UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
INSTITUTO DE QUÍMICA DE SÃO CARLOS

VICTOR FRANCISCO

Computerized pattern recognition of lung nodules in magnetic resonance imaging for lung cancer diagnosis aid

São Carlos
2019

VICTOR FRANCISCO

Computerized pattern recognition of lung nodules in magnetic resonance imaging for lung cancer diagnosis aid

VERSÃO CORRIGIDA

Dissertation presented to the São Carlos School of Engineering at the University of São Paulo, as a requirement for obtaining a Master's Degree in Bioengineering.

Advisor: Prof. Dr. Paulo Mazzoncini de Azevedo Marques

São Carlos
2019

Catalog card prepared by Patron Service at "Prof. Dr. Sergio Rodrigues Fontes" Library at EESC/USP

F819c

Francisco, Victor
Computerized pattern recognition of lung nodules in magnetic resonance imaging for lung cancer diagnosis aid / Victor Francisco; promoters Paulo Mazzoncini Azevedo-Marques. -- São Carlos, 2019.

Master (Thesis) - Graduate Program in Bioengineering and Research Area in Bioengineering of the School of Engineering of São Carlos, Ribeirão Preto Medical School and São Carlos Institute of Chemistry at University of São Paulo, 2019.

1. Pattern recognition. 2. Lung cancer. 3. Magnetic resonance imaging. 4. Machine Learning. 5. Computer-aided diagnosis. I. Title.

Elena Luzia Palloni Gonçalves – CRB 8/4464

FOLHA DE JULGAMENTO

Candidato(a): Victor Francisco

Título: "Reconhecimento computadorizado de padrões de nódulos pulmonares em imagens de ressonância magnética para auxílio ao diagnóstico do câncer de pulmão"

Data da defesa: 29/10/2019

Comissão Julgadora	Assinatura	Resultado
Prof(a). Dr(a). Paulo Mazzoncini de Azevedo Marques FMRP/USP - Orientador		Não Votante
Prof(a). Dr(a). Marcel Koenigkam Santos HCFMRP/USP		aprovado
Prof(a). Dr(a). Renato Tinós FFCLRP/USP		APROVADO
Prof(a). Dr(a). Marcelo Costa Oliveira UFAL	p/	APROVADO

ABSTRACT

FRANCISCO, V. **Computerized pattern recognition of lung nodules in magnetic resonance imaging for lung cancer diagnosis aid**. 2019. 70 f. Master (Dissertation) - São Carlos School of Engineering; Ribeirão Preto Medical School and São Carlos Chemistry Institute of the University of São Paulo, 2019.

Lung cancer is the type of cancer that takes the most victims around the world and often presents a late diagnosis. Computed tomography (CT) is currently the reference imaging test for the diagnosis and staging of lung tumors. Recent studies have shown relevance in the characterization of lung tumors by different sequences obtained with magnetic resonance imaging (MRI). MRI also has the advantage of not exposing the patient to ionizing radiation, which occurs in CT scans. This work presents an investigation about the applicability of pattern recognition methods to computer-aided diagnosis of lung cancer in MRI exams in order to classify lung nodules and masses in benign and malignant. T1-weighted contrast-enhanced (T1PC) and T2-weighted (T2) MRI images associated with lung lesions were acquired retrospectively and prospectively, then semi-automatically segmented. Quantitative features were obtained from tumor 2D and 3D segmentation models of T1PC and T2. Each segmentation model provided 75 features, totaling 150. T1PC and T2 datasets were combined creating the T1PC-T2 dataset with 300 features. Unbalancing problems were solved by synthetically oversampling the datasets and by bootstrapping. Tumor classification was based on five machine learning classifiers and leave-one-out cross-validation. Relevant feature selection was performed using Wrapper. Results showed significant performance on balanced datasets, especially after feature selection. Naïve Bayes classifying balanced T2 with selected features provided the highest area under the receiver operating characteristic (ROC) curve value of 0.944. The most selected features were extracted from gray level co-occurrence matrix and shape of the tumor, which these features might indicate good correlation with clinical and pathological data. Hence, the investigated approach demonstrates potential for computer-aided diagnosis of lung cancer in MRI.

Keywords: Pattern recognition. Lung cancer. Magnetic resonance imaging. Machine learning. Computer-aided diagnosis.

RESUMO

FRANCISCO, V. **Reconhecimento computadorizado de padrões de nódulos pulmonares em imagens de ressonância magnética para auxílio ao diagnóstico do câncer de pulmão**. 2019. 70 f. Mestrado (Dissertação) - Escola de Engenharia de São Carlos, Faculdade de Medicina de Ribeirão Preto, Instituto de Química de São Carlos, Universidade de São Paulo, São Carlos, 2019.

O câncer de pulmão é o tipo de câncer que mais faz vítimas em todo o mundo e muitas vezes apresenta diagnóstico tardio. Tomografia computadorizada (TC) é atualmente o exame de imagem referência para o diagnóstico de tumores pulmonares. Estudos recentes mostram relevância na caracterização de tumores pulmonares por diferentes sequencias obtidas por ressonância magnética (RM). A RM também tem a vantagem de não expor o paciente à radiação ionizante, como ocorre nas TC. Este trabalho apresenta uma investigação sobre a aplicabilidade dos métodos de reconhecimento de padrões ao diagnóstico auxiliado por computador de câncer de pulmão em exames de RM a fim de classificar nódulos e massas pulmonares em benigno ou maligno. Imagens de RM ponderadas em T1 pós contraste (T1PC) e em T2 associadas a lesões pulmonares foram adquiridas retrospectiva e prospectivamente, então semi-automaticamente segmentadas. Os atributos quantitativos foram extraídos a partir dos modelos 2D e 3D segmentados de T1PC e T2. Cada modelo segmentado forneceu 75 atributos, totalizando 150. T1PC e T2 foram combinados criando o conjunto de dados T1PC-T2 com 300 atributos. Problemas de desbalanceamento foram resolvidos aumentando os conjuntos de dados de forma sintética e utilizando *bootstrapping*. A classificação dos tumores foi baseada em cinco classificadores de aprendizado de máquina e validados utilizando *leave-one-out*. A seleção de atributos mais relevantes foi realizada com Wrapper. Os resultados mostraram um desempenho significativo em conjuntos de dados balanceados, especialmente após a seleção de atributos. Naive Bayes classificando imagens em T2 com atributos relevantes selecionados obteve o maior valor de área sob a curva ROC (*receiver operating characteristic*) de 0,944. Os atributos relevantes mais selecionados foram extraídos da matriz de coocorrência de nível de cinza e da forma do tumor, indicando boa correlação com características clínicas dos tumores. Além disso, o estudo demonstra potencial para o diagnóstico auxiliado por computador para câncer de pulmão em imagens de RM.

Palavras-chave: Reconhecimento de padrões. Câncer de pulmão. Ressonância magnética. Aprendizado de máquina. Diagnóstico auxiliado por computador.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Lung cancer diagnosis and clinical challenges

Considered a major public health problem, lung cancer is the most deadly type of cancer in the world (TARTAR; KILIC; AKAN, 2013a). Lung cancer is a disease where diagnosis is often late, and by the time of identifying its clinical manifestation symptoms, a poor prognosis is already assumed due to the aggressiveness of the disease. Only 15% of patients survive the first five years after diagnosis (WU et al., 2013). Generally, late diagnosis prevents curative treatment since the condition is already at an advanced stage (NOVAES et al., 2006). In the lung cancer case scenario, its diagnosis is mainly assessed by evaluation of lung nodules in computed tomography (CT) (HOLLINGS; SHAW, 2002).

Lung cancer prognosis is a difficult task because it is directly influenced by the variability of the diagnosed stage of the tumor (KOENIGKAM SANTOS et al., 2014). The stage at which the tumor is identified is a determinant for the prognosis and definition of the therapy to be applied (DETTERBECK; BOFFA; TANOUE, 2009). There is also evidence of biological parameters and visual characteristics of the tumor in computed tomography imaging that may aid the decision making process (AUSTIN et al., 2013; DUTTA; MAITY, 2007; OHNO et al., 2012; SAKAO et al., 2010; SHIMIZU et al., 2005). However, such characteristics are generally described subjectively and qualitatively (e.g. attenuation heterogeneity, spiculated contours) (AERTS et al., 2014). Adversely, qualitative visual characteristics similar to those found in benign lesions can be seen in cases of malignant tumors (BARTHOLMAI et al., 2015).

In contrast to the limitations of qualitative evaluations, the computerized image analysis allows the objective and precise extraction of quantitative descriptors which can potentially be used as a diagnostic support tool, and as predictive biomarkers for therapeutic decision making (AERTS et al., 2014). In this context, computerized pattern recognition methods have been used in CT scans of lung tumors. This approach has shown promising results in aiding lung cancer diagnosis and lung tumor characterization (EL NAQA et al., 2009; FERREIRA; OLIVEIRA; MARQUES, 2017; FERREIRA JUNIOR et al., 2018; VAIDYA et al., 2012).

Recently, in order to complement CT finding as a way of improving lung cancer diagnostic and prognostic accuracy, the use of magnetic resonance imaging (MRI) has

gradually increased in clinical practice (LIU et al., 2015). The great advantage in using MRI for the study of thoracic diseases is related to the reduction of the patient's exposure to ionizing radiation, higher resolution of contrast between different tissue types, and use of contrast media with less adverse effects. MRI also permits dynamic studies (perfusion, mobility), without adding injury to the patient (there is no greater exposure to radiation, for example) (COOLEN et al., 2014; KOENIGKAM-SANTOS et al., 2015).

1.2 Literature search

Searches were conducted in IEEE Xplore and PubMed online databases to find similar works. The following query was applied to "Full Text and Metadata" for IEEE Xplore, and "Title/Abstract" for PubMed: ("lung cancer" OR "lung" OR "nodule") AND ("mri" OR "magnetic resonance imaging") AND ("machine learning" OR "pattern recognition" OR "computer-aided diagnosis" OR "classification").

In order to include papers in the analysis, they must be published in English from January 1st 2010 until September 30th 2019, and describe the use of machine learning to classify lung cancer or lung nodules in MRI exams. IEEE Xplore showed no results using this query, and PubMed results contained 127 possible matches. Through a title and abstract screening, none of the 127 PubMed results fit into analysis requirements, showing originality in our study.

1.3 Objectives

Considering the potential and advantages of using MRI exams in lung cancer evaluation, the purpose of this study is to investigate whether the computerized pattern recognition in MRI is clinically useful in the characterization of lung cancer.

The specific objectives are:

- Creating a database with benign and malignant cases classified in both T1 post contrast and T2 MRI sequences;
- Investigating shape and texture based features to classify lung tumors in two different classes, benign and malignant;
- Investigating pattern recognition methods for lung tumors imaging classification;

- Evaluating whether the developed method is clinically useful as a computer-aided diagnosis tool.

2 MATERIAL AND METHODS

2.1 Image acquisition and segmentation

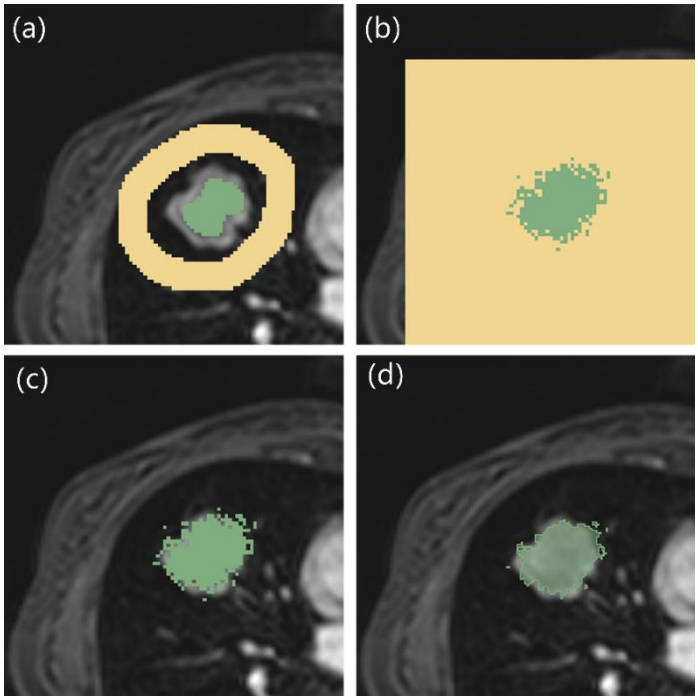
Our institutional research board with a waiver of patients' informed consent has approved this prospective study (HCRP process number: 3733/2017). The clinical chest MRI protocol included two different sequences with the aid of patient's breath hold procedure. Post-contrast T1-weighted (T1PC) images is the sequence that closely approximates post-contrast CT images, with good spatial and contrast resolution. On the other hand, the sequence with T2-weighted (T2) images provide different information from tissues where abnormal brightness could represent a disease process such as cancer. A 1.5T device (Achieva, Phillips) with chest coil obtained images of patients placed on supine position. Prior to the acquisition of subjects, we performed an evaluation in the hospital database and found 15 cases matching the adopted clinical chest MRI protocol. A senior radiologist (M.K.S) indicated the lesions on images. Diagnostic was assessed after pathological confirmation of clinical treatment/stability. The radiologist also defined a lung window level and width of 800 and 2000 respectively, to avoid any lack of pattern during his visual analysis. To ensure patients' privacy, all MRI exams were anonymized.

The full image database consists of 35 cases, 23 malignant and 12 benign. Exceptionally, there are two specific benign cases not appearing on T2-weighted images due the sequence low resolution and nodule size. Tumors have a size equal to or greater than 1 cm. Each tumor was semi-automatically segmented by the 3D region growing Fast GrowCut algorithm (ZHU et al., 2014), which is an extension of the medical imaging analysis and visualization open source platform 3D Slicer v4.7.0-2017-09-05 (r26338) (FEDOROV et al., 2012). The algorithm workflow requires the user to select two regions on each anatomical plane. As shown on Figure 1, the first region is a seed mark within the object of interest, and then another selection outside the object is defined for the second region. Chronologically, we divided our cases in training/validation and test datasets. The first 21 acquired cases (14 malignant and 7 benign) compose the training/validation, the other 14 cases (9 malignant and 5 benign) compose the test dataset. The data division has this format because the 21 first cases were acquired prior to the master's degree qualification process.

We also proposed a 3D Slicer post-segmentation correction protocol for T1PC images to evaluate the difference of performance in segmentation with less noise, calling this set as

T1PC-p. For lesions smaller than 3 cm (not masses), we performed an erode effect with 4 neighbors then a dilate effect with 4 neighbors. For lung masses (lesions equal or greater than 3 cm) we performed an erode effect with 8 neighbors, followed by two dilate effects of 4 neighbors, then repeated the whole sequence once more. No post-segmentation correction was done on T2 due to the low resolution of this sequence. Once the segmentation is done, 2D and 3D models of the tumor are created using the Model Maker module (LORENSEN; CLINE, 1987) on 3D Slicer as well. These two tumor models (2D and 3D) were used as input for the feature extraction process.

Figure 1 - Fast GrowCut segmentation of lung nodule on axial plane in T1-weighted contrast-enhanced MRI with window level of 800 and width of 2000. (a) Seed marks within the object and outside of it. (b) Fast GrowCut growing result. (c) Outside mark removal resulting the 2D segmentation. (d) 3D model boundary outline on 2D view.



