ACTIVE LEARNING FOR MAGNETIC RESONANCE IMAGE QUALITY ASSESSMENT

Annika Liebgott[⋆] *Thomas Küstner*[⋆†] *Sergios Gatidis* [†] *Fritz Schick* [‡] *Bin Yang* [⋆]

[⋆] Institute of Signal Processing and System Theory, University of Stuttgart, Stuttgart, Germany
[†] Department of Radiology, Universtity of Tübingen, Tübingen, Germany
[‡] Section on Experimental Radiology, University of Tübingen, Tübingen, Germany

ABSTRACT

In medical imaging, the acquired images are usually analyzed by a human observer and rated with respect to a diagnostic question. However, this procedure is time-demanding and expensive. Furthermore, the lack of a reference image makes this task challenging. In order to support the human observer in assessing image quality and to ensure an objective evaluation, we extend in this paper our previous no-reference magnetic resonance (MR) image quality assessment system with an active learning loop to reduce the amount of necessary labeled training data. We employ two different active learning query strategies based on uncertainty sampling. Since the classification task is performed on 2D image slices, but the human observer labels complete 3D image volumes, we present a method to select representative 3D images instead of independant 2D image slices. The performance is evaluated on *in-vivo* MR image data.

Index Terms— active learning, blind image quality assessment, machine-learning, magnetic resonance imaging

1. INTRODUCTION

Magnetic Resonance Imaging (MRI) is a widely used imaging modality in today's clinical diagnostic. It offers a variety of imaging possibilities, including different contrast mechanism or real-time imaging, by which one can visualize both anatomical structure and physiological functions inside the human body. The immense and flexible MR sequence and reconstruction parametrization makes it on one hand very tunable to specific needs and applications but demands on the other hand a profound knowledge. Besides its various advantages, MR images are often prone to artifacts originating from hardware imperfections, like magnetic field inhomogeneities, or patient variabilities, like respiratory/cardiac movement. Furthermore, nowadays enormous amounts of data are created per patient which makes the diagnostic reading a very time-demanding task.

Up to day image quality evaluation has been a manual process which mainly depends on human observers (HO) like trained physicians or experienced radiologists, to determine the underlying image quality with respect to a certain question is a very time consuming and cost-intensive task. Moreover, the considered quality criteria need to be clarified and ensured first according to specific conditions or diagnostic questions.

In order to speed up this process, an automation can be achieved by setting up a model observer (MO) [1] as a mathematical model for the HO. In the case of an existing reference image as gold-standard, several sophisticated methods have been established to quantify still natural scene images [2, 3, 4] which have shown good performance under certain quality metrics [5, 6]. Exploring known characteristics of the human visual system [7] helps to understand how the HO assess image quality [8, 9, 10]. Moreover, several approaches are specialized on the quantification of type and degree of distortion [11, 12] or are just trained on certain types of input data, i.e. the generalization ability is rather low. Hence, despite these developments an objective and accurate measure for different kinds of input images and distortions which better reflects the human perception is still missing.

For MRI, reference/gold-standard images are often hard or even impossible to acquire due to an additional required reference scan and/or the difficulty in defining an appropriate gold-standard. In addition, (intensity-based) similarity/dissimilarity measures between the to be evaluated image and the reference image cannot fully reflect complex image distortions or MRI artifacts [13] like motion ghosting or subsampling aliasing. Instead of a reference image, a supervised learning from HO labeling scores is thus a promising approach to mimic the human perception. Therefore, in the MRI environment, a blind/no-reference image quality assessment (IQA) with supervised learning is of great interest.

In medical IQA, most works focus on automatic lesion detection as a two-class problem [14, 15] or other supervised detection problems [16, 17].

We proposed in [18] an automatic blind MR IQA based on supervised learning in order to predict a HO labeling of arbitrary input images with unknown artifacts. Our system is trained on labels derived from HO and on meaningful features reflecting the image content including image distortions.

