Differentiating between giant cell arteritis and atherosclerosis on [^{18}F]FDG-PET: an explainable machine learning approach

H.S. Vries, MSc[1,2,4]; G.D. van Praagh, MSc[1]; P.H. Nienhuis, BSc[1]; O. Bouhali, PhD[2] ; R.H.J.A. Slart, MD PhD[1,3]; L. Alic, PhD[4]

1. Department of Nuclear Medicine and Molecular Imaging, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
2. Division of Arts and Science, Texas A&M University at Qatar, Doha, Qatar
3. Department of Biomedical Photonic Imaging, Technical Medical Centre, Faculty of Science and Technology, University of Twente, Enschede, The Netherlands
4. Department of Magnetic Detection and Imaging, Technical Medical Centre, Faculty of Science and Technology, University of Twente, Enschede, The Netherlands

***Abstract*—Background** This work aims to investigate the feasibility of an explainable machine learning model based on radiomics features to differentiate between giant cell arteritis (GCA) and atherosclerosis in aortic [^{18}F]FDG-PET scans. **Method** Twenty [^{18}F]FDG-PET scans (ten of patients with GCA, ten with atherosclerosis) were retrospectively included. The aorta was delineated into four segments (ascending, arch, descending, and abdominal aorta). In total, 93 radiomic features and two quantitative features were extracted from each of the 80 segments. Four different feature selection methods and four classifiers were used to identify important features for the machine learning model and determine the probability. The model's performance was evaluated using accuracy and AUC. To enhance explainability of the model, feature importance was determined, and an occlusion sensitivity map of the aorta was created. **Results** The combination of the first-order skewness, GLDM dependence non-uniformity, and GLRLM run entropy features showed the highest accuracy and AUC of, 0.90±0.08 and 0.960±0.029, respectively.
Conclusion This study demonstrated the potential of an explainable radiomics-based machine learning model for the differentiation between GCA and atherosclerosis in [^{18}F]FDG-PET scans.

Keywords— giant cell arteritis, atherosclerosis, [^{18}F]FDG-PET, explainable machine learning, radiomics

I. Introduction

Giant cell arteritis (GCA) is a systemic inflammatory condition of the arterial vessel that often cause the accumulation of immune cells within the vascular walls. This can lead to vessel stenosis and consequent occlusion. Atherosclerosis, also an inflammatory condition of the arterial vessel, causes narrowing of the arterial vessel by plaque forming with calcification and potential to rupture [1]. The two conditions have similarities as both conditions are characterized by inflammations in the arterial wall and nonspecific symptoms. However, differentiation between these two conditions is of paramount importance, since the treatment of GCA consists of high-dose glucocorticoids and the side effects of this treatment can be substantial [2]. Therefore, treatment stratification is crucial to minimize diagnostic uncertainty to ensure appropriate treatment [3]. Diagnosis of both conditions relies strongly on clinically used imaging modalities: i.e. computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (US), and 2-deoxy-2-[^{18}F]fluoro-D-glucose positron emission tomography ([^{18}F]FDG-PET). As activated inflammatory cells exhibit increased glycolytic activity, the [^{18}F]FDG-PET is an indicator of inflammation [4]. However, since GCA and atherosclerosis are both characterized by inflammations, they exhibit similar elevated [^{18}F]FDG uptake. Although atherosclerosis shows a more 'patchy' pattern and GCA looks more homogenous, visually distinguishing the two conditions using [^{18}F]FDG-PET remains challenging [5].

