Evaluating Classifiers for Atherosclerotic Plaque Component Segmentation in MRI

Arna van Engelen[1(✉)], Marleen de Bruijne[2], Torben Schneider[3], Anouk C. van Dijk[4], M. Eline Kooi[5], Jeroen Hendrikse[6], Aart Nederveen[7], Wiro J. Niessen[2], and Rene M. Botnar[1]

[1] Division of Imaging Sciences, Department of Biomedical Engineering, King's College London, London, UK
arna.van_engelen@kcl.ac.uk
[2] Biomedical Imaging Group Rotterdam, Erasmus MC, Rotterdam, The Netherlands
[3] Philips Healthcare, Guildford, UK
[4] Department of Radiology, Erasmus MC, Rotterdam, The Netherlands
[5] Department of Radiology and Nuclear Medicine, CARIM School for Cardiovascular Diseases, Maastricht University Medical Center, Maastricht, The Netherlands
[6] Department of Radiology, University Medical Center Utrecht, Utrecht, The Netherlands
[7] Department of Radiology, Academic Medical Center Amsterdam, Amsterdam, The Netherlands

Abstract. Segmentation of tissue components of atherosclerotic plaques in MRI is promising for improving future treatment strategies of cardiovascular diseases. Several methods have been proposed before with varying results. This study aimed to perform a structured comparison of various classifiers, training set sizes, and MR image sequences to determine the most promising strategy for methodology development. Five different classifiers (linear discriminant classifier (LDC), quadratic discriminant classifier (QDC), random forest (RF), and support vector classifiers with both a linear (SVM_{lin}) and radial basis function kernel (SVM_{rbf})) were evaluated. We used carotid MRI data from 124 symptomatic patients, scanned in 4 centres with 2 different MRI protocols (45 and 79 patients). Firstly, learning curves of accuracy as a function of increasing training data size showed stabilisation of performance after using ∼10–15 patients for training. Best results were found for LDC, QDC and RF. Intraplaque haemorrhage was most accurately classified in both protocols, and lowest accuracy was found for the lipid-rich necrotic core. Secondly, for LDC and RF it was shown that leaving out different MRI sequences usually negatively affects results for one or more classes. However, leaving out T2-weighted scans did not have a big impact. In conclusion, several classifiers obtain generally good results for classification of plaque components in MRI. Identification of intraplaque haemorrhage is the most promising, and lipid-rich necrotic core remains the most difficult.

M. Valdés Hernández and V. González-Castro (Eds.): MIUA 2017, CCIS 723, pp. 156–168, 2017.
DOI: 10.1007/978-3-319-60964-5_14

1 Introduction

Cardiovascular diseases are the leading cause of death and disability worldwide [1]. In ischaemic strokes and transient ischaemic attacks (TIA), atherosclerosis plays an important causal role. One of the major hazards of an atherosclerotic plaque is plaque rupture, which can result in a clinical event. One way to prevent such events is by performing carotid endarterectomy in which a high-risk plaque is removed surgically. Early diagnosis and risk stratification is therefore important to accurately select patients for treatment.

Plaque composition is considered an important determinant of plaque rupture [2]. However, due to the lack of techniques to derive high-risk plaque characteristics accurately and reproducibly in daily clinical practice, treatment still mainly relies on the degree of arterial narrowing [3]. The imaging modality that has shown most promising results in imaging of plaque composition is magnetic resonance imaging (MRI) [4,5]. Previous research has shown that MRI can visualise plaque characteristics such as a lipid-rich necrotic core and intraplaque haemorrhage [6,7], and that these MRI-derived parameters are predictive of clinical events [8–10]. However, data analysis is complex and primarily done by visual inspection and manual delineation.

Several studies have developed methods for automatic segmentation of plaque components in MRI [11–16], but these still lack accuracy for use in clinical practice. Moreover, due to differences between studies it remains unclear which segmentation methodology is most promising, how much data is required to develop stable and accurate methods, and which MRI protocol should be used for best identification of the most important plaque components. The most common approach has been to perform voxel classification trained on a ground truth dataset with either manual contours or contours from histology, using a set of imaging-derived features, including normalised intensity and Gaussian filters. Classifiers that have been used are linear discriminant classifiers [13–16], non-linear Bayesian classifiers [11,12], and support vector machines [14] with training sets ranging from 12–22 patients. The MRI-protocols all included T1-weighted and T2-weighted scans, and, in most cases proton-density (PD) weighted and bright-blood time of flight (TOF) scans, and in a number of cases included contrast-enhanced scans or a specific scan to identify intraplaque haemorrhage.

This study therefore aims to compare different classifiers and image sequences for the classification of plaque components, to provide insights to develop improved plaque characterisation methods. We evaluate these techniques on two datasets with a slightly different MRI protocol.