Source – Author

2.2 Feature extraction

Lesions were characterized based on shape and texture using 75 quantitative features extracted from both 2D and 3D models, totalizing 150 features to extract from T1PC, T1PC-p and T2. Another feature set was established combining information from both sequences at the same time, called T1PC-T2, which has 300 features (150 from T1PC and 150 from T2). The extraction process has been done via SlicerRadiomics (VAN GRIETHUYSEN et al., 2017), an encapsulated version of the pyradiomics library as 3D Slicer extension.

At the time of this work, SlicerRadiomics supported the following feature classes: First Order Statistics; Shape-based; Gray Level Co-occurence Matrix (GLCM); Gray Level Run Length Matrix (GLRLM); and Gray Level Size Zone Matrix (GLSZM). Due the inherent issue extracting some quantitative features from MRI, First Order Statistics features were not extracted, since MRI voxel intensities do not have fixed meaning. In fact, this is due to the possible contrast variance as a consequence of variations in the magnetic field and/or random presence of noise, even if the acquisition is done on the same equipment (NYÚL; UDUPA; ZHANG, 2000). Appendix A has a table with details of each shape and texture feature.

2.2.1 Shape features

Shapes' representation is possible by means of geometric features being extracted from the image or the segmented object, e.g. edges, contours, joints, curves and polygonal regions (BURAK AKGÜL et al., 2011). Some features based on shape were already used to characterize radiological lesions, such as volume, surface area, sphericity, convexity, compactness, strength, maximum extension, aspect ratio and Fourier descriptor (MARVASTI, 2013). In addition, 3D geometric features were also extracted from lung nodules imaging, e.g. sphericity index, convexity index, intrinsic and extrinsic curvature index, and surface type (SILVA; CARVALHO; GATTASS, 2005).

Features in this class are independent from the gray level intensity distribution in the ROI (region of interest), and they are descriptors of the three-dimensional size and shape of the ROI. Thus, a total of 16 features were calculated: 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, and flatness.

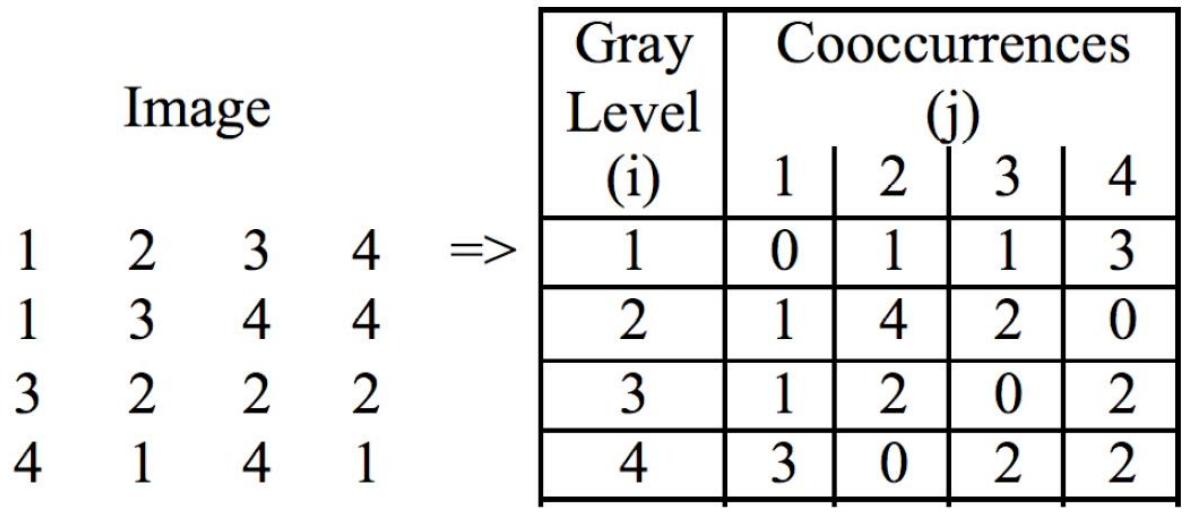
2.2.2 Gray Level Co-occurrence Matrix (GLCM) features

In the medical field, the ability of texture features to reflect details contained within a lesion on an image has been shown to be of great importance (BURAK AKGÜL et al., 2011). The statistical texture descriptors proposed by Haralick can be classified as second order features, representing one of the most used methods of texture analysis (HARALICK; SHANMUGAM; DINSTEIN, 1973). Texture characteristics based on second-order statistics

or co-occurrence matrix, obtains information about the positioning and neighborhood of pixels (OLIVEIRA; CIRNE; MARQUES, 2007). The co-occurrence matrix depends on the estimation of a discrete second-order probability function, which represents the probability of occurrence of a pixel pair with gray levels *i* and *j*, given a distance *d* and an orientation θ between the pixels in the dimensions *x* and *y*, respectively. Co-occurrence matrix calculation can be done on a volume of images as well. The three-dimensional co-occurrence matrix extends the evaluation of the second-order probability function to the *z*-axis, and then examines the probability of occurrence of pixel pairs between slices of a volume of images (MAHMOUD-GHONEIM et al., 2003). Second order statistical functions are applied in the co-occurrence matrix producing the texture features.

In this work, we have extracted 27 features from each lesion, as follows: autocorrelation, average intensity, cluster prominence, cluster shade, cluster tendency, contrast, correlation, difference average, difference entropy, difference variance, dissimilarity, energy, entropy, homogeneity 1, homogeneity 2, informal measure of correlation 1, informal measure of correlation 2, inverse difference moment, inverse difference moment normalized, inverse difference, inverse difference normalized, inverse variance, maximum probability, sum average, sum variance, sum entropy, and sum of squares.

Figure 2 – Example of the calculation of a co-occurrence matrix for a 4x4 image with 4 gray levels for $d = 1$ and $\theta = 0°$.



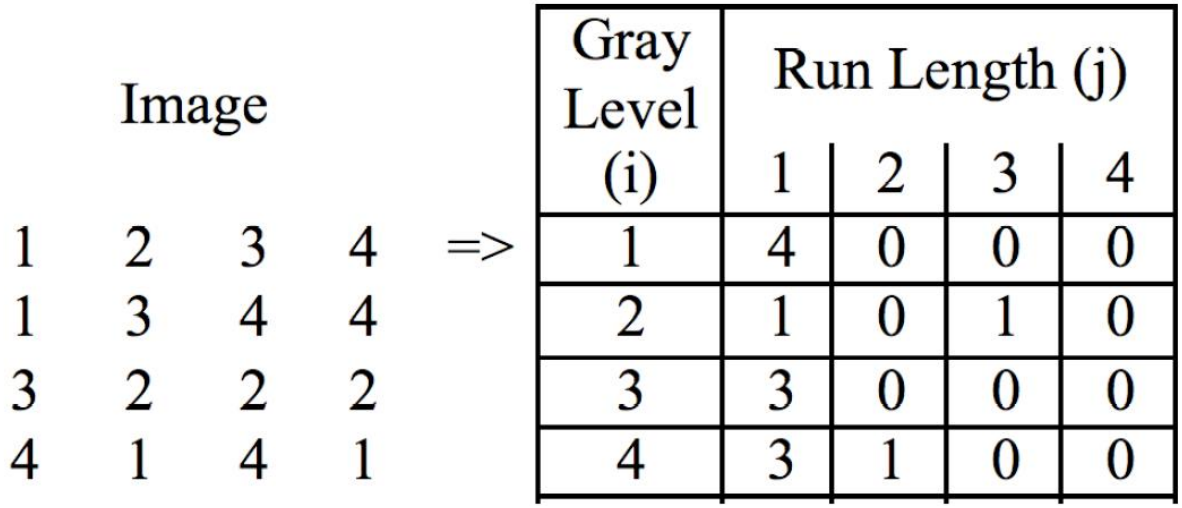
Source – Thibault et al. (2009)

2.2.3 Gray Level Run Length Matrix (GLRLM) features

Texture features extracted from a run-length matrix have demonstrated great classification results (TANG, 1998). A GLRLM computes gray level runs, which consist of a set of consecutive pixels of the same gray level value (GALLOWAY, 1975). GLRLM is formed by the number of runs with gray level *i* and length *j* that appear in the image along a given angle θ (CHU; SEHGAL; GREENLEAF, 1990; GALLOWAY, 1975). It is possible to characterize volumetric texture using run-length statistics too, since volumetric GLRLM are able to provide features capable of showing texture primitives' properties in 3D imaging (XU et al., 2004).

For this class, the following 16 features have been extracted from each lesion: short run emphasis, long run emphasis, gray level non-uniformity, gray level non-uniformity normalized, run length non-uniformity, run length non-uniformity normalized, run percentage, gray level variance, run variance, run entropy, low gray level run emphasis, high gray level run emphasis, short run low gray level emphasis, short run high gray level emphasis, long run low gray level emphasis, and long run high gray level emphasis.

Figure 3 – Example of the calculation of a run length matrix for a 4x4 image with 4 gray levels for $\theta = 0°$.



Gray Level (i)	Run Length (j)			
	1	2	3	4
1	4	0	0	0
2	1	0	1	0
3	3	0	0	0
4	3	1	0	0

Source – Thibault et al. (2009)

2.2.4 Gray Level Size Zone Matrix (GLSZM) features

The idea of a GLSZM is to quantify gray level zones of various sizes in an image, rather than calculate the number of voxels with same gray-level intensity on different

orientations as the GLRLM does (THIBAULT et al., 2009). Gray-level zone is defined by an adjoining region with voxels of same intensity value, and the enclosed voxels determine the zone size. For a GLSZM, each (*i*,*j*) element represents the number of zones with gray-level *i* and size *j*. Giving a delineated tumor, the quantity of rows is equivalent to the maximum gray level within the tumor, and the quantity of columns is equal to the largest zone size possible inside the tumor (YANG et al., 2013). Furthermore, only one matrix is calculated for every directions in the ROI, which means GLSZM is rotation independent.

Sixteen GLSZM features have been extracted from each lesion: small area emphasis, large area emphasis, gray level non-uniformity, gray level non-uniformity normalized, size-zone non-uniformity, size-zone non-uniformity normalized, zone percentage, gray level variance, zone variance, zone entropy, low gray level zone emphasis, high gray level zone emphasis, small area low gray level emphasis, small area high gray level emphasis, large area low gray level emphasis, and large area high gray level emphasis.

Figure 4 – Example of the calculation of a size zone matrix for a 4x4 image with 4 gray levels.

Image

1	2	3	4
1	3	4	4
3	2	2	2
4	1	4	1

=>

Gray Level (i)	Size Zone (j)			
	1	2	3	4
1	2	1	0	0
2	1	0	1	0
3	0	0	1	0
4	2	0	1	0

Source – Thibault et al. (2009)

2.3 Unbalanced data problem

Typically, machine learning methods' performance is evaluated by predictive accuracy. However, if the dataset is unbalanced, this method becomes unappropriated considering a resultant high accuracy if the strategy is to classify all examples as the majority class.

Acknowledging the fact that our dataset is unbalanced (23 malignant vs. 12 benign), we performed analysis based on unbalanced and balanced dataset. The balancing procedure

was done by oversampling the minority class using the synthetic minority over-sampling technique (SMOTE) (CHAWLA et al., 2002). The insertion of synthetic instances in the dataset occurs after: (i) computing the difference among a sample feature vector and its closest neighbor; (ii) given a random number between 0 and 1, multiply this number by the difference calculated previously; and (iii) adding the result to the feature vector. This method effectively turns the minority class' decision region more general (CHAWLA et al., 2002). Furthermore, a significant advantage using SMOTE has been shown in radiomics studies to balance dataset generating synthetic data (EMAMINEJAD et al., 2016). SMOTE is implemented in the data mining and machine learning Weka v3.8.0 application programming interface (API) (WITTEN et al., 2016). At the end, the unbalanced training/validation dataset stays with 21 instances for T1PC and T1PC-p (14 malignant vs 7 benign), 20 instances for T2 and T1PC-T2 (14 malignant vs. 6 benign), while all balanced datasets have 28 (14 malignant vs. 14 benign) for training purposes.

Besides oversampling techniques, another work around for unbalancing problems is by means of re-sampling methods such as Bootstrap (EFRON, 1979). This technique helps to reduce overfitting and variance of the classifier by randomly selecting with replacement *n* samples from the dataset (DUPRET; KODA, 2001; LIU et al., 2013). In our case, we are considering the Bootstrap Aggregating (Bagging) implementation, which aggregates the predictions from the *n* created models by averaging or voting their outputs (BREIMAN, 1996). As we are using Bagging, every instance has same probability of appearing in the subsampled datasets, possibly reducing noise, bias and variance. This method is implemented in Weka as a classifier, and then we select a base learner to create the models.

2.4 Feature selection

Aiming a robust analysis, and in order to increase the chance to avoid overfitting, we performed a relevant feature selection using two different algorithms. In this case, the feature vector can be reduced to the most relevant features only which produced promising results, as shown in (FERREIRA JUNIOR et al., 2018) for lung cancer histopathology and metastases classification.

First, we used the ReliefF algorithm with a ranking search function (KIRA; RENDELL, 1992; KONONENKO, 1994), both implemented in Weka (WITTEN et al., 2016). The ReliefF method works estimating features based on their values to discriminate nearest instances from each other. Instances are chosen randomly then their relevance

(weight) is updated according to the nearest neighbor instances from all classes (KONONENKO, 1994).

Then, features were selected by Wrapper (KOHAVI; JOHN, 1997) with Best First search method, both implemented in Weka (WITTEN et al., 2016) as well. The feature selection algorithm selects the most relevant features based on a specific classifier, and then Best First uses greedy hill climbing augmented algorithm with backtracking to search the feature subsets space.

2.5 Tumor classification

Tumors were classified using methods of machine learning. In artificial intelligence, machine learning is an area intending the development of systems capable of acquiring knowledge automatically, and the development of computational techniques on learning as well (REZENDE, 2003). Classification was evaluated using the leave-one-out (LOO) cross-validation method, which splits the dataset in n folds, where n is the number of instances in the dataset, then $n - 1$ folds are used for training and 1 fold is used for testing. This procedure repeats until all instances are tested, then an average performance is calculated across all folds. For balanced dataset classification, in specific, synthetic data was exclusively used during training, and all testing data represents real data. Every classifier used in this work is implemented in Weka (WITTEN et al., 2016), and they are listed as follows:

2.5.1 Naive Bayes (NB)

Naive Bayes is a probabilistic classifier based on Bayes' theorem and it assumes that numeric features are described by a unique Gaussian distribution, which provides a good estimative of real-world distributions (JOHN; LANGLEY, 1995). Notwithstanding the assumption of predictive features are independent (JOHN; LANGLEY, 1995), some radiomics studies have shown to effectively use non-independent features (EMAMINEJAD et al., 2016; WU et al., 2016). In addition, NB requires less training data to estimate each parameter (WU et al., 2016).

2.5.2 J48 Decision Tree

J48 implements the C4.5 decision tree method (QUINLAN, 1994). It is a classifier that selects the most descriptive features and use them as tree nodes, representing a function which result is used to decide what branch to follow from that node. For the same tree, branch and leaf nodes denote, respectively, the test outcome and classes (TARTAR; KILIC; AKAN, 2013b). Therefore, the method provides more information than just a view of how the machine got the results (QUINLAN, 1994).

