Deep-learning-based myocardial pathology detection

Matthias Ivantsits[1], Markus Huellebrand[1,2], Sebastian Kelle[1,3,4], Stefan Schönberg[4], Titus Kuehne[1,3,4], and Anja Hennemuth[1,2,4]

[1] Charité – Universitätsmedizin Berlin, Augustenburger Pl. 1, 13353 Berlin, Germany
[2] Fraunhofer MEVIS, Am Fallturm 1, 28359 Bremen, Germany
[3] German Heart Institute Berlin, Augustenburger Pl. 1, 13353 Berlin, Germany
[4] DZHK (German Centre for Cardiovascular Research), Berlin, Germany

Abstract. Cardiovascular diseases are the top cause of death worldwide. Commonly, physicians screen suspected pathological patients with histological examinations and blood tests. Since these clinical parameters are frequently ambiguous, they are routinely extended by delayed-enhancement magnetic resonance imaging of the myocardium.
We propose a method combining deep learning and classical machine learning to differentiate between pathological and normal cases. A convolutional neural network infers a segmentation of the left myocardium from a magnetic resonance image as a preliminary step. This segmentation is employed to determine radiomics-based features describing the morphology and texture of the myocardium. Subsequently, a multilayer perceptron deduces pathological cases from these radiomics features and clinical observations. The presented method demonstrates an accuracy of 0.96 and an F2-score of 0.98 on a nested cross-validation.

Keywords: MRI · heart · myocardial infarction · delayed-enhancement · machine learning · deep learning · classification

1 Introduction

Cardiac diseases, including myocardial infarction, are the world's leading cause of death [1] with approximately 31.9 %. Therefore, it is highly desirable to detect heart diseases early and decide about the appropriate therapy.

The diagnosis of myocardial infarction and myocarditis can be based on histological examinations [2] and blood tests. Proteins like troponin, N-terminal pro b-type natriuretic peptide (NTproBNP), and myoglobin have shown to be very reliable indicators for pathological cases [3–6]. Clinical and patient demographic parameters are sometimes ambiguous and routinely extended with delayed enhancement magnetic resonance imaging (DE-MRI). DE-MRI image data enable the assessment of anatomy and contrast agent accumulation patterns, which are indicators for pathological alterations of the heart muscle tissue [7].

Shape-based modeling has been shown to produce decisive factors to detect heart diseases in cine MRI [8,9], due to the structural change of the myocardium.

Radiomics has proven to be a very promising toolkit for medical image processing, analysis, and interpretation. It derives vast amounts of features from imaged structures describing patterns in morphology and texture that are usually hard to differentiate with the bare eye. Initially, radiomics has been employed for oncological applications but is an emerging technique in the cardiovascular field, especially with MRI. This observation has been confirmed by studies [10–13], which are extracting shape-based radiomics features and classifying diverse heart diseases.

There is an apparent lack of the usage of features describing the texture of imaged structures in the published methods. Due to highlighted myocardial infarction areas in DE-MRI, we hypothesize that features derived from the gray level co-occurrence matrix introduce valuable information to analyze pathological cases. These parameters are used along with the clinical, demographic, and shape-based parameters to distinguish normal from pathological cases for the EMIDEC classification contest [14].

2 Method and materials

2.1 Dataset

The EMIDEC dataset [14] consists of 150 cases in total, 100 for model training and 50 for model testing. Each observation includes a DE-MRI acquisition of the LV, covering the base to the apex. The training set with ground-truth segmentation of the LV myocardium is comprised of 100 cases (67 pathological cases, 33 normal cases). The testing includes 50 subjects (33 pathological cases, 17 normal cases), all different from those in the training set. The imbalance of normal to pathological cases roughly corresponds to real life managed exams. Along with the MRI, clinical parameters were provided: **sex**, **age**, **tobacco** (yes, no, and former), **overweight** (BMI over 25), **arterial hypertension**, **diabetes**, **family history of coronary artery disease**, **ECG**, **killip max** [5], **troponin** [6], **ejection fraction** (LV), and **NTproBNP** [7].

2.2 Method

Our proposed method consists of three major steps (1) LV segmentation, (2) radiomics feature extraction, (3) classification based on the features from the previous step plus the clinical parameters provided with the contest. For the LV segmentation we utilize a 2D U-Net variation proposed by Hüllebrand et al. [15], which was pre-trained on the ACDC dataset [16].

