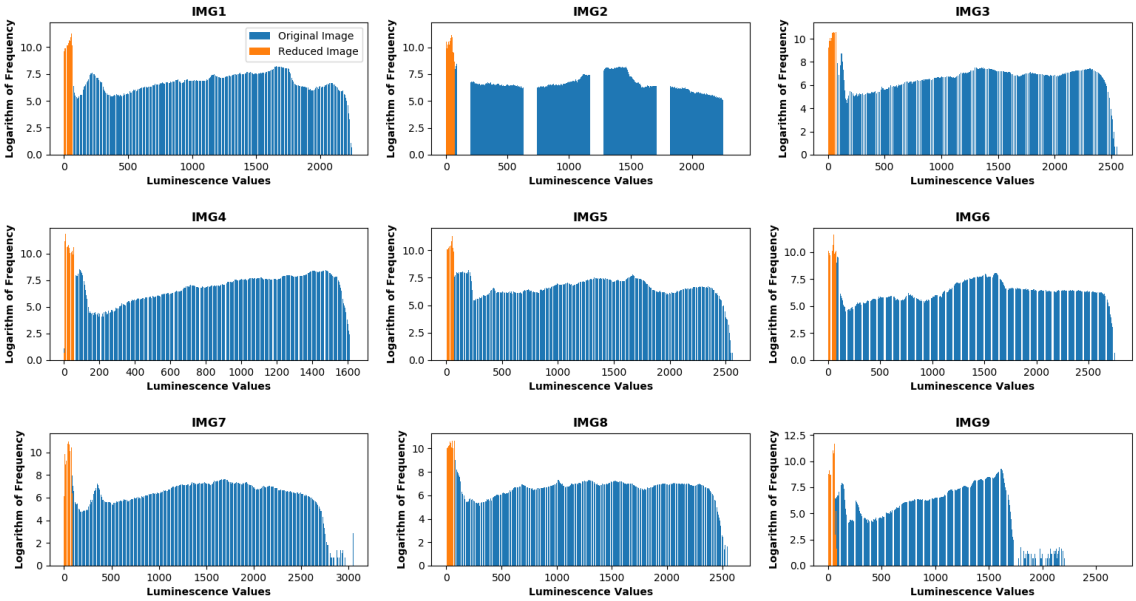


Figure 4.13: Image histograms of raw images (blue) and their corresponding images after dimensionality reduction and histogram equalization (orange). Notice the use of logarithms in the y-axis to improve the visualization.

Table 4.4 presents indicators to assess the dimensionality reduction process performed on the nine images of the subset presented at the begining of this chapter in Figure 4.3. It is of significance to highlight the percentage of reduction achieved in each image by the method (shown in the last column of the Table) where, in all cases, a reduction greater than 95% was achieved, this can be visually associated with the image histograms of Figure 4.13.

Name	Min. Value	Max. Value	Vector Length	Min. Reduced Value	Max. Reduced Value	Reduced Vector	% Reduced
img/IMG1	1027	3287	2260	0	97	97	95.71
img/IMG2.2	1281	3633	2352	0	93	93	96.05
img/IMG3.3	230	2784	2554	0	79	79	96.91
img/IMG4.6	1221	2853	1632	0	47	47	97.12
img/IMG5.2	269	2856	2587	0	84	84	96.75
img/IMG6.4	302	3079	2777	0	80	80	97.12
img/IMG7.1	127	3179	3052	0	85	85	97.21
img/IMG8.2	152	2771	2619	0	91	91	96.53
img/IMG9.4	129	2830	2701	0	59	59	97.82

Table 4.4: Performance indicators of the dimensionality reduction algorithm implemented. Data from Figure 4.13.

4.4.3 Interpretation of image

Although the scope of this project does not comprise the implementation in a clinical setting, it has been a priority to keep the results as significant as possible in this context, facilitating their introduction to a clinical practice. This is clearly stated in the fourth specific objective in Section 1.2.4.

To evaluate this, a brief survey asking 12 different physicians with varied backgrounds and training was carried out. Two images were shown in the survey, one showing a crop of the original unaltered image and the other showing the result of exposing the same image to the linear dimensionality reduction method, Figure 4.14 shows the information seen by clinicians.