2.5.3 Random Forest (RF)

By definition, RF is a method that builds a forest of decision trees randomly generated, which each tree votes for the most popular class (BREIMAN, 2001). RF can work better than other ensemble classifiers because it achieves variance reduction by means of averaging over learners, and its randomized stages decrease the correlation between distinctive learners in the set (LEE; KOUZANI; HU, 2010). In addition, RF classifier is a promising tool for lung nodule detection (TARTAR; KILIC; AKAN, 2013b).

2.5.4 K-nearest Neighbors (KNN)

KNN, known as instance-based algorithm as well, is a method for classifying objects based on training examples closer to it in the attributes' space domain (SARMENTO; DUTRA; ERTHAL, 2012). The KNN's concept involves query by similarity, which returns the *k* most similar objects of a reference object (SOUZA, 2012). This algorithm has been applied in solving image classification problems and it is one of the most effective methods already proposed (SANTOS, 2009). In this work, we used the IBk classifier on Weka (WITTEN et al., 2016), which implements the Instance-based Learning (IBL) variant of the KNN algorithm (AHA; KIBLER; ALBERT, 1991).

2.5.5 Multilayer Perceptron (MLP)

MLP is an artificial neural network composed of simple interconnected nodes arranged in a variety of layers. The first layer represents the input vector, the last one

represents the output vector (classes), and layers between those two are called hidden, which take the input and weights producing output to next layer through activation functions. The connection of nodes are weights and output signals, which suffer influence from inputs to the node that are modified by an activation function (GARDNER; DORLING, 1998). Many computer-aided diagnosis (CAD) systems have incorporated neural networks to distinguish cancerous signs from normal tissues (JIANG; TRUNDLE; REN, 2008).

The classification assessment was done by the area under the receiver operating characteristic curve (AUC), measures of sensitivity (SENS), and specificity (SPEC). Unbalanced (n = 21 for T1PC; n = 21 for T1PC-p; n = 20 for T2; n = 20 for T1PC-T2) and balanced (n = 28 for all cases) dataset classifications were evaluated using the LOO cross-validation method and predictions were done using the test dataset (n = 14). Despite bootstrap was performed using the bagging classifier in Weka, we did not use any cross-validation in this case, but we still made predictions using the test dataset.

Fine parameter tuning was made for all classifiers. In the case of MLP, Table 1, we selected the top 3 MLP performances (MLP-1, MLP-2, and MLP-3) for parameter combination where we combined values of momentum (-m) and learning rate (-l), considering training time with 500 epochs and one hidden layer with (features + classes)/2 nodes for all datasets. Momentum and learning rate varied from 0.01 to 1.00, every 0.10 and 0.05, respectively.

For others classifiers, we had RF generating 100 decision trees, J48 with confidence factor of 0.25 and no pruning process, NB without kernel estimator, and KNN assuming k equals to 1 (KNN-1), 3 (KNN-3), 5 (KNN-5), 7 (KNN-7), and 9 (KNN-9).

Table 1 – Selected MLP parameters for all datasets.

		MLP-1		MLP-2		MLP-3	
		-m	-l	-m	-l	-m	-l
T1PC	Unbalanced	0.80	0.96	0.90	0.61	0.90	0.76
	Balanced	0.70	0.96	0.70	0.81	0.20	0.01
T1PC-p	Unbalanced	0.90	0.76	0.90	0.96	0.90	0.91
	Balanced	0.30	0.71	0.50	0.46	0.60	0.41
T2	Unbalanced	1.00	0.01	0.90	0.36	0.90	0.41
	Balanced	0.20	0.06	0.30	0.06	0.04	0.06
T1PC-T2	Unbalanced	0.90	0.41	0.90	0.36	1.00	0.06
	Balanced	0.70	0.31	0.60	0.51	0.10	0.31

Source - Author

3 RESULTS AND DISCUSSION

Initial results during this work showed low performance using post segmentation correction and selecting features using ReliefF. Thus, the results related to these cases were rejected in the following analyses. T1PC-p dataset had lower performance for every classifier compared to T1PC, T2 and T1PC-T2. In the ReliefF case, results using Wrapper to select relevant features seemed more robust and reliable to execute a second classification using only the selected features since we are using 5 different classifiers and a few different configurations for some of them.

Tables 2 and 3 present the performance of every classifier according to AUC, SENS, and SPEC, for unbalanced and balanced T1PC datasets.

MLP1 had the most uniform performance, showing some of the highest values during validation and test phase considering unbalanced data. The lack of performance regarding SPEC values is noticeable for every classifier for either phase, being 0.571 and 0.600 the highest SPEC values in validation and test, respectively.

In relation to the balanced dataset, KNN-3 had the best performance with the highest values for all three measures during validation. Moreover, it achieved a fair performance on testing showing AUC and SENS above 0.750, but only 0.600 for SPEC. RF was the only classifier with more than one higher value (AUC and SENS) during the test phase, but coming up short classifying benign cases with SPEC values of 0.400.

Table 2 – Classifiers' performance for T1PC unbalanced data. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.571	0.428	**0.571**	0.444	0.889	0.000
J48	0.658	0.642	0.428	0.644	0.889	0.400
RF	0.642	0.857	**0.571**	0.800	0.889	0.200
KNN-1	0.571	0.714	0.428	0.644	0.889	0.400
KNN-3	0.647	0.785	0.285	0.800	0.778	**0.600**
KNN-5	0.505	0.714	0.142	0.711	0.778	0.200
KNN-7	0.561	0.785	0.142	**0.878**	0.889	**0.600**
KNN-9	0.520	0.857	0.000	0.856	0.889	0.200
MLP-1	**0.693**	0.785	**0.571**	0.800	0.889	**0.600**
MLP-2	0.673	0.857	0.000	0.311	**1.000**	0.000
MLP-3	0.612	**1.000**	0.142	0.311	**1.000**	0.000

Source – Author

Table 3 – Classifiers' performance for T1PC balanced data. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.704	0.500	0.857	0.444	**0.889**	0.000
J48	0.811	0.642	0.857	0.689	0.778	0.600
RF	0.765	**0.714**	0.642	**0.833**	**0.889**	0.400
KNN-1	0.750	0.642	0.857	0.644	**0.889**	0.400
KNN-3	**0.836**	**0.714**	**1.000**	0.767	0.778	0.600
KNN-5	0.729	0.571	0.714	0.767	0.778	0.600
KNN-7	0.696	0.571	0.714	0.778	0.667	0.800
KNN-9	0.688	0.500	0.857	0.789	0.667	**1.000**
MLP-1	0.816	0.642	0.785	0.689	**0.889**	0.600
MLP-2	0.811	0.571	0.785	0.756	**0.889**	0.400
MLP-3	0.801	**0.714**	0.785	0.711	**0.889**	0.400

Source – Author

Tables 4 and 5 present the performance of every classifier according to AUC, SENS, and SPEC, for unbalanced and balanced T2 datasets.

Unbalanced T2 data did not obtain good results in the validation phase. The majority of the classifiers had AUC values lower than 0.500 and for SPEC, the values did not even pass that value. Despite not having the highest AUC value, KNN-1 and MLP-2 showed interesting results with the highest SPEC value of 0.750 toward the test phase.

Balancing the T2 dataset, all MLP performances had equally higher values of 0.908 for AUC, 0.714 for SENS e 0.857 for SPEC during validation. However, their performance on testing did not follow the same path, which AUC and SPEC had lower values.

Table 4 – Classifiers' performance for T2 unbalanced data. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.369	0.357	0.333	**0.792**	**1.000**	0.250
J48	0.666	0.714	0.333	0.569	0.889	0.250
RF	0.411	0.857	0.000	0.625	0.778	0.250
KNN-1	0.380	0.428	0.333	0.764	0.778	**0.750**
KNN-3	0.261	0.785	0.000	0.708	**1.000**	0.250
KNN-5	0.523	**1.000**	0.000	0.472	0.889	0.250
KNN-7	0.380	**1.000**	0.000	0.500	**1.000**	0.000
KNN-9	0.261	**1.000**	0.000	0.694	**1.000**	0.000
MLP-1	**0.750**	0.857	**0.500**	0.639	0.778	0.500
MLP-2	0.726	0.714	**0.500**	0.778	0.778	**0.750**
MLP-3	0.714	0.857	0.333	0.750	0.889	0.500

Source – Author

Table 5 – Classifiers' performance for T2 balanced data. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.714	0.571	0.785	0.764	**1.000**	0.000
J48	0.563	0.642	0.571	0.528	0.556	0.500
RF	0.778	0.642	0.785	0.722	0.556	0.500
KNN-1	0.535	0.357	0.714	0.708	0.667	0.750
KNN-3	0.706	0.571	0.785	0.653	0.667	0.500
KNN-5	0.676	0.571	0.785	0.694	0.333	**1.000**
KNN-7	0.653	0.571	0.785	**0.806**	0.444	**1.000**
KNN-9	0.676	0.428	0.785	0.792	0.333	**1.000**
MLP-1	**0.908**	**0.714**	**0.857**	0.694	0.778	0.500
MLP-2	**0.908**	**0.714**	**0.857**	0.667	0.778	0.500
MLP-3	**0.908**	**0.714**	**0.857**	0.667	0.778	0.500

Source – Author

Tables 6 and 7 present the performance of every classifier according to AUC, SENS, and SPEC, for unbalanced and balanced T1PC-T2 datasets.

Considering unbalanced data, MLP-1 had 0.726 as the highest AUC value during validation. It is worth mentioning that the highest SPEC value in this phase is 0.500, which shows a lower performance of all classifiers during validation with this unbalanced dataset. On test phase, the highest AUC value is seen by RF (0.833), but it lacks on performance to classify benign cases with SPEC values of 0.000.

The balanced T1PC-T2 had the highest overall AUC value during validation, which is 0.959 for MLP-1. This dataset presents the highest values of SPEC as well. Surprisingly, KNN-7 had the highest performance on test phase regarding AUC and SPEC, 0.833 and 0.750, respectively. MLP-1 is just right behind, achieving SPEC of 0.750, AUC of 0.722, and

the same 0.778 SENS as KNN-7. NB had the highest SENS of 0.889, but it lacks on SPEC where it achieved 0.000.

Table 6 – Classifiers' performance for T1PC-T2 unbalanced data. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.428	0.500	0.333	0.556	0.889	0.000
J48	0.523	0.642	0.166	0.569	0.889	0.250
RF	0.559	0.928	0.333	**0.833**	0.889	0.000
KNN-1	0.571	0.642	**0.500**	0.764	0.778	0.750
KNN-3	0.440	0.714	0.166	0.736	0.889	0.250
KNN-5	0.464	0.928	0.000	0.819	**1.000**	0.250
KNN-7	0.488	**1.000**	0.000	0.750	**1.000**	0.500
KNN-9	0.494	0.928	0.000	0.806	**1.000**	0.000
MLP-1	**0.726**	0.928	0.333	0.722	0.000	**1.000**
MLP-2	0.690	0.928	0.333	0.806	0.778	0.500
MLP-3	0.589	0.928	0.166	0.500	**1.000**	0.000

Source – Author

Table 7 – Classifiers' performance for T1PC-T2 balanced data. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.750	0.571	0.857	0.444	**0.889**	0.000
J48	0.663	**0.857**	0.714	0.583	0.667	0.500
RF	0.757	0.642	0.785	0.750	0.778	0.500
KNN-1	0.571	0.500	0.642	0.764	0.778	**0.750**
KNN-3	0.732	0.500	0.857	0.750	0.778	0.500
KNN-5	0.744	0.500	0.857	0.778	0.556	**0.750**
KNN-7	0.709	0.500	0.785	**0.833**	0.778	**0.750**
KNN-9	0.719	0.428	0.785	**0.833**	0.667	**0.750**
MLP-1	**0.959**	0.785	**0.928**	0.722	0.778	**0.750**
MLP-2	0.954	0.571	**0.928**	0.722	0.778	0.500
MLP-3	0.948	0.785	**0.928**	0.722	0.778	0.500

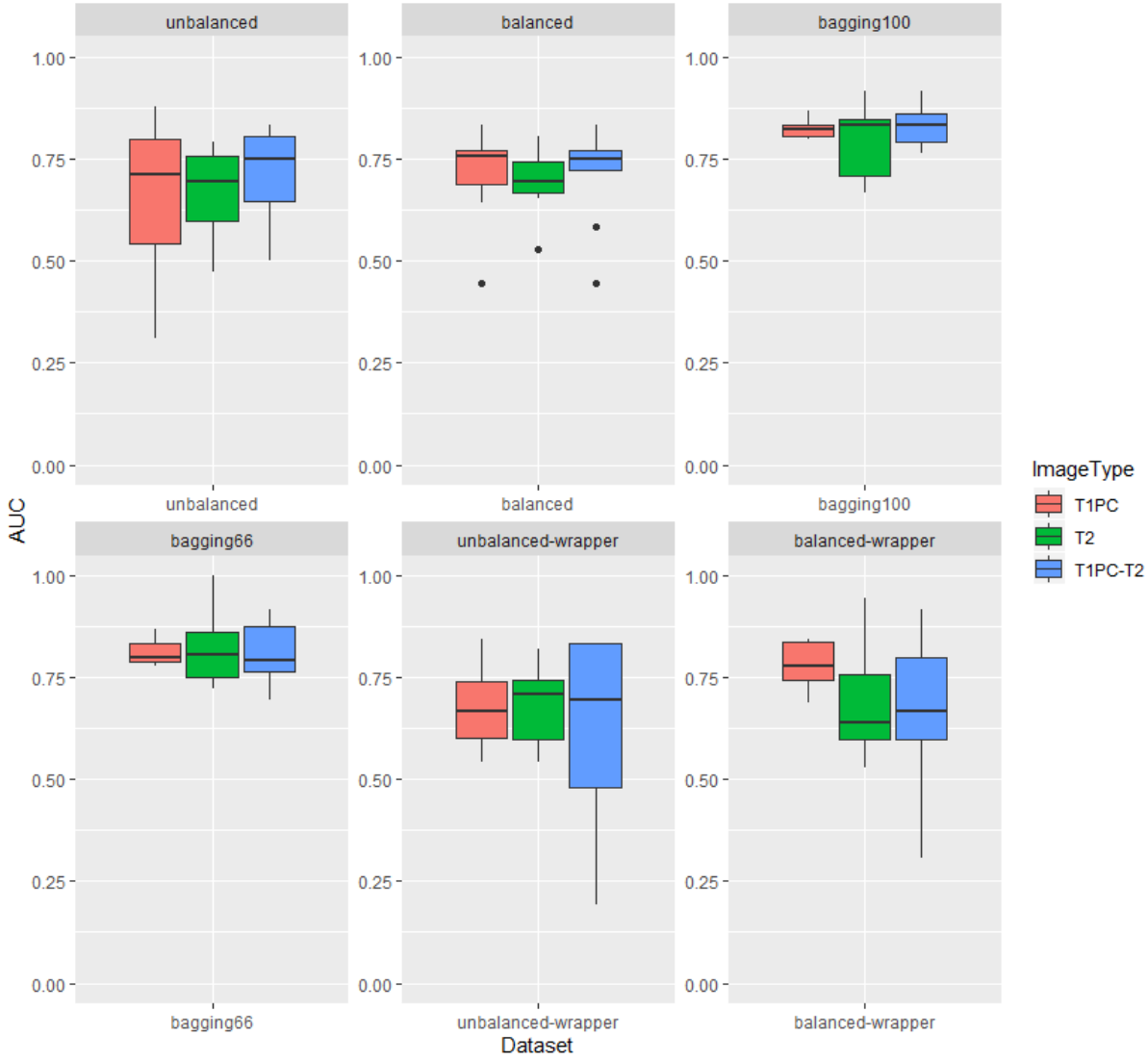
Source – Author

Since the AUC values distribution is unknown, a parametric test such as Student's *t*-test is not feasible to evaluate if our balancing solution improved classification performance. Therefore, the non-parametric Mann-Whitney Two Sample test has been done for every dataset pair (balanced vs. unbalanced) with an alternative hypothesis that balanced AUC values are greater than the unbalanced dataset. The test showed that in balancing our datasets we achieved AUC values statistically greater than unbalanced datasets for the validation phase. The test *p*-values for every dataset pair, for a significance level of 5%, are: 4.629*e*-05 (T1PC); 7.444*e*-03 (T2); and 2.496*e*-04 (T1PC-T2). On the other hand, no statistically improvement was found for AUC values on testing results as performing the same statistical test for a significance level of 5%. The *p*-values obtained in this case are 0.4869 (T1PC); 0.2345 (T2); and 0.5395 (T1PC-T2).

