# Learning from Only Positive and Unlabeled Data to Detect Lesions in Vascular CT Images

Maria A. Zuluaga[1,2], Don Hush[3], Edgar J.F. Delgado Leyton[1,2], Marcela Hernández Hoyos[2], and Maciej Orkisz[1]

[1] CREATIS; Université de Lyon; Université Lyon 1; INSA-Lyon; CNRS UMR5220; INSERM U1044; F-69621 Villeurbanne, France
[2] Grupo Imagine, Grupo de Ingeniería Biomédica, Universidad de los Andes, Bogotá, Colombia
[3] Los Alamos National Laboratory, Los Alamos, NM 87545, USA

**Abstract.** Detecting vascular lesions is an important task in the diagnosis and follow-up of the coronary heart disease. While most existing solutions tackle calcified and non-calcified plaques separately, we present a new algorithm capable of detecting both types of lesions in CT images. It builds up on a semi-supervised classification framework, in which the training set is made of both unlabeled data and a small amount of data labeled as normal. Our method takes advantage of the arrival of newly acquired data to re-train the classifier and improve its performance. We present results on synthetic data and on datasets from 15 patients. With a small amount of labeled training data our method achieved a 89.8% true positive rate, which is comparable to state-of-the-art supervised methods, and the performance can improve after additional iterations.

## 1 Introduction

As the evaluation of coronary lesions is challenging and tedious, and acquiring moderate expertise in coronary CT angiography (CTA) may take more than one year [1], a variety of methods has been proposed to perform this detection automatically. Most of the algorithms have been directed towards abnormality modeling, *i.e.* identifying the particularities of lesions. As the latter are heterogeneous by nature, and obtaining a model that copes with all possible abnormalities is difficult, most approaches have tackled the segmentation of one type of lesion: calcified plaques [2] or soft plaques [3]. However, these methods often rely on different image acquisition techniques for each task (*e.g.* non-enhanced CT for calcium quantification, contrast-enhanced CT for lumen segmentation and Dual-Source CT for soft plaque locating), which makes it difficult to combine them, in order to simultaneously tackle the automated detection of both types of plaques. Recently, a novel family of methods has made use of machine learning techniques to simultaneously detect the different forms of lesions [4]. Since such methods use supervised classification schemes, the set of examples used for training has to be highly reliable. Unfortunately, it is very expensive to collect labeled data that are accurate, as well as representative of all types of lesions.

A first attempt to use an unsupervised scheme has been proposed in [5]. However, it failed to distinguish bifurcations from actual lesions.

To overcome these shortcomings, here we propose to use a semi-supervised classification scheme that focuses in the healthy vessel sections. It permits both calcified and non-calcified lesions to be identified as the complement of the healthy sections. To do so, we introduced a classification algorithm belonging to a family named Learning from only Positive and Unlabeled data (LPU). Only a relatively small amount of healthy sections must be provided at the learning stage of this algorithm. Its use to identify vascular lesions is novel to the best of our knowledge. We also proposed a strategy that exploits new data that daily arrive in a clinical environment, in order to refine the learning and thus improve the classification performance. Our software has been made publicly available[1].

## 2   Method

In the context of lesion identification, classification methods try to differentiate between two main classes: the healthy and the diseased one. However, while the appearance of healthy vascular sections does not vary much, the appearance of the diseased ones may show a large variability. In other words, the *healthy* class is likely to form a dense cluster in the feature space, while the *diseased* class is represented by sparse points rather than by a cluster. Moreover, we consider that obtaining reliable labels of only healthy vessel sections is an easier task than obtaining representative examples of all types of diseased vessels. Based on these two statements, we addressed the vascular lesion identification through an LPU framework using support vector machines (SVM) [6]. Such an algorithm can be applied to the problems where the training input data is made up of labeled samples (the healthy class) and a large amount of unlabeled samples coming from the mixture (healthy and diseased samples).