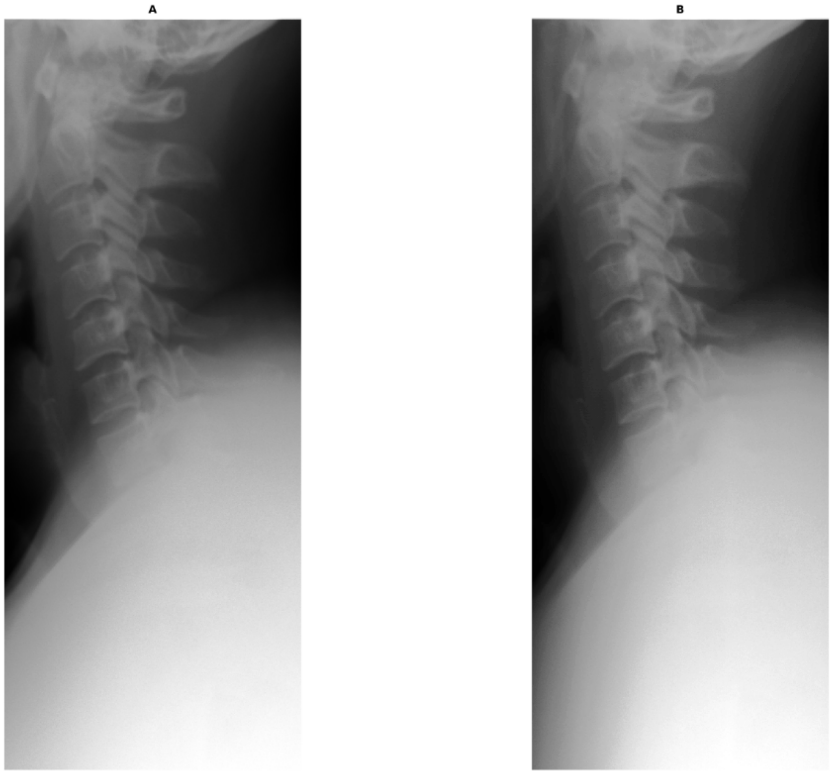


Figure 4.14: Images shown to physicians in the survey. **A** shows a crop of the original image. **B**, a crop of the same image after being exposed to the method proposed in this project.

Figure 4.15 shows the proportion of general practitioners versus specialists who answered the survey.

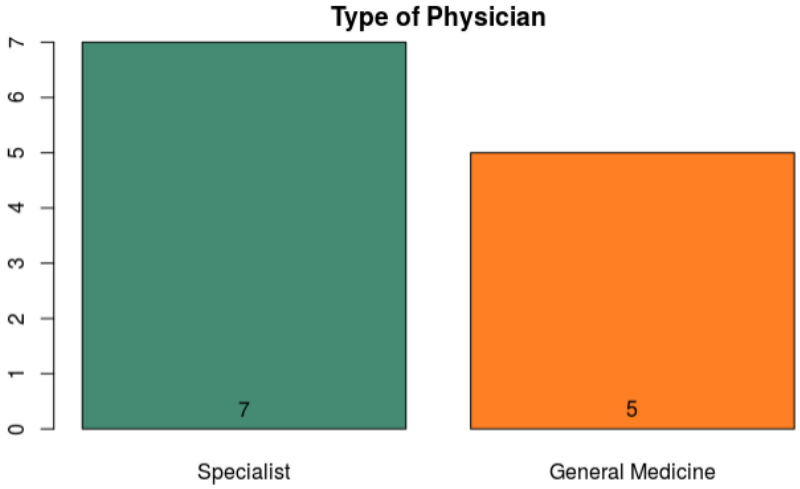


Figure 4.15: Physicians who participated in the survey stratified by general practice or specialty.

Out of the specialists, the distribution by discipline was as follows:

- 3 in internal medicine,
- 2 in traumatology and orthopedics,
- 1 in cardiology, and
- 1 in pediatrics.

After requesting the practitioners to carefully observe both images, it was asked of them to select the option that best described their opinion regarding the images. The following list shows the options available and how we coded them in the analysis.

- Image B (on the right) is better than image A (on the left), $\{B > A\}$.
- Image A (on the left) is better than image B (on the right), $\{A > B\}$.
- There is no difference. I would use both in my decision making process, $\{A = B\}$.

Figure 4.16 shows a summary of the responses obtained. As it can be appreciated, 83.3% of physicians found the reduced image (B) equally or more useful for their decision making in a clinical setting. This has to be evaluated by a larger set of physicians, but it situates the project in the direction of clinical pertinence and usability.