Following, Figures 5-7 are representing an overview of the classification performance using the test dataset according to AUC, SENS, and SPEC, respectively. These figures are composed of results using unbalanced and balanced dataset before and after Wrapper feature selection. Feature selection was done by Wrapper with Best First search method, considering LOO cross-validation for both Wrapper (28 folds) and classifier validation (27 folds). In addition, we selected for each classifier every feature considered at least in one fold. We also included the results from unbalanced datasets classified with bootstrapping technique using the Bagging classifier implemented in Weka. In this last case, we had two different configurations. For both configurations we considered the same 5 classifiers and its configurations as base learner and 100 iterations, but one configuration had bag size of 100% and another using only 66%. In this way, having a situation where bag size is 66%, we assure that at least 34% of the instances were not randomly selected in the iteration, thus iteration tests will evaluate with some instances unknown by the model. Every other table corresponding to classifiers' performance for each dataset can be found in the Appendices B and C at the end.

Figure 5 illustrates AUC values showing higher variance in the unbalanced dataset before and after feature selection. The highest AUC values with low variance is seen using the unbalanced T1PC dataset on Bagging classification with bag size of 100%. However, the highest isolated value was performed by KNN-7 classifying unbalanced T2 cases with 66% bag size bootstrapping.

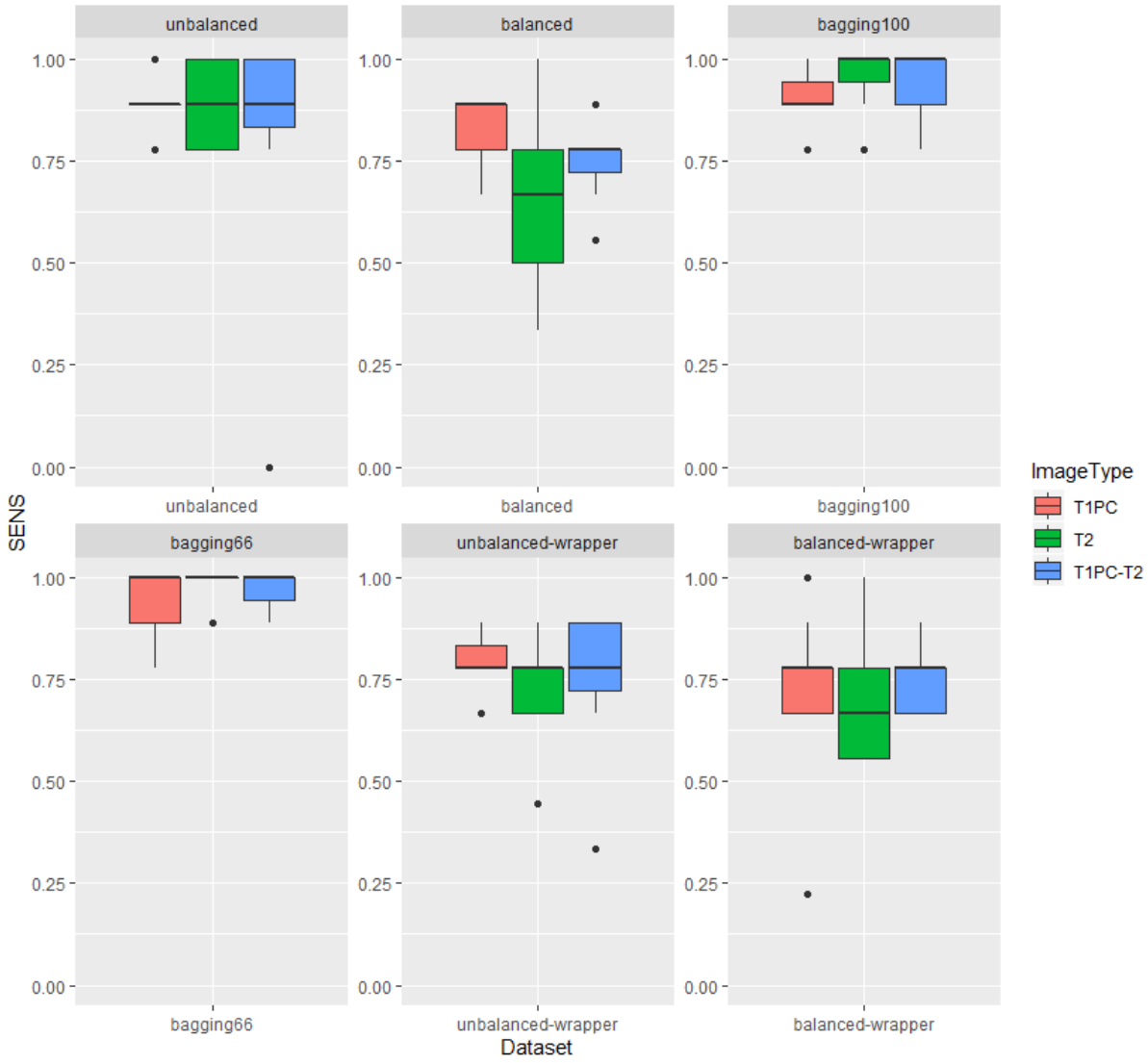
Figure 5 – Boxplot for AUC values of test dataset classification.



Source – Author

Regarding SENS (Figure 6), the highest values can be found mostly in unbalanced dataset cases, which indicates the classifiers prioritizing the majority class. This phenomenon is more expressed considering the Bagging classifier and it can be explained by its own bootstrapping method, which may create a subsample randomly picking a large number of malignant cases in many iterations.

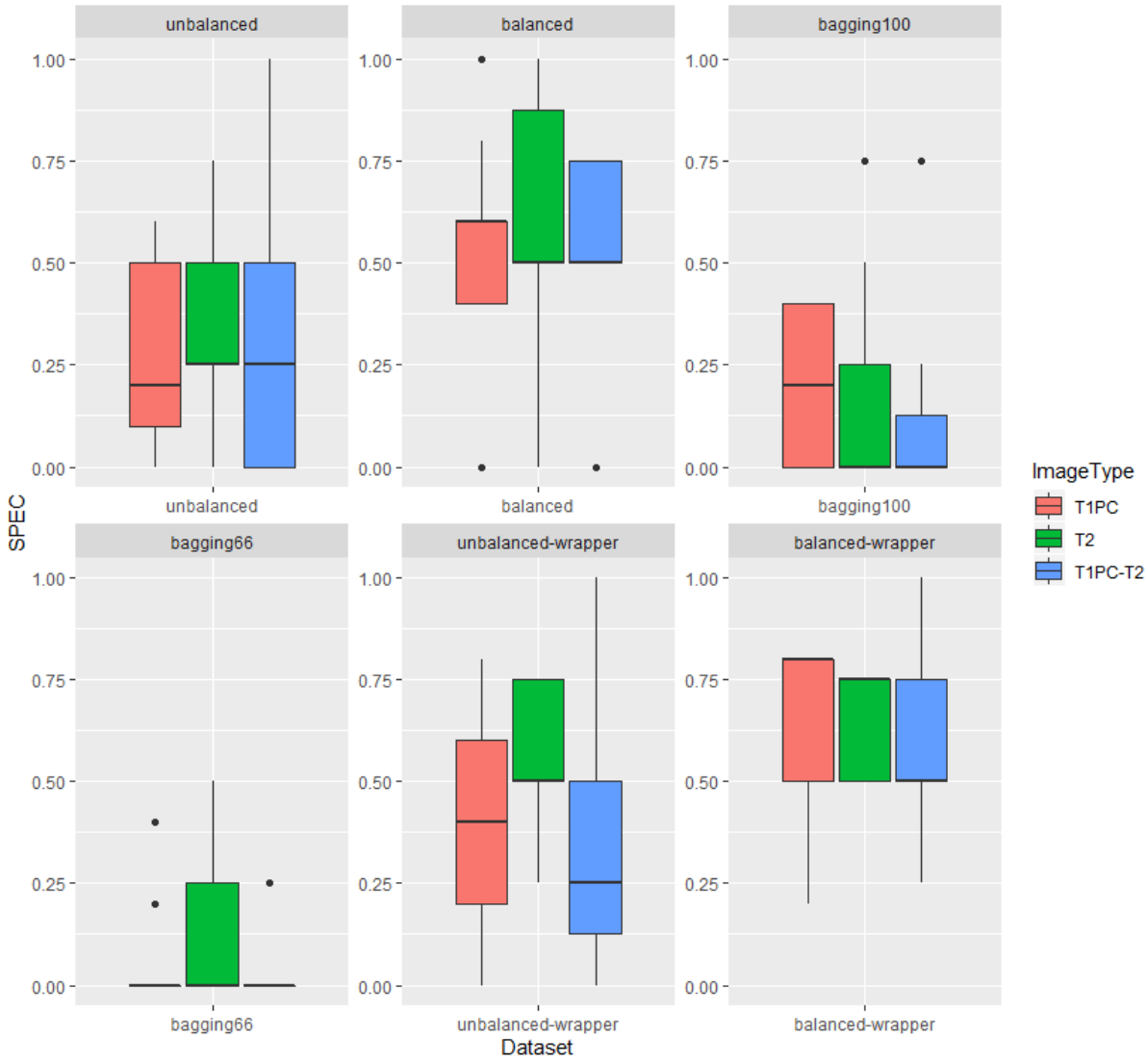
Figure 6 – Boxplot for SENS values of test dataset classification.



Source – Author

When it comes to SPEC performance (Figure 7), the bootstrap approach using Bagging showed the lowest values, especially the 66% bag size configuration. In general, most classifiers had a poor performance as regards this metric. We can see a few cases reaching values as high as 1.000, and these cases are the most consistent classifiers considering AUC and SENS as well, showing fair/good results. Thus, we can notice the importance of balancing data for machine learning and how it might affect your results.

Figure 7 – Boxplot for SPEC values of test dataset classification.



Source – Author

Considering now the three metrics (AUC, SENS and SPEC), the datasets having the highest and consistent results (assuming consistency as the fact of achieving more values above 0.500) are the balanced datasets after Wrapper feature selection. In this case, we should highlight the most notable performances. For balanced T1PC with feature selection, MLP-3 showed good results on test phase with AUC = 0.844, SENS = 0.778, and SPEC = 0.800. Balanced T2 data after feature selection had impressive results using NB for both validation (AUC = 0.929; SENS = 0.929; SPEC = 0.786) and testing (AUC = 0.944; SENS = 1.000; SPEC = 0.750). Regarding T1PC-T2 balanced data with feature selection, testing results showed MLP-1 (AUC = 0.778; SENS = 0.778; SPEC = 0.750), MLP-3 (AUC = 0.889; SENS = 0.889; SPEC = 0.750) and KNN-7 (AUC = 0.917; SENS = 0.778; SPEC = 1.000) having the highest performances. These three classifiers had high performances during validation as

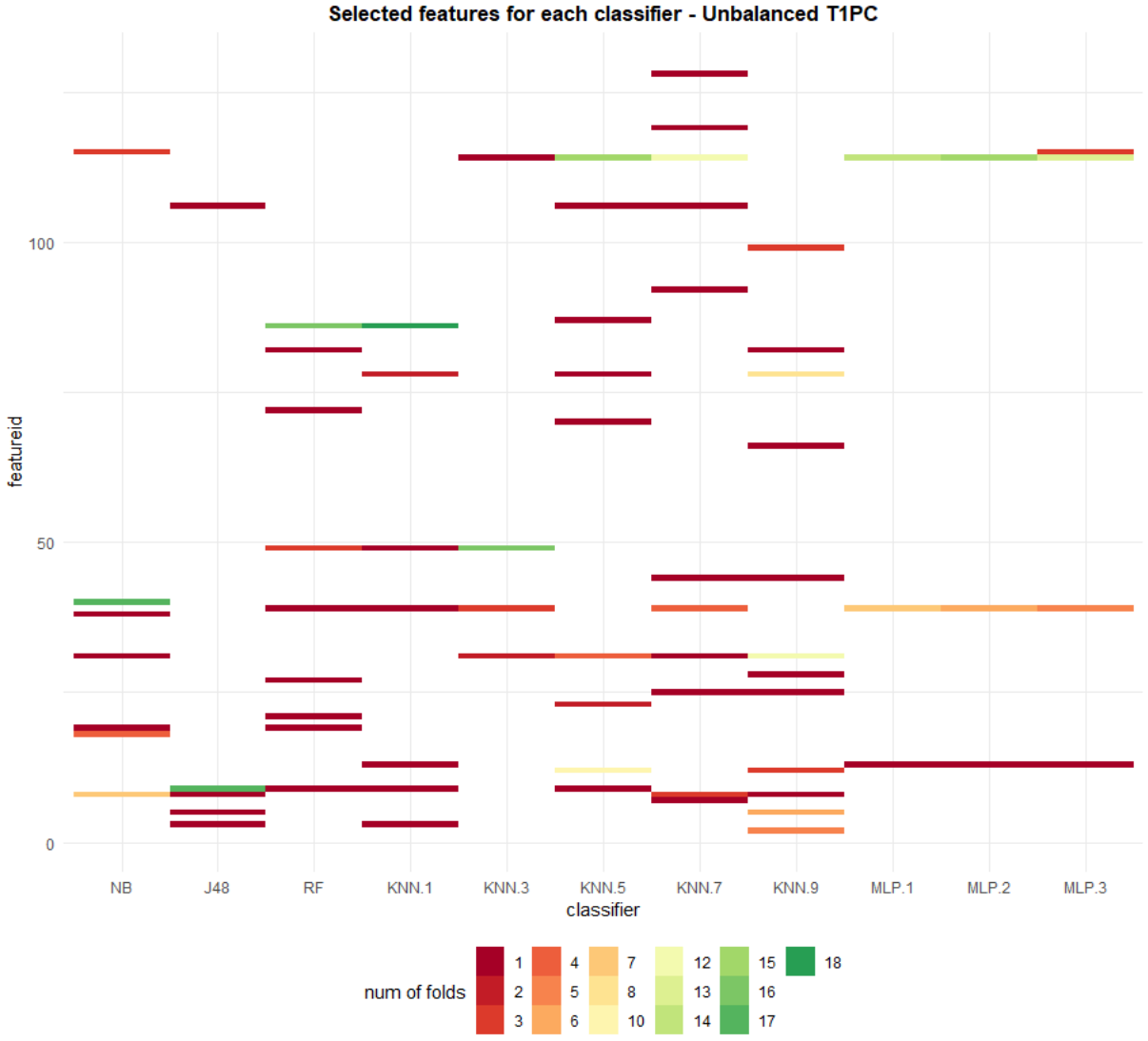
well, they achieved AUC, SENS and SPEC values above 0.850, except for KNN-7 value of 0.643 for SENS.

