Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

ISTI Technical Reports

Technical report on the development and interpretation of convolutional neural networks for the classification of multiparametric MRI images on unbalanced datasets. Case study: prostate cancer

Eva Pachetti, ISTI-CNR, Pisa, Italy

Sara Colantonio, ISTI-CNR, Pisa, Italy

ISTI-TR-2021/005

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

Technical report on the development and interpretation of convolutional neural networks for the classification of multiparametric MRI images on unbalanced datasets. Case study: prostate cancer.

Pachetti E., Colantonio S.

ISTI-TR-2021/005

This report summarized the activities carried out to define, train and validate Deep Learning models for the classification of medical imaging data.
The issue of unbalanced datasets was faced by applying some data augmentation techniques, based on transformation of the original images. Such techniques were compared to verify their impact in a frame where object morphology is relevant.
Multimodal deep learning models were defined to exploit the information contained in heterogeneous imaging data and cope with data distribution imbalance.
To verify the inner functioning of the deep learning models, the LIME algorithm was applied, thus checking that the regions that contribute to the classification were the real meaningful ones.
The case study used to was the categorization of prostate cancer aggressiveness based on Magnetic Resonance Imaging (MRI) data. The aggressiveness was determined, as a ground truth, via tissue biopsy and expressed with a score from 2 to 10 known as Gleason Score, which is obtained as the sum of two values, each one from 1 to 5, associated with the two most common patterns in the tumor tissue histological sample.

Keywords: Convolutional neural networks, Unbalanced datasets, Multimodal neural models, Deep learning interpretation.

Citation

Pachetti E., Colantonio S., *Technical report on the development and interpretation of convolutional neural networks for the classification of multiparametric MRI images on unbalanced datasets. Case study: prostate cancer*. ISTI Technical Reports 2021/005. DOI: 10.32079/ISTI-TR-2021/005.

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1
56124 Pisa Italy
http://www.isti.cnr.it

Technical report on the development and interpretation of convolutional neural networks for the classification of multiparametric MRI images on unbalanced datasets. Case study: prostate cancer

Eva Pachetti[1], Sara Colantonio[1]

[1]Institute of Information Science and Technologies, National Research Council of Italy

Abstract

This report summarized the activities carried out to define, train and validate Deep Learning models for the classification of medical imaging data.

The issue of unbalanced datasets was faced by applying some data augmentation techniques, based on transformation of the original images. Such techniques were compared to verify their impact in a frame where object morphology is relevant.

Multimodal deep learning models were defined to exploit the information contained in heterogeneous imaging data and cope with data distribution imbalance.

To verify the inner functioning of the deep learning models, the LIME algorithm was applied, thus checking that the regions that contribute to the classification were the real meaningful ones.

The case study used to was the categorization of prostate cancer aggressiveness based on Magnetic Resonance Imaging (MRI) data. The aggressiveness was determined, as a ground truth, via tissue biopsy and expressed with a score from 2 to 10 known as Gleason Score, which is obtained as the sum of two values, each one from 1 to 5, associated with the two most common patterns in the tumor tissue histological sample.

1. Introduction

Computer-aided diagnosis (CAD) is an important research field within medical imaging, which aims to support the radiologist's work in information quantification, in new biomarkers discovery or in diagnostic analysis, exploiting the wealth of information contained in imaging data. In oncology, a typical example is given by the differentiation of neoplastic lesions from benign ones or by estimating the aggressiveness of malignant lesions. In this sense, with the development of deep learning, medical image classification has made significant progress. Training deep learning models usually requires many samples belonging to different classes. However, in many clinical cases, it can be difficult to collect a balanced dataset either because of the low prevalence or the low incidence of clinically significant tumors versus indolent ones.

The work aimed to solve the problem of imbalanced datasets in the automated identification of neoplastic lesions, evaluating the application of different data augmentation techniques. The goal was to understand the impact these have in a context, such as the biomedical one, where image morphology is particularly important.

Secondly, the work focused on the interpretation of neural networks to understand the classification criterion and on this basis decide whether it is possible to trust predictions. Here, the LIME (Local Interpretable Model-agnostic Explanations) [1] algorithm was used, which intuitively highlights the image parts that increase the probability that it belongs to a certain class. The aim was to assess whether the tumor lesion is also contained within these regions and therefore if it is relevant for classification purposes.