The [^{18}F]FDG-PET images are typically evaluated using qualitative or quantitative methods. Qualitative grading is based on visual interpretation by a nuclear medicine physician [6]. The quantitative analysis involves a calculation of parameters such as the mean or maximum standardized uptake value (SUV). However, both approaches can be influenced by factors such as noise, patient-specific variables, and lack of standardized interpretation criteria [7]. Texture features derived from [^{18}F]FDG-PET images are increasingly investigated as imaging biomarkers for the quantitative description of heterogeneity [8]. These imaging biomarkers, referred to as radiomics, analyze a high number of image features containing important information that may not be appreciated by the human eye [9]. The texture analysis provides information about the distribution of specific image properties, and has therefore a high potential to capture differences in [^{18}F]FDG uptake pattern. The use of radiomics in PET-based vascular imaging has shown that texture analysis provides valuable information for plaque characterization and the diagnosis of aortitis [10][11]. While texture analysis in conjunction to machine learning (ML) has a sound technical background, using this method in clinical application is yet to be confirmed. Moreover, the ML models often lack interpretability and do not provide understandable explanations for their predictions. Since the decision-making process in ML is difficult to understand for a clinical expert, an explainable ML model is highly recommended [12]–[14].

Therefore, the aim of this study is to investigate the feasibility of an explainable radiomics-based machine

2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS) | 979-8-3503-1224-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CBMS58004.2023.00334

2372-9198/23/$31.00 ©2023 IEEE
DOI 10.1109/CBMS58004.2023.00162

learning model to differentiate between GCA and atherosclerosis in aortic [^{18}F]FDG-PET images.

II. Method

A. Study subjects and image acquisition and reconstruction

The study retrospectively included the data from twenty patients from two different cohorts: i.e., ten patients with GCA (GCA-group) [15] and ten patients with type 2 diabetes mellitus (atherosclerosis-group) [16]. People with type 2 diabetes mellitus have in general high formation of calcified atherosclerosis and as other vascular diseases were an exclusion criterium of the cohort, vascular [^{18}F]FDG update was assumed to be related to atherosclerosis [17]. To avoid interference with the atherosclerosis cohort, GCA patients with low levels of calcifications were selected. Patients with a diagnosis of GCA were included if the diagnosis of GCA stayed the same for at least 6 months after the first symptoms and a [^{18}F]FDG-PET/CT scan was performed at the time of diagnosis. Patients were excluded if they received more than three days of prednisolone therapy at the time of the [^{18}F]FDG-PET/CT scan, to prevent a false-negative [^{18}F]FDG-PET scan [18]. The inclusion and exclusion criteria of the atherosclerosis-group were previously described, but in short: patients diagnosed with diabetes mellitus type 2 were included if they had an assessable pulse wave velocity at screening and were on a stable dose of medication to regulate blood pressure- and/or lipid content [16]. Exclusion criteria were: current use of glucose-lowering drugs, diagnosis of cardiac vascular disease, and uncontrolled hypertension. All images were acquired using an integrated PET/CT system (Biograph mCT 40 or 64-slice: Siemens Healthineers, Knoxville, TN, USA). The FDG-PET/CT was imaged according to the European Association of Nuclear Medicine (EANM) procedure guidelines for imaging following EARL criteria [19]. All subjects fasted for at least six hours prior to FDG injection (3 MBq/kg). The images were acquired 60 minutes post-injection. [^{18}F]FDG-PET was acquired from skull to knee, 3 min per bed position. All images were reconstructed according to the EARL recommendations, and resulted in a voxel spacing of 3.19x3.19x2mm^3. Prior to the [^{18}F]FDG-PET, a low dose CT was acquired for attenuation correction and anatomic localization.

B. Segmentation

In the CT images, the aortas of all patients were manually annotated into volumes of interest (VOI) using Affinity Viewer (version 2.0.3; Hermes Medical Solutions, Stockholm, Sweden). Each aorta was divided into four segments: the ascending aorta, aortic arch, descending aorta, and abdominal aorta. After delineation on the low-dose CT scan, the overlaying co-registered PET image was used to ensure all arterial FDG uptake was included and any spillover uptake from neighboring tissue was excluded from the segmentation. ROIs were annotated by two clinical experts. The final dataset consisted of 80 segments: 40 segments in GCA-group and 40 segments in atherosclerosis-group.