It is worth mentioning the KNN-1 performance on Bagging cases. This classifier achieved the value of 1.000 for AUC, SENS and SPEC in almost every validation case. Its lower performance was classifying T2 cases with bag size of 66%, showing AUC of 0.952, SENS of 1.000, and SPEC of 0.500. During the test phase, KNN-1 seemed the most consistent classifier considering all three metrics. Classifying the test dataset of T1PC-T2 using Bagging with bag size of 100%, KNN-1 had AUC of 0.861. SENS of 0.778, and SPEC of 0.750.

Returning to the subject of feature selection, Figures 5-10 represent an overview of the selected features and how many times they were selected during Wrapper's iterations for each classifier. In Appendix C contains the tables for each case of feature selection classification.

Considering the unbalanced T1PC dataset, the most selected features by Wrapper are GLCM texture features. Informal Measure of Correlation 2 from both 2D and 3D models, and Correlation (2D) were selected by 7, 6 and 5 different classifiers, respectively. Assuming a threshold of 70% of the possible folds in the selection process: IMC 1 (2D) is present in NB selected feature vector; GLN (2D) is present for KNN-3 case; VOL (2D) is seen in J48 selected feature vector; MJA (3D) for both RF and KNN-1; and IMC 2 (3D) is present both MLP-2 and KNN-5 cases. The classifier that obtained the highest number of selected features was KNN-9, with 12 different features. On the other hand, MLP-1 and MLP-2 had only 3 selected features.

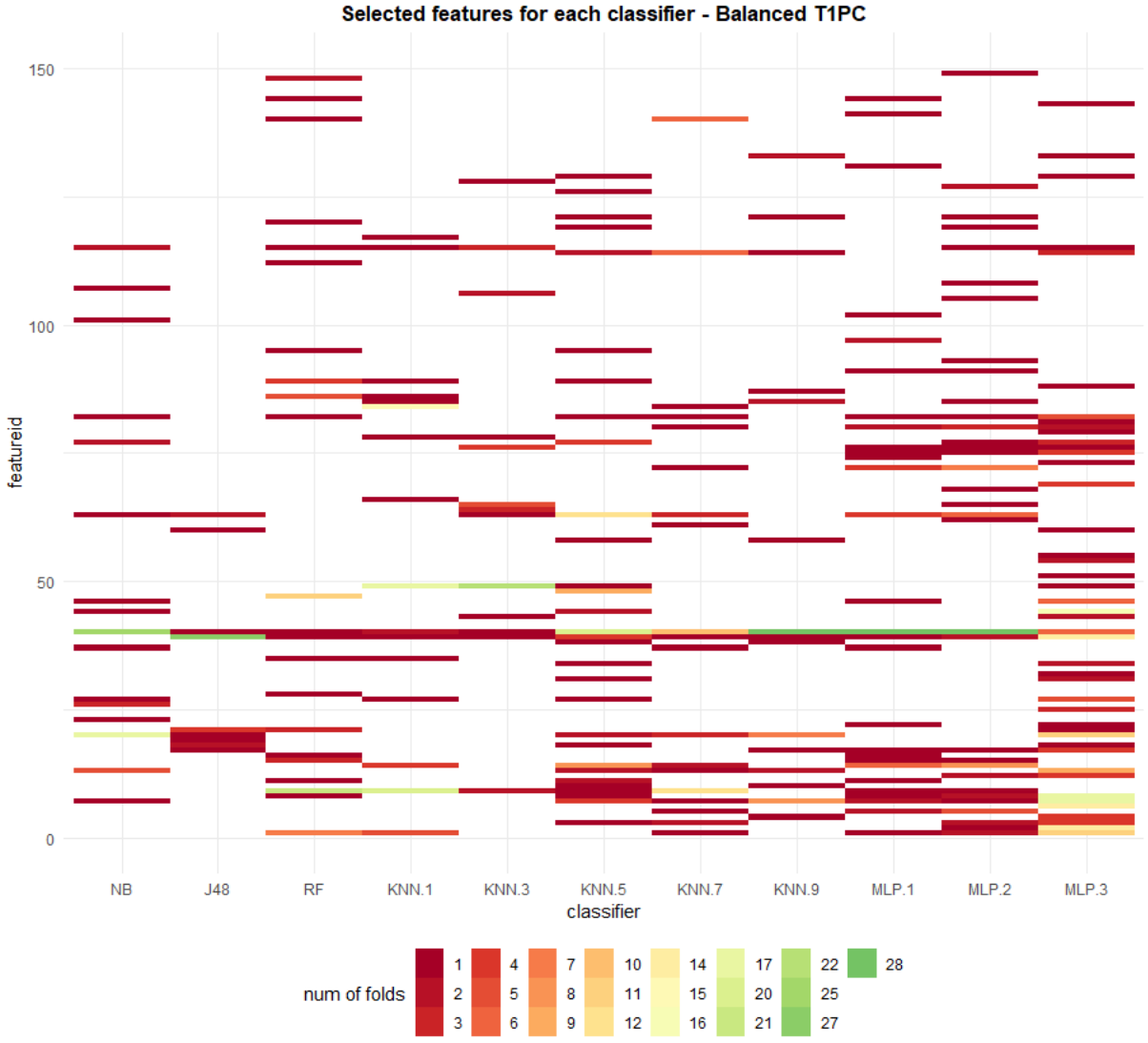
Figure 8 – Heatmap for unbalanced T1PC feature selection for each classifier according to the number of folds the feature is picked as relevant by Wrapper.



Source – Author

The balanced T1PC had predominance of texture features as well as regards frequency of selection by all classifier using Wrapper. IMC 1 (2D) was selected by all classifiers, and IMC 2 (2D) was selected by 10. We should highlight the selection of shape features as well, which ELG (2D) and VOL (2D) are present in selected feature vectors of 7 different classifiers. Considering the 70% threshold again, Wrapper for KNN-9, MLP-1 and MLP-2 selected IMC 1 (2D) in all 28 folds. The same feature was selected in 20 folds for KNN-5 and in 25 folds for NB. IMC 2 (2D) seemed relevant for Wrapper in 27 folds of J48 feature selection, and VOL (2D) was selected in more than 70% of the folds for RF and KNN-1 selection. MLP- 3 has the highest number of selected features being 44, oppositely J48 has 9 being the smaller selected feature vector in this case.

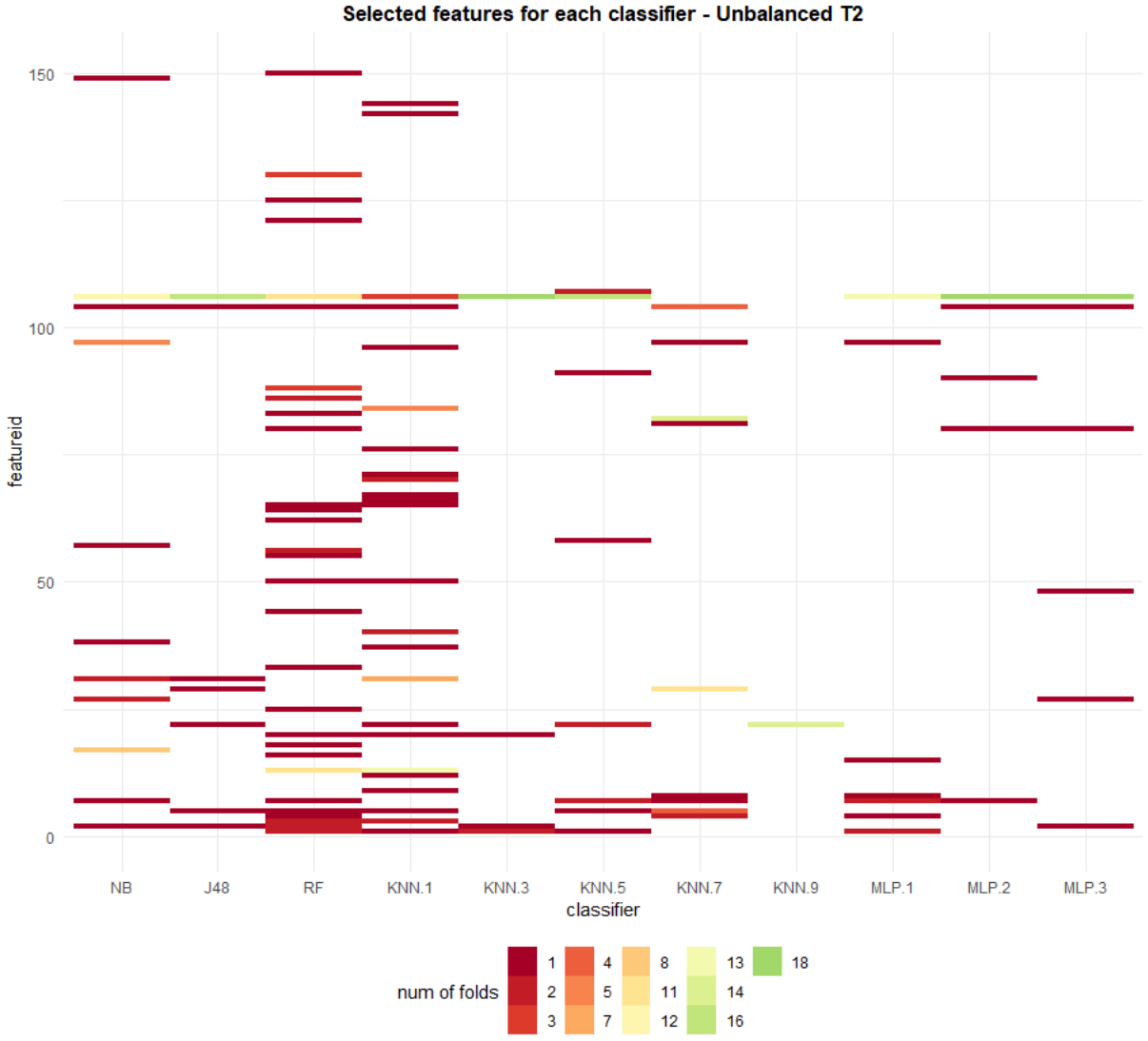
Figure 9 – Heatmap for balanced T1PC feature selection for each classifier according to the number of folds the feature is picked as relevant by Wrapper.



Source – Author

As concerns unbalanced T2 data, Wrapper showed COR (3D) and IDN (3D) being the two most relevant features across all 11 classifiers, which the first one was selected by 9 classifiers and the later one 7. Another fact in relation to COR (3D) in this case, is its predominance in number of folds it was selected regarding the 70% threshold. For KNN-3, MLP-2 and MLP-3, Wrapper selected it 18 out of 21 folds. It is worth mentioning that the same feature is present in 16 folds for J48 and KNN-5 Wrapper feature selection as well. About selected feature vector size, RF has 29 selected features as highest number against KNN-9 with only 1 feature.

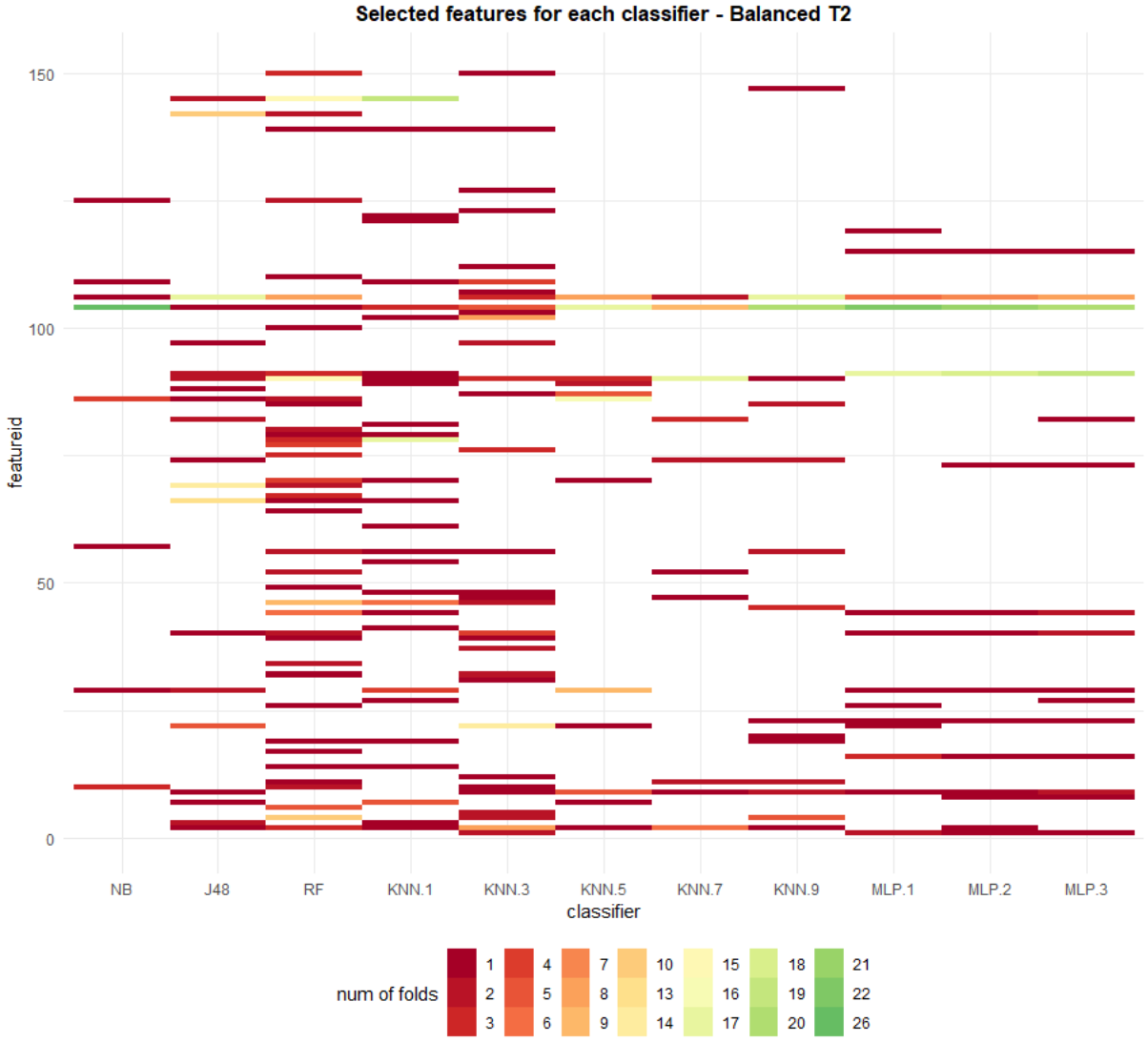
Figure 10 – Heatmap for unbalanced T2 feature selection for each classifier according to the number of folds the feature is picked as relevant by Wrapper.