The overall accuracy is mainly determined by the features and the amount of training data, i.e. HO labels. Hence, in this work we want to address the problem of keeping the amount of necessary labels low to save time and cost in the labeling procedure. We propose to include active learning (AL) into our MR IQA system to support the labeling by querying the HO with the most meaningful images.

Hoi et al. [19, 20] developed an AL framework for data with content-based categories (e.g. does an image show thorax, abdomen or foot), whereas we focus on quality-based categories of arbitrary input images. In addition, their method does not take into account that HOs often have to label 3D images composed of 2D slices and labeled single 2D images. We will address this issue in this paper as well and discuss how to select representative 3D images consisting of meaningful 2D slices. Xue et al. [21] tried to omit HO labeling completely by automatic learning the image quality from overlapping image patches for simple image degradations. Since our type of images including distortions and quality criterions are far more complex we cannot spare the HOs if we want to mimic their quality assessment. Lorente et al. [22] focused on the combination of AL with a relevance vector machine in a four-dimensional feature space whereas our framework uses a higher dimensional feature space being able to reflect more complex image distortions.

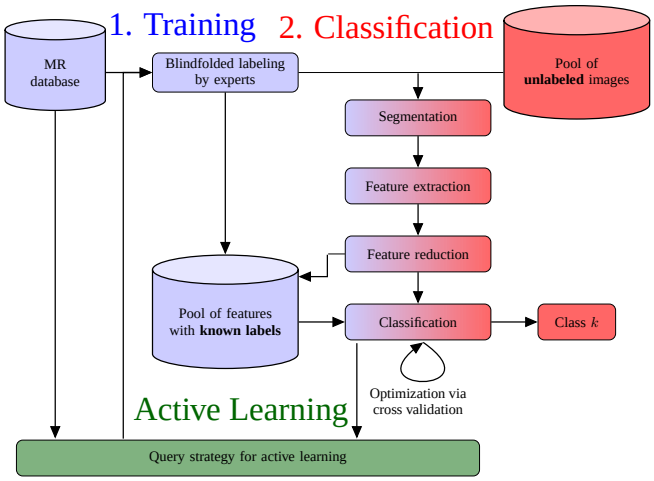


Fig. 1: Image quality assessment system including AL

2. IMAGE QUALITY ASSESSMENT SYSTEM

The proposed active learning setup is an extension of the classification system for MR IQA described in [18]. In Figure 1, the system is depicted as well as the included active learning step. As input data 2D and 3D MR images are accepted with 3D images being sliced into multiple 2D slices for further processing.

Each 2D image is represented by a feature vector $\tilde{\underline{x}}$. The features are based on image characteristics like contrast, resolution, texture and intensity. They add up to a total number of 2871 features. In order to avoid overfitting, the dimension of the feature space is reduced, e.g. by using principal component analysis, leading to a reduced feature vector $\underline{x}$. Experiments have shown that for the data used in this study a number of 36 principal components gives a good result.

In the classification step, the system uses an one-vs.-one multi-class Support Vector Machine (SVM) which is implemented utilising the LIBSVM library [23]. The employed soft-margin SVM uses a radial basis function (RBF) kernel and 10-fold cross-validation to find the optimum hyperplanes to separate $K = 5$ classes according to a 5-point Likert scale.

As stated in [18], this framework was able to achieve an overall test accuracy of 91.2%. The aim of including active learning is to achieve comparable results while using significantly less training data. For this purpose, the training of the SVM was embedded in an active learning loop.

3. ACTIVE LEARNING

The main idea of active learning is that the necessary amount of labeled training samples can be reduced significantly by selecting meaningful samples to label rather than labeling all examples. While labeling all samples often leads to adding redundant information to the training set, active learning uses certain query strategies to find those samples which are expected to have the most positive influence on the performance of the classifier. There exist different AL scenarios like membership query [24], selective sampling [25] or pool-based AL [26]. We implemented the latter one. Let $\mathcal{X}$ be a pool of samples and $\mathcal{D}$ the labeled training data. The goal is to keep $\mathcal{D}$ as small as possible. So initially only a small number of labels is used to train the classifier. The resulting low test accuracy is then iteratively increased by selecting the most meaningful samples from $\mathcal{U} = \mathcal{X} \setminus \mathcal{D}$, which are then labeled by a HO and inserted into $\mathcal{D}$. This procedure is repeated until a predefined stopping criterion is reached, e.g. when the resulting test accuracy converges. It is generally possible to either select a single sample at each iteration or a set $\mathcal{L}$ of several samples.