Finally, to increase the information content provided to the model, enhancing its generalization capabilities, a multimodal neural network was also created, which consists of several branches that process multiparametric images in parallel, combining the extracted information to make predictions.

These three aspects were studied by choosing prostate cancer as an application case, which is a typical disease example that leads to the generation of unbalanced datasets as the number of diagnoses falls, mostly, in a class of low-severity tumors. The work was conducted for classifying prostate Magnetic Resonance Imaging (MRI) data based on tumor aggressiveness, using each case's severity value as ground truth. The severity is determined via tissue biopsy and expressed with a score from 2 to 10 known as Gleason Score, which is obtained as the sum of two values, each one from 1 to 5, associated with the two most common patterns in the tumor tissue histological sample.

2. Related work

In recent years, several works have been proposed that aim to classify MRI prostate images based on tumor aggressiveness using a deep learning approach. In most cases, the classification is limited to a distinction between indolent and clinically significant tumors while only a few attempts to classify images into different Gleason Scores. The following subsections summarized the most notable of them, by distinguishing the type of classification end-point.

2.1 Binary classification

Minh Hung Le et al. [2] exploit a transfer learning approach to realize a multimodal network that carries out the classification by merging the T2W and ADC images information. An architecture belonging to the state of the art is proposed on each branch of the network. In particular, the VGG-16, GoogleNet and ResNet networks are compared. Furthermore, to increase the small size of the dataset, various data augmentation techniques are tested, among which rigid and non-rigid geometric transformations.

Abraham and Nair [3] classify images using a Sparse Autoencoder (SAE) in combination with a Random Forest classifier. In this case, in addition to the T2W and ADC images, the information of the DW images is also exploited. Furthermore, to increase the size of the dataset the ADASYN [4] methodology is used.

Finally, Yuan et al [5] uses a multimodal Convolutional Neural Network (CNN) that extracts the features respectively from axial T2W, sagittal T2W and ADC images. Also in this case, a transfer learning approach is exploited, implementing AlexNet network in each branch of the multimodal neural network. Moreover, a similarity constraint between the images in the cost function is added, which describes the relationship between the features within the same category.

2.2 Multi-label classification

Abraham and Nair [6] classify multi-parametric MRI (mpMRI) images by extracting features from T2W, ADC and DW volumes, using VGG -16 network in combination with an Ordinal Class Classifier (OCC), which allows for considering among the classification criteria, also the ordering of the various groups based on the level of tumor aggressiveness.

3. Method

3.1 Data selection and organization

Data have been provided by Careggi University Hospital and include mpMRI scans (axial plane), T2W images and ADC maps. The images were acquired on 85 patients for a total of 103 cases, considering that the same patient may have multiple lesions. From each patient's set of slices, only those containing the lesion were selected. This operation led obtaining 245 T2W images and 239 ADC maps. Since the number of acquisitions collected was particularly small, as well as unbalanced in the different aggressiveness levels, there would not have been enough examples to train the network to recognize all the Gleason Score values. For this reason, images were divided into two macro-groups: Low Grade (LG) and High Grade (HG). In particular, the LG class includes all cases with a GS $\leq 3 + 4$, while the HG one those with a GS $\geq 4 + 3$. This way, 165 LG images and 80 HG T2W images, and 159 LG and 80 HG ADC maps were obtained.

3.2 Neural network implementation

The network was developed in the scientific programming environment *Spyder*, using *Python 3.7* language and the deep learning library *Pytorch*. The choice of the number, the type, the succession of the layers within the network, as well as the parameters and the number of epochs, was guided by the creation of a network with the best classification performance possible and, above all, to avoid overfitting. For this reason, the network was developed with few layers and adopting techniques as dropout, batch normalization and early stopping.

Figure 1: Convolutional neural network architecture

3.3 Input optimal size determination

To keep the network's focus on the tumor lesion, an appropriate function was also created, which cut out the images while keeping the tumor in the center. To determine what the optimal size was, several training tests were carried out, varying the size of the cutout to understand how much the surrounding structures affected performance. The size that provides the best performance was found to be 64x64 pixels in T2W images, and 44x44 pixels in ADC maps.