2 Methods

2.1 Data

MRI Study. We used data acquired within the multi-centre Parisk study [17]. In this study patients with a recent (<3 months) ischaemic stroke or TIA and a symptomatic 30–69% carotid artery stenosis as defined on ultrasound or CT,

were prospectively recruited. MRI was performed in one of four centres within the Netherlands: Academic Medical Center Amsterdam (centre 1), Erasmus Medical Center, Rotterdam (centre 2), Maastricht University Medical Center (centre 3) and University Medical Center Utrecht (centre 4).

Five MRI sequences were used for plaque characterisation. Patients in centres 1, 3 and 4 were scanned using the same MRI protocol on an Achieva or Ingenia scanner (Philips Healthcare, Best, the Netherlands). We will refer to this protocol as *Protocol 1*. Patients in centre 2 were scanned on a Discovery MR 750 system (GE Healthcare, Milwaukee, MI, USA), (*Protocol 2*).

Both protocols contained a pre-contrast T1-weighted (T1w) and T2-weighted (T2w) scan, a post-contrast T1w scan acquired 6 min after administration of 0.1 mmol/kg body weight of a gadolinium-based contrast medium, and a pre-contrast heavily T1-weighted scan that shows high signal intensity for intraplaque haemorrhage (an 2D inversion-recovery turbo-field echo for Protocol 1, and a 3D spoiled gradient echo for Protocol 2). Protocol 1 contains a 2D bright-blood TOF scan that aims to identify calcifications bordering the lumen. In contrast, in Protocol 2 a 3D fast spoiled gradient echo scan that is specifically aimed at showing all calcifications as hypointense with little or no difference in appearance between other tissues and structures in the image, is used. Due to those differences, in our experiments the data from centres 1, 3 and 4 was taken together, and the data from centre 2 was considered separately. For details on the acquisition parameters, and more details on patient recruitment, we refer to Truijman et al. [17]. For this study we only used patients who completed the five MRI scans above with diagnostic image quality, and whose data was available. This resulted in 9 patients from centre 1, 45 patients from centre 2, 52 patients from centre 3, and 18 patients from centre 4, so in total 79 for Protocol 1 and 45 for Protocol 2.

Manual Annotation. Manual annotation of the symptomatic arteries was performed by a set of six observers, using dedicated software (VesselMASS, Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands). Firstly, the lumen centreline was manually or semi-automatically identified. Then, the other four MRI sequences were registered to the T1w precontrast scan using a region of interest around the centreline [18] and manually adjusted. The vessel wall was subsequently segmented either manually or semi-automatically using a previously published technique [19] and manually adjusted. Plaque components (lipid-rich necrotic core (LRNC), calcification (CA) and intraplaque haemorrhage (IPH)) were fully manually annotated. The remainder of the vessel wall was considered fibrous tissue. All observers received the same training, and annotations were made based on previously published criteria [20–22]. LRNC was defined as a region that shows no contrast-enhancement on the postcontrast T1w scan compared with the precontrast scan, and IPH as hyperintense signal compared to the adjacent sternocleidomastoid muscle in the bulk of the plaque on the IR-TFE or SPGR image, and is considered as part of the LRNC. Calcification was identified as hypointense on at least two sequences, where for Protocol 2 a hypointense signal on the FSPGR scan was the main criterium. For

the first 19 patients recruited in centre 3, annotations were made using MRI-Plaque View (VPDiagnostics Inc., Seattle, WA, USA) for a different study and converted for use in VesselMass as described in van Engelen et al. [15].

2.2 Experiments

Classifiers. Five commonly used classifiers were evaluated:

- Linear Discriminant Classifier (LDC): This relatively simple classifier has successfully been used for voxel classification in atherosclerotic plaques [13–16]. It determines the optimum linear boundaries between classes assuming the data is normally distributed with equal covariance matrices for each class, using the class means, class priors, and covariance matrix [23].
- Quadratic Discriminant Classifier (QDC): This classifier is similar to LDC, except that it does not assume equal covariance for all classes, and thereby allows quadratic instead of linear class boundaries.
- Random Forest (RF): Random forests have more recently become popular in medical imaging [24]. They are formed of a set of decision trees where subsets of features are randomly selected at each node. The predictions of all trees are combined for classification. This provides a data-driven way of feature selection and allows for more flexible decision boundaries. The number of trees was optimised in our experiments, with $\sqrt{}$(nr of features) features selected at each node.
- Linear Support Vector Machine (SVM_{lin}): Support Vector Machines maximise the width of the margin between classes. Therefore, the decision boundary is determined by the samples on the boundary, rather than on the distribution of all data like LDC and QDC. The parameter C that trades-off between maximising the margin and minimising misclassification is optimised in our experiments. Multiclass classification (also for SVM_{rbf}) was performed by combining different 1-vs-1 classifiers.
- Support Vector Machine with radial basis function kernel (SVM_{rbf}): The radial basis function kernel allows for non-linear decision boundaries. The kernel radius, γ, and C are optimised in our experiments.