3.1. Query strategies

The success of active learning strongly depends on the way how to choose the samples to be labeled. There are many different query strategies which have been successfully used in the past for different classification tasks like uncertainty sampling [26], query by committee [27], expected model change [28] or expected error reduction [29]. It is important to choose a strategy suitable for the implemented classifier and the data to be classified. In our study, we concentrated on two uncertainty sampling strategies to investigate whether they are useful in the MR IQA setup. Uncertainty sampling is an approach to select samples to query a HO based on how certain the classifier is in his decision. There exist various ways to measure uncertainty [30, 31, 32]. We implemented one based on the probability of the class labels $y_k \in \{1, ..., K\}$ and one which uses the distance of the sample to the SVM hyperplanes.

The probability-based approach was proposed by Joshi et al. [33]. It uses probability estimates for multi-class SVM obtained through pairwise coupling as described in [34]. Multi-class AL methods based on class probability estimates often select those samples $\underline{x}_n$ to be labeled for which the probability $P_k(\underline{x}_n)$ to belong to the most probable class k is minimal. Another way is to choose the samples for which the entropy of the distribution of class membership probability is maximal. A drawback of both methods is that they are easily influenced by the probability distribution of non-important classes. Instead, Joshi et al. proposed to minimize the difference between the probabilities $P_k(\underline{x}_n)$ of the most and $P_l(\underline{x}_n)$ of the second most probable class as a measurement for uncertainty. The set $\mathcal{L}$ of N_L samples to be labeled is created as

$$\mathcal{L} = \bigcup_{n=1}^{N_\mathrm{L}} \{\underline{x}_n | \min_n \left(P_k(\underline{x}_n) - P_l(\underline{x}_n)\right)\}. \tag{1}$$

Since their experimental results on different data sets look very promising, we decided to investigate whether this query strategy is also suitable for our application.

The second query strategy is based on the distance d between samples and the hyperplanes. The distance $d(\underline{x}_n)$ of one feature vector $\underline{x}_n$ to the hyperplane $f(\underline{x}_n) = \langle \underline{w}, \underline{x}_n \rangle + b$ is given by

$$d(\underline{x}_n) = \|\underline{w}\|_2^{-1} f(\underline{x}_n) \tag{2}$$

where $\underline{w}$ and b denote the primal parameters learned by the SVM. If an RBF kernel $k(\underline{x}_n, \underline{x})$ is used instead of a linear one, $d(\underline{x}_n)$ can be calculated by

$$f(\underline{x}_n) = \sum_{i=1}^{N_\mathrm{SV}} \alpha_i y_i k(\underline{x}_n, \underline{x}_i) + b \tag{3}$$

with dual coefficients α_i and the corresponding support vectors $\underline{x}_i$, N_SV is the number of support vectors. The distances $d(\underline{x}_n)$ are then sorted in ascending order and the first N_L samples build the set $\mathcal{L}$ of data to be labeled by an HO:

$$\mathcal{L} = \{\underline{x}_n | d(\underline{x}_n) < d(\underline{x}_m) \; \forall \; x_n \in y_k\} \setminus (\mathcal{O} \cup \mathcal{S}) \tag{4}$$

Here $\mathcal{O}$ denotes a set of outliers and $\mathcal{S}$ a group considering the slack variables which will both be described in the following.

