TITLE: Classifying brain metastases by their primary site of origin using a radiomics approach based on texture analysis: a feasibility study.

Author Names:

Rafael Ortiz-Ramón[1] ; Andrés Larroza[2] ; Silvia Ruiz-España[1] ; Estanislao Arana[3] ; David Moratal[1,*]

Author Affiliations:

[1]Centre for Biomaterials and Tissue Engineering, Universitat Politècnica de València. Camí de Vera s/n, 46022 Valencia, Spain.

[2]Department of Medicine, Universitat de València, Av. Blasco Ibáñez 15, 46010 Valencia, Spain.

[3]Department of Radiology, Fundación Instituto Valenciano de Oncología, Calle Beltrán Báguena 8, 46009 Valencia, Spain.

***Corresponding author:**

David Moratal, PhD

Centre for Biomaterials and Tissue Engineering

Universitat Politècnica de València

Camí de Vera s/n,

46022 Valencia

Spain

e-mail: dmoratal@eln.upv.es

Tel.: (+34) 96.387.70.07 (ext. 88939)

ABSTRACT

Objective:

To examine the capability of MRI texture analysis to differentiate the primary site of origin of brain metastases following a radiomics approach.

Methods:

Sixty-seven untreated brain metastases (BM) were found in 3D T1-weighted MRI of 38 patients with cancer: 27 from lung cancer, 23 from melanoma and 17 from breast cancer. These lesions were segmented in 2D and 3D to compare the discriminative power of 2D and 3D texture features. The images were quantized using different number of gray-levels to test the influence of quantization. Forty-three rotation-invariant texture features were examined. Feature selection and random forest classification were implemented within a nested cross-validation structure. Classification was evaluated with the area under receiver operating characteristic curve (AUC) considering two strategies: multiclass and one-versus-one.

Results:

In the multiclass approach, 3D texture features were more discriminative than 2D features. The best results were achieved for images quantized with 32 gray-levels (AUC = 0.873 ± 0.064) using the top four features provided by the feature selection method based on the *p*-value. In the one-versus-one approach, high accuracy was obtained when differentiating lung cancer BM from breast cancer BM (4 features, AUC = 0.963 ± 0.054) and melanoma BM (8 features, AUC = 0.936 ± 0.070) using the optimal dataset (3D features, 32 gray-levels). Classification of breast cancer and melanoma BM was unsatisfactory (AUC = 0.607 ± 0.180).

Conclusion:

Volumetric MRI texture features can be useful to differentiate brain metastases from different primary cancers after quantizing the images with the proper number of gray-levels.

KEYWORDS:

Neoplasms, Unknown Primary;

Magnetic Resonance Imaging;

Image Processing, Computer-Assisted;

Biomarkers;

Feasibility Studies

KEY POINTS:

- Texture analysis is a promising source of biomarkers for classifying brain neoplasms.
- MRI texture features of brain metastases could help identifying the primary cancer.
- Volumetric texture features are more discriminative than traditional 2D texture features.
- The number of gray-levels used to quantize images influence the results.
- Radiomics analyses merits further research in cancer studies to reduce invasive procedures.

ABBREVIATIONS AND ACRONIMS:

ANOVA = Analysis of variance

AUC = Area under receiver operating characteristics curve

BM = Brain metastases

CM = Confusion matrix

CV = Cross-validation

GLCM = Gray-level co-occurrence matrix

GLRLM = Gray-level run-length matrix

GLSZM = Gray-level size zone matrix

LGOCV = Leave-group-out cross-validation

NGL = Number of gray-levels

NGTDM = Neighborhood gray-tone difference matrix

RF = Random forest

TA = Texture analysis

INTRODUCTION

Brain metastases (BM) are the most common neoplasms of the central nervous system in adults. The prognosis of patients diagnosed with these lesions is poor: their median survival is limited to months even for patients under treatment [1–4]. The incidence of BM is unavailable although some studies reported that they occur in 9–17% of patients with cancer [1, 5]. However, these rates are currently increasing due to improved imaging techniques for diagnosis and prolonged survival from primary cancers, among other reasons [5].

The primary tumors that metastasize more frequently to the brain are those originated in lung (≥50%), breast (15–25%) and skin (melanoma) (5–20%) [3]. However, some studies indicate that there is a percentage of patients (2–14%) presenting BM as the first manifestation of an unknown primary tumor [5]. These patients are subjected to invasive neuropathological procedures and imaging evaluations, and sometimes the origin of the BM remains undiagnosed at the time of death [6–10]. Therefore, there is a clear need to detect the primary tumor in a fast, reliable and non-invasive way, as even neuropathological strategies can offer contradictory results [10].