### 2.1   Learning from Only Positive and Unlabeled Samples

Let $Q = (q_1, ..., q_k), q_i \in \mathcal{X}$ be a collection of samples generated according to a probability distribution $P_q$ and $X = (x_1, ..., x_n), x_i \in \mathcal{X}$ a set generated according to a probability distribution $P_x$. The goal is to identify the elements of $X$ that are similar to the elements of $Q$. It is assumed that elements in $Q$ belong to a particular class (our healthy class), and that elements belonging to $X$ are actually formed by two different distributions. Some are generated by $P_q$ (healthy samples) and the rest (the diseased samples) by another process $P_{other}$. Therefore, $P_x$ is a mixture distribution:

$$P_x = \beta P_q + (1 - \beta) P_{other} \qquad \text{where } 0 < \beta < 1. \tag{1}$$

Porter *et al.* [6] have defined a relative content density function $p$ that can be used as a similarity measure to quantify the relative concentration of $P_q$ with respect

**Table 1.** The LPU algorithm

| Input | : Q set- Samples of healthy vessel cross-sections |
|---|---|
| Input | : X set- Samples of unlabeled vessel cross-sections |
| 1 | : Assign y=1 labels to Q |
| 2 | : Assign y=-1 labels to X |
| 3 | : Assign the y=1 class weight of a SVM to $2\beta$ |
| 4 | : Assign the y=-1 class weight of a SVM to 1-$2\beta$ |
| 5 | : Train the SVM |
| | Select $f$ that minimizes $\mathcal{R}(f)$ |
| 6 | : Classify samples in X using the trained model M |
| **Output** | : Classified X samples |

to $P_x$, and which is given by the Radon-Nikodym derivative $p = dP_q/dP_x$. Given a threshold $\rho$, the set $\{x \in X : p(x) > \rho\}$ contains samples that are more likely to be generated by $P_q$ than by $P_{other}$.

The latter problem can be solved as a density level detection problem, where a function $f$ is constructed so that it approximates the set $\{p > \rho\}$ by means of the set $\{f > 0\}$. Steinwart *et al.* [7] have proposed a risk function $\mathcal{R}$ that can be estimated from sample data, and serves as a criterion to assess the quality of the approximation of $\{p > \rho\}$ by $\{f > 0\}$:

$$\mathcal{R}(f) = \frac{2\beta}{1 + 2\beta} \frac{1}{|Q|} \sum_{x \in Q} I(f(x) \le 0) + \frac{1}{1 + 2\beta} \frac{1}{|X|} \sum_{x \in X} I(f(x) > 0), \quad (2)$$

where $\rho = \frac{1}{2\beta}$, $I(\cdot) = 1$ if the argument is true and 0 otherwise.

The main advantage of defining such a risk function $\mathcal{R}$ is that it allows us to use a SVM in order to choose a function $f$ that minimizes $\mathcal{R}$ [7]. For this purpose, a surrogate problem is constructed by automatically assigning labels $y$ to the available data. The required steps to build the surrogate problem and solve the LPU algorithm are outlined in Table 1.

Let us note however, that the LPU formulation is theoretically valid in the infinite sample limit, *i.e.* $|X| = \infty$. Additionally, the problem we tackle is highly unbalanced, *i.e* the cardinality of one class is much larger than that of the other one (in our case, this corresponds to a true $\rho$ value close to 1). With unbalanced classes most learning methods tend to favor a response that assigns all samples to one class, which can worsen when combined with finite sample effects. Since infinite samples cannot be achieved, it is at least desired that $|X| \gg |Q|$.

## 2.2   LPU in a Clinical Environment

In the LPU formulation unlabeled data makes part of the training set. This key difference w.r.t. supervised classification approaches is an advantage in two aspects. First, the training set can be easily augmented since no labels are required for the $X$ set. Second, it is desirable and feasible to increase the training set without great risk of overfitting.
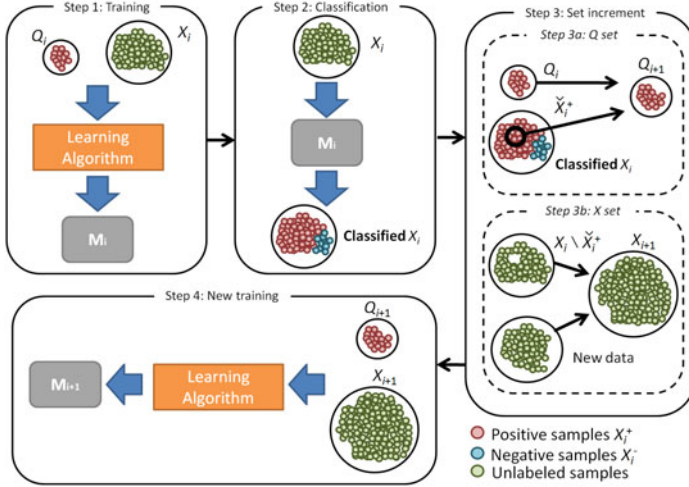
**Fig. 1.** LPU in a clinical environment. Positive samples are those identified as healthy (either by the algorithm or an expert), negative ones are those classified as diseased, and unlabeled ones are samples that belong to the $X$ set before classification.