Since outliers tend to lie close to hyperplanes or even on the wrong side, a purely distance-based criterion might be likely to select those and hence the performance of the classifier increases less or even decreases. Thus, for each class y_k we take the samples for which the classifier assigned the same label y_k and calculate the Euclidean distance $d_k(\underline{x}_n, \underline{\mu}_k) = \|\underline{x}_n - \underline{\mu}_k\|$ to the class center $\underline{\mu}_k$ to identify potential outliers. A sample $\underline{x}_n$ qualifies as outlier, if it is farther away than a predefined value ϵ compared to all other samples $\underline{x}_m$ within the same class y_k which are closer located to the class center, with $d_k(\underline{x}_m, \underline{\mu}_k) < d_k(\underline{x}_n, \underline{\mu}_k) \; \forall \; x_m \neq x_n$.

$$\begin{aligned}\mathcal{O} = &\{\underline{x}_n | d_k(\underline{x}_n, \underline{\mu}_k) - d_k(\underline{x}_m, \underline{\mu}_k) > \epsilon \wedge \\ &d_k(\underline{x}_m, \underline{\mu}_k) < d_k(\underline{x}_n, \underline{\mu}_k) \; \forall \; x_n \in y_k, x_m \neq x_n\} \end{aligned} \quad (5)$$

Outliers are discarded from $\mathcal{L}$. In SVM, the slack variables allow for a certain amount of samples near the hyperplane to be categorized into the wrong class. Since they should not be selected for labeling, we compensate for this by defining a minimum distance δ to the hyperplane. Samples with $d(\underline{x}_n) < \delta$ are also not used to query the HO

$$\mathcal{S} = \{\underline{x}_n | d(\underline{x}_n) < \delta\} \quad (6)$$

3.2. Representative 3D image selection

Another important issue is that in our MR IQA system HOs are asked to label 3D images instead of 2D image slices. Since our MR IQA system should be able to rate both 2D and 3D images, the classifier is trained using 2D images which can be partly 2D slices of 3D images. The actual selected amount for labeling of 3D images can vary a lot depending on how many of the chosen samples $\underline{x}_n$, i.e. 2D image slices, belong to different 3D images. Furthermore, the 2D slices at the beginning and end of a 3D image can be of poorer quality, due to e.g. infolding artifacts in slice direction or noise only content. They should be discarded from AL because otherwise they would be more likely selected for query due to their uncertainty. We address this issue by not only considering the uncertainty of $\underline{x}_n \in \mathcal{L}$, but also how many 2D slices of the same 3D image are selected and whether they are at the beginning or end. Images with a higher number of $\underline{x}_n \in \mathcal{L}$ are assigned a higher priority to be labeled. To account for how uncertain the classifier is regarding categorizing each $\underline{x}_n$, the slices are additionally weighted according to their corresponding uncertainty. If only slices from the beginning and/or end of a 3D image are chosen, they are discarded from $\mathcal{L}$ and the corresponding 3D image will not be part of the labeling query.

4. EXPERIMENTS AND RESULTS

The aim of this study is to reduce the labeling cost. Let N_I be the number of training samples in the initial training set $\mathcal{D}_\mathrm{I}$. For $N_\mathrm{L} = |\mathcal{L}|$ being the number of labeled samples per query and N_q being the number of queries, we get $N_\mathrm{AL} = N_q \cdot N_\mathrm{L}$ training samples for active learning. The total number of training samples in $\mathcal{D}$ is $N_\mathrm{D} = N_\mathrm{I} + N_\mathrm{AL}$. This amount is to be minimized while maintaining a high classification accuracy.

4.1. MR data sets

For our experiments we used 2D MR image slices taken from 3D images of the thorax, abdomen and pelvis of 35 patients and healthy volunteers, which were acquired using different imaging sequences, contrast weights and various subsampling strategies with their corresponding reconstruction techniques. Those images were classified into five classes according to the 5-point Likert scale by five HOs. If the experts did not agree on one class, the median of the assigned labels was used. The classes represent very high image quality (1), high quality (2), medium quality (3), poor quality (4) and very poor quality (5). The pool $\mathcal{D}$ of labeled data contains a total of 2911 2D slices taken from 100 3D images. 2038 samples were assigned to the training set $\mathcal{D}_\text{train}$ and 873 to the test set $\mathcal{D}_\text{test}$.