In the past years, radiomics analysis has been proved to be a valuable methodology to increase precision in diagnosis or to predict treatment response in cancer research [11–13]. Radiomics is defined as the analysis of a large amount of data extracted from medical images to increase the power of decision support tools [11, 12]. Radiomics involves processes like image acquisition, image segmentation or data mining, but the focus of radiomics is the extraction of features that describe quantitatively the image [14]. To this end, texture analysis (TA) has been proved to be an excellent source of imaging biomarkers. Texture analysis refers to the application of mathematical methods to evaluate the gray-level patterns and pixel interrelationships within an image [15]. The main reason behind using TA to characterize tissues in medical images is that TA quantifies the intrinsic heterogeneous properties that are usually imperceptible to the human eye. Some TA methods have been successfully applied in neurologic disorders studies, including brain lesions like BM [16–22], using MRI as the main imaging technique. Particularly, contrast-enhanced T1-weigted MRI was the main sequence in these studies as it is employed for initial brain tumor detection and contains abundant diagnostic information [22, 23].

Some considerations have to be taken before performing TA. Preprocessing of the image region should be analyzed to minimize the effects of MRI acquisition protocols. Interpolation, normalization or quantization are preprocessing techniques commonly used to improve texture discrimination [24]. In particular, the quantization process has been demonstrated to have a substantial impact on the texture profile of medical images [25, 26], so it is recommended to optimize the number of gray-levels. Also, 3D TA should be considered instead of traditional 2D TA because volumetric TA allows to capture tissue heterogeneity more accurately [27].

The purpose of this work was to identify the primary site of origin of BM using MRI texture features in combination with a random forest (RF) classifier based on the radiomics practice. Our hypothesis was that TA could help to find differences between BM from different primary sites of origin, considering that it is not possible to distinguish the primary tumor by only examining the T1-weighted image of the BM.

MATERIALS AND METHODS

Patients

This retrospective, single-center study was approved by the Institutional Review Board and all subjects provided written informed consent.

Patients showing single or multiple BM were consecutive reviewed by an expert neuroradiologist (20 years-experience). Inclusion criteria comprised: (1) pathologically confirmed lung cancer, breast cancer or melanoma and only one single primary tumor; (2) no previous treatment, biopsy or surgical resection on BM; (3) all BM confirmed by imaging and clinical follow-up and (4) no clear qualitative and/or systematic differences on T1-weighted images of the BM to identify the primary cancer (ie, hyperintense in every melanoma case). Exclusion criteria were as follows (1) small metastases (longest diameter < 9 mm) as TA cannot capture texture information properly in small regions [24]; (2) more than 3 BM per patient; (3) multiple BM were situated in different brain areas.

The first thirty-eight patients (22 men and 16 women, mean age 60.05 years, age range 24–74 years) who complied with inclusion criterion and not with exclusion criteria were selected between December 2013 and April 2016 were included. Sixty-seven baseline BM were found in these patients: 27 derived from lung cancer, 23 from melanoma and 17 from breast cancer. Figure 1 shows an example of these types of BM.

Imaging Protocol

Imaging was performed using a 1.5T MRI scanner (Optima MR450w; GE Medical Systems, Milwaukee, WI, USA). The MRI protocol included three-dimensional inversion recovery fast-spoiled gradient-echo (IR-SPGR, BRAVO) T1-

weighted brain images, according to standardized protocol [28]. Images were acquired without magnetization transfer, after intravenous administration of a single-dose of gadobenate dimeglumine (0.1 mmol/kg, MultiHance, Bracco; Milan, Italy) with a 6 minutes delay. All the BM were scanned using the same imaging parameters since changing these parameters may lead to differences in TA performance [29, 30]: repetition time/echo time (TR/TE) of 8.5/2.2 ms; flip angle of 12º; matrix size of 256×256; pixel size of 0.98×0.98 mm^2; and slice thickness of 1.3 mm. Partial bias field correction in raw data was performed via the on-scanner “pre-scan normalize” option. No on-scanner gradient distortion correction was applied. As no diffusion weighted-sequences were used in this work, post processing bias field correction was not applied

Regions of Interest

Segmentation of the BM in 2D and 3D was performed using a software tool developed specifically for this study in MATLAB (R2015b; The MathWorks Inc., Natick, MA, USA). To segment each BM in 2D, the axial slice of the 3D T1-weighted image showing the most solid lesion component was manually segmented by an expert neuroradiologist (20 years-experience). To segment each BM in 3D, all the axial slices of the 3D T1-weighted image showing tissue of the same lesion were segmented using a semiautomatic method based on the Chan-Vese algorithm [31] that takes the manually segmented 2D lesion as the initial contour. Each 3D segmented lesion was revised by the expert. The longest diameters of the volumetric lesions were normally distributed without statistical differences (One-way ANOVA *F*-test, $p>0.05$, $p=0.314$) between the three classes, with mean ± standard deviation of 24.22 ± 10.67 mm (lung cancer BM), 19.92 ± 7.93 mm (melanoma BM) and 22.08 ± 10.92 mm (breast cancer BM).