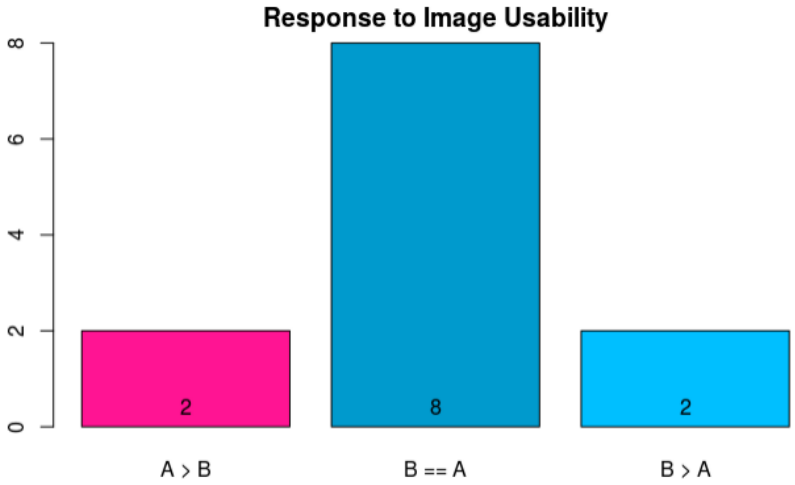


Figure 4.16: Summary of responses obtained in survey.

4.5 Precision

To evaluate this section of the method implementation, three different computers were used, their characteristics are shown in the following Table (4.5).

There was some variability in the results, specifically when using different versions of Matlab, this can be more appreciated in the next section. Regarding random error, there was no need to alter the code beyond changing working directories in the specific machines. As long as the folders and scripts needed were recognized by the path and working directory everything ran as expected.

Indicators	**Computer A**	**Computer B**	**Computer C**
Operative System	Ubuntu 16.04 LTS	Ubuntu 16.04 LTS	Windows 7
Hardware	CPU Intel Core i7 4790 @ 3.60 Ghz RAM 8GB VGA: Xeon E3-1200 v3 4th generation	CPU Intel Core i7 4700HQ @ 2.40Ghz RAM 16GB VGA: Intel Integrated Graphics Controller	CPU AMD64. Family 20 Model 1, 1600 Mhz RAM 4GB VGA:
Matlab Version	R2017a(9.2.0.538062)	R2017a(9.2.0.538062)	R2015a(8.5.0.197613)

Table 4.5: Characteristics of computers used to evaluate indicators of precision.

The time it took on each machine to load and process an image is presented in Table 4.6. *Load Time* refers to the ammount of time taken to load the DICOM file into the working environment, while the data shown in each image shows the time taken to run the lineal method in that specific image matrix, *Global Time* indicates the time taken to perform the whole process. As it can be appreciated, hardware and software have a direct impact on the computational time taken to run the process and this has to be considered when implementing the method.

	Computer A	Computer B	Computer C
Load Time	0.1145	0.1133	2.3331
Img1	0.5382	0.7439	17.443
Img2	0.5382	0.6132	18.1153
Img3	0.5458	0.6062	16.7719
Img4	0.5512	0.6126	16.7647
Img5	0.5578	0.6068	16.8686
Img6	0.5341	0.5929	16.54
Img7	0.5334	0.6099	17.621
Img8	0.5381	0.6074	16.4741
Img9	0.5418	0.6003	16.3904
Global Time	5.0241	5.7065	155.3222

Table 4.6: Computational time in seconds taken to implement the process in three different machines with software and hardware variations.

4.5.1 Error source

4.6 Limit of Detection and Limit of Quantitation

As mentioned in Section 3.1.1, the depth of luminescence of the images used for this project is in the order of 2^{16}, also called *16-bit*, which means that each cell in the image matrix, later called pixel when visualized on a screen, can take $2^{16} = 65536$ different values where 0 will translate as absolute black and 65535, as white. Nevertheless, not all images use the whole vector of luminescence values, it is more common to find an interval of values taken from a subset of the luminescence vector. This particularity becomes relevant when considering LoD and LoQ indicators.

For image matrices, the lowest value that a pixel can take will always be zero, when dealing with a grayscale color space. Still, a 16bit image can have an LoD value different than zero as mentioned in the previous paragraph.

LoQ is an indicator that will give more information when comparing the luminescence values of the original image versus the image resulting from the dimensionality reduction process.

In Table 4.8, LoQ is a calculation that takes the values from the original image (Table 4.7), where the LoQ of an image resulting from applying a dimensionality reduction linear model will have more than the unit in each step of Quantitation.

Taking IMG8.2 as example, the resulting images when applying the method with computers A and C yields a LoQ of 27, meaning that each monotonically increase in the luminescence vector of the processed images corresponds to 27 monotonically increases in the original image.

There is more space for data to be expressed in the original image, in other words

it's LoQ is smaller, nevertheless in Section 4.4.3 the argument for the clinical use of the resulting images is made, against the original ones.

	Original Image			
Name	**Max Value**	**Min Value**	**Vector**	**LoQ**
IMG8.2	2771	152	2619	1
IMG1	3287	1027	2260	1
IMG3.3	2784	230	2554	1
IMG4.6	2853	1221	1632	1
IMG9.4	2830	129	2701	1
IMG7.1	3179	127	3052	1
IMG6.4	3079	302	2777	1
IMG2.2	3633	1281	2352	1
IMG5.2	2856	269	2587	1

Table 4.7: Reference values from original images.