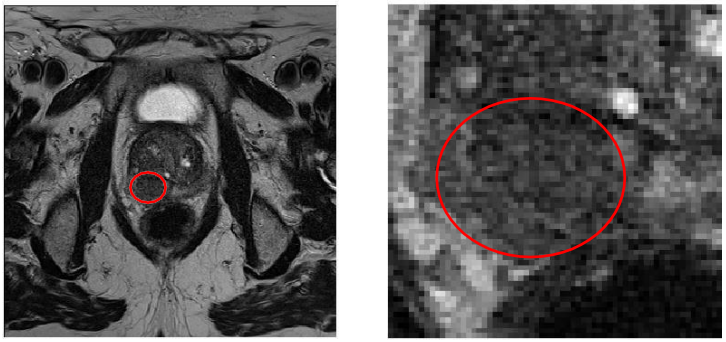

Figure 2: (a) Original image (b) Cropped image around the tumor lesion

3.4 Augmented datasets generation

At this point, augmented datasets were generated by applying the following data augmentation techniques to the original images: rotation, vertical flip, horizontal flip, translation, perspective transformation, scaling and shear. Each technique was applied to both the minority class and to both classes. In the first scenario, the dataset balancing was achieved, while in the second one the dataset was kept unbalanced but with a greater number of examples. In this way, two augmented datasets were obtained for each applied geometric transformation, both for T2W images and ADC maps.

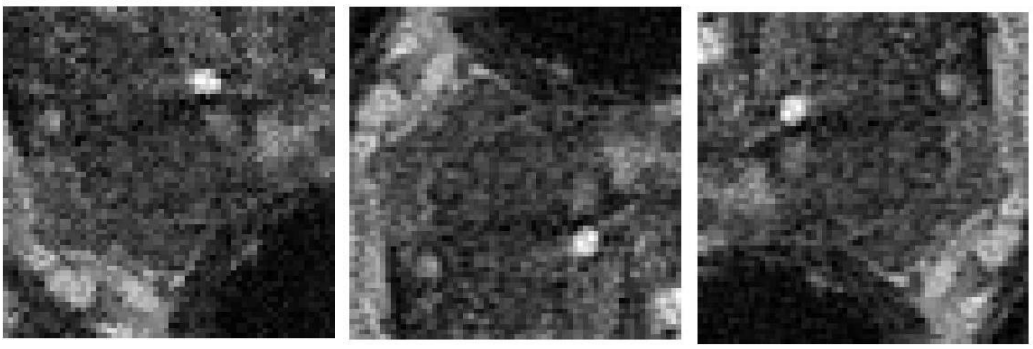

Figure 3: (a) Rotated image (b) Vertically flipped image (c) Horizontally flipped image

Then, also an augmented dataset was generated using both rotated and horizontally flipped images, to evaluate the combination of more types of augmentation techniques.

3.5 Training

Each dataset was then used to train the network. To make the results more robust, the *k-fold Cross Validation* statistical technique was used, choosing k=5. According to this, each dataset was divided into five groups, of which, in rotation, one was used as a test set and the other four as training plus validation. Five training sessions were carried out for each dataset and, for each of them, the following metrics were evaluated: specificity, sensitivity, total accuracy, and the Area Under the ROC Curve (AUC). Furthermore, the confusion matrix as well as the training and validation loss and accuracy were plotted.

3.5.1 Results

Below are the average results obtained on the five trainings for each data augmentation technique implemented, comparing the case in which it is applied to minority class's images only with the one in which it is applied to those of both classes. In Table 1, the results provided by training the network with only T2W images are reported, while in Table 2 those with only ADC maps.

Table 1: Comparison between the average results obtained with the two approaches for the network trained with T2W images only

	Data augmentation approach	**Specificity**	**Sensitivity**	**Total accuracy**	**AUC**
Rotation	LG - HG	93,6%	87,2%	91,4%	0.97
	HG only	81%	89,6%	83,8%	0.94
Vertical flip	LG - HG	89,8%	79,8%	86,4%	0.92
	HG only	84,2%	84,4%	84,2%	0.91
Horizontal flip	LG - HG	93%	82,2%	89,4%	0.94
	HG only	76,2%	89,8%	80,6%	0.93
Translation	LG - HG	83,4%	80,8%	82,6%	0.92
	HG only	88,8%	82,2%	86,6%	0.94
Perspective transformation	LG - HG	87%	83,4%	85,8%	0.93
	HG only	75%	93,4%	81%	0.95
Scale	LG - HG	91,4%	69,8%	84,2%	0.90
	HG only	63%	93,4%	73%	0.91
Shear	LG - HG	90%	82%	87,4%	0.92
	HG only	90,8%	88,6%	89,8%	0.94
Rotation + Horizontal flip	LG - HG	80,4%	89,8%	83,4%	0.91
	HG only	73,8%	91%	79,4%	0.91