C. Machine learning

The general approach for the machine learning (ML) model is illustrated in Fig. 1. The PET voxels were resampled using cubic B-Spline interpolation to an isotropic voxel of 2.0x2.0x2.0mm^3. The PET images were converted from activity (Bq/mL) into standardized uptake value normalized to lean body mass (SUL), as recommended for patients with a higher body fat percentage, which includes both cohorts [20]. This provides a more accurate assessment of FDG uptake.

$$SUL = \frac{Tissue\ activity}{\frac{Injected\ dose}{LBM}} \quad (1)$$

With a sex-wise LBM:

$$LBM^{\text{Male}} = 9270 \times \frac{body\ weight}{6680 + 216 \times BMI}$$

$$LBM^{\text{Female}} = 9270 \times \frac{body\ weight}{8780 + 244 \times BMI}$$

For each segment, the SUL was discretized with a fixed bin width of 0.5, defined by equation 2 [21]

$$X_{d,x} = \frac{X_{gl,x} - X_{gl,min}}{W_b} + 1 \quad (2)$$

Where $X_{d,x}$ is the resampled intensity of voxel x, $X_{gl,x}$ is the intensity of voxel x before resampling, $X_{gl,min}$ is the lowest intensity in the ROI and W_b is the bin width. $X_{d,x}$ is rounded to the nearest integer. The +1 is to make the lowest bin value equal to 1 [22]. A total of 93 radiomics features were extracted: i.e. 18 first-order statistic features (FOS), 24 grey level co-occurrence matrix (GLCM) features, 16 grey level size zone matrix (GLSZM) features, 16 grey level run-length matrix (GLRLM), five neighboring grey tone difference matrix features (NGRDM) and fourteen grey level dependence matrix (GLDM) features. Besides the 93 radiomics features,, SUL_{max} and SUL_{mean}, were extracted, making a total of 95 extracted features. Shape features were not included to exclude the influence of the manual delineations. Features were extracted using the standardized framework for radiomics in python, pyRadiomics, which is in compliance with the Image Biomarker Standardization Initiative (IBSI) guidelines [22]. For feature selection purposes, all 95 features were standardized using the Z-score calculated according to the following equation:

$$Z = \frac{X - \mu}{\sigma} \quad (3)$$

where X represents the individual feature value, μ the mean calculated over all 80 segments for the feature, and σ the standard deviation calculated over all 80 segments. After standardization, features with a high linear correlation – a Pearson correlation coefficient (Pearson's *r*) above 0.9 – were removed. The ML model was trained and validated using 80% (n=64) of the dataset and tested on a holdout test set which consist out of 20% (n=16) of the dataset. To ensure a randomized and balanced split, the data was split into train/validation and holdout sets based upon patient-wise assignment.