Source – Author

In turn, balanced T2 data had feature selection resulting in IDN (3D) and COR (3D) as the two most selected features like unbalanced data. However, in contrast to unbalanced data, IDN (3D) was selected at least in one fold for all 11 classifiers, while COR (3D) appears in 10 different feature vectors after selection. For this dataset, IDN (3D) had the same predominance that COR (3D) had with unbalanced data in relation to the 70% of folds. The Inverse Difference Normalized feature was selected in 26, 22, 21, and 20 folds for NB, MLP-1, MLP-2, KNN-9 and MLP-3, respectively, where both of these last two classifiers selected it in 20 folds. As we analyse the final feature vector size, RF has size of 41 features, while NB has the smallest vector with 8 features.

Figure 11 – Heatmap for balanced T2 feature selection for each classifier according to the number of folds the feature is picked as relevant by Wrapper.



Source – Author

The unbalanced T1PC-T2 results after Wrapper feature selection seems to be affected by the combination of datasets, as expected. The top 2 selected features by Wrapper for all classifiers have FeatureID 256 and 39, which means the COR (3D) of the T2 portion and the IMC 2 (2D) from T1PC portion were selected by 9 and 6 classifiers, respectively. Taking into account the features selected in more than 70% of the folds, COR (3D, Feature ID 256) was preferred in 15 folds of KNN-5 feature selection, while IMC 1 (3D, FeatureID 115) for NB, IMC 2 (2D, FeatureID 39) for MLP-2, and VOL (2D, FeatureID 9) for J48 were preferred in 15 folds during Wrapper for their respective classifier. The final feature vectors' size has RF being the larger with 15 features, and KNN-9 the smaller with 1 feature.

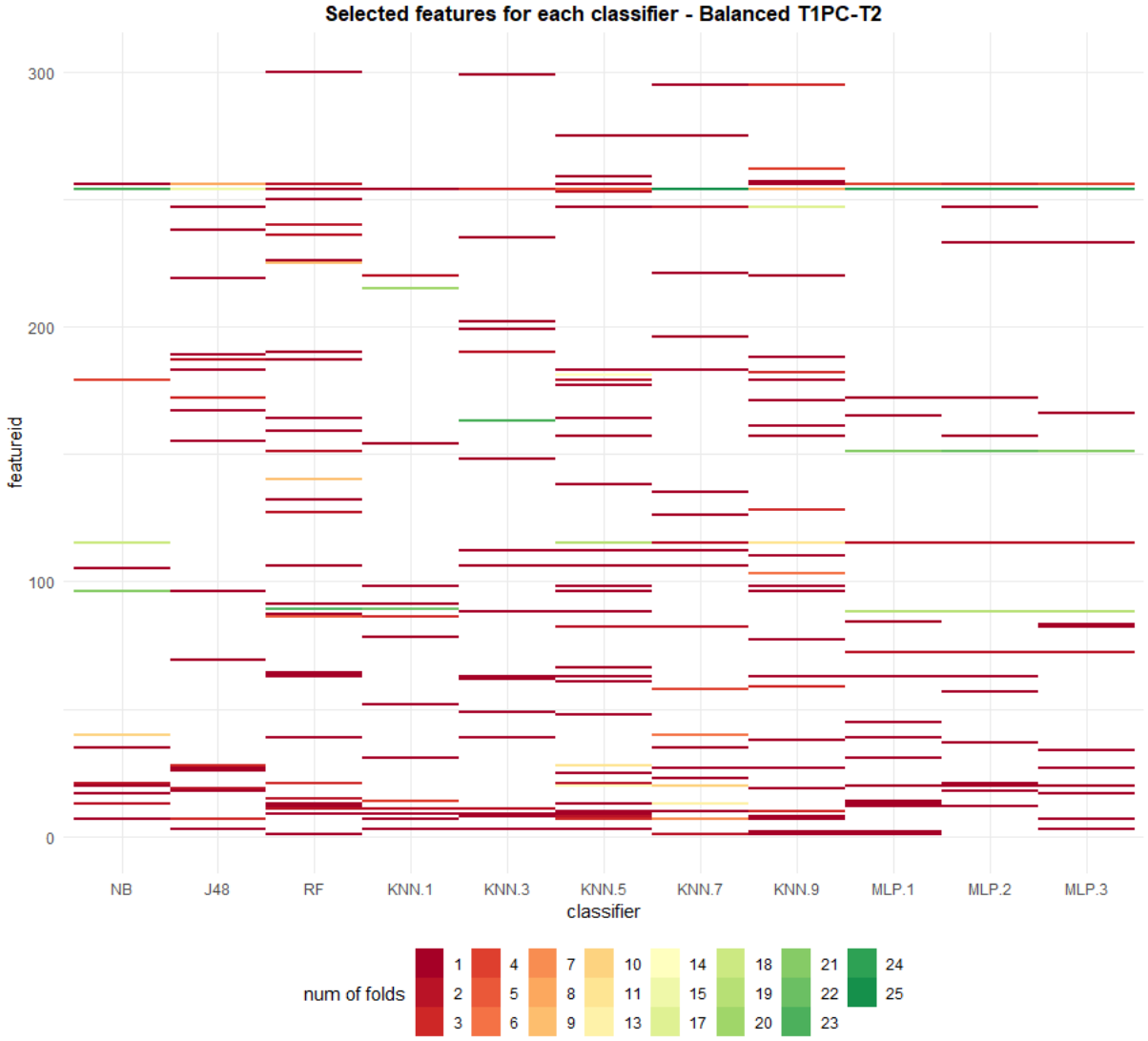
Figure 12 – Heatmap for unbalanced T1PC-T2 feature selection for each classifier according to the number of folds the feature is picked as relevant by Wrapper.

Source – Author

In the matter of feature selection from balanced T1PC-T2 data, we can see a similar aspect analysing the influence of having T1PC and T2 data at the same time during classification. The two most selected features across all classifiers are related to T2 portion of the dataset. The texture features IDN (3D, FeatureID 254) and COR (3D, FeatureID 256) were selected by Wrapper for 11 and 8 classifiers, respectively. The next features in the list are mostly from the portion corresponding to T1PC, such as ELG (2D, FeatureID 7), which was selected for 7 classifiers. The 70% threshold analysis for this case shows all MLP configurations having both IDN (3D, FeatureID 254) and M3DD (2D, FeatureID 151) being selected in more than 20 folds. In addition, IDN (3D, FeatureID 254) was more prevalent in the KNN-7 feature selection, which as picked in 25 out of 28 folds. NB also had a more than

20 folds selecting IDN (3D, FeatureID 254) and the MP (3D, FeatureID 96). Surface Area (3D, FeatureID 89) was a common feature selected between RF (23 folds) and KNN-1 (22 folds). Taking into consideration the final feature vector for theses cases, KNN-5 has the larger vector with 34 features and in the opposite side there is NB with 13 features.

Figure 13 – Heatmap for balanced T1PC-T2 feature selection for each classifier according to the number of folds the feature is picked as relevant by Wrapper.



Source – Author

Thus, after an overview on selected features, it seems interesting to correlate the most relevant features with clinical information. In general, the more aggressive the tumor, the more heterogeneous it will be, and the higher chance of necrosis and hemorrhage. In these situations, the predominance of GLCM features such as IMC1, IMC2 and COR can indicate the machine capacity to differentiate tumor tissue. On the other hand, shape features like

M3DD, MP, ELG and SA, which seemed relevant for balanced T1PC-T2, could indicate potential in mathematically representing tumor as regards its form, if it is more irregular, spiculated and infiltrative.

Another interesting point to take in consideration is the fact that selecting features for unbalanced T1PC data, which has good spatial resolution, we can see VOL and MJA as relevant shape features. These features are directly related to tumor classification when comes to medical visual analysis.

Since no similar work using MRI was found in the literature, a comparison with recent similar works using CT seems feasible. Ferreira Junior et al. (2018) uses radiomics-based features for pattern recognition of lung cancer histopathology and metastases. So we can make a parallel, we are comparing only his results regarding histopathological classification using computer features. It is worth noting that what Ferreira Junior et al. (2018) calls as testing, corresponds to our validation step, and he performed a LOO on a dataset with 52 instances. His validation process corresponds to our testing step, where he used a dataset with 16 instances to evaluate the model created by LOO with the highest performance. Then, we will adopt our nomenclature for this steps to avoid any misinterpretation. Thus, NB had the highest performances in his study showing AUC values of 0.810 for both unbalanced and oversampled (using SMOTE) datasets during test. Our MLP-2 for T1PC-T2 unbalanced dataset gets close to his results with AUC of 0.806. Now, comparing balanced data results, KNN-7 for T1PC-T2 dataset surpasses his NB results with AUC of 0.833.

We can go further in the comparison if we take in consideration our results using Bagging and feature selection. In these cases, the difference of performance increases in our favor. For example, NB in our case using T2 balanced data after feature selection showed AUC = 0.944 during test. Another example is KNN-1 as base learner of Bagging with bag size of 100% to classify T1PC-T2 images, which test results showed AUC = 0.861.

We should highlight some limitations of this work. First, the datasets' size is still small and led us to choose two different approaches to overcome this problem, the leave-one-out cross-validation and bootstrapping method. Another limitation is the lack of biopsy confirmation for some malignant cases. In future works, we suggest the use of larger datasets, the comparison between methods using CT imaging, plus the addition of malignancy confirmation from histopathological exams of biopsies.

4 CONCLUSION

In this study, we assessed quantitative features for lung tumor classification. Lesions were semi-automatically segmented in T1-weighted contrast-enhanced and T2-weighted MR images. Different classes of features, such as shape and texture, were extracted from 2D and 3D models of tumor images. Five of the most common machine learning classifiers, NB, J48, RF, KNN, and MLP were used to classify unbalanced and balanced datasets. Post-segmentation correction protocol was performed on T1PC images. Combination of T1PC and T2 feature vectors was done as well. Four different datasets were evaluated: T1PC (150 features), T1PC-p (150 features), T2 (150 features), and T1PC-T2 (300 features). Datasets were validated by leave-one-out cross-validation. The unbalancing problem was solved using the SMOTE filter in Weka. We also performed analysis by bootstrapping our unbalanced datasets.

Post-segmentation correction did not seem to improve classification, instead it showed less performance in comparison to the original segmentation. The combination of T1PC and T2 feature vectors presented interesting results with classification improvements. Prior feature selection, T1PC-T2 dataset presented the most consistent classification results in general. For this case, we should highlight the performance of KNN-7 and MLP-1. Even bootstrapping our data with Bagging did not improve our results to surpass the results of T1PC-T2 balanced dataset. Bagging results had great performance during validation, with KNN-1 being the top performer, but most classifiers lack the performance to predict the test dataset.

Relevant feature selection was done with ReliefF and Wrapper. The first method, ReliefF, showed low performance and high variation of selected features, thus we decided not to take the results into consideration. On the other hand, Wrapper seemed robust to our needs and delivered interesting results. As part of the classification with selected features, the highest AUC value predicting test data was performed by NB classifying balanced T2 (validation AUC = 0.929; SENS = 0.929; SPEC = 0.786; and testing AUC = 0.944; SENS = 1.000, SPEC = 0.750). In terms of validation phase, balanced T1PC-T2 data had the highest performances, for example MLP-3 (validation AUC = 0.995; SENS = 0.929; SPEC = 0.929; testing AUC = 0.889; SENS = 0.889; SPEC = 0.750).

In relation to most selected features, Wrapper selected features leading to some clinical correlation. The texture features selected are mostly extracted from gray level co-occurrence that might be related to pathological indicators such as necrosis and hemorrhage, which cause anomalies in tumor tissue texture. Shape features, such as volume, surface area,

and major axis were also picked as relevant by Wrapper, and they are directly correlated to malignancy of lung tumors.

After all, pattern recognition on MRI to aid lung cancer characterization seems feasible. We propose for future researches the enhancement of the semi-automatically segmentation, increase of both benign and malignant lung tumor cases, biopsy confirmation for every malignant case, the association of clinical features to improve classification performance, and the comparison of performances between methods using CT imaging.

REFERENCES

AERTS, H. J. W. L.; VELAZQUEZ, E. R.; LEIJENAAR, R. T. H.; PARMAR, C.; GROSSMANN, P.; CAVALHO, S.; BUSSINK, J.; MONSHOUWER, R.; HAIBE-KAINS, B.; RIETVELD, D.; HOEBERS, F.; RIETBERGEN, M. M.; LEEMANS, C. R.; DEKKER, A.; QUACKENBUSH, J.; GILLIES, R. J.; LAMBIN, P. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. **Nature Communications**, v. 5, 2014. DOI: 1038/ncomms5006.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. **Machine Learning**, v. 6, n. 1, p. 37–66, 1991.

AUSTIN, J. H. M.; GARG, K.; ABERLE, D.; YANKELEVITZ, D.; KURIYAMA, K.; LEE, H.-J.; BRAMBILLA, E.; TRAVIS, W. D. Radiologic implications of the 2011 classification of adenocarcinoma of the lung. **Radiology**, v. 266, n. 1, p. 62–71, 2013.

BARTHOLMAI, B. J.; KOO, C. W.; JOHNSON, G. B.; WHITE, D. B.; RAGHUNATH, S. M.; RAJAGOPALAN, S.; MOYNAGH, M. R.; LINDELL, R. M.; HARTMAN, T. E. Pulmonary nodule characterization, including computer analysis and quantitative features. **Journal of Thoracic Imaging**, v. 30, n. 2, p. 139–156, 2015.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 140, p. 123–140, 1996.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BURAK AKGÜL, C.; RUBIN, D. L.; NAPEL, S.; BEAULIEU, C. F.; GREENSPAN, H.; ACAR, B. Content-based image retrieval in radiology: current status and future directions. **Journal of Digital Imaging**, v. 24, n. 2, p. 208–222, 2011.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002.

CHU, A.; SEHGAL, C. M.; GREENLEAF, J. F. Use of gray value distribution of run lengths for texture analysis. **Pattern Recognition Letters**, v. 11, n. 6, p. 415–419, 1990.

COOLEN, J.; VANSTEENKISTE, J.; DE KEYZER, F.; DECALUWÉ, H.; DE WEVER, W.; DEROOSE, C.; DOOMS, C.; VERBEKEN, E.; DE LEYN, P.; VANDECAVEYE, V.; VAN RAEMDONCK, D.; NACKAERTS, K.; DYMARKOWSKI, S.; VERSCHAKELEN, J. Characterisation of solitary pulmonary lesions combining visual perfusion and quantitative diffusion MR imaging. **European Radiology**, v. 24, n. 2, p. 531–541, 2014.

DETTERBECK, F. C.; BOFFA, D. J.; TANOUE, L. T. The New lung cancer staging system. **Chest**, v. 136, n. 1, p. 260–271, 2009.

DUPRET, G.; KODA, M. Bootstrap re-sampling for unbalanced data in supervised learning. **European Journal of Operational Research**, v. 134, n. 1, p. 141–156, 2001.

DUTTA, P. R.; MAITY, A. Cellular responses to EGFR inhibitors and their relevance to cancer therapy. **Cancer Letters**, v. 254, n. 2, p. 165–177, 2007.

EFRON, B. Bootstrap methods: another look at the jackknife. **Annals of Statistics**, v. 7, n. 1, p. 1-26, 1979.

EL NAQA, I.; GRIGSBY, P. W.; APTE, A.; KIDD, E.; DONNELLY, E.; KHULLAR, D.; CHAUDHARI, S.; YANG, D.; SCHMITT, M.; LAFOREST, R.; THORSTAD, W. L.; DEASY, J. O. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. **Pattern Recognition**, v. 42, n. 6, p. 1162–1171, 2009.