Table 2: Comparison between the average results obtained with the two approaches for the network trained with ADC maps only

	Data augmentation approach	**Specificity**	**Sensitivity**	**Total accuracy**	**AUC**
Rotation	LG - HG	88,2%	90,8%	88,8%	0.95
	HG only	87,4%	89,6%	88,2%	0.96
Vertical flip	LG - HG	89,4%	73,2%	84%	0.93
	HG only	89,6%	87,2%	88,6%	0.96
Horizontal flip	LG - HG	87,8%	87%	87,4%	0.93
	HG only	78,2%	93,4%	83,2%	0.92
Translation	LG - HG	90,8%	88,2%	89,8%	0.94
	HG only	82%	88,6%	84%	0.94

Perspective transformation	LG - HG	85,8%	73,4%	81,6%	0.89
	HG only	81,8%	92,2%	85,2%	0.95
Scale	LG - HG	88,8%	86,2%	87,8%	0.95
	HG only	83,2%	88,4%	84,8%	0.93
Shear	LG - HG	87,2%	85,8%	86,6%	0.93
	HG only	75,6%	92,2%	81,2%	0.94
Rotation + Horizontal flip	LG - HG	92,6%	92,2%	92,2%	0.96
	HG only	74,2%	87,4%	78,6%	0.92

The best results were achieved, for T2W images, using the dataset augmented by applying rotation to both classes, which provides a specificity of 93,6%, a sensitivity of 87,2%, a total accuracy of 91,4% and an AUC of 0.97. For ADC maps, instead, vertical flip applied to the minority class only provides the best performance, giving a specificity of 89,6%, a sensitivity of 87,2%, a total accuracy of 88,6% and an AUC of 0.96.

3.6 Neural network interpretation

At this point, the single network's classification criterion was investigated by applying LIME algorithm. LIME highlights the image parts that increase the probability that it belongs to a certain class. In this way, one can determine whether the tumor lesion is contained within these regions and therefore whether it is relevant for the classification purpose.

For each test set, the average percentage of images in which the tumor is correctly displayed, compared to the total of correctly classified images was evaluated. In this way, a percentage value for each augmentation technique, both on the minority class only and on both classes has been obtained, in order to understand if and how the various techniques affect the network's ability to correctly visualize the lesion.

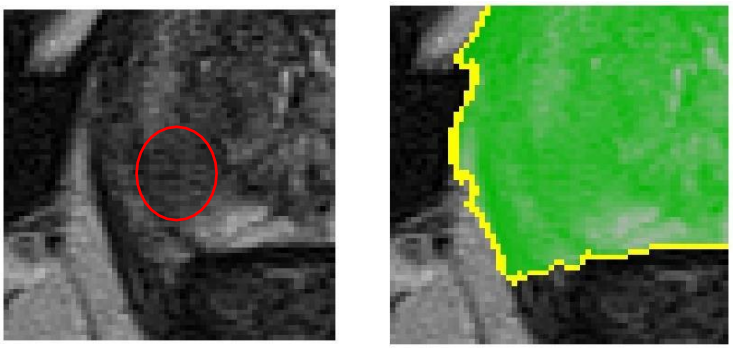

Figure 4: (a) Prostate T2W image with highlighted lesion (b) Part of the image that is most considered by the network according to LIME algorithm

3.6.1 Results

In Table 3 and Table 4, for each data augmentation technique, the average percentage of T2W images and ADC maps respectively in which the tumor is correctly visualized by the network compared to the total of images correctly classified is reported. The comparison is made, for both the augmentation approaches, against the result obtained with the non-augmented dataset.

