Universitat de les Illes Balears

Facultat de Ciències

Memòria del Treball de Fi de Grau

Caracterització de tumors amb tècniques radiòmiques

Sergi Serrano Rueda

Grau de Física

Any acadèmic 2017-18

Treball tutelat per Antoni Borrás López
Departament de Física

	Autor		Tutor	
S'autoritza la Universitat a incloure aquest treball en el Repositori Institucional per a la seva consulta en accés obert i difusió en línia, amb finalitats exclusivament acadèmiques i d'investigació	Sí	No	Sí	No
	X		X	

Paraules clau del treball:
Radiòmica, diagnosi per imatge, càncer, medicina personalitzada, anàlisi de components principals, t-SNE, regressió lineal múltiple, *random forest*

SUMARI

Resum

D'ençà que els avanços científics i tecnològics van permetre visualitzar l'interior d'un pacient sense haver de recórrer a la cirurgia, les imatges mèdiques han esdevingut una eina indispensable en Medicina. Aquestes imatges ajuden al diagnòstic, tractament i seguiment de tot tipus de malalties, tasca que sempre ha anat a càrrec d'un metge o un especialista en la seva interpretació.

En aquest treball es parteix de la hipòtesi que, a partir d'imatges mèdiques, és possible extreure dades objectives que permeten tant la caracterització de la malaltia com la realització de models matemàtics que ajuden a saber quina serà la seva evolució. Per confirmar o desmentir la hipòtesi, es disposa d'una base de dades de 422 pacients amb càncer de pulmó i l'ajuda de les eines que formen la Radiòmica.

La primera part del treball se centra en la familiarització amb les imatges mèdiques i la seva preparació per poder fer un estudi Radiòmic. El següent objectiu és estudiar les característiques que es poden extreure d'aquestes i fer anàlisis matemàtiques estàndard com són: estudiar les distribucions estadístiques o fer reduccions de dimensionalitat amb una anàlisi de components principals i mitjançant l'algoritme t-SNE. Finalment s'aprofita la informació extreta anteriorment per fer un model predictiu del temps de supervivència de cada pacient mitjançant dues tècniques: una regressió lineal múltiple i un *Random Forest*.

CAPÍTOL 1

INTRODUCCIÓ

1.1 Tomografia computada

Una Tomografia computada (TC) és una tecnologia que permet obtenir imatges de talls o seccions de l'interior d'un pacient sense haver d'accedir a l'interior del cos mitjançant algun procediment quirúrgic. Això és possible gràcies a dirigir una font de raigs X al cos del pacient, aprofitant el fet que els fotons s'atenuen de manera distinta quan travessen diferents materials. Una de les característiques principals de les tomografies i que la diferencien de, per exemple, les radiografies és que es pot reconstruir l'interior del pacient en tres dimensions, una vegada s'ha fet el processament corresponent, i que permet veure teixits i òrgans a més dels óssos.

El procediment que permet crear les imatges amb un aparell de TC (que es pot veure a la Figura 1.1), es pot dividir en tres etapes principals [1]:

Figura 1.1: Imatge d'un aparell típic emprat per fer tomografies computades. Imatge extreta de: `https://www.healthcare.siemens.com/`

1. **Producció de raigs X**

   La producció de raigs X té lloc a un tub de raigs que duu incorporat l'aparell de TC. En un extrem del tub hi ha un càtode amb un filament (generalment de tungstè o un aliatge de tungstè-reni) que s'escalfa fins a altes temperatures per aconseguir l'emissió d'electrons per efecte termoiònic. Els electrons són accelerats i reconduïts pel buit del tub mitjançant camps elèctrics intensos

fins a l'ànode, on es troba un objectiu fet de tungstè-reni i amb grafit per millorar la dissipació tèrmica.

Quan els electrons arriben a l'objectiu es poden donar tres processos diferents:

- La gran majoria dels electrons experimenten col·lisions elàstiques amb el medi, que són aquelles que conserven l'energia cinètica. Aquestes col·lisions desvien la trajectòria dels electrons, fent que augmenti l'agitació tèrmica del material, cosa que provoca l'augment de la temperatura. Aquest procés és el que fa que la producció de raigs X no tingui un rendiment del 100% i que s'hagi de tenir un sistema que refrigeri l'aparell.
- També pot passar que tinguin col·lisions inelàstiques amb els àtoms que formen l'objectiu. Aquestes col·lisions fan que els electrons transfereixin gran part de l'energia que duen. Aquesta energia és capaç d'ionitzar els àtoms que formen l'objectiu, és a dir, fer que perdin electrons. Quan electrons de les capes externes de l'àtom ocupen les vacants que han quedat, és quan s'emeten els raigs X. Aquests raigs X tenen energies discretes, per això reben el nom de "raigs X característics".
- Mentre ocorren els processos anteriors, els electrons incidents sofreixen una desacceleració important. A causa de la conservació de l'energia, l'energia cinètica que van perdent ha de transformar-se en un altre tipus d'energia. En aquest cas s'emet en forma de fotons i es coneix amb el nom de "radiació de frenada" o *bremsstrahlung*. Aquesta, és l'altra font de raigs X, amb la diferència que té un espectre d'energies continu.

Els aparells de TC normalment operen en uns potencials que van de 100 kV a 150 kV, per tant els raigs X tindran energies màximes de fins a 100-150 keV. Tot i que aquests fotons surten en totes direccions, és convenient tenir control sobre el feix per poder focalitzar-lo sobre el pacient i evitar dosis que no són estrictament necessàries. Per aquest motiu, s'utilitza un col·limador de plom que bloqueja els fotons que no van en la direcció d'interès, aconseguint un feix de pocs mil·límetres de gruix.

2. **Detecció dels raigs X**

   Un cop els raigs X surten del tub i arriben al cos del pacient, entra en joc com interaccionen els fotons amb la matèria. Si el medi no és homogeni (com és el cas del cos humà), la intensitat del feix incident es veu atenuada d'acord [2]:

$$I(x) = I_0 e^{-\int_0^x \mu dx'} \tag{1.1}$$

   On x és la profunditat penetrada, I_0 és la intensitat incident i μ és el coeficient d'atenuació lineal. Aquest coeficient depèn tant del material absorbent com de l'energia dels fotons incidents i està relacionat amb la probabilitat per unitat de longitud que un fotó interaccioni amb el medi. Es mesura amb unitats de (longitud)$^{-1}$.

   Durant el procés d'escanejar el cos, el feix de raigs X va girant al voltant del cos, a fi de passar per tots els punts de la zona que es vol examinar. Els aparells de TC tenen detectors digitals de raigs X al costat contrari d'on es troba el tub de raigs X, que van enregistrant la intensitat del feix de fotons una vegada ha travessat el cos. Un cop es té quantificat com s'ha atenuat el feix en totes les direccions possibles, es transfereixen les dades a un ordinador per passar a la darrera etapa.

3. **Construcció de les imatges**

   Amb aquestes dades es tracta d'obtenir una matriu on hi ha els valors dels coeficients μ de cada vòxel[1] del volum. Per deduir aquests valors s'utilitzen algoritmes matemàtics que no suposen

[1] Unitat mínima que forma una imatge en tres dimensions. És la generalització del concepte de píxel.

una gran càrrega computacional, molts d'ells basats en la "transformada de Radon"[2]. Aquests algoritmes aprofiten el fet que es té informació sobre l'atenuació total que ha sofert el feix un cop ha passat per diferents medis materials i direccions, i que es coneix la geometria del sistema.

En la darrera passa es tracta de construir una imatge que permeti distingir les diferents estructures que formen el cos d'una manera visual. El procés consisteix a assignar un nombre (anomenat "nombre TC") a cada vòxel relacionat amb el coeficient d'atenuació. El nombre del vòxel i es calcula [3]:

$$N_i = K\frac{\mu_i - \mu_{H_2O}}{\mu_{H_2O}} \tag{1.2}$$

On μ_i i μ_{H_2O} són els coeficients d'atenuació del vòxel i aigua respectivament, i K és una constant entera que depèn del conveni emprat per fer el càlcul.

Tot i que aquesta tècnica utilitza radiació ionitzant per prendre les imatges als pacients, les dosis que s'acumulen són petites i està molt controlada [4]. Tot i això, es té en compte que no és un examen que es pugui fer moltes vegades durant tot el procés de tractament de malalties. Per aquest motiu es planteja de tal manera que la informació que se'n pugui extreure sigui crucial per millorar el tractament. Els beneficis d'haver de passar per aquesta etapa són molt majors que els efectes perjudicials que té exposar-se a la radiació dels aparells de tomografia computada.

1.2 Radiòmica

La Radiòmica és una tècnica no invasiva que ajuda tant al diagnòstic com al tractament de tumors. La paraula es va formar a partir de la unió dels termes radio- (referit a radiació o a l'especialitat mèdica de radiologia), -om(a)- (conjunt o estructura, emprat a biologia) i -ica (estudi o tècnica) [5, 6]. El terme es va construir fent una analogia amb altres paraules (com per exemple genòmica) que es refereixen a fer un estudi complet sobre algun camp combinant diverses tècniques [7]. En aquest cas, la tècnica consisteix a fer tractament d'imatges mèdiques digitals per extreure informació quantitativa (anomenades "característiques radiòmiques" o simplement "característiques") sobre tumors i fer un estudi que permetrà: realitzar un millor diagnòstic, seguir l'evolució del tumor i poder anticipar-se, triar el tractament òptim, etc. Es complementa amb tots els altres processos que intervenen en l'estudi d'un tumor, ja que per fer un estudi Radiòmic, no és necessari realitzar exàmens que no s'hagin hagut de fer en una rutina habitual del tractament de tumors. De fet, pot arribar a disminuir el nombre de proves que s'han de dur a terme, reduint així els riscs i costs associats i accelerant el procés [8].

Tot i que les imatges mèdiques sempre han estat utilitzades com a valuoses fonts d'informació descriptiva per fer el seguiment dels tumors, no va ser fins a l'any 2010 que va néixer el camp de la Radiòmica. Fins aleshores, les anàlisis computacionals que es feien per extreure informació de les imatges (conegudes amb nom de CAD, de l'anglès: *computer-aided detection*) servien als metges com una segona opinió a l'hora de fer diagnosis o identificació de lesions [5]. La Radiòmica va néixer quan els algoritmes de *machine learning* es van popularitzar i la tecnologia va permetre la seva aplicació amb major facilitat. Per això es diferencia dels mètodes CAD, en el fet que la quantitat de característiques és molt major (centenars enfront de poques desenes) i que permet construir models predictius gràcies a combinar les característiques amb dades biològiques i mèdiques. En referència a l'ús de dades biològiques, existeix tot una branca d'investigació (anomenada Radiogenòmica) que es dedica a fer estudis combinant la Radiòmica amb informació genètica[3]. L'estudi d'algunes característiques ha demostrat que és possible identificar genomes del tumor que són diferents del teixit sa [9].

[2] Es pot trobar més informació a: T. G. Feeman, *The Mathematics of Medical Imaging*, 2nd ed. Springer International, 2015

[3] Fer un estudi d'aquest tipus no és l'objectiu d'aquest treball, però es pot trobar informació a: M. A.Mazurowski, "Radiogenomics: What it is and why it is important," *Journal of the American College of Radiology*, vol. 12 No 8, pp. 862–866, August 2015.