Based on this key feature, we proposed to progressively increase the two sets that make up the training data set: the addition of samples to $Q$ aimed at normality description improvement, whereas the increase of $X$ kept the relation $|X| \gg |Q|$ and reduced the finite sample effect. Consequently, we suggested to periodically re-train the model $M$ after new data incorporation, instead of keeping it static. Our proposal suits the clinical data workflow where new unlabeled data arrive daily. Moreover, no additional labeling is required to exploit the arriving data. The algorithm (Fig. 1) starts with an initial pair of sets, unlabeled data $X_1$ and (manually) labeled data $Q_1$, then it iterates as follows:

1. **Training:** A pair of sets $Q_i$ and $X_i$ is used to train a model $M_i$.
2. **Classification:** The model $M_i$ is used to label the data from $X_i$. A clinician validates the labels corresponding to actual lesions (subset denoted $X_i^-$).
3. **Set increase:** This step is performed for both $Q$ and $X$ sets.
   (a) Among the samples that the clinician did not consider as lesions (subset denoted $X_i^+$) a sub-subset $\check{X}_i^+$ is randomly removed from the set $X_i$ and combined with the set $Q_i$ to build up the increased set $Q_{i+1} = Q_i \cup \check{X}_i^+$.
   (b) At the arrival of newly acquired data, this is combined with the subset $X_i \setminus \check{X}_i^+$ to make up the increased set $X_{i+1}$.
4. **Loop:** The algorithm jumps to the step 1 in order to obtain a new refined model $M_{i+1}$ by using the increased sets $Q_{i+1}$ and $X_{i+1}$ for training. For this purpose, the labels previously assigned to $X_i$ are considered as unavailable.

## 2.3   Features

To select the features that fit our problem we followed some of the guidelines in [8]. We used a combination of metrics that are able to capture both the circular

shape with radially decreasing intensity profiles of healthy vessel cross-sections, and the divergence from such typical patterns. However, we used a different selection strategy based on the empirical risk $\mathcal{R}$ minimization. Four metrics calculated in cross-sections orthogonal to the vessel centerline (Concentric rings [5], Core, Hessian eigenvalues and Flux [9]) were kept.

## 3   Experiments and Results

The LPU algorithm was solved using the `LIBSVM` software [10] with a Gaussian radial basis function (RBF) kernel. The SVM regularization term $\lambda$ and the $\sigma^2$ parameter of the RBF kernel were optimized through a grid search. Different $\beta \in [0.1, 0.2, ...0.80, 0.81, 0.82, .., 0.98, 0.99]$ values were tested. The selection of the optimal $\beta$ value was included in the optimization process. For every $\{\lambda, \sigma, \beta\}$ combination, the empirical risk $\mathcal{R}$ associated with $f$, was calculated according to Equation 2. The learned decision function $f$, minimizing the empirical risk, was applied on the $X$ set to evaluate the performance of the method.

We first evaluated our method on synthetic images with known ground truth. We then applied it to 3D cardiac CT data sets, where lesions had been annotated by an expert for the purpose of evaluation only. In our experiments, lesion detection was evaluated using three measures: true positive rate $TPR = \frac{TP}{TP+FN}$, true negative rate $TNR = \frac{TN}{TN+FP}$ and balanced error rate $BER = 1 - \frac{TPR+TPN}{2}$. Here $TP$ denotes the number of correctly classified diseased cross-sections, $TN$ the number of correctly classified healthy cross-sections, while $FP$ and $FN$ respectively are the numbers of falsely classified healthy and diseased cross-section. We preferred the use of BER instead of the commonly used accuracy, since the latter is not very meaningful in highly unbalanced problems as the one we tackle.

### 3.1   Synthetic Data

We first evaluated the performance of our method on 70 artificially generated volumes containing a variety of cases typically encountered in vascular analysis. Phantoms were created using the typical Hounsfield Unit values that are found in CT images for blood, background and plaque components, as well as the typical image dimensions and voxel size. Gaussian noise was also added, resulting in a contrast-to-noise-ratio value of 10.

To demonstrate the effectiveness of the iterative increase of the training set, we simulated five iterations. The cardinality $(|Q|, |X|)$ at each iteration was as follows: (500, 15000), (1000, 19000), (1200, 22000), (1300, 22000), (1600, 23000). A sixth case aimed at illustrating the situation, where the condition $|X| \gg |Q|$ is broken, with $(|Q|, |X|) = (1600, 2000)$.