We have chosen the initial training set $\mathcal{D}_\mathrm{I}$ by randomly selecting N_I samples from $\mathcal{D}_\text{train}$. The remaining samples from $\mathcal{D}_\text{train}$ were used as the pool of unlabeled samples $\mathcal{U}$. The results are presented as the average of 10 randomly initialized runs. In this, we were able to compare the results of the classifier after training with the whole training set $\mathcal{D}_\text{train}$ (91.2% test accuracy) and with the combination of $\mathcal{D}_\mathrm{I}$ and active learning.

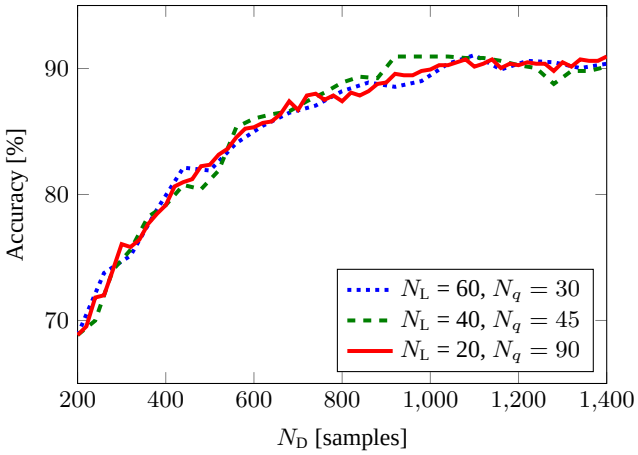


Fig. 2: Test accuracy for different N_L samples per query with initial training size $N_\mathrm{I} = 200$ and the probability-based approach. The (min/mean/max) standard deviation is for ····· $(0/0.78/2.72)$, - - - $(0.13/0.85/2.54)$ and ——— $(0/0.67/2.54)$.

4.2. Amount of data added in each iteration

The number N_L of samples selected to query the HO is an important factor for how well AL improves the classifier. On one hand, a too small value of N_L might not give a significant improvement. On the other hand, labeling too many samples on each query is expensive. It can also result in adding too much redundant information to the training set and thus decrease the efficiency of AL. We therefore investigated the optimal value of N_L. For this purpose, we trained the SVM with $N_\mathrm{I} = 200$ and different values of N_L. The results for $N_\mathrm{L} = 20$, $N_\mathrm{L} = 40$ and $N_\mathrm{L} = 60$ are depicted in Figure 2.

The results show that in terms of the total number of training samples N_D, $N_\mathrm{L} = 40$ is a good choice when using the probability-based method. For $N_\mathrm{L} = 20$, the classifier needs much more iterations to achieve a comparable accuracy which results in a high overall training time. A test accuracy over 90% is first achieved after adding a total of $N_\mathrm{AL} = 840$ samples in $N_q = 42$ iterations. This results in a total number of $N_\mathrm{D} = 1040$ labeled samples and a reduction of 49% of training data. For $N_\mathrm{L} = 40$ we need $N_q = 20$ iterations and $N_\mathrm{D} = 1000$ which reduces the needed training samples by 51%. Adding more than 40 samples per iteration does not result in a further decrease of the total number of training samples.

4.3. Amount of data used for initial training

Another important parameter is the size N_I of the initial training set $\mathcal{D}_\mathrm{I}$. Figure 3 shows the resulting test accuracy for $N_\mathrm{I} = 50$, $N_\mathrm{I} = 200$ and $N_\mathrm{I} = 500$ using the probability-based and the margin-based approach. During each iteration, $N_\mathrm{L} = 40$ samples were added.

For the probability-based strategy (see Figure 3a) and $N_\mathrm{I} = 200$, we needed a total amount of $N_\mathrm{D} = 1040$ labeled samples to achieve a test accuracy of over 90%, as mentioned in the previous section. Using $N_\mathrm{I} = 50$ leads to a total number of $N_\mathrm{D} = 1290$ examples to achieve a test accuracy $> 90\%$ (40% reduction compared to full training set). For $N_\mathrm{I} = 500$, the test accuracy after the initial training is already high, but this leads to a slower improvement. To achieve a test accuracy of over 90%, $N_\mathrm{D} = 1100$ samples have to be labeled which leads to a reduction of 46%. Similar results can be observed with the margin-based approach. The results can be seen in Figure 3b.