	Computer A				***Computer B***				***Computer C***			
Name	**Max Value**	**Min Value**	**Vector**	**LoQ**	**Max Value**	**Min Value**	**Vector**	**LoQ**	**Max Value**	**Min Value**	**Vector**	**LoQ**
IMG8.2	100	4	96	27	91	0	91	28	100	4	96	27
IMG1	116	21	95	23	72	0	72	31	116	21	95	23
IMG3.3	110	4	106	24	79	0	79	32	110	4	106	24
IMG4.6	97	27	70	23	47	0	47	34	97	27	70	23
IMG9.4	106	4	102	26	59	0	59	45	106	4	102	26
IMG7.1	115	3	112	27	85	0	85	35	115	3	112	27
IMG6.4	124	5	119	23	80	0	80	34	124	5	119	23
IMG2.2	124	27	97	24	93	0	93	25	124	27	97	24
IMG5.2	106	5	101	25	84	0	84	30	106	5	101	25

Table 4.8: Results of the implementation of LoD (or minimum values) and LoQ by the computers mentioned in Table 4.5.

4.7 Robustness

Figure 4.17 shows the indicators presented for the model evaluation, those in red are indicators that, if taken away from the process, would make it impossible to deliver the results it what was designed to generate. Those in yellow show indicators that could be excluded depending on the specific case. Steps in green represent indicators that could be ignored or not obtained as rigorous as the rest and results would still be relevant in the clinical practice.