Una altra novetat que introdueix la Radiòmica és que forma part del que es coneix amb el nom de "medicina personalitzada", és a dir, s'ha d'estudiar cada cas particular. Això fa que el procés d'aplicació de la tècnica tingui un ordre ben definit i que quasi sempre s'hagi de passar per totes les etapes. Per fer estudis que siguin validats fàcilment és molt important que, encara que cada situació és única, el mètode sigui reproduïble. En els darrers anys, s'han fet esforços per intentar estandarditzar el màxim possible la manera de treballar, la qual, en general, està dividida en cinc etapes [10] (veure Figura 1.2):

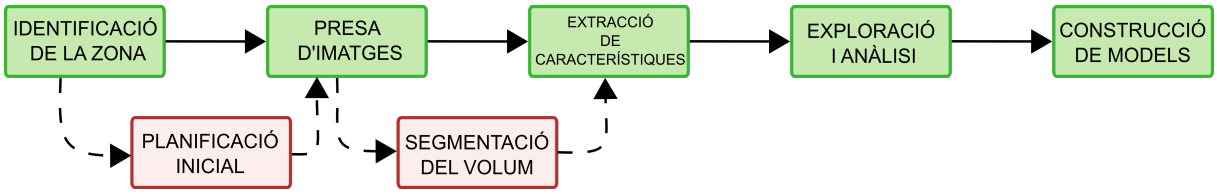

Figura 1.2: Diagrama del procés de la Radiòmica. Els requadres que apareixen en una posició inferior corresponen a etapes que no és estrictament necessari passar, ja que pot ser que sigui un problema estàndard que no necessita una planificació completament nova; o que la segmentació vengui feta per l'oncòleg i que, per tant, no formi part del procés radiòmic.

1. El procés comença amb la identificació de la zona on es vol aplicar l'anàlisi, quines característiques del volum es volen extreure, sobre quines variables es vol fer pronòstic i quina tecnologia s'utilitzarà per prendre les imatges en la següent etapa. Aquesta planificació inicial és important perquè concretaran les diferents opcions que existeixen dins cada una de les següents etapes.

2. El segon pas és la presa d'imatges del teixit afectat mitjançant una (o combinant més d'una) de les tècniques següents: una TC, una ressonància magnètica nuclear, ultrasons o una PET (tomografia per emissió de positrons). Un cop es tenen les imatges, s'ha de fer un procés de segmentació del volum. Aquest procés consisteix a triar quines regions seran analitzades i es pot fer de manera manual (per un especialista) o automàtica mitjançant tècniques de tractament d'imatges. Amb això s'aconsegueix disposar d'imatges que contenen diferents contrasts en el teixit que permetran passar al següent pas.

3. La tercera etapa és el pilar central de les tècniques radiòmiques: l'extracció de les característiques. Mitjançant una anàlisi computacional dels segments d'una imatge, és possible extreure informació quantitativa del tumor. Per tant, una característica descriu com és certa regió del teixit (sigui malalt o sa) a partir de la informació de la imatge (com la forma o la textura, per exemple) utilitzant algoritmes matemàtics.

   Una manera general de classificar les propietats és mitjançant dos grups: característiques "semàntiques" i "agnòstiques" [11]. Les primeres descriuen la zona d'interès mitjançant variables físiques com: volum, forma, densitat, textura, localització, comparació amb les regions veïnes, etc. Mentre que les segones tenen definicions de caràcter més matemàtic com: dimensió fractal, Transformada de Laplace, funcional de Minkowski, etc. I serveixen per quantificar l'heterogeneïtat de la zona.

4. Una vegada s'ha extret tot el conjunt de dades que es desitjava de les imatges, es passa a la seva exploració i anàlisis. La varietat de característiques que es poden extreure d'una imatge pot ser immensa i estudiar-les totes pot ser un procés computacional exigent. D'aquí ve la importància de conèixer bé les propietats que es volen estudiar per evitar redundàncies en la informació obtinguda i agilitzar el procés. Per fer això, se cerquen relacions entre característiques que donin els mateixos resultats clínics o que donin un pronòstic del tumor semblant. També entra en joc l'objectiu final de l'estudi, hi pot haver propietats que són útils per fer certs tipus de previsions (com el temps de supervivència, per exemple) i altres que són millors per fer estudis de com serà l'evolució de les propietats del tumor en si o l'evolució del tumor cap a zones amb teixit sa.

5. Finalment es passa a la creació d'un model. Per tal que tot el procés resulti útil és important estudiar en cada cas quines característiques han estat les que han permès reproduir amb més exactitud l'evolució del teixit malalt. A l'hora de crear els models hi ha diferents opcions, com per exemple:

    - Una manera més tradicional seria construir una base de dades gràcies a la realització de diversos estudis i aprofitar l'experiència per actuar de la millor manera possible davant cada cas: a mesura que es van incorporant diversos casos de diferents pacients i es conjunten amb informació clínica, les tècniques radiòmiques agafen força.
    - Una altra opció, que està guanyant popularitat en els darrers anys, és la utilització de models computacionals: gràcies a tècniques com *machine learning*, s'acaba disposant d'una xarxa neuronal de coneixement que fa que les anàlisis siguin cada vegada més refinades.

    També és possible combinar les dues possibilitats, a fi d'aconseguir el model més robust possible.

Els models permetran donar valor als resultats obtinguts i fer que la radiòmica sigui molt important tant en el diagnòstic com en el tractament de tumors. S'han realitzat estudis recents que demostren la utilitat de l'aplicació de les tècniques radiòmiques per fer previsions sobre el càncer [12]. Concretament, aquest estudi, afirma que l'ús de 23 característiques ha permès preveure regions del pulmó que podrien tornar canceroses en un any o dos amb un 80 % i 79 % de fiabilitat respectivament.

Tot i això la Radiòmica no és útil en totes les situacions, és important conèixer les seves limitacions. Hi ha característiques que no tenen una fiabilitat tan alta a causa de la seva sensibilitat a com se segmenten i es parametritzen les imatges. Si la segmentació es realitza manualment, es pot veure afectada per la subjectivitat que això suposa. Per altra banda, el problema de fer-ho de manera automàtica amb ordinadors, és que els estudis en Radiòmica encara no estan estandarditzats. Això és una limitació molt important que fa que la reproductibilitat es vegi afectada. Per aquesta raó es destinen molts esforços a intentar que tots els grups d'investigació utilitzin protocols equivalents.

El problema més important de fer models utilitzant xarxes neuronals és que encara que es tingui una xarxa neuronal amb moltes correlacions, és molt difícil interpretar els resultats. A més, el fet que s'hagi seguit una metodologia ordenada i rigorosa per fer l'estudi, no fa que identificar les causes dels resultats observats sigui una tasca senzilla.

També hi ha vegades que l'anàlisi dóna falsos positius: després de la realització de diversos estudis [8, 12], s'ha observat que són necessaris entre 10 i 15 pacients per característica per reduir aquest nombre. Per altra banda és molt difícil fer un estudi d'un tumor on es faci ús exclusiu de tècniques radiòmiques. És un camp que s'ha posat en pràctica en els darrers anys i encara està en els primers estadis del seu desenvolupament.

CAPÍTOL 2

METODOLOGIA

2.1 Cas d'estudi

El primer objectiu del treball és aconseguir extreure característiques radiòmiques d'imatges mèdiques per un conjunt de pacients. A la figura 1.2 es pot veure com abans d'això hi ha quatre etapes. En aquest cas, es parteix d'una base de dades amb imatges mèdiques i les corresponents segmentacions; per aquest motiu, passar per les etapes de "presa d'imatges" i "segmentació inicial" no serà necessari. El que sí que serà molt important és fer la planificació, ja que s'ha de fer un procés de familiarització amb la base de dades i de preparació de les imatges per fer l'estudi.

La base de dades [13] correspon a una col·lecció d'imatges de càncer pulmonar de "cel·les no petites" (carcinoma pulmonar no microcític (CPNM), el càncer de pulmó més comú) de 422 pacients. Per això en aquest cas, l'etapa de "identificació de la zona" és molt senzilla, ja que per tots és la mateixa.

Les dades incloses, que ocupen 25 GB en total, són les següents:

- El conjunt d'imatges preses abans del tractament mitjançant una tomografia computada.
- La segmentació del volum en tres dimensions del tumor (feta manualment per un oncòleg radioteràpic).
- Dades clíniques de cada pacient amb informació concreta relacionada amb el tipus de càncer que pateixen.

Cada pacient té associat una sèrie d'imatges de tipus DICOM (de l'anglès, *Digital Imaging and Communications in Medicine*) corresponent a cada secció bidimensional del tòrax. DICOM és un estàndard utilitzat en Medicina per emmagatzemar imatges, que està pensat per facilitar la seva transmissió, visualització, manipulació, impressió, etc [14]. En aquest format, la informació està organitzada en conjunts de dades entre les quals es poden trobar: la imatge del pacient, la marca de l'escàner de TC o quin dia es va prendre la imatge, entre d'altres. L'avantatge d'aquest estàndard és que aquesta informació no es pot separar de la imatge, evitant així que les dades es perdin per error. Existeixen diverses modalitats de les quals en aquesta base dades s'inclouen dues [15]:

- Per una banda hi ha la modalitat "CT" (*Computed Tomography*) que són els arxius que contenen les imatges de les seccions del tòrax i permeten formar el volum tridimensional.
- L'altra modalitat és "RTSTRUCT" (*Radiotherapy Structure Set*) que són arxius associats als CT i que contenen la segmentació manual del metge de cada tumor (un sol arxiu ja conté la segmentació de cada secció del tòrax).

El primer pas a realitzar és seleccionar aquells pacients que són vàlids per realitzar un estudi radiòmic, ja que dels 422 pacients de la base de dades, no tots tenen associada una segmentació. Sense aquesta informació, no es poden extreure les característiques radiòmiques, per això s'han descartat els 104 pacients sense segmentació del tumor.

El següent pas és entendre com funciona el paquet de python *pyradiomics*, que és el que s'utilitzarà per extreure les característiques radiòmiques dels tumors [16]. Per fer una extracció de totes les característiques amb la configuració dels paràmetres que ve per defecte, es necessita un arxiu que contengui una imatge o més d'una (segons si es fa un estudi en 2D o 3D) de la zona on està el tumor i un altre arxiu on es trobi la segmentació associada a aquell tumor. *Pyradiomics* és capaç de llegir els arxius en format DICOM però la modalitat "RTSTRUCT" no la suporta [1]. Per altra banda, com que es volen extreure les característiques del tumor complet, és necessari combinar totes les imatges de les seccions en un sol arxiu. Per aquests dos motius és imprescindible fer una conversió de formats, tant per les imatges com per les segmentacions. Resulta convenient convertir els dos tipus d'arxius al format "NRRD" (de l'anglès, *nearly raw raster data*), ja que és un dels formats recomanats per la documentació per treballar amb imatges 3D.

Per realitzar la conversió, s'utilitza el programari lliure 3DSlicer [17], ja que no només permet treballar amb aquests tipus de dades i visualitzar-les (vegeu figura 2.1), sinó que permet exportar-les a altres formats. En concret, farem ús d'un script de python [2] que crida 3DSlicer i converteix automàticament tant les imatges com la segmentació corresponent que es trobin a un directori determinat.