All experiments were performed in Matlab. LDC and QDC were implemented using the PrTools toolbox [25], SVM_{lin} and SVM_{rbf} using libsvm [26], and RF using a toolbox based on the implementation described in [27].

Data Preparation. Preprocessing and feature computation was carried out in a similar fashion as in [15]. A bias field due to coil inhomogeneity was present in the data from Protocol 2, which was corrected by N4 inhomogeneity correction [28]. Intensity in all MRI sequences for both protocols was normalised by scaling the 5th and 95th percentile in a region of 4×4 cm around the lumen centre between 0 and 1000, on a per-volume basis scaling all slices with the same values.

Image features for classification were based on previous studies [11,13,15]: (1) the normalised intensities for each MRI sequence, (2) the images blurred with

a Gaussian filter ($\sigma = 0.3$ mm), (3) First order (gradient magnitude) and second order (Laplacian) derivatives at the same scale, (4) the Euclidean distance to the lumen and to the outer wall, and the product of those two distances.

Learning Curves. To compare the five different classifiers, learning curves were made to determine the performance with increasing size of training data for each classifier. The aim of this was firstly, to compare the accuracy between different classifiers, and secondly, to establish the amount of required training data and to compare the this between the classifiers.

The data of each protocol was randomly split in three groups. For Protocol 1 two groups of 30 patients were used for training and testing, and the remaining 19 were used to optimise classifier-specific parameters. For Protocol 2 two groups of 20 patients were used for training and testing, and 5 patients for optimisation. The distribution of all four classes was kept similar between the three groups (overall, Protocol 1 had 86% F, 3% LRNC, 5% CA and 6% IPH. Protocol 2 had 82% F, 3% LRNC, 8% CA and 7% IPH.). The two non-optimisation groups for each protocol served both as training and testing data in a two-fold cross-validation. For parameter optimisation, classifiers were trained on the full set of 20 or 30 patients, and the average best accuracy on the optimisation data was determined. For *RF*, the number of decision trees evaluated was 10, 25, 50, 100, 250, 500, 750 and 1000. For SVM, C and (for SVM_{rbf}) γ were evaluated for 0.001, 0.01, 0.1, 1, 10, 100 and 1000.

To create the learning curves, the size of the training set was increased from 1 to 30 (Protocol 1) or 20 (Protocol 2). For training set sizes of 1, 2 and the maximum minus 1, all possibilities were evaluated (so 20 or 30 repeats). For all sizes in between 25 randomly selected combinations of patients were used for training. Those were the same for every classifier. For the maximum training set size only 1 combination, using all training data, was possible. When training on the first fold, patients in the second fold were used for testing and the other way around. Results are presented by averaging over all 60 (Protocol 1) or 40 (Protocol 2) datasets. Overall voxelwise accuracy was determined. Moreover, sensitivity, intra-class correlation coefficients of volumes, and Cohen's kappa (presence/absence in ground truth vs. the result) were determined for all four classes. Since fibrous tissue was present and detected in all cases, its kappa is always 1 and not further presented. Standard deviations were calculated over all repetitions.

MRI Sequences. To determine which sequences are most important in plaque component classification, and to determine whether this is classifier-dependent, classification using the full training sets was repeated after leaving all features of single MRI sequences out. This was performed for two classifiers that were deemed most successful based on the learning curves. Again, sensitivity, ICC and Cohen's kappa were determined and presented for LRNC, CA and IPH. The voxelwise accuracy of classification without each sequence was statistically compared with the full dataset using a Wilcoxon signed ranks test, a non-parametric test for paired differences.

3 Results

3.1 Learning Curves

Parameter optimisation for RF resulted in 250 trees used for Protocol 1, and 100 or 750 trees for Protocol 2 for the two training sets. For SVM_{lin}, for Protocol 1 C was 0.01 for both training sets, and for Protocol 2 C was 0.1 and 0.01. For SVM_{rbf}, for Protocol 1, the parameters for the first training set were $C = 1$ and $\gamma = 0.01$, and for the second set $C = 10$ and $\gamma = 0.01$. For Protocol 2 those were $C = 10$ and $\gamma = 0.001$, and $C = 1$ and $\gamma = 0.01$.