Table 3: Percentage of T2W images in which the tumor is correctly displayed for each data augmentation technique following both the augmentation approaches

	Data augmentation approach	Average percentage of images correctly displayed by the network
Not augmented dataset		63%
Rotation	LG - HG	57%
	HG only	62%
Vertical flip	LG - HG	70%
	HG only	60%
Horizontal flip	LG - HG	56%
	HG only	67%
Translation	LG - HG	59%
	HG only	59%
Perspective transformation	LG - HG	62%
	HG only	64%
Scale	LG - HG	65%
	HG only	60%
Shear	LG - HG	63%
	HG only	57%
Rotation + Horizontal flip	LG - HG	71%
	HG only	56%

Table 4: Percentage of ADC maps in which the tumor is correctly displayed for each data augmentation technique, following both the augmentation approaches

	Data augmentation	Average percentage of images correctly displayed by the network
Not augmented dataset		85%
Rotation	LG - HG	78%
	HG only	77%
Vertical flip	LG - HG	86%
	HG only	69%
Horizontal flip	LG - HG	75%
	HG only	75%
Translation	LG - HG	73%
	HG only	79%

Perspective transformation	LG - HG	73%
	HG only	72%
Scale	LG - HG	78%
	HG only	69%
Shear	LG - HG	78%
	HG only	75%
Rotation + Horizontal flip	LG - HG	78%
	HG only	73%

The dataset in which the tumor lesion is most correctly displayed was found to be the one augmented by vertical flipping applied to both classes, in which the network, both for T2W images and ADC maps, correctly displays the tumor in a larger number of images with respect to the non-augmented dataset. For T2W images, 70% was correctly visualized, compared to 63% obtained with the non-augmented dataset, while in ADC maps 86% versus 85%.

3.7 Multimodal neural network implementation

The information contained in the two image modes was then combined in the multimodal network, which consists of two branches that process the T2W image and the ADC map in parallel for the same acquisition slice. The same architecture used for the separate classification of T2W and ADC images has been re-proposed on the two branches, to which is added a common fully connected layer that, after the features concatenation, makes the prediction. This approach, besides providing a complete characterization of the tumor lesion, also makes it possible to compensate for the limited size of the dataset, providing the network with a double amount of information.

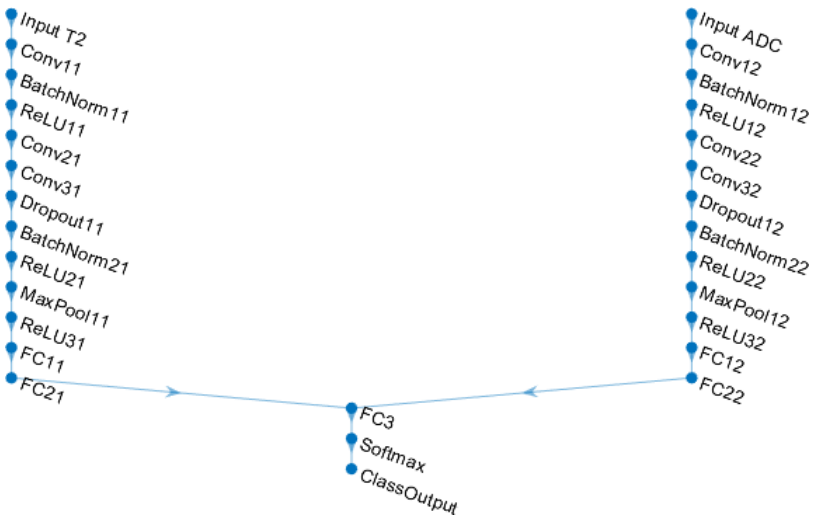

Figure 5: Multimodal network architecture

3.7.1 Results

Table 5 shows the average results achieved by applying the multimodal network to each augmented dataset on the minority class only and on both classes.