Durant la realització d'aquest procés, ha resultat que alguns pacients tenien la seva segmentació mal referenciada, és a dir, no concordava perfectament la segmentació amb les imatges. Per evitar que totes les anàlisis que farem posteriorment quedin invalidades i tenint en compte que el nombre de dades que es podran estudiar segueix sent considerable, s'ha decidit descartar els 25 pacients que tenien aquest problema. Un altre factor important a tenir en compte és que, per com està dissenyat l'algoritme de conversió, la segmentació perd la geometria que tenia originalment, per tant deixa de quadrar amb les imatges de referència. Per corregir aquest problema s'utilitza un altre script de Python dissenyat per corregir aquests tipus de divergències automàticament [3].

Quan ja es tenen tant les imatges com les segmentacions de tots els pacients en el format i geometria adequades, es pot passar a l'extracció de les característiques amb *pyradiomics*. Tenint en compte que són 293 casos, hem escrit un programa amb Python que fa totes les passes que s'han explicat fins ara de manera automàtica. El programa ha invertit entre uns 20 i 30 segons per pacient (amb un processador de 8 nuclis a 3.6 GHz), fent que el temps total sigui d'unes dues hores.

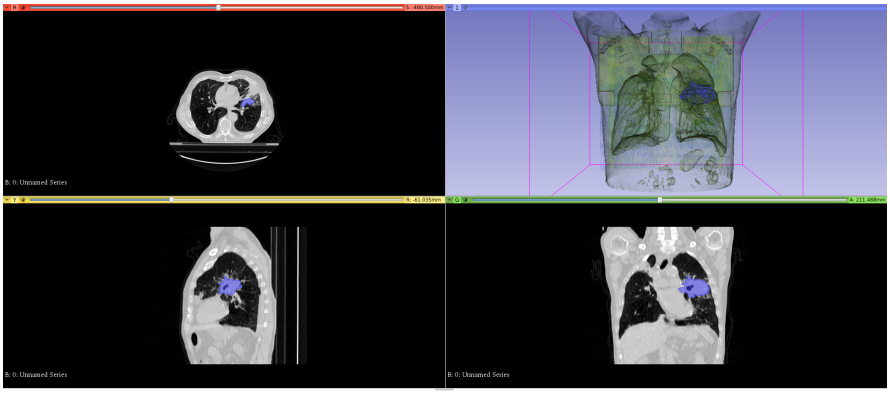

Figura 2.1: Visualització de les imatges de la tomografia computada d'un pacient de la base de dades. D'esquerra a dreta i de dalt a baix, es veu: secció en el pla axial, reconstrucció tridimensional del tòrax, secció en el pla sagital i, finalment, secció en el pla coronal. La regió de color blau correspon a la segmentació.

[1] Comprovació feta el març de 2018, amb la versió de *pyradiomics* 1.3.0.

[2] Codi font disponible a: `https://github.com/SlicerRt/SlicerRT/tree/master/BatchProcessing`

[3] Aquest script de Python ve amb la llibreria i s'explica en els exemples de la documentació, pàgina 71 "resampleMask.py".

2.2 Característiques radiòmiques

Les característiques que es poden extreure de les imatges amb *pyradiomics* estan basades en les definicions que donen a l'article [18] de IBSI (de l'anglès, *Imaging Biomarker Standardization Initiative*). Les definicions matemàtiques de les 113 característiques que hi ha en total es poden trobar a la documentació de *pyradiomics*. A la taula 2.1 es poden veure els noms de totes les característiques. Estan dividides en set categories diferents, les definicions de les quals s'expliquen a continuació:

- Característiques de Primer Ordre

  Les característiques de primer ordre descriuen com és la distribució de la intensitat, és a dir, el valor en l'escala de grisos dels píxels o vòxels (segons sigui 2D o 3D) que estan en la zona delimitada per la segmentació.

- Característiques basades en la forma

  En aquesta categoria s'hi inclouen les característiques que quantifiquen valors relacionats amb la mida i la forma de la regió segmentada. Cal dir que els valors calculats són completament independents de l'escala de grisos de la imatge.

- Característiques de la matriu de coocurrència dels nivells de grisos (GLCM)

  La matriu GLCM (de l'anglès, *Gray Level Co-occurrence Matrix*) ajuda a conèixer la textura d'una imatge descrivint la probabilitat que es trobin dos píxels amb el mateix valor de nivell de grisos en la regió delimitada de la imatge. Si la zona d'interès de la imatge té N_g nivells diferents de grisos, la matriu serà de $N_g \times N_g$.

  Un element de matriu de la matriu GLCM en la posició (i, j) representa el nombre de vegades que la combinació (i, j) de nivell de grisos apareix en dos píxels que estan separats una distància de δ píxels en la direcció de l'angle θ.

  Aquesta distància δ està definida d'acord amb la norma de L_∞ [4]. En cas que es parli de vòxels, la distància es mesura entre els centres dels dos vòxels.

- Característiques de la matriu de la mida de les zones amb un cert nivell de grisos (GLSZM)

  La matriu GLSZM (de l'anglès, *Gray Level Size Zone Matrix*) proporciona informació sobre si hi ha zones que comparteixen el mateix nivell de grisos dins la regió d'interès. Per detectar una zona es compara el nivell de gris d'un vòxel amb tots els seus vòxels veïns. Un vòxel veí és, per definició, aquell vòxel que es troba a distància 1 del vòxel considerat (amb la mateixa definició de distància que abans).

  Un element en la posició (i, j) quantifica el nombre de zones amb nivell de gris i i mida j que apareixen en la imatge. Una diferència fonamental amb les altres matrius, és que aquesta és independent de rotacions. Això vol dir que no s'ha de calcular una matriu per cada direcció considerada, calcular una única matriu és suficient.

- Característiques de la matriu de la llargària de sèries de nivells de grisos (GLRLM)

  La matriu GLRLM (de l'anglès, *Gray Level Run Length Matrix*) incorpora informació relativa a les sèries de nivells de grisos dins la segmentació. Una sèrie de nivell de grisos es defineix com la longitud (en píxels) de píxels consecutius que tenen el mateix nivell de gris.

  Un element en la posició (i, j) descriu el nombre de sèries amb nivell de gris i i longitud j en la direcció θ.

[4] La norma de L_∞ es defineix de la següent manera: sigui un vector $\mathbf{x} = (x_1, x_2, ..., x_i, ..., x_n)$ on cada $x_i \in \mathbb{C}$, la seva norma és $\|\mathbf{x}\|_\infty = \max_i(|x_i|)$.

- Característiques de la matriu de diferències dels nivells de grisos entre veïns (NGTDM)

  La matriu NGTDM (de l'anglès, *Neigbouring Gray Tone Difference Matrix*) quantifica la diferència entre un nivell concret de gris amb el valor mitjà que hi ha entre els veïns que es troben dins una distància de δ píxels. Cada element de la matriu conté la suma de diferències absolutes del nivell de gris i.

  Sigui $\mathbf{X}_{gl}$ un conjunt de vòxels dins la segmentació i sigui $x_{gl}(j_x, j_y, j_z) \in \mathbf{X}_{gl}$ el nivell de gris d'un vòxel a la posició (j_x, j_y, j_z), el nivell de gris mitjà es calcula:

$$\overline{A}_i = \overline{A}(j_x, j_y, j_z) = \frac{1}{W} \sum_{k_x=-\delta}^{\delta} \sum_{k_y=-\delta}^{\delta} \sum_{k_z=-\delta}^{\delta} x_{gl}(j_x + k_x, j_y + k_y, j_z + k_z) \tag{2.1}$$

  On s'ha de complir que $(k_x, k_y, k_z) \neq (0,0,0)$ i $x_{gl}(j_x + k_x, j_y + k_y, j_z + k_z) \in \mathbf{X}_{gl}$. W és el nombre de vòxels veïnats que també pertanyen a $\mathbf{X}_{gl}$.

- Característiques de la matriu de dependències dels nivells de grisos (GLDM)

  La matriu GLDM (de l'anglès, *Gray Level Dependence Matrix*) quantifica les dependències dels nivells de grisos de la imatge. La dependència de nivell de grisos es defineix com el nombre de vòxels que estan dins una distància δ i depenen del vòxel central de referència. Un vòxel amb un nivell de gris j es considera dependent d'un veí seu amb valor de gris i si: $\left|i - j\right| \leq \alpha$. On α és un cert llindar que s'ha d'especificar.

  Un element en la posició (i, j) descriu el nombre de vegades que un vòxel amb un nivell de gris i i amb j vòxels veïns dependents apareix en la regió d'interès de la imatge.

Grup	**Nombre**	**Llista de característiques**
Primer Ordre	19	Energy, Total Energy, Entropy, Minimum, Maximum, Mean, Median, 10th percentile, 90th percentile, Interquartile Range, Range, MAD, rMAD, RMS, Standard Deviation, Skewness, Kurtosis, Variance i Uniformity
Forma	16	Volume, Surface Area, Surface Area to Volume ratio, Sphericity, Compactness 1, Compactness 2, Spherical Disproportion, Maximum 3D diameter, Maximum 2D diameter (Slice), Maximum 2D diameter (Column), Maximum 2D diameter (Row), Major Axis, Minor Axis, Least Axis, Elongation i Flatness.
GLCM	26	Autocorrelation, Joint Average, Cluster Prominence, Cluster Shade, Cluster Tendency, Contrast, Correlation, Difference Average, Difference Entropy, Difference Variance, Dissimilarity, Joint Energy, Joint Entropy, Homogeneity 1, Homogeneity 2, IMC 1, IMC 2, IDM, IDMN, ID, IDN, Inverse Variance, Maximum Probability, Sum Average, Sum Entropy i Sum of Squares
GLSZM	16	SAE, LAE, GLN, GLNN, SZN, SZNN, ZP, GLV, ZV, ZE, LGLZE, HGLZE, SALGLE, SAHGLE, LALGLE i LAHGLE
GLRLM	16	SRE, LRE, GLN, GLNN, RLN, RLNN, RP, GLV, RV, RE, LGRE, HGLRE, SRLGLE, SRHGLE, LRLGLE i LRHGLE
NGTDM	5	Coarseness, Contrast, Busyness, Complexity i Strength
GLDM	15	SDE, LDE, GLN, GLNN, DN, DNN, GLV, DV, DE, LGLE, HGLE, SDLGLE, SDHGLE, LDLGLE i LDHGLE

Taula 2.1: Taula on s'inclouen la llista de les 113 característiques utilitzades i al grup que pertanyen. Els noms o sigles estan en anglès per respectar els mots originals. Per més obtenir més informació, es pot consultar la documentació [16].

CAPÍTOL 3

RESULTATS

3.1 Propietats de les característiques radiòmiques

Sempre que es fa un estudi amb moltes dades o resultats la primera passa és saber com es distribueixen aquestes dades al llarg dels diferents casos o repeticions d'un experiment. Si es tracta d'un estudi on no es coneixen *a priori* les distribucions de probabilitat (com és aquest cas), fer gràfiques de les distribucions és molt interessant, ja que ajudarà a tenir una comprensió més visual de cada variable i també servirà per escollir els millors mètodes per analitzar aquestes dades.