[5] a parameter based on a physical assessment, quantifying the risk of mortality
[6] a protein released in large quantities in the event of damage to the heart muscle cells
[7] a protein released in large quantities when the heart needs to work harder

After the segmentation of the LV, we extract radiomics features based on the LV shape and texture. The MRI can be used as direct input to a classification via a CNN or similar architectures, but since the MRI is a very high dimensional input, it is hard to train and prone to overfit if not parameterized correctly. Therefore, we argue that using a lower-dimensional input, by deriving radiomics features, produces a more reliable classification of normal and pathological cases. Generally, radiomics features can be grouped into morphological and textural descriptors. They can be further divided into the following seven categories:

1. **First Order**, describes the voxel intensity distribution within a masked region.
2. **Shape (3D)**, describes the overall size and shape of a structure. This includes volume, surface area, sphericity, etc.
3. **Gray Level Co-occurrence Matrix (GLCM)**, describes the second-order joint probability function of an image region. This includes cluster tendency, cluster shape, contrast, etc.
4. **Gray Level Size Zone Matrix (GLSZM)** quantifies the gray level zones, which is defined as the number of connected voxels that share the same intensity. This includes gray level variance, zone entropy, zone variance, etc.
5. **Gray Level Run Length Matrix (GLRLM)** describes the length in several consecutive voxels that share the same intensity. This includes run percentage, run variance, run entropy, etc.
6. **Neighbouring Gray Tone Difference Matrix (NGTDM)** describes the difference between the intensity and the average intensity of its neighbors. This includes coarseness, contrast, busyness etc.
7. **Gray Level Dependence Matrix (GLDM)** describes the intensity dependency, which is defined as the number of connected voxels within a distance δ that depend on the center voxel. This includes dependence entropy, dependence variance, dependence non-uniformity, etc.

Shape-based features have shown to be useful for the classification of pathological cases in cine echocardiography and MRI [8–13]. We hypothesize that due to the injected contrast agent in DE-MRI, highlighted infarction areas can be described by the radiomics texture features as shown in similar approaches for T1- and T2-mapping [17,18]. The extracted radiomics features result in 108 variables for the LV, plus 12 clinical parameters per patient, summing up to 120 features. Due to this high dimensional input combined with only 100 patient observations, we propose an 8-fold nested cross-validation (CV) to select relevant radiomics feature classes. Moreover, during the CV, we perform a model selection of five classifiers and their respective hyperparameters, which are illustrated in table 1.

Support vector machines (SVM) are particularly effective in high dimensional spaces with a clear margin between the classes and have shown to work well on pathological case detection by Cetin et al. [10]. Random forests are very robust to outliers and comparatively little impacted by noise. They have shown to be effective in medical settings, as illustrated by [11–13]. Multilayer perceptrons

Model name	Hyperparameters
Support vector machine	C: 0.01, 0.05, 0.1, 1, 10, 100, 1000 kernel: linear, poly, rbf, sigmoid gamma: auto, scale
Random forest	n_estimators: 100, 200, 300, 500 max_depth: 90, 100, 110 max_features: 2, 3 min_samples_leaf: 3, 4, 5
Gradient Boosting	loss: deviance, exponential n_estimators: 10, 20, 50 max_depth: 3, 5 max_features: log2, sqrt criterion': friedman_mse, mae
Multilayer perceptron	hidden_layer_sizes: (100,), (100, 50), (50,), (50, 25) max_iter: 200, 300, 500, 700, 1000
K-nearest neighbors	n_neighbors: 1, 3, 5, 10 weights: uniform, distance

Table 1: Five models and their respective hyperparameters used during the 8-fold nested cross-validation.

(MLP) have been thoroughly studied, and Isensee et al. [13] have illustrated that they can be successfully applied for pathological case detection. Gradient boosting is one of many ensemble methods, which utilizes a collection of weakly trained classifiers. It has been shown to be useful in individual treatment estimations [19]. The K-nearest neighbors (KNN) algorithm does not need any training, and therefore new data can be seamlessly added. Akhil [20] has demonstrated medical applicability by predicting various heart diseases.

3 Results

We used an 8-fold CV during all experiments, performed on an Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz with 16 GBs RAM. Since the contest is evaluated on the predictions' accuracy, we evaluate all models based on this metric. However, we argue that a false-negative classification is more harmful to the screening process, and since the dataset is imbalanced, we additionally evaluate the F2-score for each model. Where the F2-score combines sensitivity and precision and considers sensitivity twice as important as precision.