The learning curves for all classifiers are shown in Figs. 1 (Protocol 1) and 2 (Protocol 2). It can be seen that the average overall voxelwise accuracy (bottom-right in the figures) varies little between classifiers, except for a slightly lower performance for QDC, and stabilises after using about ∼10 patients for training. The other curves stabilise after 10–15 datasets (Protocol 1) or 5–10 datasets (Protocol 2), which may be related to all data being acquired in the same centre for Protocol 2. Kappa, which only looks at presence or absence of tissue components, stabilises the quickest. Only for the ICC for LRNC, the smallest class, the plateau might not always be reached. LRNC is better identified in Protocol 1 than in Protocol 2, although the largest differences between classifiers is seen here. LDC, QDC and, for Protocol 1, RF, perform best for LRNC. SVM shows much lower performance, and particularly fails to identify CA in Protocol 1 compared with the other classifiers. For both protocols best results are achieved for IPH, with high ICC and kappa values and reasonable sensitivity. While voxelwise sensitivity for calcification is low in both protocols, ICC is good, with reasonable Kappa values.

3.2 MRI Sequences

LDC and RF were repeated after leaving out each MRI sequence individually. Results are presented in Tables 1 (LDC) and 2 (RF). Leaving out one single sequence has a minor effect on overall classification accuracy for both protocols and classifiers, but more relevant effects on plaque components are seen. For Protocol 1, results for LRNC decrease mostly when the postcontrast (LDC and RF) or precontrast T1w scan (only for LDC) are left out. For Protocol 2, LRNC classification is generally low, but leaving out the SPGR sequence has the biggest effect. Calcification is for both classifiers mostly dependent on the FSPGR scan (Protocol 2) or a combination of several sequences (Protocol 1), for which the TOF scan may be the most important one. Classification of IPH is not strongly affected by leaving out any single sequence, though leaving out the heavility T1-weighted IR-TFE or SPGR has the biggest effect. For both classifiers and protocols, no reduction in performance was seen after leaving out T2w scans.

4 Discussion and Conclusion

We have shown a comparison between five common classifiers on two different MRI datasets. Generally, learning curves show stabilisation of results after

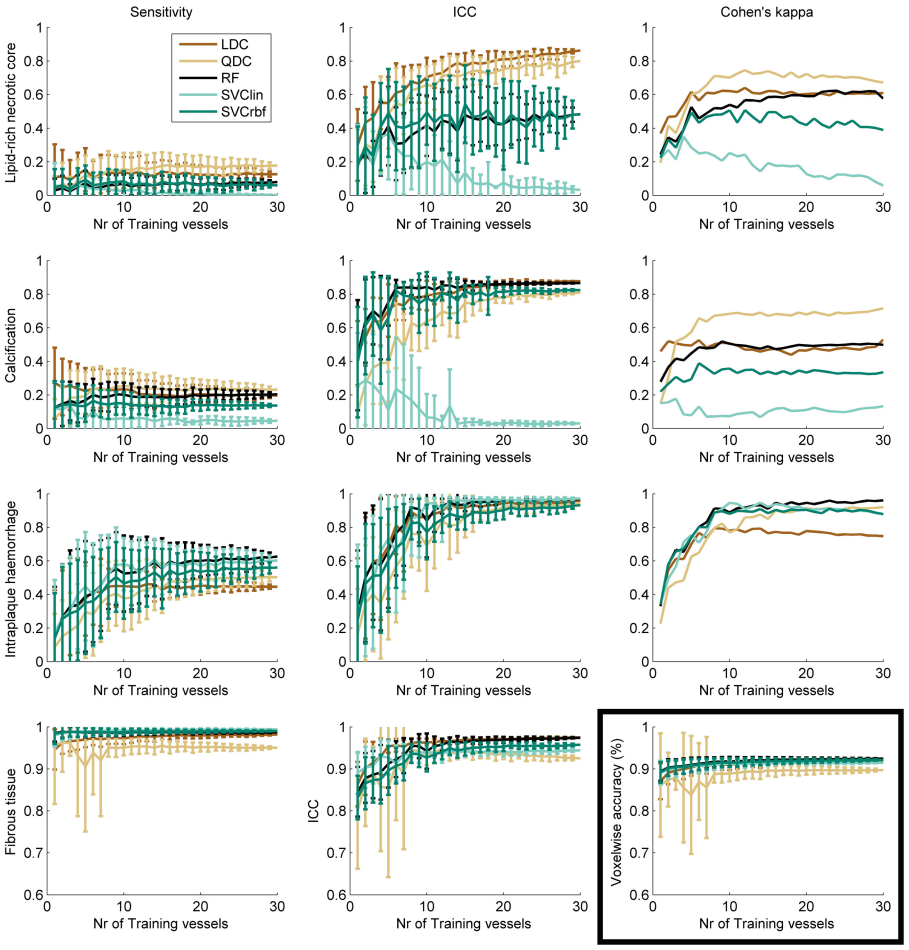

Fig. 1. Learning curves for the data of **Protocol 1**. Cohen's kappa is not shown for fibrous tissue, since it is 1 in all cases. Instead the voxelwise accuracy is shown in the bottom right figure. Note that the figures in the bottom row are scaled from 0.6 to 1 instead of 0 to 1. The error bars indicate the standard deviation for the average of all patients over the number of repetitions.