Mirant la matriu de resultats, és fàcil adonar-se'n que els valors són semblants per una mateixa característica, mentre que els ordres de magnitud són molt diferents si es comparen amb valors d'una altra. En aquesta situació el més recomanable és fer una "estandarditzación" dels valors (veure la secció A.1.3 de l'apèndix), ja que així es podran comparar totes les magnituds encara que originalment siguin molt diferents.

Amb els valors de totes les característiques estandarditzats, es tracta de comprovar si les distribucions dels valors corresponen a distribucions gaussianes, també conegudes amb el nom de distribucions normals:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.1}$$

La raó d'intentar veure si els perfils s'ajusten a una distribució d'aquest tipus, és que la distribució normal és una de les distribucions que apareix amb més freqüència en estadística, per tant molts models d'anàlisi i ajust de dades, suposen que les dades estan caracteritzades amb aquesta distribució. Per fer aquesta comprovació, s'han representat histogrames dels valors de cada característica radiòmica, amb l'objectiu de tenir una manera visual de saber si les dades s'ajusten a la funció.

Mirant les gràfiques una a una, es veu com moltes característiques segueixen una distribució normal, però que la gran majoria no, sent les distribucions exponencials les que més predominen. A la primera columna de la Figura 3.1, hi ha dos exemples.

Vist que no es pot fer la suposició que les dades segueixen una distribució gaussiana, es farà servir una eina matemàtica molt útil en aquestes situacions: la transformació Box-Cox (vegeu la secció A.1.4 de l'apèndix), amb l'objectiu que totes les característiques segueixin una distribució normal.

Coneixent en quines situacions és única la transformació Box-Cox, el procediment a seguir per cada característica consta de dues passes:

1. S'aplica la transformació Box-Cox amb un valor del paràmetre λ òptim.
2. Quan ja es té la matriu transformada, s'estandarditzen els valors d'una mateixa característica seguint l'explicació del principi.

A continuació es comprova que totes les característiques segueixin una distribució més semblant a la gaussiana, com es veu a la Figura 3.1:

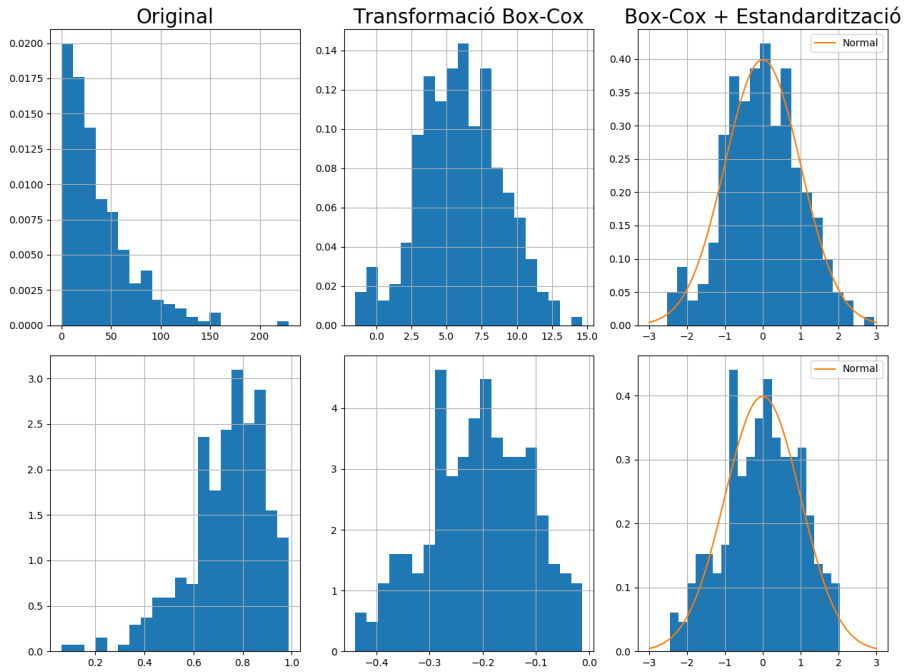

Figura 3.1: Representació de les successives transformacions que s'han aplicat a les dades. La primera fila correspon a la característica *Contrast* del grup "GLCM" i la segona fila, a *Elongation* del grup basat en la textura. A la primera columna hi ha els valors originals, a la segona només s'ha aplicat la transformació Box-Cox i a la darrera, els valors de Box-Cox "estandarditzats".

3.2 Correlacions i clusterització

Una vegada s'està segur que els valors es poden comparar entre si, l'objectiu és veure quina relació hi ha entre les variables. Per això es construeix una matriu amb els coeficients de correlació de Pearson de totes les característiques entre si (vegeu la secció A.1.1).

Un mmètode molt ràpid de poder veure aquests resultats és mitjançant unes representacions gràfiques que es coneixen amb el terme anglès de *heatmaps*. La representació consisteix a assignar un color a cada valor del coeficient de correlació, que anirà de -1 a 1. D'aquesta forma es té una manera visual de representar la matriu que fa possible reconèixer característiques molt correlacionades més fàcilment que una matriu de números.

No obstant això, encara no és útil per formar grups de característiques, ja que la matriu de la qual es parteix no té cap ordre particular. Per poder aconseguir això, s'ha d'aplicar un algoritme d'agrupament o de clusterització, en aquest cas es fa servir un mètode que es coneix amb el nom d'"agrupament jeràrquic acumulatiu". Això vol dir que es parteix d'un estat on cada variable representa un grup i posteriorment es van mesclant els grups amb l'objectiu d'anar formant grups cada vegada més grans amb la jerarquia corresponent. El resultat és que els valors similars, s'han resituat a una mateixa zona

de la matriu de tal manera que es vegin agrupacions de píxels del mateix color. Per poder veure les agrupacions, també s'hi representa una ajuda que consisteix a posar un diagrama tipus arbre per l'exterior de la matriu que indica les diferents agrupacions que es tenen. Amb aquesta guia també es pot distingir el grau de similitud, ja que com més branques es consideren, vol dir que més semblants són els valors; en canvi, si només es volen saber quants grups raonablement similars es tenen, mirant dues o tres branques és suficient.

Aplicant l'algoritme per veure les agrupacions queda una gràfica com la de la Figura 3.2:

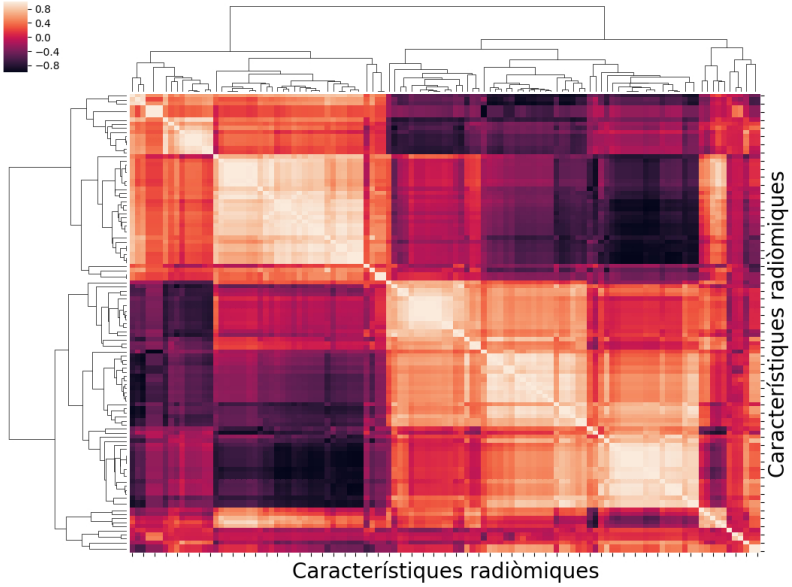

Figura 3.2: *Heatmap* on s'ha aplicat clusterització per identificar grups. Aquesta figura correspon a un mapa amb les correlacions de les característiques radiòmiques entre si.

Mirant les estructures rectangulars que són del mateix color, es poden identificar els grups de característiques que resulten estar directament o inversament correlacionades (els colors clars són per correlacions directes i els colors foscos són per correlacions inverses). Per exemple hi ha característiques que tenen una correlació d'1, tot i que, en principi, el nom i la definició són diferents. En concret s'han trobat les següents parelles:

- *Dissimilarity* i *Difference Average* del grup GLCM.
- *Homogeneity 1* i *Inverse Difference* (ID) del grup GLCM.
- *Homogeneity 2* i *Inverse Difference Moment* (IDM) del grup GLCM.
- *Gray Level Non-Uniformity Normalized* (GLNN) del grup GLDM i *Uniformity* del grup de Primer Ordre.

L'explicació del fet que aquestes parelles de característiques radiòmiques estiguin tan fortament correlacionades, es troba a la documentació de la llibreria [16]. Resulta que, posteriorment, es va demostrar que aquestes característiques són matemàticament equivalents, per això, només és necessari quedar-se amb una de cada parella. A les versions posteriors de la llibreria, ja s'han eliminat les característiques matemàticament equivalents a altres.

El mapa de correlacions no sols serveix per identificar característiques radiòmiques iguals, sinó que també permet conèixer quines són semblants. Encara que les altres característiques no siguin matemàticament equivalents, hi ha molts grups que tenen correlacions elevades (vegeu els grans rectangles que hi ha en el mapa la Figura 3.2). És aquí on es veu la importància de fer una anàlisi d'aquest tipus, ja que permet eliminar moltes variables similars per aconseguir reduir la càrrega computacional,

sigui de possibles càlculs posteriors o de futurs estudis radiòmics.

Un altre *heatmap* que serà especialment útil en el següent apartat és aquell on es representa el valor de les característiques segons el pacient. L'algoritme d'agrupament servirà per ajuntar pacients amb valors de característiques semblants. D'aquesta manera, es podrà tenir una primera idea de quants grups de pacients tenen característiques similars i quins són els integrants d'aquests grups. El mapa obtingut es pot veure a la Figura 3.3:

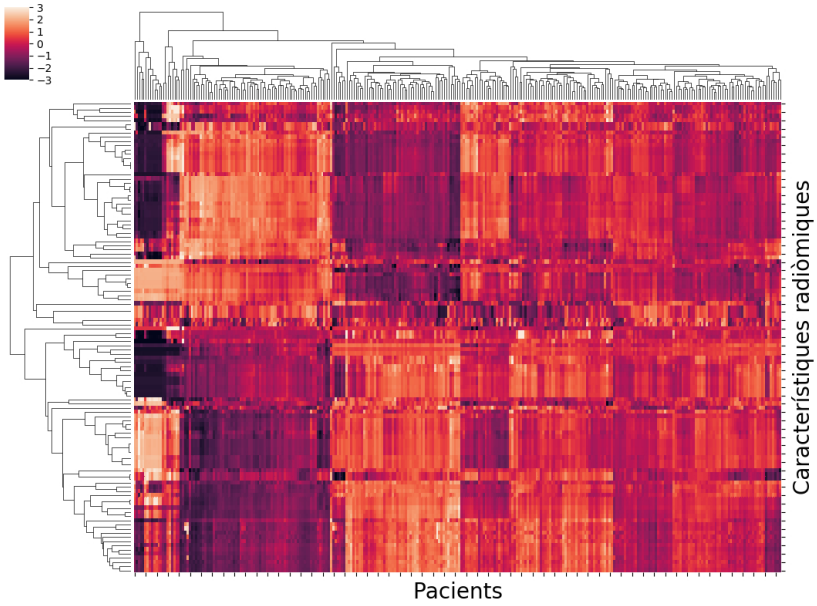

Figura 3.3: *Heatmap* on s'ha aplicat clusterització per identificar grups. Aquesta figura correspon a un mapa amb els valors de les característiques segons els pacients.