Table 5: Comparison between the two approaches for the multimodal network

Dataset	Data augmentation approach	Specificity	Sensitivity	Total accuracy	AUC
Rotation	LG - HG	92%	87%	89,8%	0.95
	HG only	94,4%	85,6%	90,2%	0.95
Vertical flip	LG - HG	90%	89,4%	89%	0.95
	HG only	91,4%	85,8%	89,4%	0.96
Horizontal flip	LG - HG	88,4%	84,8%	87,4%	0.96
	HG only	85%	93,4%	87,6%	0.95
Translation	LG - HG	88,4%	94,8%	90,6%	0.96
	HG only	83,8%	89,8%	85,8%	0.96
Perspective transformation	LG - HG	88,8%	93,2%	89,4%	0.95
	HG only	87%	93,4%	89%	0.96
Scale	LG - HG	94%	67,2%	85%	0.92
	HG only	92,6%	89,8%	91,6%	0.97
Shear	LG - HG	92,6%	78,2%	87,6%	0.92
	HG only	92,2%	87,2%	90,2%	0.95
Rotation + Horizontal flip	LG - HG	89,4%	82%	86,8%	0.95
	HG only	95,2%	87%	92,2%	0.98

In this case, the best results were obtained on the augmented dataset using the combination of rotation and horizontal flipping to both LG and HG classes, which provides a specificity of 95.2%, a sensitivity of 87%, a total accuracy of 92.2% and an AUC of 0.98.

3.8 Testing the networks with PROSTATEX-2 dataset

Ultimately, it was also interesting to validate the networks on a completely unknown dataset, consisting of images acquired with a different protocol with respect to the one the models were trained on. For this purpose, both the single and the multimodal networks were tested on images of the open source dataset of PROSTATEx-2 challenge. Only the images containing peripheral tumors were selected, which is the only kind the networks were trained on, obtaining a total of 43 LG and 6 HG images.

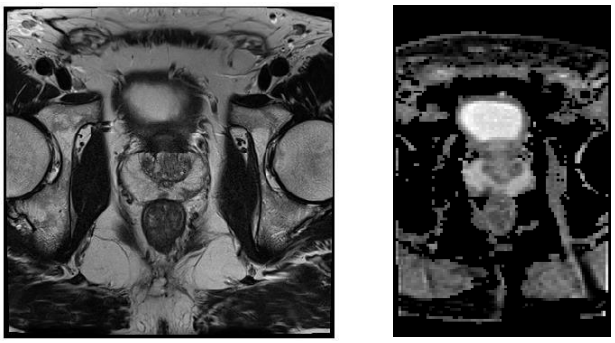

Figure 6: Images from PROSTATEx-2 challenge dataset. (a) T2W image (b) ADC map

The LIME algorithm was also applied, in order to understand if the results obtained with Careggi's dataset are also confirmed by the PROSTATEx-2 one.

3.8.1 Results

Below, for each of the five networks trained using a specific data augmentation technique, applied both to the minority class only and to both classes, the one that provided the best results and their values are reported. Table 6, Table 7, and Table 8 consider the network trained on T2W images only, on ADC maps only and the multimodal network respectively.

Table 6: Best results obtained for each data augmentation technique by testing the network on the T2W PROSTATEx-2 image dataset

Dataset	Data augmentation approach	Specificity	Sensitivity	Total accuracy	AUC	Training
Not augmented dataset	-	55% (24/43)	66% (4/6)	57%	0.66	1
Rotation	LG - HG	32% (14/43)	83% (5/6)	38%	0.60	5
	HG only	25% (11/43)	83% (5/6)	32%	0.66	3
Vertical flip	LG - HG	76% (33/43)	50% (3/6)	73%	0.62	2
	HG only	37% (16/43)	66% (4/6)	40%	0.65	2
Horizontal flip	LG - HG	62% (27/43)	50% (3/6)	61%	0.55	1
	HG only	39% (17/43)	66% (4/6)	42%	0.64	1

Translation	LG - HG	65% (28/43)	66% (4/6)	65%	0.67	2
	HG only	58% (25/43)	50% (3/6)	57%	0.63	1
Perspective transformation	LG - HG	65% (28/43)	66% (4/6)	53%	0.65	1
	HG only	53% (23/43)	50% (3/6)	53%	0.54	3
Scale	LG - HG	41% (18/43)	83% (5/6)	46%	0.74	4
	HG only	25% (11/43)	83% (5/6)	34%	0.62	3
Shear	LG - HG	51% (22/43)	83% (5/6)	55%	0.67	4
	HG only	44% (19/43)	66% (4/6)	46%	0.64	3
Rotation + Horizontal flip	LG - HG	44% (19/43)	66% (4/6)	46%	0.60	5
	HG only	44% (19/43)	66% (4/6)	46%	0.60	5