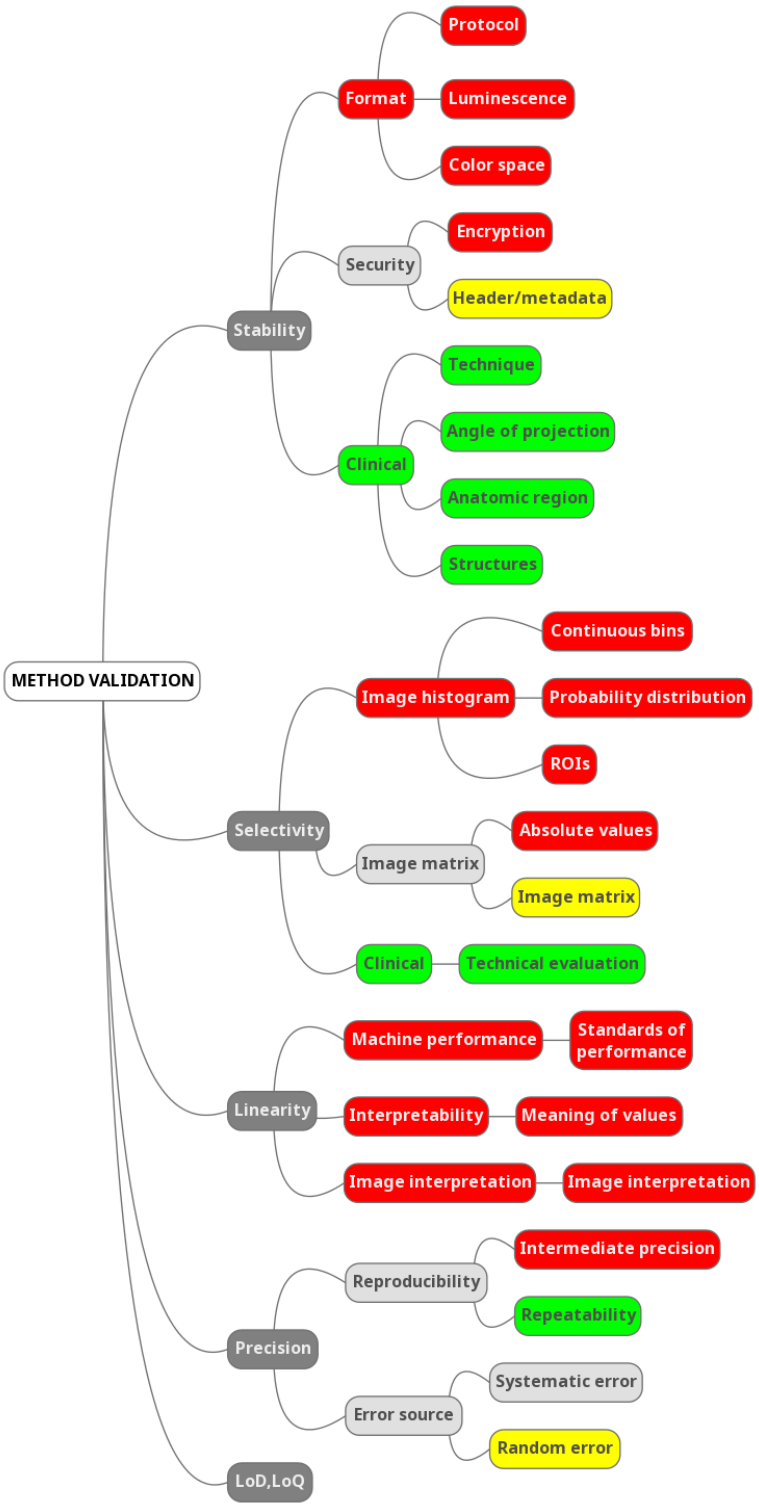


Figure 4.17: Evaluation of the model robustness.

It is important to notice that the indicators in green are mostly clinical, meaning, in this project, that even when the clinical technique of the X-ray images is not ideal, the method still delivers results within the expected quality. This is evidence of the robustness of the method given that it is a tool to implement Computer Assisted Diagnosis.

Physicians could eventually significantly benefit from this approach, given that it seems to be unaltered or minimally affected by clinical variability.

Chapter 5

Discussion and Conclusions

This chapter presents conclusions achieved based on the results reported in Chapter 4, and future work derived from the this project.

5.1 Discussion

Medical images comprise a significant proportion of the modern clinical day to day practice. As mentioned in Section 2 of this thesis, the amount of data available in the medical context has exploded since the last decade. This phenomenon is not particular to the medical field, it has been happening in other disciplines as well, nevertheless the clinical areas have been rather reticent to the adoption of novel techniques to turn the amount of data generated into significant information.

Two reasons why this adoption might be slower with health care are privacy and clarity in the automation of the decision making process. The use of computational resources to process data and turn it into clinically significant information is a challenging task. On one side, high levels of clinical skills are needed to understand medical information and what makes it relevant, on the other, data analytics competences are equally relevant. Teams wanting to tackle these issues should consider a transdisciplinary approach with medical and computational competences. This is one of the core reasons why the results of this project are relevant.

Dimensionality reduction in image matrices has been achieved with a method implemented from the data analytics perspective and has yielded results that impact a digital image processing approach, while keeping the resulting image clinically relevant for physicians. This opens the possibilities for a Computer Assisted Diagnosis method that allows algorithms to deal with less data, therefore, yield faster and more accurate results.

While the method is technically robust from an algorithmic and image processing perspective, it produces medical images that can still be used in the day to day practice and, according to the results shown in Section 4.4.3, the majority of physicians would not even realize that they are working with dimensionally reduced images.

It was a priority for this project to function under the reproducible research framework, I adapted the method validation proposal from [9], they work with liquid chromatography mass spectrometry (LC-MS) and this work focuses on medical images, still both subjects of study can be understood as signals and this is what enabled the adaptation of their method.

The resulting method validation process is a tool that delivers a workflow to assure reproducibility and diminish, to minimum or nonexistent,low quality results.

5.2 Conclusions

Going back to Section 1.2.4 of this document to consider the specific objectives of this project. Section 2.5 presents the results of a literature review on the most published algorithms for dimensionality reduction in medical and biomedical data, it also shows the tendency of publication from 2007 to 2017. Three algorithms dominated the publications in that time interval:

- Principal Components Analysis (PCA),
- Artificial Neural Networks (ANN), and
- Support Vector Machines (SVM).

PCA being the most published within the time interval and ANNs being the one with the highest rates of publishing in the last four years. The decision to focus on PCA was taken because it uses a lineal approximation to yield results (eigenvalues and eigenvectors) having therefore, by nature, high explainability while ANNs are consider "black box" algorithms [60]. The previous paragraphs answer the first specific objective and the first half of the second one, given the rationale under the decision to use PCA for the implementation of the model in this project.

The method validation steps adapted from Kruve and collaborators [9] facilitated the implementation of specific objectives 2 (Section 4.3) and 3 (Section 4.4), while a consultation with physicians provided the data from clinical practitioners, shown in Section 4.4.3 and fulfilled objective 4.

The transdisciplinary approach to this project from the beginning has allowed the results to be clinically pertinent by nature, the method was developed with the objective to be translational and "close the gap between bench and bedside" [61].

Even though the concept of translational medicine has been used more widely in molecular medicine, Steven H. Woolf defines it as the "enterprise of harnessing knowledge from basic science to produce new drugs, devices, and treatment options for patients" [62]. Under this definition, this project can be labeled as translational research for clinical purposes.