3.3 Reducció de dimensionalitat i agrupament

En la secció anterior, s'ha vist que és convenient reduir el nombre de característiques que es tenen. Per aquesta raó, s'aplicarà una reducció de dimensionalitat per poder construir gràfiques que siguin visibles en dues o tres dimensions. L'objectiu és veure si, fent anàlisis del tipus "components principals o CP" i "t-SNE", les noves característiques obtingudes permeten ser interpretades com a grups de pacients amb dades mèdiques en comú, en aquest espai de menor dimensionalitat. Concretament, es comprovarà si els grups es poden identificar la informació mèdica que es té dels pacients, que és la següent [19]:

- Estadi T: fa referència al tumor principal. S'assigna un valor numèric del 0 al 4, segons la mida i l'extensió del tumor.
- Estadi N: descriu com s'ha estès als ganglis limfàtics. S'assigna un valor numèric que va del 0 al 3.
- Estadi M: dóna informació de si el tumor s'ha estès per la resta del cos. Si s'ha estès, s'assigna el número 1; si no, el 0.
- Estadi general: és una manera alternativa de descriure el desenvolupament d'un càncer, però sense donar tants detalls com el sistema TNM anterior. S'assigna un nombre romà que va de 0 a IV segons l'evolució del càncer.
- Histologia: descriu el tipus de càncer a partir d'informació sobre les cèl·lules i els teixits afectats.

3.3.1 Anàlisi de components principals

Es comença fent una anàlisi de components principals (o PCA, vegeu secció A.2.1), ja que és un dels procediments conceptualment més simples i molt utilitzat actualment. S'ha decidit mantenir les primeres 3 components principals per poder fer una representació tridimensional i veure si hi ha alguna direcció on s'acumulen els punts. Aquestes 3 components representen el 80 % de la informació de les dades originals.

Després de mirar les gràfiques per totes les dades mèdiques comentades anteriorment, el cas que sembla tenir una interpretació més clara és la gràfica corresponent a l'estadi T. A la Figura 3.4 es poden veure els resultats:

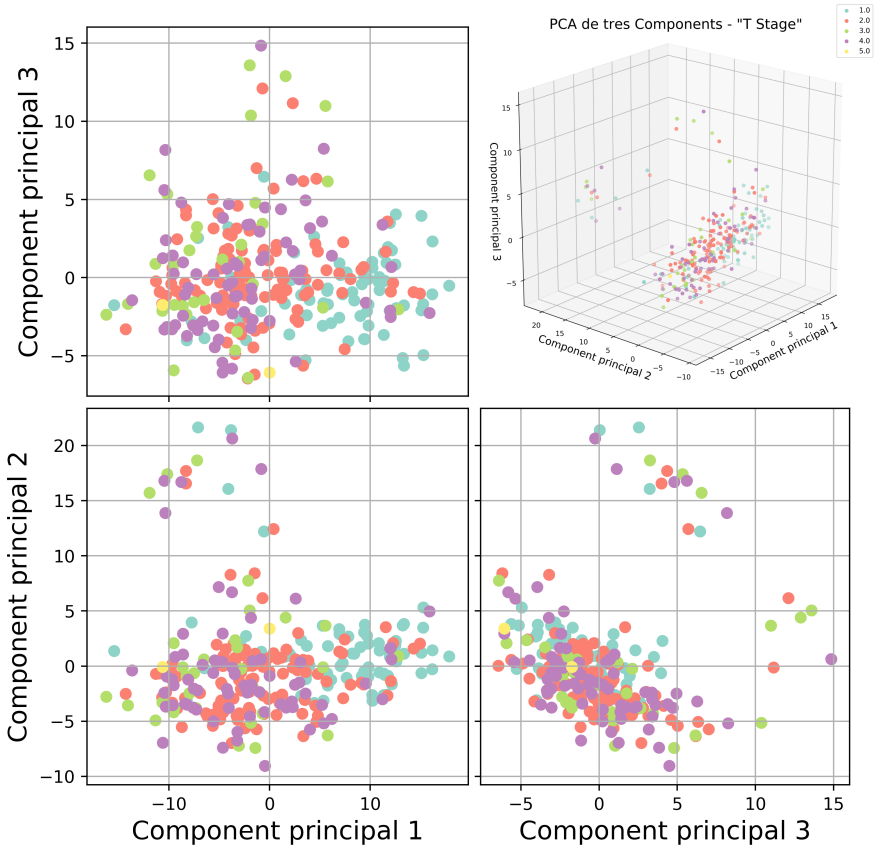

Figura 3.4: Representació en 3D de l'anàlisi de components principals. També es poden veure les projeccions en dues dimensions sobre els eixos més importants. Els pacients tenen un color associat segons l'estadi T al qual està el seu tumor.

Si es miren les dues gràfiques bidimensionals de la part esquerra, es pot observar que hi ha una direcció on s'acumulen els casos que estan en l'estadi 1. Això pot fer pensar que existeix un eix privilegiat on estan els pacients que formen part d'aquest grup, però que pel fet de ser una projecció en dues dimensions no es veu. Per aquest motiu s'ha representat la figura en tres dimensions del cantó superior dret i, així, tenir una perspectiva més amplia del que està passant.

Aquesta figura mostra clarament que, en realitat, la majoria dels punts estan acumulats en un núvol, fent que les sospites que les components principals es puguin interpretar com alguna dada mèdica, quedin descartades. De totes maneres, encara es pot extreure informació d'aquests diagrames:

- Aquest nombre de components principals no porta la suficient quantitat d'informació per a

formar estructures més separades.

- Tot i això, pacients que estan en el mateix estadi, tendeixen a estar propers dins el núvol. Això vol dir que hi ha característiques radiòmiques que reflecteixen d'alguna manera en quin estadi es troba el càncer.

- També es poden observar grups de pacients que se surten del núvol. Mirant de quins pacients es tracta, s'ha trobat que correspon a la primera franja vertical de la Figura 3.3. Allà es pot veure com aquesta franja destaca una mica de la resta de tonalitats, per tant, es tracta de pacients amb característiques molt similars entre ells, però diferents de la resta al mateix temps.

3.3.2 *T-distributed Stochastic Neighbor Embedding*

En l'apartat anterior, s'ha vist que s'ha de ser prudent a l'hora d'interpretar l'anàlisi de components principals, ja que no s'han format clústers de dades mèdiques massa clars. Aquesta secció tracta sobre la utilització de la tècnica *t-distributed Stochastic Neighbor Embedding* (o t-SNE, veure secció A.2.2) per intentar formar estructures més clares. En aquest cas, es combina la tècnica de PCA amb t-SNE (s'utilitzen 25 components, que representen un 99 % de la variància original). Això es fa per aconseguir que l'algoritme no sigui tan pesat des de la perspectiva computacional i, així, ajudar a construir clústers més diferenciats.

Un concepte molt important dins el t-SNE és la perplexitat, la qual fa que es tingui una mica de control sobre el tipus de gràfiques que es volen generar. En concret, permet triar, en certa manera, la mida dels clústers, ja que està relacionada amb el nombre de punts que han d'estar propers. Per fer aquesta secció, s'ha experimentat amb les perplexitats: 5, 20, 35, 50 i 70; però només es representa una gràfica de 5 i una altra de 35, ja que s'ha considerat que són les més il·lustratives.

Aquesta tècnica està pensada per afavorir la clusterització, però és molt important assegurar-se que els clústers que es formen són rígids i no són fruit de coincidències. Per aquest motiu s'han construït diversos diagrames per cada una de les perplexitats i dades mèdiques i així poder comparar les gràfiques. En aquest cas ha resultat que les gràfiques eren molt similars, només canviava la distribució espacial dels punts, un fet que no és important en t-SNE.

Analitzant tots els casos, ha resultat que, igual que en PCA, les gràfiques corresponents a l'estadi T, són les que tenen les estructures més clares. A la figura 3.5 es poden veure dos exemples de gràfiques que s'obtenen:

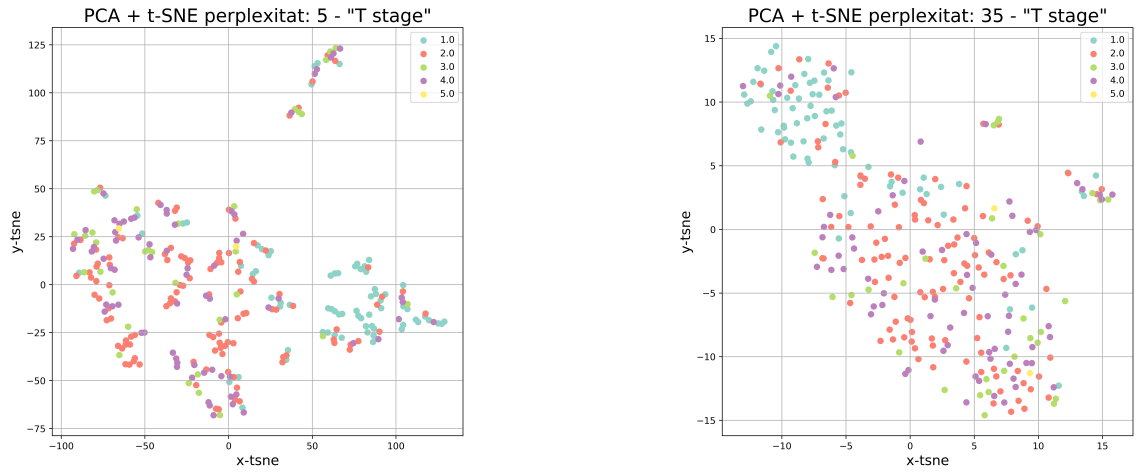

Figura 3.5: Distribucions dels pacients una vegada s'ha aplicat t-SNE combinat amb PCA. A la figura esquerra s'ha imposat una perplexitat de 5; a la figura dreta, de 35. Els punts tenen un color associat segons l'estadi T al qual està el tumor del pacient corresponent.

Si es compara la Figura 3.5 amb la Figura 3.4, es veu com ara les estructures són més clares. Per exemple, el clúster que inclou la majoria dels pacients a l'estadi 1 ara està molt més diferenciat que en el cas PCA. Abans, encara que els punts estiguessin al mateix lloc, no deixaven de pertànyer al mateix núvol de punts; ara es veu una separació més apreciable.

Tot i això, la distribució general es manté: s'observa un gran clúster al costat del clúster de l'estadi T, també hi ha dos clústers més petits a la perifèria. Vista la dificultat de separar el núvol principal en altres grups, s'arriba a la conclusió que hi ha un gran nombre de pacients amb moltes coses en comú, independentment dels diferents estadis del càncer. Per altra banda, els clusters de mida petita, tornen a correspondre als mateixos pacients de la primera franja de la Figura 3.3.

Un altre fet destacable és que, encara que en la figura de la perplexitat 5 té els grups més petits, no es perd l'estructura habitual que té l'altra gràfica o la Figura 3.4. Per tant, en aquest cas, la perplexitat 5 resulta ser molt útil perquè permet extreure més informació d'un cop d'ull: a part de veure la tendència general, també permet conèixer grups de pacients reduïts molt similars. Si s'observa la gràfica amb atenció, es pot veure com la majoria de clústers tenen punts del mateix color, és a dir, que reuneix pacients en les mateixes condicions mèdiques de manera relativament eficaç.