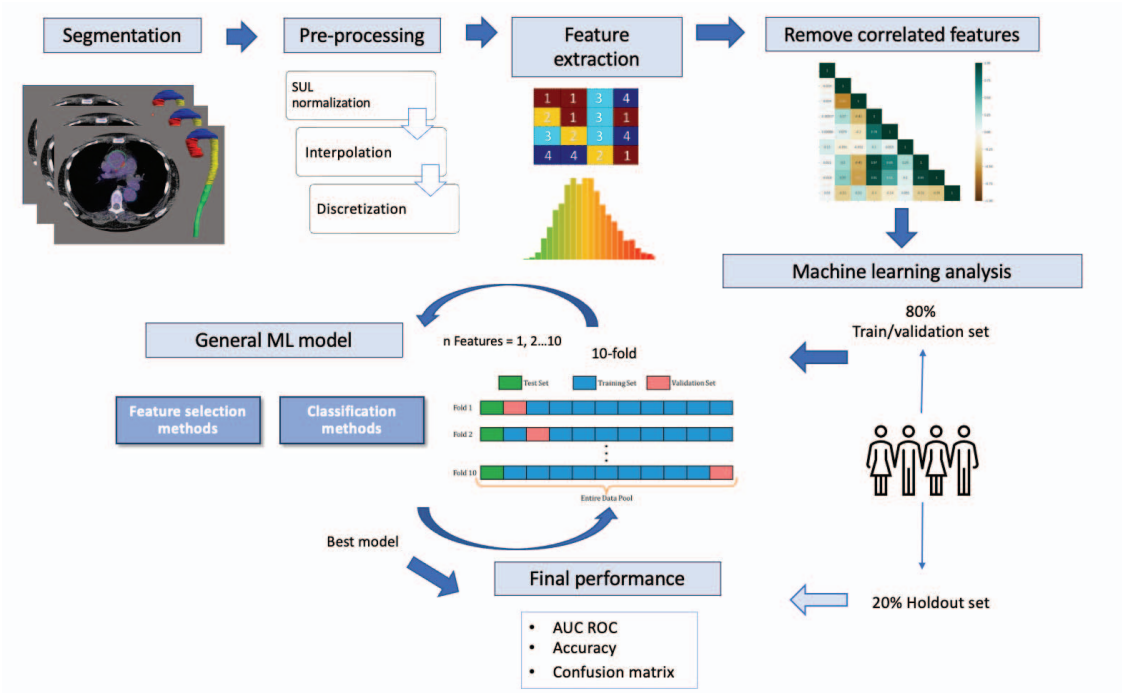


Figure 1: Flowchart of the different steps. Ninety-three radiomic features and SUL_{mean} and SUL_{max} were extracted from the 80 segments of the PET images. Feature selection and classification is trained on the training cohort(n=64), with a ten-fold cross validation. Predictive performance is evaluated on the holdout set (n=16).

The feature selection and classification model was trained by ten-fold cross-validation on the training data, randomly assigned to the segments. Four feature selection methods were used to perceive the most important features [23]: i.e., least absolute shrinkage and selection operator (LASSO), support vector machines-recursive feature elimination (SVM-RFE), ReliefF, and minimum redundancy and maximum relevance ensemble (MRMRe). These feature selection methods were chosen for their popularity in literature, computational efficiency, and publicly available implementations which increases reusability[23]. To assess the optimal number of features necessary to differentiate between GCA and atherosclerosis, ten feature selection models were built (containing an increasing number of features from one to ten)[24]. That resulted in 10x4 = 40 feature selection models. Four different classifiers were used to assess the classification quality for each of these 40 ML models: i.e. logistic regression (LR), support vector machine-radial basis function kernel (SVM-RBF), extreme gradient boosting (XGBoost), and neural network (NN), resulting in sixteen different models per ML model. For all classifiers, the following hypermeters were tuned: estimator (25, 100, 250) and learning rate (0.001, 0.05. 0.1, 0.4, 0.8). All others were set on default. Every combination of hyperparameters were tested and the combination of number of features, selection method, classifier and hyperparameter with the highest area under the receiver operating characteristic curve (AUC-ROC) was used to train the models.

To make the ML model more interpretable, feature importance of the selected features was determined using the scores of the features calculated by the feature selection methods. Moreover, an occlusion sensitivity map of the aorta segments was created using a kernel of (10x10x5), to set specific regions of the segment to zero. For every region the change in prediction probability was measured and mapped to visualize the influence per region [13], [25], [26]. The difference between the probability of the whole segment and the probability after setting regions to zero, was plotted in an occlusion map. Doing so, a heatmap from the probability differences of the aortic segment was created, depicting the importance of these regions.

All steps in feature selection and classification were executed using the Sci-kit Learn package (Version 1.1.3) of Python (Version 3.9.12).

D. Performance of ML model

The performance of all combinations: i.e. increasing number of features from one to ten, four feature selection methods and four classifiers, (10*4*4 = 160) was evaluated by the holdout set (n=16). The evaluation was performed by accuracy, positive prediction value (PPV), negative prediction value (NPV) and the AUC. For the model with the highest AUC, a ROC curve was created and evaluated.

E. Statistical analysis

Normality of the data was tested using the Shapiro-Wilk test. Depending on the normality, the baseline patient characteristics of the two groups were compared using the independent student's t-test or the Mann-Whitney U test. A p-value of < 0.05 was considered statistically significant. Continuous variables are displayed as mean with standard

deviation (mean±std) or median with range (median [range]) dependent on the normality distribution of the variable. The statistical analyses were performed using Statistical Package for the Social Sciences version 27.0 software (SPSS Inc., Chicago, IL, USA).