Image Preprocessing

Prior to feature extraction, some preprocessing techniques were applied to improve texture discrimination. Firstly, the MRI regions were normalized using the $\mu \pm 3\sigma$ method to enhance the differences between classes [32].

Gray-level quantization (reduction of the levels of gray used to represent the image) was also applied to reduce the computational time and to improve the signal-to-noise ratio of the texture outcome [33]. In particular, different numbers of gray-levels (NGL) were tested (8, 16, 32, 64 and 128) to study the influence of the quantization process in the discriminative power of the features.

Finally, volumetric regions were isotropically resampled to the in-plane resolution (voxel size = $0.98 \times 0.98 \times 0.98$ mm^3) using cubic interpolation to ensure the conservation of scales and directions when extracting the 3D features [27].

Feature Extraction

Feature extraction was performed using the *Radiomics* MATLAB package [34]. Forty-three texture-based features derived from five statistical methods were computed. Three features were extracted from the intensity histogram (first-order statistics) and the other 40 features were extracted from the following higher-order statistical methods: gray-level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size-zone matrix (GLSZM) and neighborhood gray-tone difference matrix (NGTDM). Table 1 summarizes the features described in [34].

The proposed features met the criterion of rotation invariance to achieve texture parameters that are not dependent on the orientation of the brain in the images. To this end, only one GLCM, GLRLM, GLSZM and NGTDM per lesion was computed. For 2D TA, the neighboring properties of pixels in the 4 directions of the

2D space (0°, 45°, 90°, 135°) were averaged equally. For 3D TA, the neighboring properties of voxels in the 13 directions of the 3D space were averaged differently to take into account discretization length differences [34].

Finally, 10 different datasets of texture features were obtained: five datasets, one per NGL, extracted from the 2D regions and five datasets, one per NGL, from the 3D regions. All features were standardized to zero mean and unit variance to avoid model building being affected by the differences in the feature scales [35]. A summary of the procedure followed to obtain the datasets is illustrated in Figure 1.

Strategies for Multiclass Classification

As mentioned before, three classes of BM were considered according to the primary site of origin (lung cancer, breast cancer and melanoma). Random forest (RF) is a well-known ensemble learning method of the decision trees family that usually provides excellent classification results, especially when dealing with multiclass problems [36, 37]. Therefore, the RF classifier was chosen as the predictive model to evaluate the discriminative power of features. The number of trees in the RF model was set to 250 and the number of random variables used as candidates at each split (*mtry*) was chosen from $mtry \in \{2, 3, 4, ..., 14, 15\}$ in the parameter tuning process.

In the first stage of the study, the 10 datasets were analyzed separately using a purely multiclass approach with RF. The resulting statistical metrics derived from the model performance of each dataset were compared to identify the dataset of features that provided the best classification results. Afterwards, the optimal dataset was evaluated using a one-versus-one strategy to examine the capability of these features to differentiate between individual types of BM.

Model Performance

Considering the small sample size of our datasets, we decided to evaluate the performance of the classifier within a nested cross-validation (CV) structure (Figure 2). Good estimates of the model performance can be achieved using the validation data when the number of samples is not large [38]. The outer resampling loop of the nested CV structure was used to optimize the number of features and the inner resampling loop was used to tune the model parameter (*mtry*).

Leave-group-out CV (LGOCV) was applied in the outer resampling loop. This resampling method randomly divides each dataset into training and test sets *N* times, forming *N* groups. Each group is examined independently: the samples of the training set of a group are used to build the model and then this model is evaluated using the samples of the test set of the same group. Then, the classification results provided by the estimates of all groups are averaged. A total of N=100 groups were used to reduce the variance of the CV results [38]. In each group, 25% of the samples were randomly selected as test set and the remaining 75% were used as training set.

Brain metastases from the same patient were treated indistinctively in the resampling step to avoid selection bias. To support this decision, a Pearson correlation test was performed to measure the linear dependence between random pairs of vectors of texture features from BM of the same patient ($|r|$ = 0.431 ± 0.296) and BM from different patients ($|r|$ = 0.424 ± 0.248). No statistical difference was found between the two groups (Welch’s *t*-test: p=0.917), suggesting that BM from the same patient are correlated in the same way that BM of different patients can be.

For the feature selection step, a filter method based on the *p*-value was employed to obtain a ranking of features with the most discriminative power. This method evaluates the statistical significance of each feature independently, without analyzing the relation between features [39]. The *p*-values were obtained with the One-way Analysis of Variance (ANOVA) *F*-test for the multiclass strategy and the Welch's *t*-test for the one-versus-one strategy. The RF variable importance computed in the training process was also tested as a feature selection method to compare the results obtained with this method and our proposed filter method. To avoid overfitting, feature selection was implemented within the model-building process, that is, a different ranking of features was obtained in each group using only the training samples of each group [40]. The ranked features were progressively added one by one from most to least important and then each feature subset was used to tune the model parameter (inner 10-fold CV loop), to train the model and to compute the metrics on the test samples of the same group. At the end, a total of *F*=43 sets of metrics were obtained in each group evaluation, one per each feature subset.