5.3 Future work

Once dimensionality reduction has been achieved in medical images without altering the spatial pixel relations and staying clinically relevant, the natural step forward is to work with series of images from the same region to generate hypercubes and 3D models derived from the reduced images. This approach would be pertinent to apply to Computed Axial Tomogrpahy (CAT) and Magnetic Resonance Imaging (MRI) series, where the a given set of image matrices from a common anatomical region are taken by the machine and a hypercube is generated to produce a 3D image of the region.

Various questions arise from this idea, will the hypercube stay as relevant as the reduced images are after being processed by the method? How much will computational time be reduced by applying this method? Will it stay as relevant as a "non-reduced" series of images for clinical practice?

On the other hand, the method validation process to ensure reproducibility is a tool that should be tested with larger and more varied subsets of images to understand and adjust the variability in the indicators proposed.

5.4 Academic production

1. Published article: *Oportunidades que presenta el aprendizaje automatizado para la integración y análisis predictivo del ámbito de datos médicos.* 2o Congreso de Investigación Universitaria de la División Interamericana. ISBN: 978-607-8001-14-9. Gener José Avilés Rodríguez, María de los Ángeles Cosío León.

2. Accepted article: *Efecto de la calibración de parámetros mediante un diseño Taguchi en el algoritmo GRASP resolviendo el problema de rutas de vehículos con restricciones de tiempo.* Revista Computación y Sistemas ISSN-2007-9737, 2017. Alma Danisa Romero Ocaño, María de los Ángeles Cosío León, Víctor Manuel Valenzuela Alcaraz, Gener Avilés Rodríguez, Anabel Martínez Vargas.

3. Accepted article: *Reducción de Dimensionalidad en los Datos, Una Aproximación desde la Función Cerebral.* Komputer Sapiens. ISSN 2007-0691,2018. Gener José Avilés Rodríguez, María de los Ángeles Cosío León, Arturo Jesús Laflor Hernández.

Bibliography

[1] Shadi Alzubi, Naveed Islam, and Maysam Abbod. 3D Multiresolution Analysis for Reduced Features Segmentation of Medical Volumes Using PCA. *Ieee*, (December):6–9, 2010.

[2] Marco Loog, Bram Van Ginneken, and Robert PW Duin. Dimensionality reduction by canonical contextual correlation projections. In *European Conference on Computer Vision*, pages 562–573. Springer, 2004.

[3] Benjamin M Kandel, Danny JJ Wang, James C Gee, and Brian B Avants. Eigenanatomy: Sparse dimensionality reduction for multi-modal medical image analysis. *Methods*, 73:43–53, 2015.

[4] Irem Ersöz Kaya, Ayça Çakmak Pehlivanlı, Emine Gezmez Sekizkardeş, and Turgay Ibrikci. Pca based clustering for brain tumor segmentation of t1w mri images. *Computer Methods and Programs in Biomedicine*, 140:19–28, 2017.

[5] Safaa E Amin and MA Megeed. Brain tumor diagnosis systems based on artificial neural networks and segmentation using mri. In *Informatics and Systems (INFOS), 2012 8th International Conference on*, pages MM–119. IEEE, 2012.

[6] Yuxuan Huang, Zhongpan Qiu, and Zhijun Song. 3d reconstruction and visualization from 2d ct images. In *IT in Medicine and Education (ITME), 2011 International Symposium on*, volume 2, pages 153–157. IEEE, 2011.

[7] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380, 2000.

[8] Oge Marques. *Practical image and video processing using MATLAB*. John Wiley & Sons, 2011.

[9] Anneli Kruve, Riin Rebane, Karin Kipper, Maarja-Liisa Oldekop, Hanno Evard, Koit Herodes, Pekka Ravio, and Ivo Leito. Tutorial review on validation of liquid chromatography–mass spectrometry methods: Part ii. *Analytica chimica acta*, 870:8–28, 2015.

[10] Peter Densen. Challenges and opportunities facing medical education. *Transactions of the American Clinical and Climatological Association*, 122:48, 2011.

[11] Eugene C Nelson, Mark E Splaine, Paul B Batalden, and Stephen K Plume. Building measurement and data collection into medical practice. *Annals of internal medicine*, 128(6):460–466, 1998.

[12] Saint John Walker. Big data: A revolution that will transform how we live, work, and think, 2014.

[13] John A Swets, Robyn M Dawes, and John Monahan. Better decisions through science. *Scientific American*, 283(4):82–87, 2000.

[14] Peter Mildenberger, Marco Eichelberg, and Eric Martin. Introduction to the dicom standard. *European radiology*, 12(4):920–927, 2002.