3.4 Models predictius

Un cop s'han acabat de fer les anàlisis que ajuden a entendre com són les dades que es manegen, és moment de passar a la darrera etapa de la Radiòmica: la construcció de models. Arribats a aquest punt hi ha bastant llibertat a l'hora de triar el tipus de model. Segons els objectius de l'estudi radiòmic, pot interessar fer un model classificador que ajudi a dividir els pacients en grups de manera automàtica, o models predictius que ajudin a entendre com serà l'evolució de la malaltia. En aquest cas, es construirà un model predictiu del temps de supervivència de cada pacient, aprofitant que és una informació que ve inclosa en el conjunt de dades mèdiques.

Es provaran dos models molt diferents: es començarà provant una tècnica tradicional com és un ajust lineal múltiple i, finalment, s'utilitzarà una tècnica molt més moderna com és el *Random Forest*.

3.4.1 Regressió lineal múltiple

A l'hora de fer models de qualsevol tipus, sempre s'intenta trobar el model que reprodueixi de la millor manera possible allò que s'està modelant, de la forma més simple possible. Per aquesta raó, una de les primeres propostes que es poden provar és la regressió lineal múltiple. L'expressió matemàtica, no és més que la generalització d'una recta en un espai de N dimensions, és a dir, un hiperplà:

$$t_{sup} = a_0 + \sum_{i=1}^{N} a_i x_i \tag{3.2}$$

Les variables predictores són les x_i, mentre que a_i són els coeficients que donen la importància necessària a cada variable. En aquest cas, les variables predictores seran les components principals, aprofitant el fet que són variables que no tenen cap correlació entre elles, evitant així utilitzar variables redundants. L'objectiu és quedar-se amb el nombre de variables més petit possible. Per triar quines seran, s'utilitzarà un procés iteratiu basat en el valor p estadístic i que consisteix en el següent:

1. Es calculen totes les components principals possibles, en aquest cas són 113.
2. S'ajusta el temps de supervivència utilitzant les 113 variables. A continuació es calcula el coeficient de correlació i s'emmagatzema fins al final del procés.
3. Es calcula el valor p de cada variable predictora (vegeu la secció A.1.2) i s'elimina aquella que tingui el valor més alt, ja que això indica quina és la variable menys significativa.
4. Es repeteixen les passes anteriors fins que quedi només una component principal.

5. De tots els models realitzats, es tria aquell que dóna millors resultats, mirant el coeficient de correlació que s'ha guardat anteriorment.

Després d'haver fet el procés anterior, s'ha trobat que el millor model és aquell que utilitza 69 components principals, amb un coeficient de correlació $R^2 = 0.795$ com es pot veure a la Figura 3.6:

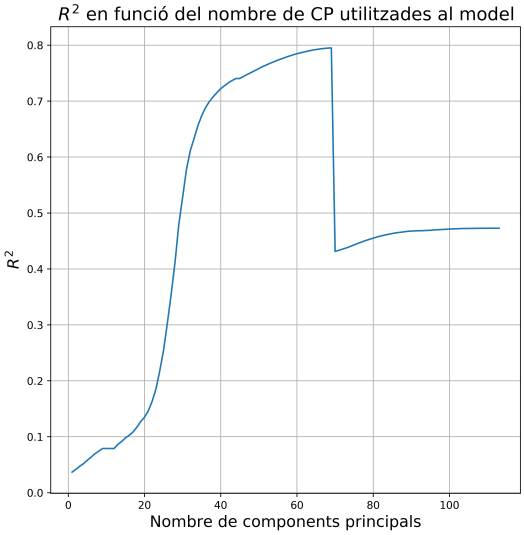

Figura 3.6: Representació de com evoluciona el coeficient de correlació del model en funció del nombre de components principals utilitzades.

Una vegada s'ha triat el model, només queda utilitzar-lo per predir tots els valors de supervivència i comparar-los amb els valors reals. Per facilitar la visualització de resultats, s'han ordenat els pacients utilitzant el temps de supervivència en ordre ascendent, a la Figura 3.7 hi ha els resultats:

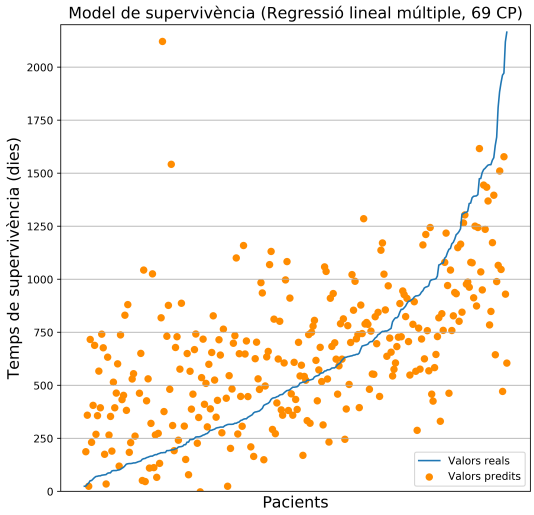

Figura 3.7: Representació dels resultats del millor model predictiu basat en una regressió lineal múltiple. En l'eix x hi ha els pacients ordenats segons el temps de supervivència. Els punts són els valors predits i la línia contínua correspon als valors reals.

Tot i que s'ha construït el millor model possible, els resultats estan molt allunyats dels esperats. A excepció d'alguns punts (que fins i tot podrien ser coincidències), la visió general és que els valors no

estan predits de manera satisfactòria. A més, els errors són molt grans (molts passen del 100 %), tant per valors subestimats com pels sobreestimats. Per aquest motiu, és necessària la recerca d'un model més sofisticat que ajudi a millorar aquests resultats.

3.4.2 *Random Forest*

Després de llegir articles recents que intenten fer models d'aquests tipus (com per exemple [20, 21]), resulta que una de les tècniques que dóna millors resultats és el *Random Forest*. Es pot consultar informació sobre el seu funcionament a la secció A.3.

Seguint la mateixa dinàmica dels darrers apartats, s'aprofitarà l'anàlisi de components principals per ser utilitzades com a variables predictores. Com en l'apartat 3.3.2, s'utilitzaran les primeres 25 components principals. La novetat respecte al model de regressió lineal, és que també s'inclouran algunes dades mèdiques com a variables predictores. Específicament, s'inclouran: els estadis T, N i M; el gènere (assignat un 0 si es tracta d'un home i un 1 si es tracta d'una dona) i si el pacient estava viu o no en el moment de construir la base de dades. La resta de dades mèdiques no s'han pogut incloure, ja que estaven incompletes per alguns dels pacients.

S'ha decidit construir el model amb 1000 arbres de decisió i entrenar-lo amb el 80 % dels pacients. Les prediccions d'aquest model es poden veure a la Figura 3.8:

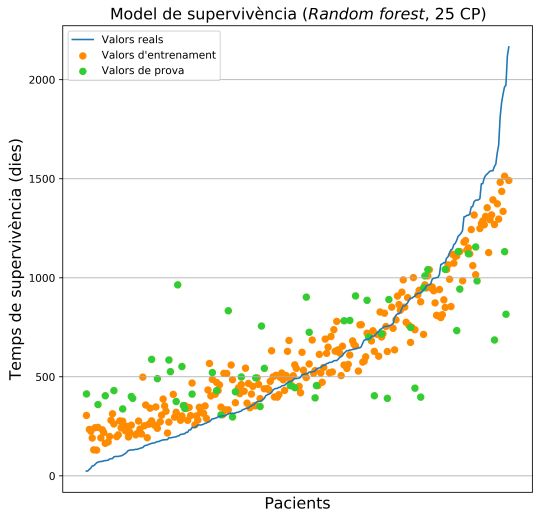

Figura 3.8: Representació dels resultats del model predictiu basat *Random Forest*. En l'eix x hi ha els pacients ordenats segons el temps de supervivència. La corba contínua són els valors reals, mentre que els punts (que es tan separats segons si són d'entrenament o no) són els valors que prediu el model.

Les millores en els resultats són més que evidents, si es comparen amb la Figura 3.7. Els valors estan molt més propers a la corba i no hi ha tanta dispersió: aquí es veu reproduïda, d'una manera més clara, la forma de la corba. També és fàcil identificar com es comporta el model segons l'ordre de magnitud del model: per valors de supervivència baixos, el model sobreestima el temps i per valors de supervivència alts, subestima. Allà on es comporta millor és a la zona central: entre els 500 i 1000 dies d'esperança de vida.

Un altre fet que indica que no és un model perfecte és que entre 10 i 15 pacients dels 59 que s'han usat per provar el rendiment del model, són els que tenen els pitjors resultats. Les subestimacions i sobreestimacions més elevades corresponen a pacients d'aquest grup. Això, no és més que un reflex de la dificultat de construir un model predictiu, encara que es triï la tècnica adequada.

CAPÍTOL 4

CONCLUSIONS

Durant aquest treball, s'han realitzat tot una sèrie de càlculs i anàlisis per esbrinar si és possible obtenir informació objectiva a partir d'imatges mèdiques.

El primer que s'ha notat és, com bé adverteixen alguns dels articles de les referències, que encara hi ha feina a fer per tractar d'estandarditzar el procés de la Radiòmica. Bona part de la primera part del treball ha consistit a preparar les imatges per poder fer l'estudi, ja que hi ha molts problemes d'incompatibilitat. També s'ha de fer més recerca per tractar d'evitar utilitzar característiques radiòmiques que aportin informació equivalent dins el mateix estudi. Sigui perquè resulten ser matemàticament iguals o perquè hi ha característiques que, en casos concrets, s'assemblen molt a alguna altra.

Mentre s'estudia per millorar aquests aspectes, les tècniques d'estandardització i *clustering* resulten molt útils per realitzar aquesta tasca de manera automàtica. Han donat la seguretat de poder utilitzar el mateix conjunt de dades durant tot l'estudi. També han ajudat a interpretar algunes de les estructures que es troben quan les dades es representen amb les tècniques de reducció de dimensionalitat. Pel que fa a aquestes tècniques, s'ha arribat a una conclusió molt important: PCA és molt útil per reduir el nombre de dades que es manegen, reduint la càrrega computacional: s'ha aconseguit reduir un conjunt de 113 variables a només 25, mantenint el 99 % d'informació original. Per altra banda, no ha resultat tan útil per visualitzar, a causa de la limitació d'haver d'utilitzar només 3 components per poder visualitzar les gràfiques. El que sí que ha proporcionat més ajuda per visualitzar, ha estat la tècnica t-SNE: Ha aconseguit passar d'un conjunt de 113 dimensions a un altre conjunt de només 2, que permet veure similituds entre pacients d'una manera molt còmoda.

Finalment, s'ha comprovat que fer models radiòmics no és una tasca gens senzilla. Els models simples no donen bons resultats, s'ha de recórrer a tècniques més complexes per poder arribar a resultats relativament satisfactoris.

En definitiva, la Radiòmica sí que permet extreure informació objectiva de les imatges mèdiques, però és important conèixer les seves limitacions actuals. També s'ha de tenir en compte que és un camp molt recent, gràcies als grans avanços que s'estan produint en tècniques modernes com el *Machine learning*, la intel·ligència artificial i les xarxes neuronals, és possible que la Radiòmica esdevingui una eina indispensable en l'àmbit de la medicina personalitzada.