The first experiment we conducted was only using the clinical parameters to train the models described in table 1. The results of this analysis are illustrated in figure 1. Based on the accuracy score, the random forest is the best performing metric with 0.89 ± 0.10, closely followed by the KNN with 0.86 ± 0.14. When taking the F2-score into account, the random forest with 0.93 ± 0.08 is considered the best choice, followed by the MLP with 0.92 ± 0.06.

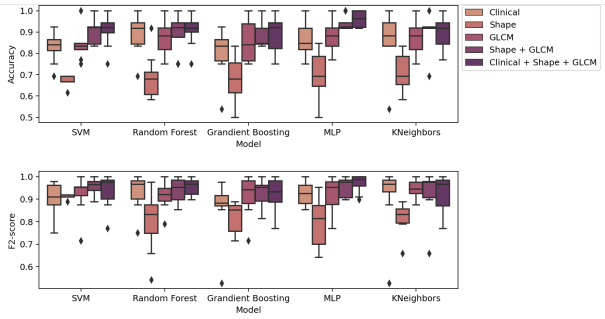

Fig. 1: An illustration of the models accuracy and F2-score produced by an 8-fold nested cross-validation. All models tested on clinical, shape, GLCM, shape + GLCM, and clinical + shape + GLCM.

In the next setup, we experimented with sole shape, GLCM, and shape+GLCM features. The results of this procedure are illustrated in figure 1. We observed similar accuracy compared to the clinical features when only taking the GLCM features into account. The best performing model (MLP) trained on the GLCM variables achieves an accuracy of 0.88 ± 0.07 and an F2-score of 0.92 ± 0.08. After combining the shape and GLCM features, the best performing model improves to 0.93 ± 0.05 and an F2-score of 0.94 ± 0.07.

Due to the improved accuracy with models trained on the shape and GLCM features, we combined them with the clinical observations. Clinical, morphological, and textural features sum up to 50 variables. We assume the number of features can be further compressed and subsequently improve the models' predictive power. Consequently, we performed a feature importance analysis as proposed by Breiman [21]. This analysis can be performed on any fitted model by calculating a base score produced by the model on the training or test set. This is followed by a random shuffle to one of the features and compared to the baseline's predictive power. This procedure is then repeated and applied to all features to come up with an importance score. Collinear features in the input result in lower importance scores when permuted. This can be counteracted by clustering highly correlated features and keeping only one feature per cluster.

To estimate the effect of a not perfectly segmented myocardium, we conducted experiments with different levels of structural changes to the myocardium segmentation. We performed contour distortions utilizing increasing degrees of random affine and elastic deformations [22]. An overview of the changes is illustrated in table 2. By introducing a light structural change to the myocardial segmentation, the accuracy of the trained model stays constant (see figure 3). After increasing the distortion, we observed a drop in the models' performance to 0.93. The deformed myocardium shows a dice coefficient of 0.92 and a Hausdorff distance of 1.4 mm. Further decreasing the dice coefficient to 0.87 showed

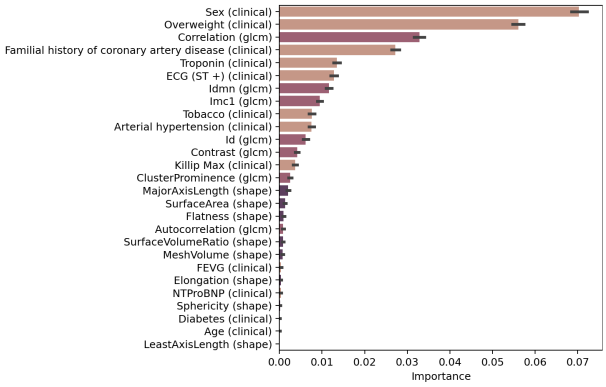

Fig. 2: The feature importance scores on the finally trained MLP, where the importance is defined by the difference of the models' baseline and the score after a feature permutation.

similar accuracy, only after decreasing the dice to 0.75 with a Hausdorff distance of 4.3 mm shrinks the models' predictive capability to 0.87.