The LPU algorithm had a good performance in terms of TPR, which is desirable for disease detection (Fig. 2). As more samples were included in the sets, the TNR improved. Since the TPR remained unchanged, the total error (BER) decreased. The 6-th case confirmed the expected poor performance of LPU when the condition $|X| \gg |Q|$ is broken.
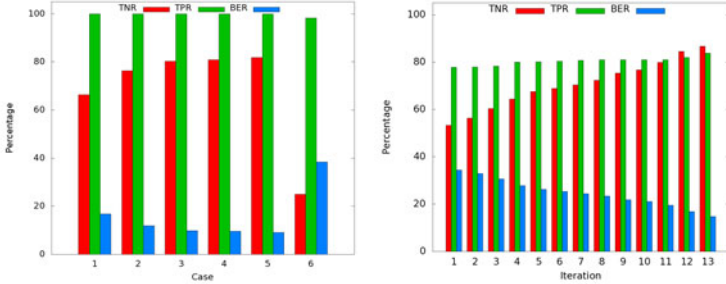
**Fig. 2.** Evolution of TNR, TPR and BER as a consequence of the modification of $Q$ and $X$ datasets. *Left*, 5 iterations in synthetic data, and a 6-th case, where $|X| \gg |Q|$ does not hold. *Right*, 13 iterations using patient data. Final TNR, TPR and BER values were 86.7%, 83.8% and 14.2%, respectively.

## 3.2   Patient Data

Fifteen cardiac CT datasets, with centerlines available in a total of 53 arteries, were obtained from two different sources: Hôpital Louis Pradel (Bron, France) and Rotterdam Coronary Artery Algorithm Evaluation Framework [11], the latter containing data acquired at the Erasmus Medical Center (Rotterdam, The Netherlands). Cross-sections orthogonal to the centerlines were calculated, from which the required features were extracted. Additionally, each cross-section was labeled as normal or abnormal by an observer. These annotations were used to evaluate the performance of the classifier and a small percentage of annotated normal cross-sections were used to build up the initial set $Q_1$. At the first iteration, we used normal data from a healthy subject ($|Q_1| = 426$) and unlabeled cross-sections ($|X_1| = 1148$) from two patients. Due to a limited amount of annotated data available for evaluation, only one patient was added at every new iteration. This allows us to demonstrate the behavior of our method in real data, although in clinical practice several patients may arrive daily. At the final iteration, the cardinalities of the sets were: $|Q_{13}| = 886, |X_{13}| = 12180$.

The evolution of the TNR, TPR and BER as a function of the iterations is presented in Figure 2. Similarly to the synthetic data, the error decreased as new samples were added to $Q$ and $X$. LPU showed a tendency to overestimate the lesions, possibly because the number of labeled samples $|Q| = 886$ used for training was not yet sufficiently representative of all normal configurations. However, despite a small number of training data, the TNR increased from 0.53 to 0.86 as new samples were added, *i.e.* the false alarm rate (1-TNR) decreased from 0.47 to 0.14. From the tendency of TPR, TNR and BER (Fig. 2), we believe that lesion overestimation can be reduced by using LPU with an increased training set. Figure 3 shows examples of classified cross-sections in several coronary arteries.

The same definition of TP has been used in [5], so a direct comparison of the results can be done. In that work, a TPR=0.860 and TNR=0.812 have been reported. From this information we have computed the BER=0.164. While, our new proposal gives a lower TPR=0.838, it has a higher TNR=0.867 that can be
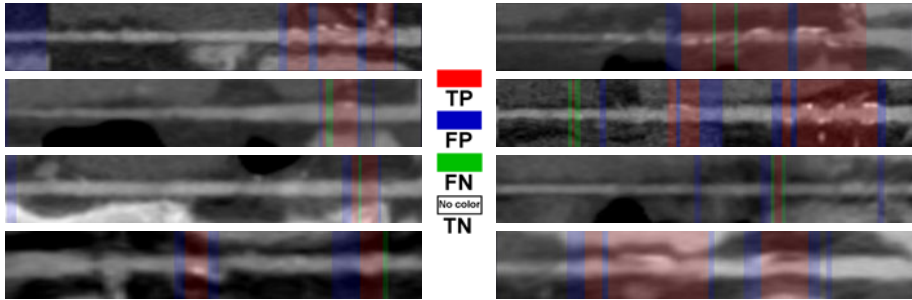
**Fig. 3. Lesion detection results.** Color-coded labels on a stretched curved planar reformatted view of different coronary arteries.

explained by the fact that our novel method does not misclassify bifurcations, contrary to the one in [5]. Our lower BER=0.142 confirms that our overall performance is better. Another related work [4] lacks a clear definition of the evaluation measurement unit, making direct comparison uneasy. Their evaluation has been performed in terms of detected lesions, *i.e.* a lesion is correctly detected (TP) if at least one cross-section within the extent of the lesion is detected as diseased. The available information permitted us to compute their TPR=0.890. Using the latter definition of TP, our approach gave a slightly better TPR=0.898.