Our results show that the positive effect of AL occurs already with a small amount of labeled data for initial training. But both too small and too big values for N_I lead to a slower improvement of the test accuracy.

4.4. Probability-based vs. distance-based query

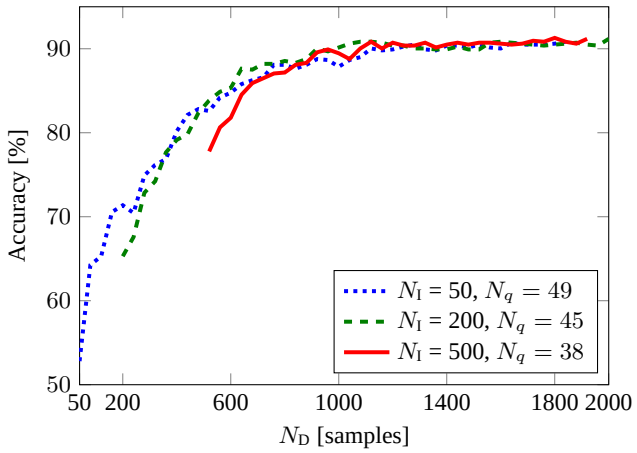
During our experiments, we came to the conclusion that the margin-based query strategy slightly outperforms the probability-based one (see Figure 3). Both of them give clearly better results than using random samples from $\mathcal{U}$ to query the HO. Therefore, both methods help to reduce the labeling effort significantly.

In addition, we evaluated both strategies with respect to the distribution of the image class labels and their content (e.g. thorax, pelvis, abdomen, etc.). Regarding the distribution of the classes of the selected samples, our experiments showed that the margin-based approach strongly favours samples belonging to the most dominant classes, whereas the probability-based method selects them more equally from all classes. The impact of the correction sets for outlier $\mathcal{O}$ and slack variables $\mathcal{S}$ is for $N_\mathrm{I} = 200, N_\mathrm{L} = 40$ in the average range of $|\mathcal{O}| = 11, |\mathcal{S}| = 8$ samples per query and has thus a positive impact on the selection.

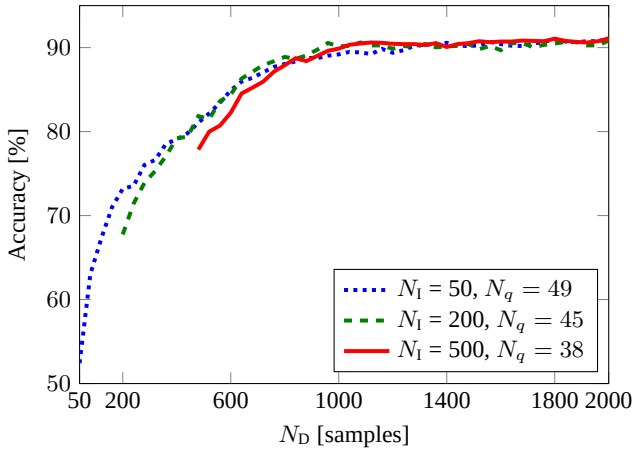
In terms of image content, our experiments led to the conclusion that both query strategies are not influenced by the content of an image. We analyzed the chosen images for several different values of N_I and N_L. Neither of both query strategies resulted in a different distribution of the image content than choosing random samples.

5. CONCLUSION

In this paper, we propose a way to reduce the amount of labeled training data for a MR image quality assessment system by using active learning. We implemented two strategies based on uncertainty (one probability-based and one margin-based) to select samples to query the human observer. Testing the system on *in-vivo* MR data revealed that both strategies reduce the training data by roughly 50% while achieving comparable classification results to the previous system setup without active learning. The margin-based approach slightly outperforms the probability-based one. For the case of labeling a complete 3D image, we also presented a selection strategy for choosing the most meaningful 2D slices belonging to a few significant 3D images. Overall, using active learning for an automatic magnetic resonance image quality assessment system results in a reduced labeling effort for the human observer, saving time and cost.