[15] Philip Greenland, Joseph S Alpert, George A Beller, Emelia J Benjamin, Matthew J Budoff, Zahi A Fayad, Elyse Foster, Mark A Hlatky, John McB Hodgson, Frederick G Kushner, et al. 2010 accf/aha guideline for assessment of cardiovascular risk in asymptomatic adults: a report of the american college of cardiology foundation/american heart association task force on practice guidelines developed in collaboration with the american society of echocardiography, american society of nuclear cardiology, society of atherosclerosis imaging and prevention, society for cardiovascular angiography and interventions, society of cardiovascular computed tomography, and society for cardiovascular magnetic resonance. *Journal of the American College of Cardiology*, 56(25):e50–e103, 2010.

[16] N Pipitone, A Versari, and C Salvarani. Role of imaging studies in the diagnosis and follow-up of large-vessel vasculitis: an update. *Rheumatology*, 47(4):403–408, 2008.

[17] Sally F Barrington, N George Mikhaeel, Lale Kostakoglu, Michel Meignan, Martin Hutchings, Stefan P Müeller, Lawrence H Schwartz, Emanuele Zucca, Richard I Fisher, Judith Trotman, et al. Role of imaging in the staging and response assessment of lymphoma: consensus of the international conference on malignant lymphomas imaging working group. *Journal of clinical oncology*, 32(27):3048–3058, 2014.

[18] Julie A Hides, Carolyn A Richardson, and Gwendolen A Jull. Use of real-time ultrasound imaging for feedback in rehabilitation. *Manual therapy*, 3(3):125–131, 1998.

[19] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2(1):3, 2014.

[20] Krishan Soni, Mythreyi Bhargavan, Howard P Forman, and Jonathan H Sunshine. Who's underworked and who's overworked now? an update on radiologist shortage and surplus. *American Journal of Roentgenology*, 194(3):697–703, 2010.

[21] Bram van Ginneken, Cornelia M Schaefer-Prokop, and Mathias Prokop. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology*, 261(3):719–732, 2011.

[22] Shijun Wang and Ronald M Summers. Machine learning and radiology. *Medical image analysis*, 16(5):933–951, 2012.

[23] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10:66–71, 2009.

[24] Johann Summhammer. Maximum predictive power and the superposition principle. *International Journal of Theoretical Physics*, 33(1):171–178, 1994.

[25] Imola K Fodor. A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Lab., CA (US), 2002.

[26] Markus Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3):303–304, 2008.

[27] Ian T Jolliffe. Principal component analysis and factor analysis. *Principal component analysis*, pages 150–166, 2002.

[28] Lindsay I Smith et al. A tutorial on principal components analysis. *Cornell University, USA*, 51(52):65, 2002.

[29] Matthan WA Caan, Koen A Vermeer, Lucas J van Vliet, Charles BLM Majoie, BD Peters, GJ den Heeten, and Frans M Vos. Shaving diffusion tensor images in discriminant analysis: A study into schizophrenia. *Medical image analysis*, 10(6):841–849, 2006.

[30] Ciprian M Crainiceanu, Brian S Caffo, Sheng Luo, Vadim M Zipunnikov, and Naresh M Punjabi. Population value decomposition, a framework for the analysis of image populations. *Journal of the American Statistical Association*, 106(495):775–790, 2011.

[31] Yong Rui, Alfred C She, and Thomas S Huang. Modified fourier descriptors for shape representation-a practical approach. In *Proc of First International Workshop on Image Databases and Multi Media Search*, pages 22–23. Citeseer, 1996.

[32] Zhang Jing and Kang Bao Sheng. A novel medical freehand sketch 3d model retrieval method by dimensionality reduction and feature vector transformation. *Computational and mathematical methods in medicine*, 2016, 2016.

[33] Jérémie Bouttier, Philippe Di Francesco, and Emmanuel Guitter. Geodesic distance in planar graphs. *Nuclear Physics B*, 663(3):535–567, 2003.

[34] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[35] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[36] Boaz Nadler, Stéphane Lafon, Ronald R Coifman, and Ioannis G Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

[37] A Schenone, F Firenze, F Acquarone, M Gambaro, F Masulli, and L Andreucci. Segmentation of multivariate medical images via unsupervised clustering with "adaptive resolution". *Computerized medical imaging and graphics*, 20(3):119–129, 1996.

[38] Shadi Al Zu'bi, Naveed Islam, and Maysam Abbod. 3d multiresolution analysis for reduced features segmentation of medical volumes using pca. In *Circuits and Systems (APCCAS), 2010 IEEE Asia Pacific Conference on*, pages 604–607. IEEE, 2010.

[39] Tobias Heimann and Hans-Peter Meinzer. Statistical shape models for 3d medical image segmentation: a review. *Medical image analysis*, 13(4):543–563, 2009.

[40] Nilanjan Dey, Vikrant Bhateja, and Aboul Ella Hassanien. Medical imaging in clinical applications.