including 10–20 patients for training. Good results can be obtained for IPH and CA, however, accurate classification of LRNC was shown to be more difficult. The largest differences between classifiers were also seen for LRNC, and, for Protocol 1, CA. LDC, QDC and RF generally showed best performance, while SVM had lower performance.

Lower performance for SVM could be related with more difficult optimisation of the classifier parameters. SVM is likely to suffer more from the considerable class imbalance that was present. Moreover, optimisation on accuracy tends to

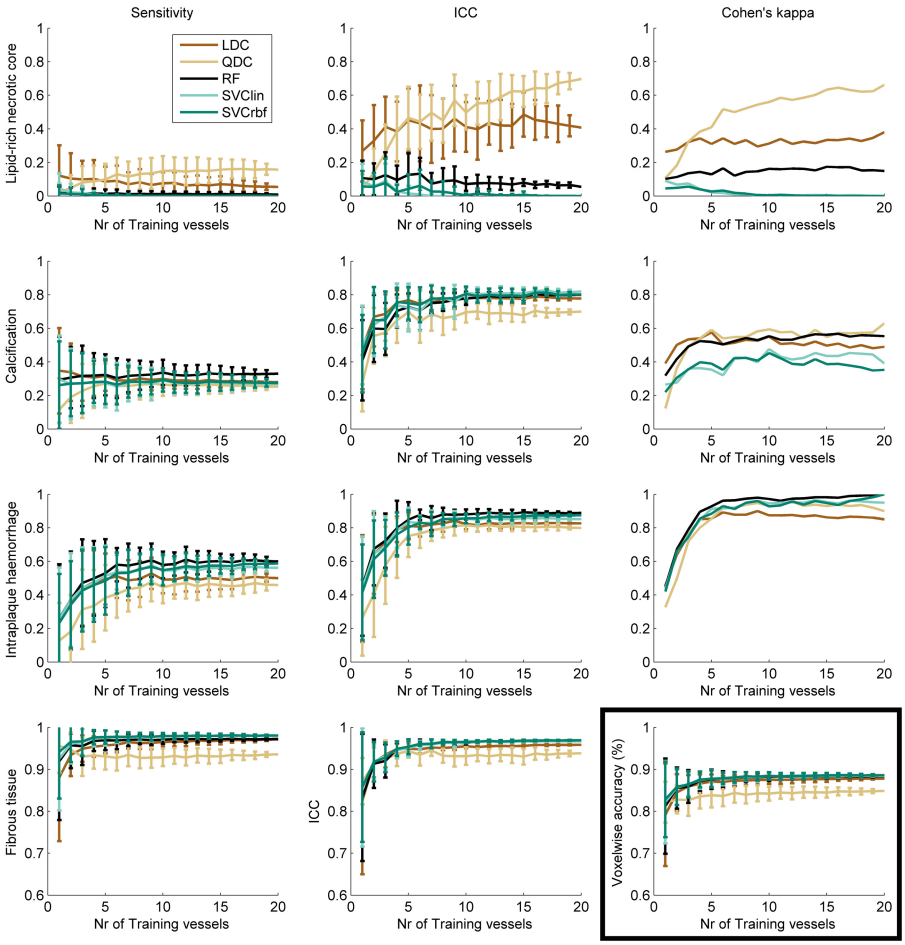

Fig. 2. Learning curves for the data of **Protocol 2**. Cohen's kappa is not shown for fibrous tissue, since it is 1 in all cases. Instead the voxelwise accuracy is shown in the bottom right figure. Note that the figures in the bottom row are scaled from 0.6 to 1 instead of 0 to 1. The error bars indicate the standard deviation for the average of all patients over the number of repetitions.

be biased towards correct classification of fibrous tissue, rather than the other, smaller, classes. A more optimal way for feature selection may be to look at average sensitivity over the four classes, or the F-score, which balances between sensitivity and precision. Furthermore, more differences between the optimised parameters for the two training sets were seen for Protocol 2. This could be due to only five patients being used for optimisation. However, this was chosen to have as much data as possible available to evaluate classification performance.