(a) probability-based approach



(b) margin-based approach

Fig. 3: Test accuracy for different initial training set sizes N_I with $N_\mathrm{L} = 40$ samples per query and both query strategies. The (min/mean/max) standard deviation of (a) is for ····· $(0.07/1.03/5.17)$, - - - $(0.13/0.85/2.54)$, ——— $(0.07/0.56/2.07)$ and of (b) is for ····· $(0.07/1.02/5.17)$, - - - $(0.13/0.82/2.8)$, ——— $(0.07/0.52/1.58)$.

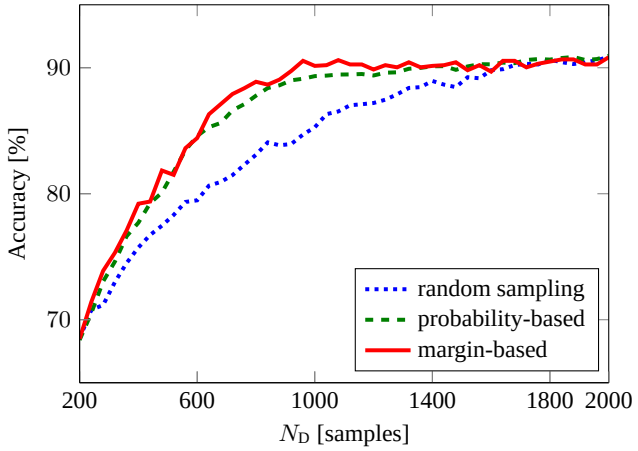


Fig. 4: Test accuracy using active learning with random, probability-based and margin-based selection approach for initial training size of $N_\mathrm{I} = 200$, $N_\mathrm{L} = 40$ samples per query and $N_q = 45$ queries.

6. REFERENCES

[1] H.H. Barrett, J. Yao, J.P. Rolland, and K.J. Myers, "Model observers for assessment of image quality," in *Proc. Natl. Acad. Sci. USA*, Nov 1993, vol. 90, pp. 9758–9765.

[2] Z. Wang and A.C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, March 2002.

[3] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE T. Image Process.*, vol. 13, no. 4, pp. 600–612, April 2004.

[4] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE T. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb 2006.

[5] M. Eckert and A. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Process.*, vol. 70, no. 3, pp. 177 – 200, 1998.

[6] R. Borse and P. Markad, "Competitive analysis of existing image quality assessment methods," in *Int'l Conf. Adv. Comp. Comm. Inform.*, Sept 2014, pp. 1440–1444.

[7] H.H. Barrett and K.J. Myerrs, *Foundations of Image Science*, John Wiley and Sons, Inc., Hoboken, New Jersey, USA, 2004.

[8] G. Ivkovic and R. Sankar, "An algorithm for image quality assessment," in *IEEE Int'l Conf. Acoust. Spee. Signal Process.*, May 2004, vol. 3, pp. 713–16.

[9] Y. Gavet, M. Fernandes, and J.-C. Pinoli, "Automatic quantitative evaluation of image registration techniques with the epsilon dissimilarity criterion in the case of retinal images," in *Int'l Conf. Quality Control by Artifcial Vision*, 2011, pp. 8000–23.

[10] Y. Zhang and D.M. Chandler, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *J. Electron Imaging*, vol. 22, no. 4, 2013.

[11] A.K. Moorthy and A.C. Bovik, "Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec 2011.

[12] L. Xu, J. Li, W. Lin, Y. Zhang, L. Ma, Y. Fang, Y. Zhang, and Y. Yan, "Multi-task rank learning for image quality assessment," in *IEEE Int'l Conf. Acoust. Spee. Signal Process.*, April 2015, pp. 1339–1343.