[41] Ahmad Alzu'bi, Abbes Amira, and Naeem Ramzan. Semantic content-based image retrieval: A comprehensive study. *Journal of Visual Communication and Image Representation*, 32:20–54, 2015.

[42] Luka Šajn and Matjaž Kukar. Image processing and machine learning for fully automated probabilistic evaluation of medical images. *Computer methods and programs in biomedicine*, 104(3):e75–e86, 2011.

[43] Reyer Zwiggelaar, Susan M Astley, Caroline RM Boggis, and Christopher J Taylor. Linear structures in mammographic images: detection and classification. *IEEE transactions on medical imaging*, 23(9):1077–1086, 2004.

[44] R Lazcano, D Madroñal, R Salvador, K Desnos, M Pelcat, R Guerra, H Fabelo, S Ortega, S Lopez, GM Callico, et al. Porting a pca-based hyperspectral image dimensionality reduction algorithm for brain cancer detection on a manycore architecture. *Journal of Systems Architecture*, 77:101–111, 2017.

[45] Ryo Jimbo, Julia Fernandez-Rodriguez, Young-Taeg Sul, and Carina B Johansson. Principal component analysis: A novel analysis to evaluate the characteristics of osseointegration of different implant surfaces. *Implant dentistry*, 20(5):364–368, 2011.

[46] Alireza Akhbardeh and Michael A Jacobs. Comparative analysis of nonlinear dimensionality reduction techniques for breast mri segmentation. *Medical physics*, 39(4):2275–2289, 2012.

[47] Benjamin Doerr, Carola Doerr, and Franziska Ebel. From black-box complexity to designing new genetic algorithms. *Theoretical Computer Science*, 567:87–104, 2015.

[48] Anil K Jain. *Fundamentals of digital image processing*. Prentice-Hall, Inc., 1989.

[49] Chris Solomon and Toby Breckon. *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. John Wiley & Sons, 2011.

[50] Rafael C Gonzalez and Richard E Woods. *Digital Image Processing*. Prentice Hall, 3rd edition, 2008.

[51] Andreas Thiel, Johannes Bernarding, Johachim Hohmann, Domagoi Cosic, and Thomas Tolxdorff. Security concepts in clinical applications using dicom. In *Medical Imaging 1998: PACS Design and Evaluation: Engineering and Clinical Issues*, volume 3339, pages 11–23. International Society for Optics and Photonics, 1998.

[52] Anneli Kruve, Riin Rebane, Karin Kipper, Maarja-Liisa Oldekop, Hanno Evard, Koit Herodes, Pekka Ravio, and Ivo Leito. Tutorial review on validation of liquid chromatography–mass spectrometry methods: Part i. *Analytica chimica acta*, 870:29–44, 2015.

[53] Gerald De Lacey, Simon Morley, and Laurence Berman. *The Chest X-Ray: A Survival Guide E-Book*. Elsevier Health Sciences, 2012.

[54] Im Deok Jung, Min Kook Cho, Sang Min Lee, Kong Myeong Bae, Phill Gu Jung, Chi Hoon Lee, Jae Min Lee, Seungman Yun, Ho Kyung Kim, Seong Sik Kim, et al. Flexible gd2o2s: Tb scintillators pixelated with polyethylene microstructures for digital x-ray image sensors. *Journal of Micromechanics and Microengineering*, 19(1):015014, 2008.

[55] Paul R Halmos. What does the spectral theorem say? *The American Mathematical Monthly*, 70(3):241–247, 1963.

[56] Bertil Magnusson. *The fitness for purpose of analytical methods: a laboratory guide to method validation and related topics (2014)*. Eurachem, 2014.

[57] Bertil Magnusson, Teemu Näykki, Håvard Hovind, and Mikael Krysell. Handbook for calculation of measurement uncertainty in environmental. 1.

[58] Michael Thompson, Stephen LR Ellison, and Roger Wood. Harmonized guidelines for single-laboratory validation of methods of analysis (iupac technical report). *Pure and Applied Chemistry*, 74(5):835–855, 2002.

[59] EC Dg-sanco. Guidance document on analytical quality control and validation procedures for pesticide residues analysis in food and feed. *SANCO/12571*, 2013.

[60] Julian D Olden and Donald A Jackson. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, 154(1-2):135–150, 2002.

[61] Joanne S Luciano, Bosse Andersson, Colin Batchelor, Olivier Bodenreider, Tim Clark, Christine K Denney, Christopher Domarew, Thomas Gambet, Lee Harland, Anja Jentzsch, et al. The translational medicine ontology and knowledge base: driving personalized medicine by bridging the gap between bench and bedside. In *Journal of biomedical semantics*, volume 2, page S1. BioMed Central, 2011.

[62] Steven H Woolf. The meaning of translational research and why it matters. *Jama*, 299(2):211–213, 2008.