Although several metrics were obtained, the relevance of the classification results was estimated using the area under receiver operating characteristic curve (AUC) averaged over groups' estimates (mean ± standard deviation). In the multiclass strategy, AUC was computed by averaging the one-versus-all statistics, as it is a simple way to extend the AUC computation to multiple classes problems [41].

The model evaluation process was implemented with the Caret package [42] in R language, version 3.2.5 (R Development Core Team, Vienna, Austria).

RESULTS

Multiclass strategy

In general, 3D features provided better classification accuracy than 2D features in the multiclass strategy, but the number of gray-levels used for quantization affected the model performance considerably. As it is shown in Figure 3, 3D features from the MRI lesions quantized with NGL of 8, 16 and 32 gray-levels provided better AUC than the equivalent 2D features. However, for NGL=64, the resulting AUC was similar for 3D and 2D features, and for NGL=128, 2D features were more discriminative than 3D features. Therefore, 3D features were more influenced by the quantization of the MRI regions than 2D features, losing discriminative power when increasing NGL. The exact AUC values obtained for the 10 datasets are shown in Table 1 of the Supplementary Material.

The highest AUC was achieved using 3D features from the lesions quantized with NGL=32, obtaining an AUC = 0.873 ± 0.064 using only the top four features ranked with the *p*-value feature selection method. When using the RF variable importance in the classification of this dataset, the features were ranked similarly and the results were slightly worse, but comparable (AUC = 0.841 ± 0.074, with 12 features).

Table 2 shows that features derived from the GLCM, GLRLM and GLSZM topped the ranking with significant *p*-value ($p<10^{-3}$). However, the *p*-value obtained with the ANOVA *F*-test indicated that there was a significant difference between at least two of the three classes of BM primary cancers, so additional evaluation of the difference between individual groups was needed.

One-versus-one strategy

An overall confusion matrix (CM) was obtained for the dataset presenting the highest results in the multiclass strategy (3D features, NGL=32) by summing up all confusion matrices obtained in every group's estimate (Table 3). The overall CM revealed that lung cancer BM were classified correctly most of the time (82%), but breast cancer and melanoma BM were often misclassified.

The one-versus-one analysis revealed that it would be possible to differentiate precisely lung cancer BM from breast cancer BM (AUC = 0.963 ± 0.054) and melanoma BM (AUC = 0.936 ± 0.070) using few features of the optimal dataset (4 and 8 features respectively). However, poor accuracy was achieved when discriminating BM from breast cancer and melanoma (AUC = 0.607 ± 0.180), thus indicating that these features are not suitable for classifying those types of BM. These results are shown in Figure 4. Additional statistical metrics were computed to validate the results (Table 4).

Regarding the top ranked features, Table 5 shows that the ranking of features provided by the multiclass strategy mostly coincided with the rankings computed to classify lung cancer BM from breast cancer and melanoma BM. Furthermore, the top ten features of both rankings showed significant average *p*-values ($10^{-8} < p < 10^{-2}$). However, none of the features showed significant average *p*-value (p>0.2) when classifying BM from breast cancer and melanoma.

The ranking computed with the RF feature selection method was similar to that obtained with our filter method when classifying lung cancer BM from breast cancer and melanoma BM, especially in the top-ranked features. On the contrary the ranking obtained when classifying breast cancer from melanoma BM was different. When performing the classification analysis with this RF ranking we found that the results did not differ very much from those results obtained with the *p*-value method (lung cancer versus breast cancer BM: AUC = 0.966 ± 0.052 with 5 features; lung cancer versus melanoma BM: AUC = 0.922 ± 0.069 with 10 features; melanoma versus breast cancer BM: AUC = 0.608 ± 0.186 with 42 features).

DISCUSSION

The radiomics approach used in this study showed that 3D texture features were more suitable than 2D features for classifying lung cancer BM from breast cancer and melanoma BM, achieving an average AUC > 0.9 in both cases. Random Forest provided better accuracy results when limiting the number of features. The results showed that, with further research, TA could help in the identification of the primary site of origin in patients with BM from an unknown primary cancer. Also, patients with two known primary tumors could benefit from this methodology to find which tumor has metastasize to the brain.

This study extends and improves the preliminary results exposed in [43] and [44]. In [43] we analyzed the potential of 2D MRI texture features to classify lung cancer BM from breast cancer BM, focusing on the influence of the NGL. In [44], we studied the classification of lung cancer and melanoma BM by comparing the discrimination power of 2D and 3D MRI texture features and by testing several classifiers. In the present study we extended these works by studying the differentiation of lung cancer, breast cancer and melanoma BM all together, establishing a robust methodology to perform a multiclass classification applicable to other primary sites of origin. We also evaluated two feature selection methods, studied the influence of the NGL and compared more exhaustively the performance of 2D and 3D TA.