EMAMINEJAD, N.; QIAN, W.; GUAN, Y.; TAN, M.; QIU, Y.; LIU, H.; ZHENG, B. Fusion of quantitative image and genomic biomarkers to improve prognosis assessment of early stage lung cancer patients. **IEEE Transactions on Biomedical Engineering**, v. 63, n. 5, p. 1034–1043, 2016.

FEDOROV, A.; BEICHEL, R.; KALPATHY-CRAMER, J.; FINET, J.; FILLION-ROBIN, J. C.; PUJOL, S.; BAUER, C.; JENNINGS, D.; FENNESSY, F.; SONKA, M.; BUATTI, J.; AYLWARD, S.; MILLER, J. V.; PIEPER, S.; KIKINIS, R. 3D Slicer as an image computing platform for the quantitative imaging network. **Magnetic Resonance Imaging**, v. 30, n. 9, p. 1323–1341, 2012. DOI: https://dx.doi.org/10.1016/j.mri.2012.05.001.

FERREIRA, J. R.; OLIVEIRA, M. C.; MARQUES, P. M. A. Characterization of pulmonary nodules based on features of margin sharpness and texture. **Journal of Digital Imaging**, v.31, n.4, p.451-463, 2018.

FERREIRA JUNIOR, J. R.; KOENIGKAM-SANTOS, M.; CIPRIANO, F. E. G.; FABRO, A. T.; MARQUES, P. M. A. Radiomics-based features for pattern recognition of lung cancer histopathology and metastases. **Computer Methods and Programs in Biomedicine**, v. 159, p. 23–30, 2018.

GALLOWAY, M. M. Texture analysis using gray level run lengths. **Computer Graphics and Image Processing**, v. 4, n. 2, p. 172–179, 1975.

GARDNER, M. W.; DORLING, S. R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. **Atmospheric Environment**, v. 32, n. 14–15, p. 2627–2636, 1998.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. Textural features for image classification**. IEEE Transactions on Systems, Man and Cybernetics**, v. SMC-3, n. 6, p. 610-621, 1973.

HOLLINGS, N.; SHAW, P. Diagnostic imaging of lung cancer. **European Respiratory Journal**, v. 19, n. 4, p. 722–742, 2002.

JIANG, J.; TRUNDLE, P.; REN, J. Medical imaging analysis with artificial neural networks. **Psychotherapy Research**, v. 22, p. 1753–1759, 2008.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, 11, 1995. Proceedings… [S.l.:s.n.]. p. 338–345.

KIRA, K.; RENDELL, L. A. A Practical approach to feature selection. **Machine Learning Proceedings 1992**, p. 249–256, 1992.

KOENIGKAM-SANTOS, M.; OPTAZAITE, E.; SOMMER, G.; SAFI, S.; HEUSSEL, C. P.; KAUCZOR, H. U.; PUDERBACH, M. Contrast-enhanced magnetic resonance imaging of pulmonary lesions: description of a technique aiming clinical practice. **European Journal of Radiology**, v. 84, n. 1, p. 185–192, 2015. DOI: http://dx.doi.org/10.1016/j.ejrad.2014.10.007.

KOENIGKAM SANTOS, M.; MULEY, T.; WARTH, A.; DE PAULA, W. D.; LEDERLIN, M.; SCHNABEL, P. A.; SCHLEMMER, H. P.; KAUCZOR, H. U.; HEUSSEL, C. P.; PUDERBACH, M. Morphological computed tomography features of surgically resectable pulmonary squamous cell carcinomas: Impact on prognosis and comparison with adenocarcinomas. **European Journal of Radiology**, v. 83, n. 7, p. 1275–1281, 2014.

KOHAVI, R.; JOHN, G.H. Wrappers for feature subset selection. **Artificial Intelligence**, v. 97, n. 1/2, p. 273–324, 1997.

KONONENKO, I. Estimating attributes: analysis and extensions of RELIEF. **European Conference on Machine Learning**, p. 171–182, 1994.

LEE, S. L. A.; KOUZANI, A. Z.; HU, E. J. Random forest based lung nodule classification aided by clustering. **Computerized Medical Imaging and Graphics**, v. 34, n. 7, p. 535–542, 2010.

LIU, K. Y.; SMITH, M. R.; FEAR, E. C.; RANGAYYAN, R. M. Biomedical signal processing and control evaluation and amelioration of computer-aided diagnosis with artificial neural networks utilizing small-sized sample sets. **Biomedical Signal Processing and Control**, v. 8, n. 3, p. 255–262, 2013. DOI: http://dx.doi.org/10.1016/j.bspc.2012.11.001.

LIU, H.; LIU, Y.; YU, T.; YE, N.; WANG, Q. Evaluation of apparent diffusion coefficient associated with pathological grade of lung carcinoma, before therapy. **Journal of Magnetic Resonance Imaging**, v. 42, n. 3, p. 595–601, 2015.

LORENSEN, W. E.; CLINE, H. E. Marching cubes: a high resolution 3D surface construction algorithm. ACM SIGGRAPH Computer Graphics, v. 21, n. 4, p. 163–169, 1987.

MAHMOUD-GHONEIM, D.; TOUSSAINT, G.; CONSTANS, J. M.; DE CERTAINES, J. D. Three dimensional texture analysis in MRI: a preliminary evaluation in gliomas. **Magnetic Resonance Imaging**, v. 21, n. 9, p. 983–987, 2003.

MARVASTI, N. B. Clinical experience sharing by similar case retrieval categories and subject descriptors. In: ACM INTERNATIONAL WORKSHOP ON MULTIMEDIA INDEXING AND INFORMATION RETRIEVAL FOR HEALTHCARE, 2013, Barcelona. **Proceedings…** New York: ACM, 2013. p. 67–74.

NOVAES, F. T.; CATANEO, D. C.; LOPES, R.; JUNIOR, R.; DEFAVERI, J.; MICHELIN, O. C.; JOSÉ, A.; CATANEO, M. Câncer de pulmão: histologia, estádio, tratamento e sobrevida. **Jornal Brasileiro de Pneumologia**, v. 34, n. 8, p. 595–600, 2006.

NYÚL, L. G.; UDUPA, J. K.; ZHANG, X. New variants of a method of MRI scale standardization. **IEEE Transactions on Medical Imaging**, v. 19, n. 2, p. 143–150, 2000.

OHNO, Y.; KOYAMA, H.; YOSHIKAWA, T.; MATSUMOTO, K.; AOYAMA, N.; ONISHI, Y.; SUGIMURA, K. Diffusion-weighted MRI versus 18F-FDG PET/CT: Performance as predictors of tumor treatment response and patient survival in patients with non-small cell lung cancer receiving chemoradiotherapy. **American Journal of Roentgenology**, v. 198, n. 1, p. 75–82, 2012.

OLIVEIRA, M. C.; CIRNE, W.; MARQUES, P. M. A. Towards applying content-based image retrieval in the clinical routine. **Future Generation Computer Systems**, v. 23, n. 3, p. 466–474, 2007.

QUINLAN, J. R. **Book review** : C4 . 5 : programs for machine learning. San Mateo: California: Morgan Kaufmann, 1994. v. 240.

REZENDE, S. O. **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Manole, 2003.

SAKAO, Y.; OKUMURA, S.; MUN, M.; UEHARA, H.; ISHIKAWA, Y.; NAKAGAWA, K. Prognostic heterogeneity in multilevel N2 non-small cell lung cancer patients: importance of lymphadenopathy and occult intrapulmonary metastases. **Annals of Thoracic Surgery**, v. 89, n. 4, p. 1060–1063, 2010.

SANTOS, F. C. **Variações do método kNN e suas aplicações na classificação automática de textos**. 2009. 96p. Dissertação (Mestrado) - Universidade Federal de Goiás, Goiás, 2010.

SARMENTO, P. L.; DUTRA, L. V; ERTHAL, G. J. Redução do conjunto de dados de treinamento para melhorar a eficiência do classificador SVM. In: WORKSHOP DE COMPUTAÇÃO APLICADA, 12., 2012, São José dos Campos. **Anais...** São José dos Campos: INPE, 2012. p. 1–8.

SHIMIZU, K.; YOSHIDA, J.; NAGAI, K.; NISHIMURA, M.; ISHII, G.; MORISHITA, Y.; NISHIWAKI, Y. Visceral pleural invasion is an invasive and aggressive indicator of non-small cell lung cancer. **Journal of Thoracic and Cardiovascular Surgery**, v. 130, n. 1, p. 160–165, 2005.

SILVA, A. C.; CARVALHO, P. C. P.; GATTASS, M. Diagnosis of lung nodule using semivariogram and geometric measures in computerized tomography images. **Computer Methods and Programs in Biomedicine**, v. 79, n. 1, p. 31–38, 2005.

SOUZA, J. P. **Modelo de qualidade para desenvolvimento e avaliação da viabilidade clínica de sistemas de recuperação de imagens médicas baseada em conteúdo**. 2012. 217p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2012.

TANG, X. Texture information in run-length matrices. **IEEE Transactions on Image Processing**, v. 7, n. 11, p. 1602–1609, 1998.

TARTAR, A.; KILIC, N.; AKAN, A. Classification of pulmonary nodules by using hybrid features. **Computacional and Mathematical Methods in Medicine**, v. 2013, ID 148363. 2013a.

TARTAR, A.; KILIC, N.; AKAN, A. A new method for pulmonary nodule detection using decision trees. In: ANNUAL INTERNATIONAL CONFERENCE OF THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY, 35., 2013, Osaka. **Proceedings…** Piscataway: IEEE, 2013b. p. 7355–7359.

THIBAULT, G.; FERTIL, B.; NAVARRO, C.; PEREIRA, S.; CAU, P.; LEVY, N.; SEQUEIRA, J.; MARI, J. Texture indexes and gray level size zone matrix application to cell nuclei classification. In: INTERNATIONAL CONFERENCE PATTERN RECOGNITION AND INFORMATION PROCESSING, 10., 2009, Tous. **Proceedings…** London: IET, 2009. p. 140–145.

VAIDYA, M.; CREACH, K. M.; FRYE, J.; DEHDASHTI, F.; BRADLEY, J. D.; EL NAQA, I. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. **Radiotherapy and Oncology**, v. 102, n. 2, p. 239–245, 2012.

VAN GRIETHUYSEN, J. J. M.; FEDOROV, A.; PARMAR, C.; HOSNY, A.; AUCOIN, N.; NARAYAN, V.; BEETS-TAN, R. G. H.; FILLON-ROBIN, J. C.; PIEPER, S.; AERTS, H. J. W. L. Computational radiomics system to decode the radiographic phenotype. **Cancer Research**, v. 77, n. 21, p. e104–e107, 2017.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. The WEKA workbench online appendix. In: _____. **Data mining**: practical machine learning tools and techniques. 4th ed. Amsterdam: Morgan Kaufmann, 2016.

WU, H.; SUN, T.; WANG, J.; LI, X.; WANG, W.; HUO, D.; LV, P.; HE, W.; WANG, K.; GUO, X. Combination of radiological and gray level co-occurrence matrix textural features used to distinguish solitary pulmonary nodules by computed tomography. **Journal of Digital Imaging**, v. 26, n. 4, p. 797–802, 2013.

WU, W.; PARMAR, C.; GROSSMANN, P.; QUACKENBUSH, J.; LAMBIN, P.; BUSSINK, J.; MAK, R.; AERTS, H. J. W. L. Exploratory study to identify radiomics classifiers for lung cancer histology. **Frontiers in Oncology**, v. 6, p. 1–11, 2016.

XU, D.-H.; KURANI, A. S.; FURST, J. D.; RAICU, D. S. Run-length encoding for volumetric texture. In: INTERNATIONAL CONFERENCE ON VISUALIZATION, IMAGING AND IMAGE PROCESSING, 7P., 2004, Marbella. **Proceedings…** Calgary: Acta Press, 2004. p. 452–458.

YANG, F.; THOMAS, M. A.; DEHDASHTI, F.; GRIGSBY, P. W. Temporal analysis of intratumoral metabolic heterogeneity characterized by textural features in cervical cancer. **European Journal of Nuclear Medicine and Molecular Imaging**, v. 40, n. 5, p. 716–727, 2013.

ZHU, L.; KOLESOV, I.; GAO, Y.; KIKINIS, R.; TANNENBAUM, A. An Effective interactive medical image segmentation method using fast growCut. In: INTERNATIONAL CONFERENCE ON MEDICAL IMAGE COMPUTING AND COMPUTER ASSISTED INTERVENTION, 17., Boston. **Proceedings…** Berlin: Springer International, 2014.

APPENDIX A

Appendix A contains Table 8 regarding features information.

Table 8 – Full list of shape and texture features with their respective Feature ID, Acronym and class. (Table continues in the next page)

Feature Class	Feature Name	Acronym	T1PC-T2 (T1PC portion)		T1PC-T2 (T2 portion)	
			Feature ID (2D)	**Feature ID (3D)**	**Feature ID (2D)**	**Feature ID (3D)**
shape	Maximum 3D Diameter	M3DD	1	76	151	226
shape	Compactness 2	C2	2	77	152	227
shape	Maximum 2D Diameter Slice	M2DDS	3	78	153	228
shape	Sphericity	SPT	4	79	154	229
shape	Minor Axis	MNA	5	80	155	230
shape	Compactness 1	C1	6	81	156	231
shape	Elongation	ELG	7	82	157	232
shape	Surface-Volume Ratio	SFR	8	83	158	233
shape	Volume	VOL	9	84	159	234
shape	Spherical Disproportion	SD	10	85	160	235
shape	Major Axis	MJA	11	86	161	236
shape	Least Axis	LA	12	87	162	237
shape	Flatness	FTN	13	88	163	238
shape	Surface Area	SA	14	89	164	239
shape	Maximum 2D Diameter Column	M2DDC	15	90	165	240
shape	Maximum 2D Diamenter Row	M2DDR	16	91	166	241
glcm	Sum Variance	SV	17	92	167	242
glcm	Homogeneity 1	HM1	18	93	168	243
glcm	Homogeneity 2	HM2	19	94	169	244
glcm	Cluster Shade	CS	20	95	170	245
glcm	Maximum Probability	MP	21	96	171	246
glcm	Inverse Difference Moment Normalized	IDMN	22	97	172	247
glcm	Contrast	CNT	23	98	173	248
glcm	Difference Entropy	DE	24	99	174	249
glcm	Inverse Variance	IV	25	100	175	250

Source – Author

Table 8 – Full list of shape and texture features with their respective Feature ID, Acronym and class. (Table continues in the next page)