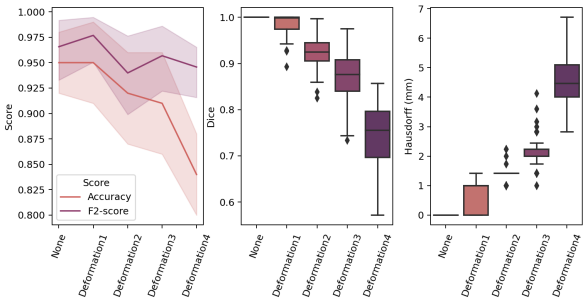

Fig. 3: An illustration of different levels of distortions performed to the myocardium segmentation. **Left**: highlights the change of performance with increasing levels of deformations. **Center**: illustrates the dice coefficient of the original myocardium segmentation and the deformed one. **Right**: shows the Hausdorff distance of the original myocardium segmentation the the distorted structure in mm.

Finally, we cross-validate an MLP with optimal hyperparameters obtained in the previous experiments. The model includes one hidden layer with 100 neurons and the ReLU activation function. The network was trained with Adam (with standard parameters) and a learning rate of 0.001 for 500 iterations. This CV

Name	Structural changes
Mask Transformation 1	Random affine transformation with scaling $[0.97, 1.03]$
Mask Transformation 2	Random elastic deformation with $(7, 7, 5)$ control points and $(3, 3, 1)$ maximum displacement.
Mask Transformation 3	Random elastic deformation with $(7, 7, 5)$ control points and $(5, 5, 1)$ maximum displacement.
Mask Transformation 4	Random elastic deformation with $(15, 15, 5)$ control points and $(10, 10, 1)$ maximum displacement.

Table 2: The four levels of the distortion we performed on the myocardium segmentation. The deformations are illustrated in increasing strength.

results in a network with an accuracy of 0.96 $\pm$ 0.05 and an F2-score of 0.98 $\pm$ 0.04. The feature importance scores of this model are illustrated in figure 2. This analysis shows very high importance for most clinical variables, followed by some GLCM and morphological features. For the clinical integration of the proposed method, we utilized LIME [23], which is an explanatory framework for any black-box classifier.

4 Discussion and Conclusion

We have illustrated a pathological classification pipeline, with comparable accuracy scores to state-of-the-art solutions. While clinical parameters provide an excellent baseline to distinguish normal from pathological cases, this can be further improved by including morphological and textural features extracted from DE-MRI. Our proposed method includes an automatic segmentation of the LV, an extraction of large amounts of morphological and textural radiomics features followed by classification into normal and pathological cases. A nested CV was performed to deal with the high dimensional data produced by the radiomics feature extraction. During this process, an optimal model from a pool of classifiers and hyperparameters was chosen. Moreover, the input space was further reduced by clustering highly correlated features, resulting in 27 clinical and MRI-based features. These features were ranked according to the final models' importance for clinical interpretation. The final MLP achieves an accuracy score of 0.96 and an F2-score of 0.98 on the performed nested cross-validation. We hypothesize that the utilization of GLCM features derived from DE-MRI enables the differentiation between categories of pathological cases like myocardial infarction and myocarditis. Furthermore, we illustrated the generalization capability of the proposed method by introducing distortion of the myocardium segmentation. The model exhibits invariance in performance on small to medium imperfections of the segmentation. On more intense structural changes, the model still illustrates an accuracy of 0.87 (see figure 3).

Interestingly, the importance of the Killip and diabetes observations are almost irrelevant to the combined model as shown in figure 2. A potential expla-

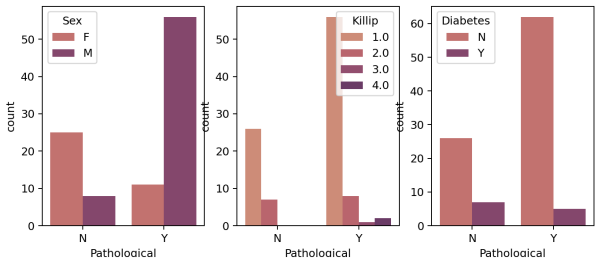

Fig. 4: **Left**: an illustration of pathological cases grouped by the patients' sex. **Center**: an overview of pathological cases grouped by the Killip parameter. **Right**: a plot of pathological cases grouped by diabetes.

nation for this phenomenon is illustrated in figure 4. The diabetes distribution of pathological and normal cases does not show any significant difference. Moreover, the Killip classes one and two do not seem to be good predictors for pathological cases, since they mostly emerge in this class. Only classes three and four seem to be useful for predicting pathological cases, although these classes' frequency seems negligible. Notable is the high importance of the patients' sex to the prediction of pathological cases. This observation can be explained by the vast difference of men with myocardial infarction or myocarditis showing up to the emergency room compared to women. Finally, the proposed method's probable advancement is training with randomly elastically deformed segmentations of the myocardium, which should help the model to a superior generalization.