Our work is not the first attempt to differentiate BM by its primary site of origin using texture features. Beres et al. [45] studied the statistical significance of 2D and 3D texture features from the histogram and the GLCM to identify the differences between lung and breast cancer BM. Our work enhances this study by exploring more texture features, including melanoma patients and considering a machine learning approach. With our results, we support the conclusions of Beres et al. that TA may help in the discrimination of BM from different primary tumors.

We based our work on other similar studies that showed the potential of MRI texture features combined with machine learning techniques to classify different brain lesions, including BM. Larroza et al. [21] used texture features to distinguish between BM and radiation necrosis using a LGOCV structure and support vector machine classifier (AUC>0.9). Li et al. [22] used texture features to differentiate BM from different pathological types of lung cancers using K-nearest neighbor and back-propagation artificial neural network classifiers in a one-versus-one approach (AUC≥0.9 when differentiating small cell lung carcinoma from other types of lung cancers). Both studies showed promising result and were very influential to our work. However, we tried to go beyond by including 3D texture features and taking into account rotation invariance.

Several studies have addressed the problem of classifying different brain tumor types by analyzing the potential of 3D MRI texture features in comparison with 2D features [17, 18, 20]. These studies showed an improvement in classification accuracy when using 3D TA. The conclusions in these works are clear: 3D texture descriptors capture more information about the lesion heterogeneity than 2D descriptors. In particular, the study of Fetit et al. [18] is very conclusive on this matter. This study mainly compares 2D and 3D texture features with several predictive models to classify different childhood brain tumors. All the models worked better with 3D features: for example, the neural network classifier showed 12% improvement in AUC and 19% in overall accuracy when using 3D TA instead of 2D TA. Nevertheless, 3D TA presents some drawbacks. Firstly, the 3D segmentation of the lesion can be more complex and time-consuming than the segmentation of a single slice.

Additionally, 3D TA requires MRI scans as isotropic as possible to reduce the effect of the image interpolation, and the acquisition process of these scans can be very slow.

The influence of the NGL used in the quantization of MRI has been analyzed in some studies with mixed results. No difference was reported by several studies [34, 46] when comparing the effect of changing NGL on the texture outcome. However, other studies showed that the discriminative power of texture-based features were affected by the gray-level quantization. Chen et al. [47] found that the optimal results for characterizing breast lesions were achieved for NGL=32. Leite et al. [25] observed that quantizing with NGL=16 allowed to identify the etiology of brain white matter lesions more accurately. Mahmoud-Ghoneim et al. [26] analyzed the impact of varying NGL on GLCM features of brain white matter: they concluded that their classification results were influenced significantly by the NGL chosen and they obtained better results with NGL=128 for both 2D and 3D TA. Our results support the fact that the NGL should be optimized for each specific application because it can lead to better classification results.

Our study showed several limitations. The main limitation was the reduced set of BM; more samples would be needed to build and test a final predictive model. Also, we only considered metastases derived from the most common primary sites of origin; other types of BM like those from renal or colorectal cancer should be considered in further analyses because it is necessary to consider all possible sites of origin to build a reliable final predictive model. Moreover, we only included MR images acquired with the same scanner and imaging parameters since TA can be affected by differences in scan parameters; a multicenter study on this specific application should be performed to evaluate this limitation. Finally, our study failed to classify breast cancer and melanoma BM, so further investigation will be performed by exploring other texture methods like Local Binary Patterns or transform methods (Wavelets, Gabor filters…) or other MRI sequences that could capture differences between BM from different primary sites of origin. To our knowledge, a genetic or pathologic link between breast cancer and melanoma that could be related to these TA results is unclear at this point, and the study of this association goes beyond the objective of this work.

In conclusion, our results show that TA on T1-weighted MRI in combination with a RF classifier allows differentiating accurately BM of lung cancer origin from those of breast cancer and melanoma origin when the proper features are chosen. These results are promising but further research is expected to consolidate this methodology. Our results support the conclusions derived from other studies to encourage radiologists to use TA as a new tool to improve precision in diagnosis.