Table 7: Best results obtained for each data augmentation technique by testing the network on the ADC PROSTATEx-2 image dataset

Dataset	Data augmentation approach	Specificity	Sensitivity	Total accuracy	AUC	Training
Not augmented dataset		76% (33/43)	16% (1/6)	69%	0.39	2
Rotation	LG - HG	55% (24/43)	33% (2/6)	53%	0.43	2
	HG only	34% (15/43)	66% (4/6)	38%	0.42	5
Vertical flip	LG - HG	44% (19/43)	33% (2/6)	42%	0.35	3
	HG only	27% (12/43)	66% (4/6)	32%	0.34	1
Horizontal flip	LG - HG	48% (21/43)	33% (2/6)	46%	0.42	3

	HG only	11% (5/43)	66% (4/6)	18%	0.34	4
Translation	LG - HG	58% (25/43)	50% (3/6)	57%	0.47	4
	HG only	72% (31/43)	16% (1/6)	65%	0.47	4
Perspective transformation	LG - HG	72% (31/43)	16% (1/3)	26%	0.45	2
	HG only	18% (8/43)	50% (3/6)	22%	0.34	5
Scale	LG - HG	83% (36/43)	0% (0/6)	73%	0.37	3
	HG only	39% (17/43)	33% (2/6)	38%	0.27	2
Shear	LG - HG	30% (13/43)	83% (5/6)	36%	0.37	3
	HG only	11% (11/43)	66% (4/6)	18%	0.35	3
Rotation + Horizontal flip	LG - HG	39% (17/43)	50% (3/6)	40%	0.41	3
	HG only	0% (0/43)	100% (6/6)	12%	0.29	1

Table 8: Best results obtained for each data augmentation technique by testing the multimodal network on the PROSTATEx-2 dataset

Dataset	**Data augmentation approach**	**Specificity**	**Sensitivity**	**Total accuracy**	**AUC**	**Training**
Not augmented dataset		72% (31/43)	66% (4/6)	71%	0.79	4
Rotation	LG - HG	58% (25/43)	83% (5/6)	61%	0.82	2
	HG only	41% (18/43)	83% (5/6)	46%	0.71	1
Vertical flip	LG - HG	27% (12/43)	66% (4/6)	32%	0.69	1
	HG only	44% (19/43)	66% (4/6)	46%	0.71	3

Horizontal flip	LG - HG	53% (23/45)	100% (6/6)	59%	0.74	4
	HG only	34% (15/43)	100% (6/6)	42%	0.70	3
Translation	LG - HG	55% (24/43)	83% (5/6)	59%	0.80	1
	HG only	25% (11/43)	83% (5/6)	32%	0.79	3
Perspective transformation	LG - HG	55% (24/43)	83% (5/6)	59%	0.72	3
	HG only	37% (16/43)	83% (5/6)	42%	0.83	1
Scale	LG - HG	79% (34/43)	50% (3/6)	75%	0.72	5
	HG only	27% (12/43)	83% (5/6)	36%	0.73	3
Shear	LG - HG	65% (28/43)	83% (5/6)	67%	0.77	1
	HG only	60% (26/43)	66% (4/6)	61%	0.70	2
Rotation + Horizontal flip	LG - HG	*67% (29/43)*	*100% (6/6)*	*71%*	*0.78*	5
	HG only	*32% (14/43)*	*100% (6/6)*	*40%*	*0.69*	3

Regarding the interpretation aspect, Table 9, and Table 10 show the average percentages of the correctly displayed images with respect to the total amount of the correctly classified ones, for T2W images and ADC maps respectively.