III. RESULTS

A. Patient characteristics

The median age of the GCA and atherosclerosis group were 63 [56 – 74] years and 64 [44 – 76] years (p=0.57), respectively. The average body mass index of the two groups was 27.3±5.4 and 32.0±7.8 (p=0.14).
Table I shows the baseline characteristics of the study population.

B. Feature extraction, selection, and classification

Of the 95 extracted features, 53 features had a Pearson's r value higher than 0.9 and were therefore removed from the dataset. The remaining dataset consisted of 42 features. The holdout set (containing 16 aorta segments) was used to assess the predictive value of the ML models for all feature selection methods and classifiers.

The AUC was used to assess the classification quality for each individual number of features, and for all four feature selection methods and all four classifier combinations Fig. 2 illustrates the highest AUC as a function of the increasing number of features used in the holdout set. Table II shows the feature selection method and classifier used for every number of features. Three and four features show the highest, identical AUC with accuracy slightly higher for three features. Table II provides the performance for different numbers of features, where a model with three features was the most optimal: i.e., accuracy=0.90±0.08, PPV=0.90±0.05, and NPV=0.90±0.05. In order of importance, these features included FOS-feature skewness, GLDM non-uniformity, and GLRLM run entropy. For FOS skewness and GLRLM run entropy, the GCA group showed significantly higher values (p<0.001). For the GLDM non-uniformity feature, there was no significant difference between the two groups (p=0.13).

Fig. 3a illustrates the mean AUC for ten-fold validation of feature selection methods (rows) and classifiers (columns) for a ML model containing three features. The ROC of this model is shown in Fig. 3b. Among all combinations, the feature selection method MRMRe and classifier NN showed the highest performance with an AUC of 0.960±0.029. MRMRe in combination with NN, LR, SVM-RBF, or XGB showed the highest AUC values in all of the cases, while the feature selection method SVM-RFE resulted in the lowest AUC values, regardless of the classifier used. The influence of the step-wise blocked 10x10x5 kernel to the ML model is illustrated as a occlusion map for the GCA aorta. The occlusion map from one of the GCA ascending aorta segments is illustrated in Fig. 4. The lighter regions in the occlusion map indicate a greater influence on the prediction outcome for GCA, while the darker regions indicate a greater influence on the prediction outcome for atherosclerosis.

TABLE I: Baseline characteristics of the two groups

Characteristics	*GCA group (n=10)*	*Atherosclerosis group (n=10)*	*p-value*
Age (years (median [range]))	63[56 – 74]	64 [44 – 76]	0.57
Sex (female)	6 (60%)	5 (50%)	0.66
BMI (kg/m^2 (mean±SD))	27.3±5.4	32.0±7.8	0.14
Diabetes mellitus type II (yes)	1 (10%)	10 (100%)	**<0.001**
FBGL (mmol/L (median [range]))	6.5 [5.2– 10.8]	6.3 [5.6 -7.2]	0.27

SD = standard deviation; BMI = body mass index; FBGL = fasting blood glucose level .

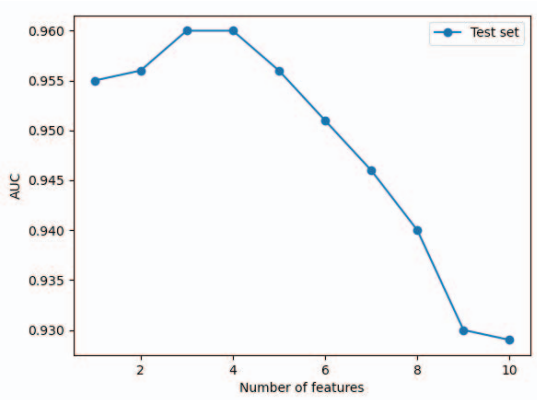
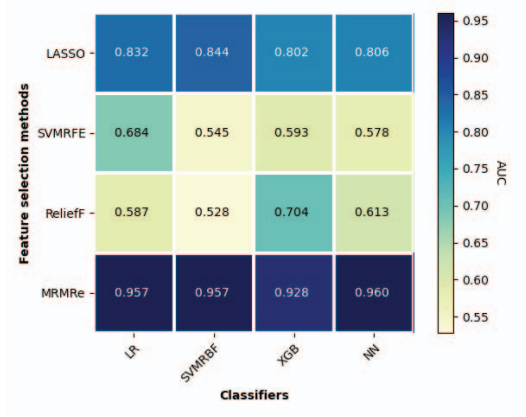