REFERENCES

1. Gavrilovic IT, Posner JB (2005) Brain metastases: epidemiology and pathophysiology. J Neurooncol 75:5–14
2. Stelzer KJ (2013) Epidemiology and prognosis of brain metastases. Surg Neurol Int 4:S192--202
3. Soffietti R, Cornu P, Delattre JY, et al (2006) EFNS Guidelines on diagnosis and treatment of brain metastases: report of an EFNS Task Force. Eur J Neurol 13:674–681
4. Kaal ECA, Taphoorn MJB, Vecht CJ (2005) Symptomatic management and imaging of brain metastases. J Neurooncol 75:15–20
5. Nayak L, Lee EQ, Wen PY (2012) Epidemiology of brain metastases. Curr Oncol Rep 14:48–54
6. Bartelt S, Lutterbach J (2003) Brain metastases in patients with cancer of unknown primary. J Neurooncol 64:249–53
7. Agazzi S, Pampallona S, Pica A, et al (2004) The origin of brain metastases in patients with an undiagnosed primary tumour. Acta Neurochir (Wien) 146:153–157
8. Pekmezci M, Perry A (2013) Neuropathology of brain metastases. Surg Neurol Int 4:245
9. Zakaria R, Das K, Bhojak M, et al (2014) The role of magnetic resonance imaging in the management of brain metastases: diagnosis to prognosis. Cancer Imaging 14:1–8
10. Bekaert L, Emery E, Levallet G, Lechapt-Zalcman E (2017) Histopathologic diagnosis of brain metastases: current trends in management and future considerations. Brain Tumor Pathol 34:8–19

11. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: images are more than pictures, they are data. Radiology 278:563–77

12. Lambin P, Rios-Velazquez E, Leijenaar R, et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48:441–446

13. Yip SSF, Aerts HJWL (2016) Applications and limitations of radiomics. Phys Med Biol 61:R150-66

14. Kumar V, Gu Y, Basu S, et al (2012) Radiomics: the process and the challenges. Magn Reson Imaging 30:1234–1248

15. Castellano G, Bonilha L, Li LM, Cendes F (2004) Texture analysis of medical images. Clin Radiol 59:1061–9

16. Kassner A, Thornhill RE (2010) Texture analysis: a review of neurologic MR imaging applications. AJNR Am J Neuroradiol 31:809–816

17. Mahmoud-Ghoneim D, Toussaint G, Constans JM, De Certaines JD (2003) Three dimensional texture analysis in MRI: a preliminary evaluation in gliomas. Magn Reson Imaging 21:983–987

18. Fetit AE, Novak J, Peet AC, Arvanitis TN (2015) Three-dimensional textural features of conventional MRI improve diagnostic classification of childhood brain tumours. NMR Biomed 28:1174–1184

19. Zacharaki EI, Wang S, Chawla S, et al (2009) Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. Magn Reson Med 62:1609–1618

20. Georgiadis P, Cavouras D, Kalatzis I, et al (2009) Enhancing the discrimination accuracy between metastases, gliomas and meningiomas on brain MRI by volumetric textural features and ensemble pattern recognition methods. Magn

Reson Imaging 27:120–130

21. Larroza A, Moratal D, Paredes-Sánchez A, et al (2015) Support vector machine classification of brain metastasis and radiation necrosis based on texture analysis in MRI. J Magn Reson Imaging 42:1362–8

22. Li Z, Mao Y, Li H, et al (2016) Differentiating brain metastases from different pathological types of lung cancers using texture analysis of T1 postcontrast MR. Magn Reson Med 76:1410–1419

23. Fink KR, Fink JR (2013) Imaging of brain metastases. Surg Neurol Int 4:S209-19

24. Larroza A, Bodí V, Moratal D (2016) Texture analysis in magnetic resonance imaging: review and considerations for future applications. In: Assessment of cellular and organ function and dysfunction using direct and derived MRI methodologies. InTech, Rijeka, Croatia, pp 75–106

25. Leite M, Rittner L, Appenzeller S, et al (2015) Etiology-based classification of brain white matter hyperintensity on magnetic resonance imaging. J Med Imaging 2:14002

26. Mahmoud-Ghoneim D, Alkaabi MK, De Certaines JD, Goettsche F-M (2008) The impact of image dynamic range on texture classification of brain white matter. BMC Med Imaging 8:1–8

27. Depeursinge A, Foncubierta-Rodriguez A, Van De Ville D, Müller H (2014) Three-dimensional solid texture analysis in biomedical imaging: review and opportunities. Med Image Anal 18:176–196

28. Ellingson BM, Bendszus M, Boxerman J, et al (2015) Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials. Neuro Oncol 17:1188–1198

29. Mayerhoefer ME, Breitenseher MJ, Kramer J, et al (2005) Texture analysis for tissue discrimination on T1-weighted MR images of the knee joint in a multicenter study: Transferability of texture features and comparison of feature selection methods and classifiers. J Magn Reson Imaging 22:674–680

30. Waugh SA, Lerski RA, Bidaut L, Thompson AM (2011) The influence of field strength and different clinical breast MRI protocols on the outcome of texture analysis using foam phantoms. Med Phys 38:5058–5066

31. Chan TF, Vese LA (2001) Active contours without edges. IEEE Trans Image Process 10:266–277

32. Collewet G, Strzelecki M, Mariette F (2004) Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magn Reson Imaging 22:81–91