To improve results for small classes, future research can also investigate the effect of using (more) balanced classes in training, however, measures should be

Table 1. Classification results for varying image protocols for **LDC**. Acc. = accuracy, TOF = time of flight, IR-TFE = inversion-recovery turbo field echo, (F)SPGR = (fast) spoiled gradient echo. *Statistically different from using all features ($p < 0.05$)

	Voxelwise acc. (%)	Sensitivity			ICC			Kappa		
		LRNC	CA	IPH	LRNC	CA	IPH	LRNC	CA	IPH
Protocol 1										
All 5 sequences	92	0.13	0.20	0.45	0.86	0.88	0.95	0.61	0.53	0.75
No postcontrast	91*	0.05	0.16	0.45	0.76	0.84	0.95	0.48	0.40	0.75
No T2w	92	0.12	0.19	0.44	0.86	0.87	0.95	0.61	0.50	0.75
No TOF	92*	0.13	0.12	0.45	0.86	0.79	0.95	0.61	0.30	0.75
No T1w	92*	0.08	0.20	0.45	0.77	0.88	0.95	0.55	0.42	0.75
No IR-TFE	91*	0.10	0.15	0.35	0.81	0.74	0.89	0.58	0.44	0.75
Protocol 2										
All 5 sequences	88	0.05	0.28	0.50	0.41	0.78	0.83	0.38	0.49	0.85
No postcontrast	88	0.05	0.27	0.51	0.35	0.77	0.83	0.23	0.44	0.85
No T2w	88	0.06	0.26	0.51	0.32	0.77	0.83	0.25	0.49	0.90
No FSPGR	87*	0.04	0.12	0.51	0.40	0.77	0.83	0.30	0.16	0.90
No T1w	88	0.04	0.28	0.46	0.30	0.78	0.83	0.25	0.49	0.85
No SPGR	87*	0.03	0.28	0.36	0.25	0.78	0.79	0.17	0.44	0.80

taken to prevent overclassification of small classes in this case. Another reason for suboptimal results for certain classes is that the classes may not be separable with the evaluated features. We have used the ones that have commonly been used for this application in more recent previous studies [13–16]. Other features, such as Gaussian filters at more scales, or texture features, may be interesting to study as well. Instead of evaluating the effect of leaving out individual features, we have evaluated them on a per-MRI-sequence basis. This was chosen because eliminating features on a per-MRI-sequence basis would be advantageous in clinical practice, since it could reduce scan time. In presence of the four other available sequences, leaving out the T2w scan had the smallest effect in our study.

Much more classifiers than the ones evaluated here exist. We have chosen to use the most commonly used ones. Currently, deep learning techniques, using deep neural networks, have gained enormous popularity in image analysis. These techniques need to be considered in future research. Furthermore, both MRI protocols have been considered separately in this study, as previous research has shown that considerable differences exist between them [15]. In future research it could be interesting to see whether some classifiers are better at handling all MRI data combined.

Table 2. Classification results for varying image protocols for **RF**. Acc. = accuracy, TOF = time of flight, IR-TFE = inversion-recovery turbo field echo, (F)SPGR = (fast) spoiled gradient echo. *Statistically different from using all features ($p < 0.05$)

	Voxelwise acc. (%)	Sensitivity			ICC			Kappa		
		LRNC	CA	IPH	LRNC	CA	IPH	LRNC	CA	IPH
Protocol 1										
All 5 sequences	92	0.08	0.20	0.63	0.48	0.87	0.97	0.58	0.50	0.96
No postcontrast	92*	0.02	0.17	0.62	0.25	0.81	0.97	0.61	0.63	0.96
No T2w	92	0.08	0.21	0.62	0.47	0.87	0.97	0.74	0.50	0.96
No TOF	92*	0.09	0.18	0.63	0.53	0.85	0.97	0.61	0.42	0.96
No T1w	92*	0.07	0.20	0.60	0.46	0.87	0.97	0.61	0.50	0.96
No IR-TFE	91*	0.07	0.18	0.33	0.41	0.77	0.78	0.51	0.50	0.79
Protocol 2										
All 5 sequences	88	0.01	0.33	0.60	0.05	0.80	0.89	0.15	0.55	1.00
No postcontrast	88*	0.02	0.34	0.60	0.05	0.79	0.89	0.27	0.55	1.00
No T2w	88*	0.02	0.33	0.60	0.10	0.79	0.89	0.23	0.55	1.00
No FSPGR	87*	0.02	0.15	0.60	0.08	0.73	0.89	0.25	0.39	0.95
No T1w	88*	0.01	0.34	0.60	0.07	0.78	0.89	0.15	0.55	1.00
No SPGR	88*	0.00	0.35	0.36	0.06	0.84	0.83	0.05	0.63	0.95

The most accurate results were obtained for IPH, which also has been considered as one of the most promising imaging characteristics for use in clinical practice due to its high predictive value for future events [9]. This study confirms that also automated techniques can identify IPH well in MRI. Leaving out one single sequence did not have a large effect for IPH. This is probably because two T1-weighted sequences were available in both protocols, so when one if left out the other still provides enough information.