Pel que fa a possibles extensions del treball, hi ha diverses possibilitats en cada un dels apartats del treball. Per exemple, es podrien realitzar les següents propostes:

- Estandardització i automatització de la segmentació del tumor.

  Com s'ha comentat, la part més subjectiva de la Radiòmica es troba a l'hora de segmentar el tumor. Es podria corregir el problema, utilitzant alguna de les tècniques que existeixen de

tractament d'imatges. D'aquesta manera es podrien segmentar tots els tumors fent servir la mateixa tècnica i no haver de deprendre d'un especialista.

- Incorporació de més dades mèdiques o genètiques.

  Disposar d'una major quantitat de dades mèdiques ajudaria a poder triar millor quines són les més rellevants a l'hora de fer agrupacions o prediccions. Per altra banda, incorporar dades genètiques per fer un estudi Radiogenòmic, tal vegada facilitaria la tasca de fer prediccions.

- Altres tècniques per fer models predictius.

  Encara que en aquest treball s'hagi intentat trobar una tècnica eficaç com és el *Random forest*, existeixen moltes altres tècniques per fer aquests tipus de models. Seria interessant posar a prova altres tècniques de *Machine learning* com les xarxes neuronals artificials. Es podria fer un petit estudi que compari aquestes tècniques amb les utilitzades aquí, per poder trobar la que tingui la millor relació precisió/cost computacional.

APÈNDIX A

FONAMENTS MATEMÀTICS

A.1 Estadística

A.1.1 Coeficient de correlació de Pearson

Per un conjunt de parelles de valors $\{x_i, y_i\}$ $(i = 1, ..., n)$ associades a un conjunt de dades, es defineix el coeficient de correlació lineal o de Pearson de la següent manera [22]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{A.1}$$

On $\overline{x}$ i $\overline{y}$ són les mitjanes aritmètiques de cada variable.

Aquest coeficient serveix per quantificar quina linealitat existeix entre dues variables, prenent el següent rang de valors: $-1 < r < 1$. Si r és proper a 1, vol dir que hi ha una gran dependència lineal directa entre les dues variables. Per altra banda, si r té un valor proper a -1, la seva relació lineal és gran, però inversa. Finalment si r està al voltant de 0, indica que existeix poca relació lineal entre les variables.

A.1.2 Valor p

En el camp de l'estadística, el valor p és una mesura de significació estadística [23]. Això vol dir que dóna informació sobre si algun resultat d'un estudi estadístic és produït per atzar o no. És una mesura de probabilitat que està relacionada amb la "hipòtesi nul·la"; la qual afirma que no existeix cap relació entre dos fenòmens mesurables. La hipòtesi no es descarta mentre el valor p sigui superior a cert valor triat de manera arbitrària segons el criteri de qui fa l'estudi.

Per trobar el valor p s'utilitza alguna taula que, a partir d'alguna altra variable estadística prèviament calculada (relacionada amb el nombre de graus de llibertat de les variables aleatòries), diu quin valor p correspon a cada resultat. El nombre serà un valor entre 0 i 1, que indicarà quina és la probabilitat d'obtenir un valor igual o encara més extrem que el valor observat, suposant que es compleix la hipòtesi nul·la.

A.1.3 Variable estandarditzada

Si x és un valor procedent d'una mostra caracteritzada per una distribució gaussiana que té mitjana μ i desviació estàndard σ, el valor de x estandarditzat, z, es defineix com:

$$z = \frac{x - \mu}{\sigma} \tag{A.2}$$

Transformant tots els valors d'un mateix conjunt a la seva z corresponent, s'aconsegueix que la distribució tingui mitjana $\mu = 0$ i desviació estàndard $\sigma = 1$. D'aquesta manera, comparar dades procedents de diferents variacions es pot fer de manera directa.

Per altra banda, si es coneixen la mitjana i la desviació estàndard de la variable original es pot calcular la transformació inversa fàcilment:

$$x = z\sigma + \mu \tag{A.3}$$

A.1.4 Transformació Box-Cox

Quan s'estan fent anàlisis de dades, es pot donar la situació que un conjunt d'observacions y_1, y_2, ...y_n no compleixi algun dels requisits que es tenien presents abans de començar l'estudi, com per exemple:

- Les dades no segueixen una distribució normal amb variància constant.
- Les observacions no són independents.
- No hi ha una correlació lineal entre elles.
- La variància de l'error no és consistent, és a dir, tenen diferències en les variàncies dels errors o asimetries en les distribucions dels valors dels errors.

Molts dels procediments més habituals per analitzar dades suposen que es compleixen els punts anteriors, ja que moltes vegades se cerca fer els estudis més simples possibles. Per aquest motiu hi ha cert interès a trobar algun tipus de transformació matemàtica que corregeixi tots (o alguns) dels problemes anteriors.

Per aquest cas resulta adequat el procediment que van proposar els matemàtics George E. P. Box i David Cox l'any 1964 [24]. Ells van proposar una transformació no lineal amb l'objectiu d'aconseguir un nou conjunt de valors que s'acostessin més al compliment de totes les suposicions comentades anteriorment i així evitar fer anàlisis invàlides i prediccions errònies. Posteriorment es va demostrar que la transformació pot ser usada fins i tot en situacions on cap transformació basada en potències pot aconseguir una distribució normal exacta [25].

La transformació Box-Cox és en realitat una família de transformacions que passen del conjunt y_i al conjunt $y_i^{'}$ i que depenen d'un paràmetre lambda a determinar:

$$y_i^{'} = \frac{(y_i)^{\lambda} - 1}{\lambda} \tag{A.4}$$

L'expressió és vàlida per λ diferent de zero, per aplicar la transformació al cas que sigui zero, es pot fer el límit aplicant la regla de L'Hôpital, per així aconseguir una transformació contínua:

$$\lim_{\lambda \to 0} \frac{(y_i)^{\lambda} - 1}{\lambda} = \ln y_i \tag{A.5}$$

Tant en un cas com en l'altre, les dades han de ser positives. En cas que no sigui així, es pot afegir una constant additiva a tot el conjunt i després aplicar la transformació.

Per trobar la transformació inversa, senzillament s'ha d'aïllar y_i de les expressions anteriors:

$$y_i = \begin{cases} (\lambda y_i^{'} + 1)^{1/\lambda} & \lambda \neq 0 \\ e^{y_i^{'}} & \lambda = 0 \end{cases} \tag{A.6}$$

És important adonar-se'n que les dades transformades i les dades originals no tenen les mateixes unitats.

En aquest treball el mètode que s'utilitza per trobar el valor de lambda òptim és diu "MLE" (de l'anglès, *Maximum Likehood Estimation*) i consisteix a maximitzar la funció "log Box-Cox" que, llevat d'una constant additiva, és:

$$\text{llf} = (\lambda - 1)\sum_i \ln(y_i) - \frac{N}{2}\ln\left(\sum_i \frac{(y_i^{'} - \bar{y^{'}})^2}{N}\right) \tag{A.7}$$

A.2 Tècniques de reducció de dimensionalitat

A.2.1 Anàlisi de components principals

Una anàlisi de components principals (PCA, de l'anglès *principal component analysis*) serveix per reduir la dimensionalitat d'un conjunt de dades que està format per un gran nombre de variables [26]. L'objectiu és partir d'aquestes variables (les quals tenen certa correlació) i arribar a un nou conjunt de variables de dimensió menor que l'original i que no tinguin cap correlació entre elles. Aquestes noves variables reben el nom de "components principals" o CP i s'ordenen de tal manera que, quedant-se amb les primeres components, s'aconseguirà reproduir la major part de la variabilitat de les variables originals.

Suposem que aquestes dades estan emmagatzemades en una matriu $\mathbf{X}$ de $n \times d$, on n és el nombre de mostres que es tenen i d és el nombre de variables que té cada mostra. Cada fila es pot veure com un vector $\mathbf{a}$ de d dimensions, ja que té d valors associats.

Hi ha diverses maneres d'obtenir les components principals, una d'elles és mitjançant la "matriu de covariància" Σ, que té dimensions $d \times d$. L'element σ_{ij} d'aquesta matriu representa la covariància entre l'element i-èssim i el j-èssim del vector $\mathbf{a}$, sempre que $i \neq j$:

$$\sigma_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{\mathbf{a}}_i)(x_{kj} - \bar{\mathbf{a}}_j) \tag{A.8}$$

Si $i = j$, queda la fórmula de la variància

Una vegada es té construïda la matriu de covariància es tracta de fer la seva descomposició en autovectors i autovalors. Els autovectors representen les "direccions" dels nous eixos, les "components principals". Per aconseguir reduir la dimensionalitat, s'ha de decidir quins d'aquests autovectors es poden rebutjar. Per això es mira l'autovalor associat: els autovectors amb els autovalors més petits són aquells que tenen menys informació associada, per tant són aquests que es poden eliminar.

A continuació, es tracta de decidir el nombre de components principals que seran suficients per realitzar l'anàlisi desitjada. Per aquesta tasca, resulta molt útil calcular la "variància explicada", que es pot calcular a partir dels autovalors, λ_k:

$$v_k = \frac{\lambda_k}{\sum_{i=1}^{d}\lambda_i} \times 100 \tag{A.9}$$

Amb això es podrà saber quanta "variància" de les dades originals representa cada CP.

La darrera etapa per completar el càlcul consisteix a transformar les dades originals al nou subespai. Triant k components principals (on $k < d$) es tindrà una nova matriu de dades $\mathbf{Y}$ que serà de dimensió $n \times k$. Per fer la transformació es necessita la matriu de projecció $\mathbf{P}$ que serà de $d \times k$ i es construeix posant els autovectors que s'han triat anteriorment en columnes.

Finalment, queda la transformació lineal següent:

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{P} \tag{A.10}$$

A.2.2 *T-distributed Stochastic Neighbor Embedding*

La tècnica *t-distributed stochastic neighbor embedding* (normalment abreujat com t-SNE) és un algoritme de reducció de dimensionalitat no lineal, que pertany al camp del *machine learning*. La tècnica permet visualitzar dades, que originalment tenen un nombre elevat de dimensions, en diagrames de dues o tres dimensions. Es parteix d'un conjunt de N dades multidimensionals $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, i es transforma en un altre conjunt $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ de dues o tres dimensions cada component.

L'objectiu és formar diagrames de tal manera que les dades que tenen coses en comú apareguin molt properes, mentre que els conjunts que no comparteixen similituds es trobin separats. Es caracteritza per ser una tècnica que es basa en la probabilitat, no és un procés matemàtic que sempre doni el mateix resultat. A continuació es donarà una idea general de com es transformen les dades mitjançant t-sne, per més detalls, vegeu l'article [27].

El procés comença calculant probabilitats condicionals entre els punts del conjunt de dades $\mathbf{X}$, ja que aquests valors són proporcionals al grau de similitud entre els punts:

$$p_{j|i} = \begin{cases} \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)} & i \neq j \\ 0 & i = j \end{cases} \tag{A.11}$$

La interpretació de l'equació (A.11) és la següent: la similitud entre el punt $\mathbf{x}_j$ i el punt $\mathbf{x}_i$ és la probabilitat condicionada ($p_{j|i}$) que $\mathbf{x}_i$ s'esculli com a veïnat de $\mathbf{x}_j$, si els veïnats s'escullen segons la seva densitat de probabilitat sota una distribució Gaussiana, centrada en el punt $\mathbf{x}_i$, de les distàncies euclidianes entre els valors. Això fa que punts propers en el conjunt original, tenen alta similitud i punts llunyans tenen una similitud molt petita.