33. Gibbs P, Turnbull LW (2003) Textural analysis of contrast-enhanced MR images of the breast. Magn Reson Med 50:92–98

34. Vallières M, Freeman CR, Skamene SR, El Naqa I (2015) A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med Biol 60:5471–96

35. Kuhn M, Johnson K (2013) Data pre-processing. In: Applied predictive modeling, 1st ed. Springer, New York, NY, pp 27–59

36. Fernández-Delgado M, Cernadas E, Barro S, et al (2014) Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15:3133–3181

37. Caruana R, Karampatziakis N, Yessenalina A (2008) An empirical evaluation of supervised learning in high dimensions. In: Proceedings of the 25th

international conference on Machine learning - ICML ’08. ACM Press, Helsinki, Finland, pp 96–103

38. Kuhn M, Johnson K (2013) Over-fitting and model tuning. In: Applied predictive modeling, 1st ed. Springer, New York, NY, pp 61–92

39. Kuhn M, Johnson K (2013) An introduction to feature selection. In: Applied predictive modeling, 1st ed. Springer, New York, NY, pp 487–519

40. Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A 99:6562–6566

41. Provost F, Domingos P (2003) Tree induction for probability-based ranking. Mach Learn 52:199–215

42. Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28:1–26

43. Ortiz-Ramon R, Larroza A, Arana E, Moratal D (2017) Identifying the primary site of origin of MRI brain metastases from lung and breast cancer following a 2D radiomics approach. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). Melbourne, VIC, pp 1213–1216

44. Ortiz-Ramon R, Larroza A, Arana E, Moratal D (2017) A radiomics evaluation of 2D and 3D MRI texture features to classify brain metastases from lung cancer and melanoma. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Seogwipo, pp 493–496

45. Béresová M, Larroza A, Arana E, et al (2017) 2D and 3D texture analysis to differentiate brain metastases on MR images: proceed with caution. Magn Reson Mater Phy 1–10

46. Ahmed A, Gibbs P, Pickles M, Turnbull L (2013) Texture analysis in assessment and prediction of chemotherapy response in breast cancer. J Magn Reson Imaging 38:89–101

47. Chen W, Giger ML, Li H, et al (2007) Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images. Magn Reson Med 58:562–571

Table 1

List of texture features used in this study

Method	Features	Number of Features
Histogram	Variance, Skewness and Kurtosis	3
GLCM	Energy, Contrast, Correlation, Homogeneity, Variance, Sum Average, Entropy and Autocorrelation	9
GLRLM	Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray-level Non-uniformity (GLN), Run-Length Non-uniformity (RLN), Run Percentage (RP), Low Gray-level Run Emphasis (LGRE), High Gray-level Run Emphasis (HGRE), Short Run Low Gray-level Emphasis (SRLGE), Short Run High Gray-level Emphasis (SRHGE), Long Run Low Gray-level Emphasis (LRLGE), Long Run High Gray-level Emphasis (LRHGE), Gray-level Variance (GLV) and Run-Length Variance (RLV)	13
GLSZM	Small Zone Emphasis (SZE), Large Zone Emphasis (LZE), Gray-level Non-uniformity (GLN), Zone-Size Non-uniformity (ZSN), Zone Percentage (ZP), Low Gray-level Zone Emphasis (LGZE), High Gray-level Zone Emphasis (HGZE), Small Zone Low Gray-level Emphasis (SZLGE), Small Zone High Gray-level Emphasis (SZHGE), Large Zone Low Gray-level Emphasis (LZLGE), Large Zone High Gray-level Emphasis (LZHGE), Gray-level Variance (GLV) and Zone-Size Variance (ZSV)	13
NGTDM	Coarseness, Contrast, Busyness, Complexity and Strength	5

GLCM Gray-level co-occurrence matrix, *GLRLM* Gray-level run-length matrix, *GLSZM* Gray-level size zone matrix, *NGTDM* Neighborhood gray-tone difference matrix

Table 2

Top ten features of the dataset with the highest accuracy (3D features, NGL = 32 gray-levels) ranked according to their average *p*-value computed with the ANOVA *F*-test in the multiclass analysis.

Method	Feature	Average Ranking	Average *p*-value
GLCM	**Variance**	**1,02**	**$< 10^{-8}$**
GLSZM	**Low Gray-level Zone Emphasis**	**2,72**	**$< 10^{-6}$**
GLCM	**Sum Average**	**3,02**	**$< 10^{-6}$**
GLSZM	**Small Zone Low Gray-level Emphasis**	**3,73**	**$< 10^{-6}$**
GLRLM	Short Run Low Gray-level Emphasis	5,36	$< 10^{-5}$
GLRLM	Low Gray-level Run Emphasis	6,72	$< 10^{-5}$
GLRLM	High Gray-level Run Emphasis	6,86	0.00001
GLSZM	High Gray-level Zone Emphasis	7,37	0.00001
GLCM	Autocorrelation	8,52	0.00004
GLSZM	Gray-level Non-uniformity	10,32	0,00062

The subset of features highlighted in bold provided the highest classification accuracy.