References

1. Gregory A Roth et al., Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017, The Lancet, Volume 392, Issue, 2017 10159, 1736 - 1788
2. E. B. Lieberman, G. M. Hutchins, A. Herskowitz, N. R. Rose, and K. L. Baughman, Clinicopathoiogic description of myocarditis, 1991
3. Y. Li, F. Zhang, X. Wang, D. Wang, Expression and clinical significance of serum follistatin-like protein 1 in acute myocardial infarction, 2017
4. J. Kottwitz, K. A. Bruno, J. Berg, G. R. Salomon, D. Fairweather, M. Elhassan, N. Baltensperger, C. K. Kissel, et al., Myoglobin for Detection of High-Risk Patients with Acute Myocarditis, 2020
5. S. Sachdeva, X. Song, N. Dham, D. M. Heath, and R. L. DeBiasi, Analysis of clinical parameters and cardiac magnetic resonance imaging as predictors of outcome in pediatric myocarditis, 2015
6. A. S. V. Shah, D. A. McAllister, R. Mills, K. K. Lee, A. M. D. Churchhouse, K. M. Fleming, E. Layden, A. Anand, O. Fersia et al., Sensitive Troponin Assay and the Classification of Myocardial Infarction, 2015
7. E. Jackson, N. Bellenger, M. Seddon, S. Harden, C. Peebles, Ischaemic and non-ischaemic cardiomyopathies—cardiac MRI appearances with delayed enhancement, 2007

8. M. Tabassian, M. Alessandrini, L. Herbots, O. Mirea, E. D. Pagourelias, R. Jasaityte, J. Engvall, L. De Marchi, G. Masetti, and J. D'hooge, Machine learning of the spatio-temporal characteristics of echocardiographic deformation curves for infarct classification, 2017
9. A. Suinesiaputra, J. Dhooge, N. Duchateau, J. Ehrhardt, A. F. Frangi, A. Gooya, V. Grau, K. Lekadir, et al., Statistical Shape Modeling of the Left Ventricle: Myocardial Infarct Classification Challenge, 2018
10. I. Cetin, G. Sanroma, S. E. Petersen, S. Napel, O. Camara, M.-A. G. Ballester, and K. Lekadir, A Radiomics Approach to Computer-Aided Diagnosis with Cardiac Cine-MRI, 2017
11. J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, Automatic Segmentation and Disease Classification Using Cardiac Cine MR Images, 2017
12. M. Khened, V. Alex, and G. Krishnamurthi, Densely Connected Fully Convolutional Network for Short-Axis Cardiac Cine MR Image Segmentation and Heart Diagnosis Using Random Forest, 2017
13. F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features, 2017
14. EMIDEC Classification Contest, http://emidec.com/classification-contest, accessed: 2020-09-12
15. Markus Hüllebrand et al., ..., 2020
16. Automated Cardiac Diagnosis Challenge, https://www.creatis.insa-lyon.fr/Challenge/acdc/, accessed: 2020-09-12
17. B. Baessler, M. Mannil, S. Oebel, D. Maintz, H. Alkadhi, and R. Manka, Subacute and Chronic Left Ventricular Myocardial Scar: Accuracy of Texture Analysis on Nonenhanced Cine MR Images, 2018
18. B. Baessler, C. Luecke, J. Lurz, K. Klingel, A. Das, M. von Roeder, S. de Waha-Thiele, C. Besler, et al., Cardiac MRI and Texture Analysis of Myocardial T1 and T2 Maps in Myocarditis with Acute versus Chronic Symptoms of Heart Failure, 2019
19. S. Sugasawa and H. Noma, Estimating individual treatment effects by gradient boosting trees, 2019
20. Jabbar Akhil, Prediction of heart disease using k-nearest neighbor and particle swarm optimization, 2017
21. Leo Breiman, Random Forests, 2001
22. F. Pérez-García, R. Sparks, S. Ourselin, TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning, 2020
23. M. Túlio Ribeiro, S. Singh, C. Guestrin, Ẅhy Should I Trust You?̈: Explaining the Predictions of Any Classifier, 2016