## 4    Conclusions

We proposed a new semi-supervised algorithm to detect coronary artery lesions in CTA images. The method can achieve a high detection rate (TPR) even when a small amount of labeled training data is available. Its false alert rate (1-TNR) can be substantially reduced by increasing the training set, as new data arrive and new normal samples are validated. The improvement is guaranteed since adding new samples reduces the finite sample problem. However, recovery is not guaranteed if the user erroneously validates misclassified samples. The computational time increases with the size of the sets $Q$ and $X$. After a number of iterations, the classification error may not significantly decrease despite the inclusion of additional samples. Therefore, the optimal trade-off between computational time and capacity of improvement is to be carefully evaluated. Let us note that the LPU formalism uses only one type of labels (here 'normal') to train the classifier. It is not straightforward to use the diseased samples correctly labeled in previous iterations, in order to train the model in the next iterations. Future work should attempt to overcome this limitation.

# References

1. Pugliese, F., Hunink, M.G.M., Gruszczynska, K., Alberghina, F., Malagó, R., van Pelt, N., Mollet, N.R., Cademartiri, F., Weustink, A.C., Meijboom, W.B., Witteman, C.L.M., de Feyter, P.J., Krestin, G.P.: Learning curve for coronary CT angiography: what constitutes sufficient training? Radiology 251, 359–368 (2009)
2. Saur, S.C., Alkadhi, H., Desbiolles, L., Székely, G., Cattin, P.C.: Automatic detection of calcified coronary plaques in computed tomography data sets. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 170–177. Springer, Heidelberg (2008)
3. Makrogiannis, S., Bhotika, R., Miller, J.V., Skinner Jr., J., Vass, M.: Nonparametric intensity priors for level set segmentation of low contrast structures. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part I. LNCS, vol. 5761, pp. 239–246. Springer, Heidelberg (2009)
4. Tessmann, M., Vega-Higuera, F., Fritz, D., Scheuering, M., Greiner, G.: Multi-scale feature extraction for learning-based classification of coronary artery stenosis. In: Proc. SPIE Med. Imaging, vol. 7260 (2009), doi:10.1117/12.811639
5. Zuluaga, M.A., Magnin, I.E., Hernández Hoyos, M., Delgado Leyton, E.J.F., Lozano, F., Orkisz, M.: Automatic detection of abnormal vascular cross-sections based on Density Level Detection and Support Vector Machines. Int. J. Computer Assisted Radiol. Surg. 6, 163–174 (2011)
6. Porter, R., Ruggiero, C., Hush, D.: Density-based similarity measures for content based search. In: 43rd Asilomar Conf. Signals Systems Computers, pp. 390–394 (2009)
7. Steinwart, I., Hush, D., Scovel, C.: A classification framework for anomaly detection. J. Mach. Learning Research 6, 211–232 (2005)
8. Zuluaga, M.A., Delgado Leyton, E.J.F., Hernández Hoyos, M., Orkisz, M.: Feature selection for SVM-based vascular anomaly detection. In: Menze, B., Langs, G., Tu, Z., Criminisi, A. (eds.) Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging. LNCS, vol. 6533, pp. 141–152. Springer, Heidelberg (2010)
9. Lesage, D., Angelini, E.D., Bloch, I., Funka-Lea, G.: A review of 3D vessel lumen segmentation techniques: models, features and extraction schemes. MedIA 13(6), 819–845 (2009)
10. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines (2004), http://www.csie.ntu.edu.tw/~cjlin/libsvm/
11. Schaap, M., Metz, C., van Walsum, T., van der Giessen, A.G., Weustink, A.C., Mollet, N.R., Bauer, C., Bogunovic, H., Castro, C., Deng, X., Dikici, E., O'Donnell, T., Frenay, M., Friman, O., Hernández Hoyos, M., Kitslaar, P.H., Krissian, K., Kühnel, C., Luengo-Oroz, M.A., Orkisz, M., Smedby, Ö., Styner, M., Szymczak, A., Tek, H., Wang, C., Warfield, S.K., Zambal, S., Zhang, Y., Krestin, G.P., Niessen, W.J.: Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms. MedIA 13(5), 704–715 (2009)