GLCM Gray-level co-occurrence matrix, *GLRLM* Gray-level run-length matrix, *GLSZM* Gray-level size zone matrix, *NGTDM* Neighborhood gray-tone difference matrix

Table 3

Overall CM extracted from the RF model performance using the dataset with the best results in the multiclass strategy (3D features, NGL = 32 gray-levels).

		Predicted Class		
		Breast Cancer	Lung Cancer	Melanoma
Actual Class	Breast Cancer	235 (58.75%)	44 (11%)	121 (30.25%)
	Lung Cancer	55 (9.17%)	492 (82%)	53 (8.83%)
	Melanoma	95 (19%)	66 (13.20%)	339 (67.80%)

Table 4

Additional metrics obtained in the one-versus-one analysis using the RF model on the best dataset (3D features, NGL = 32 gray-levels).

Primary Site of Origin	Lung Cancer vs. Breast Cancer	Lung Cancer vs. Melanoma	Breast Cancer vs. Melanoma
Number of Features	4	8	42
Sensitivity [a]	0,895 ± 0,163	0,867 ± 0,153	0,600 ± 0,258
Specificity [a]	0,880 ± 0,132	0,856 ± 0,153	0,496 ± 0,224
Overall Accuracy	0,862 ± 0,091	0,861 ± 0,092	0,560 ± 0,146
Kappa Index	0,711 ± 0,192	0,722 ± 0,185	0,097 ± 0,298

Values are shown as mean ± standard deviation as a result over groups' estimates.

[a] Sensitivity and specificity were computed according to the optimal cutoff point of the ROC curve computed with the "closest-to-(0,1)" criterion.

Table 5

Top ten features of the best dataset (3D features, NGL = 32 gray-levels) ranked according to their average *p*-value computed with the Welch's *t*-test in the one-versus-one analysis.

Lung Cancer vs. Breast Cancer		Lung Cancer vs. Melanoma		Breast Cancer vs. Melanoma	
Feature	Average *p*-value	Feature	Average *p*-value	Feature	Average *p*-value
Variance [a]	$< 10^{-7}$	**Variance** [a]	$< 10^{-6}$	**Autocorrelation**	0,25991
Sum Average	$< 10^{-5}$	**Low Gray-level Zone Emphasis**	$< 10^{-5}$	**Sum Average**	0,27571
Low Gray-level Zone Emphasis	$< 10^{-5}$	**Small Zone Low Gray-level Emphasis**	$< 10^{-5}$	Gray-level Variance [b]	0,30066
Small Zone Low Gray-level Emphasis	$< 10^{-5}$	**Short Run Low Gray-level Emphasis**	$< 10^{-5}$	Entropy	0,30861
High Gray-level Zone Emphasis	0.00005	**Low Gray-level Run Emphasis**	0.00001	Strength	0,32501
Short Run Low Gray-level Emphasis	0,00011	**Sum Average**	0.00003	Coarseness	0,33500
Autocorrelation	0,00014	**High Gray-level Run Emphasis**	0.00007	**High Gray-level Zone Emphasis**	0,33575
High Gray-level Run Emphasis	0,00016	**High Gray-level Zone Emphasis**	0,00020	**Gray-level Non-uniformity** [b]	0,34701
Low Gray-level Run Emphasis	0,00035	Long Run Low Gray-level Emphasis	0,00048	Energy	0,34283
Gray-level Non-uniformity [b]	0,00503	**Gray-level Non-uniformity** [b]	0,00075	**High Gray-level Run Emphasis**	0,36704

The features highlighted in bold are in accordance with those features ranked in the multiclass analysis.

[a] These features are computed from the GLCM (Gray-level co-occurrence matrix)

[b] These features are computed from the GLSZM (Gray-level size zone matrix)

Figure 1. Procedure for obtaining the 10 different datasets of features. Examples of T1-weighted MRI axial slices showing the most solid area of brain metastases from lung cancer origin (a), breast cancer origin (b) and melanoma origin (c) are presented. Images were segmented in 2D and 3D, normalized with the $\mu \pm 3\sigma$ method and quantized using 5 different gray-levels. Then, features were extracted and standardized.

Figure 2. Structure of the nested CV method used to evaluate the different datasets of features. All the samples of each dataset were randomly separated in training and test sets N=100 times to evaluate the RF model with the AUC, examining different subsets of features.

Figure 3. Comparison between RF model performance using 2D and 3D features for all the number of gray-levels considered in this study. The numbers on the curves indicate the number of features used to achieve the maximum AUC.

Figure 4. Average receiver operating characteristics curves obtained in the one-versus-one analysis. The highlighted points on the curves indicate the optimal cutoff points that weighs both sensitivity and specificity equally computed with the “closest-to-(0,1)” criterion.