Feature Class	Feature Name	Acronym	T1PC-T2 (T1PC portion)		T1PC-T2 (T2 portion)	
			Feature ID (2D)	**Feature ID (3D)**	**Feature ID (2D)**	**Feature ID (3D)**
glcm	Dissimilarity	DSS	26	101	176	251
glcm	Sum Avarage	SAVG	27	102	177	252
glcm	Difference Variance	DV	28	103	178	253
glcm	Inverse Difference Normalized	IDN	29	104	179	254
glcm	Inverse Difference Moment	IDM	30	105	180	255
glcm	Correlation	COR	31	106	181	256
glcm	Autocorrelation	ACOR	32	107	182	257
glcm	Sum Entropy	SE	33	108	183	258
glcm	Avarage Intensity	AVGI	34	109	184	259
glcm	Energy	ENG	35	110	185	260
glcm	Sum Squares	SS	36	111	186	261
glcm	Cluster Prominence	CP	37	112	187	262
glcm	Entropy	ETP	38	113	188	263
glcm	Informal Measure of Correlation 2	IMC2	39	114	189	264
glcm	Informal Measure of Correlation 1	IMC1	40	115	190	265
glcm	Difference Average	DAVG	41	116	191	266
glcm	Inverse Difference	ID	42	117	192	267
glcm	Cluster Tendency	CT	43	118	193	268
glrlm	Short Run Low Gray Level Emphasis	SRLGLE	44	119	194	269
glrlm	Gray Level Variance	GLV	45	120	195	270
glrlm	Low Gray Level Run Emphasis	LGLRE	46	121	196	271
glrlm	Gray Level Non Uniformity Normalized	GLNN	47	122	197	272
glrlm	Run Variance	RV	48	123	198	273
glrlm	Gray Level Non Uniformity	GLN	49	124	199	274
glrlm	Long Run Emphasis	LRE	50	125	200	275
glrlm	Short Run High Gray Level Emphasis	SRHGLE	51	126	201	276
glrlm	Run Length Non Uniformity	RLN	52	127	202	277

Source – Author

Table 8 – Full list of shape and texture features with their respective Feature ID, Acronym and class. (End of table)

Feature Class	Feature Name	Acronym	T1PC-T2 (T1PC portion)		T1PC-T2 (T2 portion)	
			Feature ID (2D)	**Feature ID (3D)**	**Feature ID (2D)**	**Feature ID (3D)**
glrlm	Short Run Emphasis	SRE	53	128	203	278
glrlm	Long Run High Gray Level Emphasis	LRHGLE	54	129	204	279
glrlm	Run Percentage	RP	55	130	205	280
glrlm	Long Run Low Gray Level Emphasis	LRLGLE	56	131	206	281
glrlm	Run Entropy	RE	57	132	207	282
glrlm	High Gray Level Run Emphasis	HGLRE	58	133	208	283
glrlm	Run Length Non Uniformity Normalized	RLNN	59	134	209	284
glszm	Gray Level Variance	GLVSZ	60	135	210	285
glszm	Small Area High Gray Level Emphasis	SAHGLE	61	136	211	286
glszm	Gray Level Non Uniformity Normalized	GLNNSZ	62	137	212	287
glszm	Size Zone Non Uniformity Normalized	SZNN	63	138	213	288
glszm	Size Zone Non Uniformity	SZN	64	139	214	289
glszm	Gray Level Non Uniformity	GLNSZ	65	140	215	290
glszm	Large Area Emphasis	LAE	66	141	216	291
glszm	Zone Variance	ZV	67	142	217	292
glszm	Zone Percentage	ZP	68	143	218	293
glszm	Large Area Low Gray Level Emphasis	LALGLE	69	144	219	294
glszm	Large Area High Gray Level Emphasis	LAHGLE	70	145	220	295
glszm	High Gray Level Zone Emphasis	HGLZE	71	146	221	296
glszm	Small Area Emphasis	SAE	72	147	222	297
glszm	Low Gray Level Zone Emphasis	LGLZE	73	148	223	298
glszm	Zone Entropy	ZE	74	149	224	299
glszm	Small Area Low Gray Level Emphasis	SALGLE	75	150	225	300

Source – Author

APPENDIX B

Appendix B contains Tables 9-14 representing the classifiers' performance for each unbalanced dataset using Bagging as bootstrapping method.

Table 9 – Classifiers' performance for T1PC unbalanced data on Bagging method with bag size of 100%. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.806	0.643	0.857	0.811	0.889	0.000
J48	**1.000**	**1.000**	0.857	0.800	0.778	0.200
RF	**1.000**	**1.000**	0.857	0.800	0.889	0.200
KNN-1	**1.000**	**1.000**	**1.000**	0.822	0.889	**0.400**
KNN-3	0.857	0.786	0.714	0.822	0.778	**0.400**
KNN-5	0.796	0.929	0.429	0.800	0.889	**0.400**
KNN-7	0.735	0.929	0.286	0.844	0.889	0.000
KNN-9	0.755	**1.000**	0.000	0.822	**1.000**	0.000
MLP-1	**1.000**	0.929	**1.000**	**0.867**	0.889	**0.400**
MLP-2	0.990	**1.000**	0.143	0.822	**1.000**	0.000
MLP-3	0.990	**1.000**	0.000	0.844	**1.000**	0.000

Source – Author

Table 10 – Classifiers' performance for T1PC unbalanced data on Bagging method with bag size of 66%. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.878	0.857	0.714	0.844	**1.000**	0.000
J48	0.990	**1.000**	0.714	0.800	0.889	0.000
RF	0.980	**1.000**	0.714	0.800	0.889	0.000
KNN-1	**1.000**	**1.000**	**1.000**	0.822	0.889	**0.400**
KNN-3	0.827	0.929	0.429	0.778	0.778	0.200
KNN-5	0.745	**1.000**	0.143	0.778	0.889	0.000
KNN-7	0.776	**1.000**	0.000	**0.867**	**1.000**	0.000
KNN-9	0.827	**1.000**	0.000	0.800	**1.000**	0.000
MLP-1	**1.000**	**1.000**	0.857	0.822	**1.000**	0.000
MLP-2	0.980	**1.000**	0.000	0.778	**1.000**	0.000
MLP-3	0.980	**1.000**	0.000	0.844	**1.000**	0.000

Source – Author

Table 11 – Classifiers' performance for T2 unbalanced data on Bagging method with bag size of 100%. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.881	0.857	0.667	**0.917**	**1.000**	0.000
J48	**1.000**	**1.000**	0.667	0.694	0.778	0.250
RF	**1.000**	**1.000**	0.833	0.722	0.889	0.500
KNN-1	**1.000**	**1.000**	**1.000**	0.833	0.778	**0.750**
KNN-3	0.833	0.929	0.333	0.722	**1.000**	0.250
KNN-5	0.738	**1.000**	0.167	0.667	**1.000**	0.250
KNN-7	0.750	**1.000**	0.000	0.694	**1.000**	0.000
KNN-9	0.714	**1.000**	0.000	0.861	**1.000**	0.000
MLP-1	**1.000**	**1.000**	0.667	0.861	**1.000**	0.000
MLP-2	**1.000**	**1.000**	0.500	0.833	**1.000**	0.000
MLP-3	**1.000**	**1.000**	0.500	0.833	**1.000**	0.000

Source – Author

Table 12 – Classifiers' performance for T2 unbalanced data on Bagging method with bag size of 66%. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.952	**1.000**	0.500	0.833	**1.000**	0.000
J48	**0.988**	**1.000**	**0.667**	0.722	0.889	0.250
RF	0.976	**1.000**	**0.667**	0.722	0.889	0.250
KNN-1	0.952	**1.000**	0.500	0.806	**1.000**	**0.500**
KNN-3	0.726	**1.000**	0.167	0.750	**1.000**	0.250
KNN-5	0.738	**1.000**	0.000	0.750	**1.000**	0.000
KNN-7	0.762	**1.000**	0.000	**1.000**	**1.000**	0.000
KNN-9	0.762	**1.000**	0.000	0.944	**1.000**	0.000
MLP-1	0.952	**1.000**	0.500	0.778	**1.000**	0.000
MLP-2	0.952	**1.000**	0.500	0.833	**1.000**	0.000
MLP-3	0.976	**1.000**	0.500	0.889	**1.000**	0.000

Source – Author

Table 13 – Classifiers' performance for T1PC-T2 unbalanced data on Bagging method with bag size of 100%. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.917	0.786	0.833	0.764	0.889	0.000
J48	**1.000**	**1.000**	0.833	0.778	0.778	0.000
RF	**1.000**	**1.000**	**1.000**	0.778	0.889	0.000
KNN-1	**1.000**	**1.000**	**1.000**	0.861	0.778	**0.750**
KNN-3	0.940	**1.000**	0.500	0.889	**1.000**	0.250
KNN-5	0.833	**1.000**	0.167	0.861	**1.000**	0.250
KNN-7	0.821	**1.000**	0.000	0.833	**1.000**	0.000
KNN-9	0.762	**1.000**	0.000	**0.917**	**1.000**	0.000
MLP-1	**1.000**	**1.000**	0.000	0.833	**1.000**	0.000
MLP-2	**1.000**	**1.000**	0.000	0.833	**1.000**	0.000
MLP-3	0.589	0.928	0.166	0.806	**1.000**	0.000

Source – Author

Table 14 – Classifiers' performance for T1PC-T2 unbalanced data on Bagging method with bag size of 66%. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.964	**1.000**	0.500	0.792	**1.000**	0.000
J48	**1.000**	**1.000**	0.833	0.722	0.889	0.000
RF	0.988	**1.000**	0.833	0.778	0.889	0.000
KNN-1	**1.000**	**1.000**	**1.000**	0.861	0.889	**0.250**
KNN-3	0.845	**1.000**	0.167	0.833	**1.000**	0.000
KNN-5	0.774	**1.000**	0.000	**0.917**	**1.000**	0.000
KNN-7	0.762	**1.000**	0.000	0.889	**1.000**	0.000
KNN-9	0.655	**1.000**	0.000	0.778	**1.000**	0.000
MLP-1	**1.000**	**1.000**	0.000	0.750	**1.000**	0.000
MLP-2	0.988	**1.000**	0.000	0.694	**1.000**	0.000
MLP-3	**1.000**	**1.000**	0.000	0.889	**1.000**	0.000

Source – Author

APPENDIX C

Appendix C contains Tables 15-20 representing the classifiers' performance for each dataset after feature selection using Wrapper.

Table 15 – Classifiers' performance for T1PC unbalanced data after Wrapper feature selection. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.796	0.929	**0.714**	0.689	0.667	**0.800**
J48	0.383	0.786	0.429	0.544	**0.889**	0.200
RF	**0.847**	0.929	**0.714**	**0.844**	**0.889**	0.200
KNN-1	0.643	0.714	0.571	0.589	0.778	0.400
KNN-3	0.709	**1.000**	0.571	0.544	0.778	0.200
KNN-5	0.679	**1.000**	0.429	0.611	0.778	0.200
KNN-7	0.587	0.714	0.286	0.767	0.778	0.400
KNN-9	0.520	0.857	0.000	0.833	**0.889**	0.000
MLP-1	0.724	0.929	0.571	0.667	0.778	0.600
MLP-2	0.704	0.929	0.571	0.667	0.778	0.600
MLP-3	0.714	0.929	0.571	0.711	0.778	0.600

Source – Author

Table 16 – Classifiers' performance for T1PC balanced data after Wrapper feature selection. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.867	**0.929**	0.714	0.689	0.222	**0.800**
J48	0.883	0.786	0.857	0.689	0.778	0.600
RF	**0.916**	**0.929**	0.786	0.833	0.778	**0.800**
KNN-1	0.821	0.714	**0.929**	0.733	0.667	**0.800**
KNN-3	0.842	0.786	0.786	**0.844**	0.778	0.400
KNN-5	0.885	0.714	0.857	0.778	0.778	**0.800**
KNN-7	0.786	0.643	0.857	0.756	0.667	**0.800**
KNN-9	0.839	0.571	**0.929**	0.756	0.667	**0.800**
MLP-1	0.883	0.786	0.786	**0.844**	**1.000**	0.400
MLP-2	0.898	0.786	**0.929**	0.800	0.889	0.200
MLP-3	0.867	0.714	0.857	**0.844**	0.778	**0.800**

Source – Author

Table 17 – Classifiers' performance for T2 unbalanced data after Wrapper feature selection. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	**0.738**	0.857	**0.667**	0.722	0.667	0.750
J48	0.405	0.857	0.333	0.569	**0.889**	0.250
RF	0.571	0.857	0.333	0.722	0.778	0.500
KNN-1	0.583	0.500	**0.667**	0.708	0.667	**0.750**
KNN-3	0.655	**1.000**	0.500	0.542	**0.889**	0.500
KNN-5	0.613	**1.000**	0.167	0.778	0.778	0.500
KNN-7	0.601	**1.000**	0.333	**0.819**	0.778	**0.750**
KNN-9	0.726	**1.000**	0.333	0.639	0.667	0.500
MLP-1	0.732	0.929	0.500	0.764	0.778	**0.750**
MLP-2	0.619	0.643	**0.667**	0.583	0.444	**0.750**
MLP-3	0.571	0.786	0.500	0.611	0.667	0.500

Source – Author

Table 18 – Classifiers' performance for T2 balanced data after Wrapper feature selection. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.929	0.929	0.786	**0.944**	**1.000**	**0.750**
J48	0.587	0.643	0.786	0.528	0.556	0.500
RF	0.870	0.857	0.786	0.694	0.889	0.500
KNN-1	0.750	0.643	0.857	0.639	0.778	0.500
KNN-3	0.793	0.643	0.857	0.583	0.556	0.500
KNN-5	0.852	**1.000**	0.786	0.750	0.778	**0.750**
KNN-7	0.781	0.857	0.786	0.764	0.778	**0.750**
KNN-9	0.770	0.786	0.786	0.819	0.667	**0.750**
MLP-1	**0.985**	0.929	**0.929**	0.556	0.556	**0.750**
MLP-2	0.949	0.786	0.857	0.639	0.556	**0.750**
MLP-3	0.959	0.857	0.857	0.611	0.556	**0.750**

Source – Author

Table 19 – Classifiers' performance for T1PC-T2 unbalanced data after Wrapper feature selection. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.810	0.857	0.667	0.194	0.333	0.250
J48	0.149	0.786	0.167	0.444	**0.889**	0.000
RF	0.875	**1.000**	0.500	**0.833**	**0.889**	0.000
KNN-1	**0.917**	**1.000**	**0.833**	0.514	0.778	0.250
KNN-3	0.655	**1.000**	0.167	0.389	**0.889**	0.250
KNN-5	0.845	**1.000**	0.167	0.722	**0.889**	0.250
KNN-7	0.619	0.786	0.333	0.694	0.778	0.000
KNN-9	0.726	**1.000**	0.333	0.639	0.667	0.500
MLP-1	0.857	**1.000**	0.333	**0.833**	0.778	0.500
MLP-2	0.845	**1.000**	0.333	**0.833**	0.778	0.500
MLP-3	0.702	0.929	0.167	**0.833**	0.667	**1.000**

Source – Author

Table 20 – Classifiers' performance for T1PC-T2 balanced data after Wrapper feature selection. Highest observations for AUC, SENS and SPEC are highlighted as bold values.

	VALIDATION			TEST		
	AUC	SENS	SPEC	AUC	SENS	SPEC
NB	0.959	**0.959**	0.857	0.306	0.778	0.250
J48	0.689	0.929	0.714	0.583	0.667	0.500
RF	0.934	0.786	0.857	0.667	0.778	0.500
KNN-1	0.964	0.929	**1.000**	0.458	0.667	0.250
KNN-3	0.926	0.643	**1.000**	0.611	0.667	0.750
KNN-5	0.888	0.714	0.857	0.764	0.778	0.500
KNN-7	0.895	0.643	**1.000**	**0.917**	0.778	**1.000**
KNN-9	0.898	0.500	0.929	0.819	0.667	0.750
MLP-1	0.985	0.929	0.857	0.778	0.778	0.750
MLP-2	0.980	0.857	0.929	0.639	0.778	0.500
MLP-3	**0.995**	0.929	0.929	0.889	**0.889**	0.750

Source – Author