In conclusion, for the evaluated classifiers training set sizes of 15–20 patients are sufficiently large. A simple classifier such as LDC, but also QDC and RF, yields good results. However, improvements can still be made. Especially classification of LRNC remains difficult. Classification of IPH is possible with high accuracy and therefore most promising for implementation into clinical practice.

Acknowledgments. This research has been supported by an EPSRC Technology Strategy Board CR&D Grant (EP/L505304/1). The PARISK study was performed within the framework of the Center for Translational Molecular Medicine (www.ctmm.nl), project PARISK (Plaque At RISK; grant 01C-202) and supported by the Dutch Heart Foundation. This research was also partly funded by the Netherlands Organisation for Scientific Research (NWO).

The Division of Imaging Sciences also receives support from the Centre of Excellence in Medical Engineering (funded by the Welcome Trust and EPSRC; grant number

WT 088641/Z/09/Z) and the Department of Health through the National Institute for Health Research (NIHR) Biomedical Research Centre award to Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London, and by the NIHR Healthcare Technology Co-operative for Cardiovascular Disease at Guys and St Thomas NHS Foundation Trust. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. World Health Organisation: WHO global burden of disease 2000–2015 (2015)
2. Finn, A.V., Nakano, M., Narula, J., Kolodgie, F.D., Virmani, R.: Concept of vulnerable/unstable plaque. Arterioscler. Thromb. Vasc. Biol. **30**(7), 1282–1292 (2010)
3. Paraskevas, K.I., Mikhailidis, D.P., Veith, F.J.: Comparison of the five 2011 guidelines for the treatment of carotid stenosis. J. Vasc. Surg. **55**(5), 1504–1508 (2012)
4. Sanz, J., Fayad, Z.A.: Imaging of atherosclerotic cardiovascular disease. Nature **451**(7181), 953–957 (2008)
5. Huibers, A., de Borst, G., Wan, S., Kennedy, F., Giannopoulos, A., Moll, F., Richards, T.: Non-invasive carotid artery imaging to identify the vulnerable plaque: current status and future goals. Eur. J. Vasc. Endovasc. Surg. **50**(5), 563–572 (2015)
6. Yuan, C., Mitsumori, L.M., Ferguson, M.S., Polissar, N.L., Echelard, D., Ortiz, G., Small, R., Davies, J.W., Kerwin, W.S., Hatsukami, T.S.: In vivo accuracy of multispectral magnetic resonance imaging for identifying lipid-rich necrotic cores and intraplaque hemorrhage in advanced human carotid plaques. Circulation **104**(17), 2051–2056 (2001)
7. Saam, T., Ferguson, M., Yarnykh, V., Takaya, N., Xu, D., Polissar, N., Hatsukami, T., Yuan, C.: Quantitative evaluation of carotid plaque composition by in vivo MRI. Arterioscler. Thromb. Vasc. Biol. **25**(1), 234–239 (2005)
8. Gupta, A., Baradaran, H., Schweitzer, A.D., Kamel, H., Pandya, A., Delgado, D., Dunning, A., Mushlin, A.I., Sanelli, P.C.: Carotid plaque MRI and stroke risk. Stroke **44**(11), 3071–3077 (2013)
9. Saam, T., Hetterich, H., Hoffmann, V., Yuan, C., Dichgans, M., Poppert, H., Koeppel, T., Hoffmann, U., Reiser, M.F., Bamberg, F.: Meta-analysis and systematic review of the predictive value of carotid plaque hemorrhage on cerebrovascular events by magnetic resonance imaging. J. Am. Coll. Cardiol. **62**(12), 1081–1091 (2013)
10. Sun, J., Zhao, X.Q., Balu, N., Neradilek, M.B., Isquith, D.A., Yamada, K., Canton, G., Crouse, J.R., Anderson, T.J., Huston, J., OBrien, K., Hippe, D.S., Polissar, N.L., Yuan, C., Hatsukami, T.S.: Carotid plaque lipid content and fibrous cap status predict systemic CV outcomes: the MRI substudy in AIM-HIGH. JACC: Cardiovasc. Imaging **10**(3), 241–249 (2017)
11. Liu, F., Xu, D., Ferguson, M.S., Chu, B., Saam, T., Takaya, N., Hatsukami, T.S., Yuan, C., Kerwin, W.S.: Automated in vivo segmentation of carotid plaque MRI with morphology-enhanced probability maps. Magn. Reson. Med. **55**(3), 659–668 (2006)
12. Hofman, J., Branderhorst, W., ten Eikelder, H., Cappendijk, V., Heeneman, S., Kooi, M., Hilbers, P., ter Haar Romeny, B.: Quantification of atherosclerotic plaque components using in vivo MRI and supervised classifiers. Magn. Reson. Med. **55**(4), 790–799 (2006)