Figure 2: AUC as a function of number of features for the holdout set

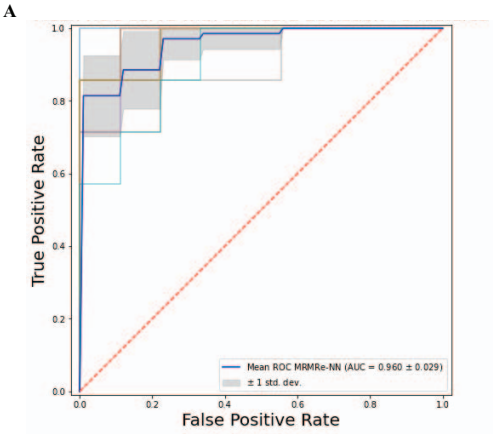
TABLE II: Performance evaluation for different number of features, holdout set (n=16), mean of 10-fold. The sum of TP, TN, FP, FN = 16.

Number of features	TP	TN	FP	FN	Accuracy (SD)	AUC (SD)	FS method - classifier
1	6.3	7.3	1.7	0.7	0.85±0.05	0.955 ±0.022	MRMRe – NN
2	5.3	8.4	0.6	1.7	0.86±0.05	0.956 ±0.012	MRMRe – NN
3	6.1	8.3	0.7	0.9	**0.90±0.08**	**0.960** ±0.029	MRMRe – NN
4	6.1	7.4	1.6	0.9	0.84±0.07	**0.960** ±0.023	MRMRe – SVM-RBF
5	6.3	7.3	1.7	0.7	0.85±0.05	0.956 ±0.024	MRMRe - SVM-RBF
6	6.2	7.1	1.9	0.8	0.83±0.04	0.951 ±0.019	MRMRe - LR
7	6.1	6.9	2.1	0.9	0.81±0.04	0.946 ±0.026	MRMRe - LR
8	6.1	6.7	2.3	0.9	0.80±0.05	0.940 ±0.028	MRMRe - LR
9	5.9	6.6	2.4	1.1	0.78±0.05	0.930 ±0.031	MRMRe - LR
10	5.4	8	1	1.6	0.84±0.07	0.929 ±0.037	MRMRe – XGB

TP = true positive; TN = true negative; FP = false positive; FN = false negative; AUC = area under the operator curve.



A



B

Figure 3: A) Heatmap presenting performance in terms of AUC for different feature selection methods and classifiers. The combination of MRMRe and NN shows the highest predictive performance for three selected features .B) ROC curve the best performing combination of feature selection method (MRMRe) and classifier (NN) for ten folds, number of features = 3.

IV. DISCUSSION

In this study, we demonstrated the feasibility for radiomic analysis to differentiate between GCA and atherosclerosis in the aorta based upon [^{18}F]FDG-PET images. The results of our study indicated that among all the combinations tested, the feature selection method MRMRe in conjunction with the classifier NN exhibited the best performance. Furthermore, the ML model with three features had a highest AUC of 0.96±0.029. This model employed the following three features to distinguish GCA from atherosclerosis: FOS-skewness, GLDM non-uniformity, and GLRLM run entropy. Positive values for skewness indicate that there is a substantial number of bright pixels within the ROI, indicating higher [^{18}F]FDG uptake[22]. This is in accordance with different studies showing a higher intensity for GCA compared with atherosclerosis [5], [27], [28], [29]. GLDM and GLRLM are both textural features that quantify heterogeneity by analyzing the spatial arrangement of voxel values [30]. GLDM non-uniformity measures the similarity of grey-level intensity values, whereas GLRLM run entropy assesses the randomness in the grey levels. For both, a higher value indicating a more heterogeneous texture pattern [22]. In our study, there was not a significant difference between the groups for the GLDM non-uniformity feature (p=0.13). The value for GLRLM run entropy is significantly higher in the GCA group compared with the atherosclerosis group (p<0.001). This finding indicates a higher randomness of texture in GCA, which is not in accordance with visual findings, stating a more homogenous pattern in GCA [5], [6]. The outcome from our study suggest that high values are observed in both groups, only values for the GCA group are higher. Visually it can be difficult to notice the difference, the ML model, on the other hand, can distinguish these differences using the extracted features.