[13] T. Rohlfing, C.R. Maurer, W.G. O'Dell, and J. Zhong, "Modeling liver motion and deformation during the respiratory cycle using intensity-based nonrigid registration of gated MR images," *Med. Phys.*, vol. 31, no. 3, pp. 427–432, 2004.

[14] J.G. Brankov, Y. Yang, L. Wei, I. El Naqa, and M.N. Wernick, "Learning a nonlinear channelized observer for image quality assessment," *IEEE T. Med. Imaging*, vol. 28, no. 7, pp. 991–999, 2009.

[15] L. Zhang, C. Cavaro-Menard, P. Le Callet, and D. Ge, "A multi-slice model observer for medical image quality assessment," in *IEEE Int'l Conf. Acoust. Spee. Signal Process.*, April 2015, pp. 1667–1671.

[16] Y. Jiang, D. Huo, and D.L. Wilson, "Methods for quantitative image quality evaluation of MRI parallel reconstructions: detection and perceptual difference model," *Magn. Reson. Imaging*, vol. 25, no. 5, pp. 712–721, June 2007.

[17] T. Marin, M.M. Kalayeh, F.M. Parages, and J.G. Brankov, "Numerical surrogate of a human observer for cardiac motion defect detection in SPECT imaging," *IEEE T. Med. Imaging*, vol. 33, no. 1, pp. 38–47, 2014.

[18] T. Küstner, P. Bahar, C. Würslin, S. Gatidis, P. Martirosian, NF. Schwenzer, H. Schmidt and B. Yang, "A new approach for automatic image quality assessment," in *Proc. Annual Meeting ISMRM 2015*, 2015.

[19] S.C.H. Hoi, R. Jin, J. Zhu, and M.R. Lyu, "Batch mode active learning and its application to medical image classification," in *Proc. Int'l Conf. Mach. Learn.*, 2006.

[20] S.C.H. Hoi, R. Jin, J. Zhu, and M.R. Lyu, "Semi-supervised SVM batch mode active learning for image retrieval," in *IEEE Conf. Comp. Vis. Pattern Recogn.*, June 2008, pp. 1–7.

[21] W. Xue, L. Zhang, and X. Mou, "Learning without Human Scores for Blind Image Quality Assessment," in *IEEE Conf. Comp. Vis. Pattern Recogn.*, June 2013, pp. 995–1002.

[22] I. Lorente and J.G. Brankov, "Active Learning for Image Quality Assessment by Model Observer," in *IEEE Int'l Symp. Biomed. Imaging*, April 2014.

[23] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *T. Intell. System. Tech.*, vol. 2, pp. 1–27, 2011.

[24] D. Angluin, "Queries and concept learning," *Mach. Learn.*, vol. 2, no. 4, pp. 319–342, 1988.

[25] L. Atlas, D. Cohn, R. Ladner, M. A. El-Sharkawi, and R. J. Marks, II, *Advances in Neural Information Processing Systems 2*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[26] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," in *ACM Int'l Conf. Res. Dev. Inform. Ret.*, 1994, pp. 3–12.

[27] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Int'l Conf. Comp. Learn. Theory*, 1992, pp. 287–294.

[28] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Adv. Neural Inform. Process. Systems*, 2008, vol. 20, pp. 1289–1296.

[29] N. Roy and A. Mccallum, "Toward optimal active learning through sampling estimation of error reduction," in *Proc. Int'l Conf. Mach. Learn.*, 2001, pp. 441–448.

[30] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *Int'l Conf. Intell. Data Anal.*, 2001, pp. 309–318.

[31] A. Culotta and A. McCallum, "Reducing labeling effort for structured prediction tasks," in *Proc. Nat. Conf. Artif. Intell.*, 2005, pp. 746–751.

[32] D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proc. Int'l Conf. Mach. Learn.*, 1994, pp. 148–156.

[33] A. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *IEEE Conf. Comp. Vis. Pattern Recogn.*, 2009, pp. 2372–2379.

[34] T. Wu, C. Lin, and R. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, 2003.