σ_i és la variància de la distribució gaussiana, la qual es tria de tal manera que la distribució de probabilitat (P_i) sigui òptima per tots els punts. Concretament, se cerca aquella distribució que s'ajusta a la perplexitat [1] que ha triat l'usuari. En aquest cas, la perplexitat es defineix de la següent manera:

$$\text{Perp}(P_i) = 2^{H(P_i)} \tag{A.12}$$

On $H(P_i)$ és "l'entropia de Shannon" de la distribució P_i i està relacionada amb la variància de la distribució: com més alta és σ, més alta és l'entropia. $H(P_i)$ es calcula segons la següent equació:

$$H(P_i) = -\sum_j p_{j|i} \log_2(p_{j|i}) \tag{A.13}$$

Com els càlculs estan basats a calcular distàncies euclidianes amb $\|\mathbf{x}_i - \mathbf{x}_j\|^2$, la fórmula (A.11) dóna problemes per determinar la posició dels punts en el nou mapa quan un punt està molt allunyat dels altres. Aquest problema es corregeix calculant la probabilitat conjunta de la següent manera:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{A.14}$$

Una vegada acabat el procés anterior, es construeix el conjunt $\mathbf{Y}$, amb el nombre de dimensions al qual es vol arribar, a partir de les dades originals i tenint en compte el nombre de veïns que s'ha triat. Inicialment, els punts no tenen cap ordre particular (aquest és un dels motius que fa que sigui un algoritme amb cert atzar), serà quan s'apliqui el següent procediment quan s'obtinguin els valors que formaran el mapa final.

[1] La perplexitat es pot entendre com el nombre de veïnats que té un punt.

Es calculen les similituds entre els punts mitjançant les probabilitats, amb la diferència que s'utilitza una distribució de t de Student per veure la distribució de les distàncies. El càlcul es fa seguint la següent expressió:

$$q_{ij} = \begin{cases} \frac{(1+\|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq l}(1+\|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}} & i \neq j \\ 0 & i = j \end{cases} \tag{A.15}$$

La raó de triar aquesta distribució en lloc de la gaussiana, és que fa que la construcció del mapa final quasi no es vegi afectada per canvis d'escala.

A partir d'aquest moment es tracta d'anar movent els punts del conjunt **Y**, de manera que les similituds entre els punts reprodueixin, amb la màxima fidelitat possible, les similituds p_{ij} del conjunt original **X**. Amb aquest algoritme s'arribarà al conjunt **Y** òptim, que servirà per construir les gràfiques. S'ha de dir que, encara que cada vegada surti un resultat diferent, només afecta a com estan distribuïts els clústers per l'espai, no al nombre d'estructures que hi ha ni als seus integrants.

A.3 *Random Forest*

El mot anglès *Random Forest* fa referència a una tècnica que pertany al *Machine Learning* per fer models classificadors i/o predictius, a partir de combinar múltiples arbres de decisió [28].

Un arbre de decisió és un mètode per fer prediccions que parteix d'un conjunt de dades per construir un diagrama format per nodes i branques. Cada node representa un punt en el diagrama on s'ha de prendre una decisió sobre algun atribut de les dades. Possibles decisions serien, per exemple: mirar si el valor d'una variable és major o menor que un cert llindar; o suposar que existeix algun tipus d'agrupació, i decidir si el conjunt de dades hi pertany o no. Els elements que uneixen els diferents nodes entre si són les branques, i representen cada una de les diferents opcions que hi ha en les decisions. Els arbres de decisió, parteixen d'un node inicial i, d'una manera lògica (és a dir, fent les preguntes adequades, mirant correlacions entre variables, etc.), el van dividint fins a arribar a totes les diferents opcions possibles, les quals representen les prediccions finals. Un factor important a tenir en compte, és que les decisions són excloents, això vol dir que des del node inicial només hi ha un camí possible fins a un dels nodes finals.

Una de les característiques principals dels *Random Forests*, és que s'han d'"entrenar" amb un subconjunt de les dades totals, juntament amb els resultats reals associats a aquest subconjunt. La idea és construir un gran nombre d'arbres de decisió que van "aprenent" a partir del subconjunt, per intentar guanyar una intuïció que permetrà fer les preguntes adequades comentades anteriorment. L'objectiu és esbrinar com predir els valors desitjats quan se li presenti un conjunt de dades completament nou. Per aconseguir un aprenentatge de qualitat, és important disposar d'un conjunt de dades d'entrenament diversificat i amb les variables predictores que puguin ser més rellevants, en cas que aquestes es coneguin.

Com el seu nom del mètode suggereix, té certa aleatorietat. En aquest cas, l'atzar es troba a l'hora de triar els criteris per construir els arbres de decisió. Normalment, es comença amb un subconjunt de variables predictores i es construeix un arbre de decisió. Però per tenir en compte altres possibilitats, el model va permutant de manera aleatòria les variables dins el subconjunt. Això ho va repetint amb totes les mides del subconjunt possibles: si el conjunt de dades consta de M variables predictores, l'algoritme construeix arbres amb $1, 2, \ldots, M-1, M$ variables. Quan s'ha acabat el procés, s'estudia quines variables produeixen les prediccions més acurades, aprofitant la informació que es té sobre la relació entre elles.

Matemàticament, el mètode es construeix de la següent manera:

1. Es calculen totes les prediccions de tots els arbres utilitzant una funció de predicció $f(\mathbf{x})$:

$$f(\mathbf{x}) = c_0 + \sum_{m=1}^{M} c(\mathbf{x}, m) \tag{A.16}$$

On c_0 és el valor del qual es parteix en el node inicial i $c(\mathbf{x}, m)$ és la contribució de la variable m, del subconjunt de variables $\mathbf{x}$ ($\mathbf{x}$ és un vector). Notar que c depèn de $\mathbf{x}$ i m: això vol dir que es tenen en compte les branques que s'han seguit dins l'arbre, no només el valor de la variable m.

2. A continuació s'aplica la clau del *Random forest*: fer una mitjana de tots els arbres per calcular les prediccions finals. Matemàticament es calculen les mitjanes aritmètiques de les prediccions dels arbres. Si hi ha T arbres en el *Random forest*, la predicció final $F(\mathbf{x})$ és:

$$F(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} c_0(t) + \sum_{m=1}^{M}\left(\frac{1}{T}\sum_{t=1}^{T} c(\mathbf{x}, m, t)\right) \tag{A.17}$$

Un problema que apareix amb freqüència quan es fan models matemàtics, és el fet de sobreestimar les dades. És a dir, s'aconsegueix un model molt precís per un conjunt concret de dades, però falla quan es troba un conjunt nou. Els models basats en *Random Forests* intenten evitar el problema, fent que el nombre d'arbres T sigui gran. Així es poden beneficiar de la Llei dels grans nombres: a mesura que s'afegeixen arbres, els valors predits i els diferents errors associats, convergeixen a valors límit. L'explicació és que com més arbres s'utilitzen, més s'aprèn de les dades, fent que puguin aparèixer patrons que no estaven en les dades utilitzades per entrenar. Per tant, serà capaç de predir valors d'un conjunt nou on sí existeixi algun d'aquests patrons.

REFERÈNCIES

[1] G. Michael, "X-ray computed tomography," *Physics Education*, vol. 36 No 6, pp. 442–451, September 2001.

[2] E. B. Podgoršak, *Radiation Physics for Medical Physicists*, 3rd ed. Springer International, 2016.

[3] L. W. Goldman, "Principles of CT and CT Technology," *Journal of Nuclear Medicine Technology*, vol. 35, pp. 115–128, 2007.

[4] ——, "Principles of CT: Radiation Dose and Image Quality," *Journal of Nuclear Medicine Technology*, vol. 35, pp. 213–225, 2007.

[5] M. Avanzo, J. Stancanello, and I. El Naqa, "Beyond imaging: The promise of radiomics," *Physica Medica*, vol. 38, pp. 122–139, June 2017.

[6] "Diccionario médico-biológico, histórico y etimológico (Universidad de Salamanca)." [Online]. Available: https://dicciomed.usal.es/

[7] E. Florez et al., "Emergence of Radiomics: Novel Methodology Identifying Imaging Biomarkers of Disease in Diagnosis, Response, and Progression," *SM Journal of Clinical and Medical Imaging*, vol. 4 No.1, p. 1019, March 2018.

[8] S. Hawkins et al., "Predicting Malignant Nodules from Screening CT Scans," *Journal of Thoracic Oncology*, vol. 11 No. 12, pp. 2120–2128, July 2016.

[9] R. Tawani et al., "Radiomics and radiogenomics in lung cancer: A review for the clinician," *Lung Cancer*, vol. 115, pp. 34–41, January 2018.

[10] P. Lambin et al., "Radiomics: the bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, vol. 14, pp. 749–762, 2017.

[11] R. J. Gillies, P. E Kinahan, and H. Hedvig, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278 No 2, pp. 563–577, 2016.

[12] S. S. F. Yip and H. J. L. W. Aerts, "Applications and limitations of radiomics," *Physics in Medicine & Biology*, vol. 61 No 13, pp. R150–R166, June 2016.

[13] H. J. W. L. Aerts et al., "NSCLC-Radiomics," *The Cancer Imaging Archive*, 2015. [Online]. Available: http://doi.org/10.7937/K9/TCIA.2015.PF0M9REI

[14] "DICOM." [Online]. Available: https://www.dicomstandard.org/

[15] "DICOM Library." [Online]. Available: https://www.dicomlibrary.com/dicom/modality/

[16] J. J. M. van Griethuysen et al., "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77 No 21, pp. e104–e107, 2017.

[17] "3D Slicer." [Online]. Available: https://www.slicer.org/

[18] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, "Image biomarker standardisation initiative - feature definitions. arXiv:1612.07003v6," 2016. [Online]. Available: https://arxiv.org/pdf/1612.07003.pdf

[19] "National Cancer Institute (NCI)." [Online]. Available: https://www.cancer.gov/

[20] A. Chaddad et al., "Predicting survival time of lung cancer patients using radiomic analysis," *Oncotarget*, vol. 8, No 61, pp. 104 393–104 407, 2017.

[21] Y. Zhang et al., "Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer," *Scientific Reports*, vol. 7, 2017.

[22] M. H. Kutner et al., *Applied Linear Statistical Models*, 5th ed. Mc Graw- Hill/Irwin, 2005.

[23] R. L. Wasserstein and N. A. Lazar, "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, vol. 70, pp. 129–133, 2016.

[24] G. E. P. Box and D. R. Cox, "An Analysis of Transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 26 No 2, pp. 211–252, 1964.

[25] ——, "On Distributions and Their Transformation to Normality," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 31 No 3, pp. 472–476, 1964.

[26] A. Hervé and L. J. Williams, "Principal component analysis," *WIREs Computational Statistics*, vol. 2 No 4, pp. 433–459, 2010.

[27] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[28] L. Brieman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.