13. van't Klooster, R., Naggara, O., Marsico, R., Reiber, J., Meder, J.F., van der Geest, R., Touz, E., Oppenheim, C.: Automated versus manual in vivo segmentation of carotid plaque MRI. Am. J. Neuroradiol. **33**, 1621–1627 (2012)
14. van Engelen, A., Niessen, W.J., Klein, S., Groen, H.C., Verhagen, H.J.M., Wentzel, J.J., van der Lugt, A., de Bruijne, M.: Atherosclerotic plaque component segmentation in combined carotid MRI and CTA data incorporating class label uncertainty. PLOS ONE **9**(4), 1–14 (2014)
15. van Engelen, A., van Dijk, A., Truijman, M., van't Klooster, R., van Opbroek, A., van der Lugt, A., Niessen, W., Kooi, M., de Bruijne, M.: Multi-center MRI carotid plaque component segmentation using feature normalization and transfer learning. IEEE Trans. Med. Imaging **34**(6), 1294–1305 (2015)
16. Gao, S., van't Klooster, R., van Wijk, D.F., Nederveen, A.J., Lelieveldt, B.P.F., van der Geest, R.J.: Repeatability of in vivo quantification of atherosclerotic carotid artery plaque components by supervised multispectral classification. Magn. Reson. Mater. Phys., Biol. Med. **28**(6), 535–545 (2015)
17. Truijman, M., Kooi, M., van Dijk, A., de Rotte, A., van der Kolk, A., Liem, M., Schreuder, F., Boersma, E., Mess, W., van Oostenbrugge, R., Koudstaal, P., Kappelle, L., Nederkoorn, P., Nederveen, A., Hendrikse, J., van der Steen, A., Daemen, M., van der Lugt, A.: Plaque At RISK (PARISK): prospective multicenter study to improve diagnosis of high-risk carotid plaques. Int. J. Stroke **9**, 747–754 (2013)
18. van't Klooster, R., Staring, M., Klein, S., Kwee, R.M., Kooi, M.E., Reiber, J.H.C., Lelieveldt, B.P.F., van der Geest, R.J.: Automated registration of multispectral MR vessel wall images of the carotid artery. Med. Phys. **40**(12), 121904 (2013)
19. van't Klooster, R., de Koning, P.J., Dehnavi, R.A., Tamsma, J.T., de Roos, A., Reiber, J.H., van der Geest, R.J.: Automatic lumen and outer wall segmentation of the carotid artery using deformable three-dimensional models in MR angiography and vessel wall images. J. Magn. Reson. Imaging **35**(1), 156–165 (2012)
20. Cai, J., Hatsukami, T.S., Ferguson, M.S., Kerwin, W.S., Saam, T., Chu, B., Takaya, N., Polissar, N.L., Yuan, C.: In vivo quantitative measurement of intact fibrous cap and lipid-rich necrotic core size in atherosclerotic carotid plaque: Comparison of high-resolution, contrast-enhanced magnetic resonance imaging and histology. Circulation **112**(22), 3437–3444 (2005)
21. Cappendijk, V.C., Heeneman, S., Kessels, A.G., Cleutjens, K.B., Schurink, G.W.H., Welten, R.J., Mess, W.H., van Suylen, R.J., Leiner, T., Daemen, M.J., van Engelshoven, J.M., Kooi, M.E.: Comparison of single-sequence T1w TFE MRI with multisequence MRI for the quantification of lipid-rich necrotic core in atherosclerotic plaque. J. Magn. Reson. Imaging **27**(6), 1347–1355 (2008)
22. Kwee, R.M., van Engelshoven, J.M., Mess, W.H., ter Berg, J.W., Schreuder, F.H., Franke, C.L., Korten, A.G., Meems, B.J., van Oostenbrugge, R.J., Wildberger, J.E., Kooi, M.E.: Reproducibility of fibrous cap status assessment of carotid artery plaques by contrast-enhanced MRI. Stroke **40**(9), 3017–3021 (2009)
23. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning, Corrected edn. Springer, New York (2003)
24. Criminisi, A., Shotton, J., Konukoglu, E.: Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. NOW Publishers, Breda (2012)
25. Duin, R., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D., Tax, D., Verzakov, S.: PRTools4.1, A Matlab Toolbox for Pattern Recognition. Delft University of Technology (2007)

26. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011)
27. Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **2**(3), 18–22 (2002)
28. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging **29**(6), 1310–1320 (2010)