Table 9: Percentage of T2W images from PROSTATEx-2 dataset in which the tumor is correctly displayed for each data augmentation technique, following both the augmentation approaches

	Data augmentation approach	**Average percentage of images correctly displayed by the network**
Not augmented dataset		64%
Rotation	LG - HG	53%
	HG only	69%
Vertical flip	LG - HG	58%
	HG only	60%

Horizontal flip	LG - HG	53%
	HG only	67%
Translation	LG - HG	61%
	HG only	71%
Perspective transformation	LG - HG	59%
	HG only	62%
Scale	LG - HG	48%
	HG only	50%
Shear	LG - HG	52%
	HG only	52%
Rotation + Horizontal flip	LG - HG	61%
	HG only	61%

Table 10: Percentage of ADC maps from PROSTATEx-2 dataset in which the tumor is correctly displayed for each data augmentation technique, following both the augmentation approaches

Dataset	**Data augmentation approach**	**Average percentage of images correctly displayed by the network**
Not augmented dataset		100%
Rotation	LG - HG	88%
	HG only	74%
Vertical flip	LG - HG	81%
	HG only	69%
Horizontal flip	LG - HG	91%
	HG only	78%
Translation	LG - HG	79%
	HG only	78%
Perspective transformation	LG - HG	84%
	HG only	90%
Scale	LG - HG	94%
	HG only	89%
Shear	LG - HG	100%
	HG only	78%
Rotation + Horizontal flip	LG – HG	85%
	HG only	67%

The best classification performance was provided by applying the multimodal network to the augmented dataset with only the rotation applied to both LG and HG classes. Results include a specificity of 58%, a sensitivity of 83%, a total accuracy of 61% and an AUC of 0.82. The case of augmented dataset with horizontal flipping and rotation to both classes should also be mentioned, which in the face of a lower AUC (0.78) provides a perfect sensitivity (100%) and good specificity (67%).

For what concerns LIME algorithm application instead, the results obtained with Careggi's dataset were not confirmed by PROSTATEx-2 dataset. In fact, in this case the augmentation technique that most increases the network's ability of visualizing the tumor lesion is the translation applied to the minority class only for T2W images, and the shear applied to both classes for ADC maps.

4. Conclusion

The aim of the work was to develop and interpret neural models that correctly classify datasets characterized by an imbalance between classes, evaluating different data augmentation techniques to understand the impact these have in the biomedical context. A single network that separately classifies T2W and ADC images was initially tested, and its decision criterion was investigated using LIME algorithm. Subsequently, a multimodal network was implemented, which combines the two image modalities to increase generalization capabilities.

Relying on the results, no data augmentation technique absolutely prevails over the others, since much depends on the application context. However, in most cases, the best results are obtained with the rotation technique alone or combining it with the horizontal flipping, applied to both classes. On the other hand, these techniques are also those that, by examining the results obtained with the LIME algorithm, in most cases worsen the network's ability to correctly visualize the tumor. The techniques that, instead, increase this ability are the vertical flipping applied to both classes for both T2W images and ADC maps in Careggi's dataset, and translation applied to HG class only, and shear applied to both classes, for T2W images and ADC maps respectively in the PROSTATEx-2 dataset.

In any case, it can be stated that the multimodal network, compared to the single network, is the one that provides the best classification performance, as well as generalization, obtaining good results also on the PROSTATEx-2 dataset.

Ideas for future developments may concern: experimentation of different balancing techniques; training with datasets including images belonging to different acquisition protocols; repetition of the experiments using a larger dataset; use of other architectures for classification; modification of the LIME algorithm in order to apply it to multimodal networks.

References

[1] M. T. Ribeiro, S. Singh e C. Guestrin, «"Why Should I Trust You?" Explaining the Predictions of Any Classifier,» in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, 2016.

[2] M. H. Le, Chen e Wang, «Automated diagnosis of prostate cancer in multiparametric MRI based on multimodal convolutional neural networks,» *Physics in Medicine & Biology,* 2017.

[3] B. Abraham e M. S. Nair, «Computer-aided diagnosis of clinically significantprostate cancer from MRI images using sparseautoencoder and random forest classifier,» *Biocybernetics and Biomedical Engineering,* 2018.

[4] H. He, B. Yang, E. A. Garcia e S. Li, «ADASYN: Adaptive synthetic sampling approach for imbalanced learning,» in *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence),* Hong Kong, China, 2008.

[5] Y. Yuan, W. Qin, M. Buyyounouski, B. Ibragimov, S. Hancock, B. Han e L. Xing, «Prostate cancer classification with multiparametric MRI transfer learning model,» *Medical Physics,* pp. 756-765, 2019.

[6] B. Abraham e M. S. Nair, «Automated grading of prostate cancer using convolutional neural network and ordinal class classifier,» *Informatics in Medicine Unlocked,* n. 17, 2019.