In radiomics, there is no consensus on the final number of features. Some studies suggest that the final number of features should be approximately ten percent of the number of patients [31], while other studies suggest any number regardless of the number of patients [30]. To decrease the chance of overfitting we set the maximum number of features to ten.

To the best of our knowledge, only two previous studies, both from Duff et al. used radiomics to diagnose GCA, showing high performance with AUC scores ranging from 0.80 to 1.00 [11], [32]. However, the dataset was pruned to exclude GCA-patients without imaging evidence of active GCA, and to exclude control subjects with atherosclerosis.

The occlusion map of the aorta segment showed that the lighter the regions, the higher the positive probability difference. This suggest that these regions contain important information for GCA prediction. On the other hand, darker regions, with a negative probability difference, contain most likely information for atherosclerosis prediction. Overlaying the occlusion map allows for the nuclear medicine physician to identify the regions of importance for decision-making.

There are some limitations to this study. Firstly, the small dataset potentially limits the generalizability of the results and the ability of the model to perform well on new data [33]. Therefore, a larger dataset is preferable to validate the model. Moreover, the limited sample size increases the risk of overfitting. Although feature selection methods and a holdout set were used to avoid overfitting, there was still a possibility of overfitting. Secondly, this is a single-center, single-scanner retrospective study. To further develop, validate and generalize the predictive model, a prospective multi-center study is needed. Furthermore, the aorta is manually segmented by two observers. However, a (semi)-automated segmentation method would be preferred to increase reproducibility and decrease human error.

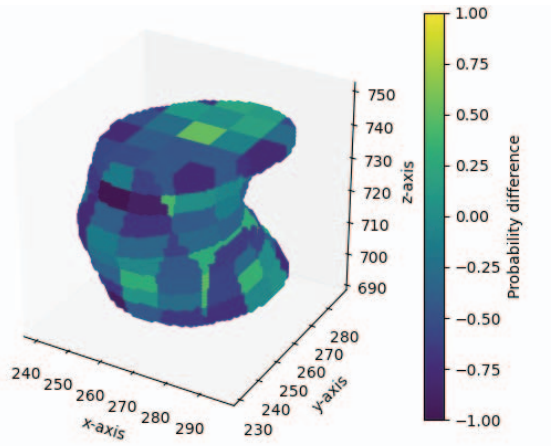


Figure 4: Occlusion map of aorta segment (kernel size 10x10x5), showing which regions contain important information for the decision-making process of the ML model.

V. CONCLUSION

In conclusion, our findings demonstrate that the utilization of the feature selection method MRMRe in combination with the classifier NN resulted in a high AUC and accuracy for differentiation between GCA and atherosclerosis in [^{18}F]FDG-PET images. Moreover, feature importance and an occlusion map of the aorta segments can help to understand the decision-making process of the ML model.

REFERENCES

[1] J. C. Jennette, "Overview of the 2012 revised International Chapel Hill Consensus Conference nomenclature of vasculitides," *Clin Exp Nephrol*, vol. 17, no. 5, pp. 603–606, 2013, doi: 10.1007/s10157-013-0869-6.

[2] P. H. Nienhuis, G. D. van Praagh, A. W. J. M. Glaudemans, E. Brouwer, and R. H. J. A. Slart, "A review on the value of imaging in differentiating between large vessel vasculitis and atherosclerosis," *J Pers Med*, vol. 11, no. 3, p. 236, 2021.

[3] K. S. M. van der Geest, M. Sandovici, E. Brouwer, and S. L. Mackie, "Diagnostic Accuracy of Symptoms, Physical Signs, and Laboratory Tests for Giant Cell Arteritis: A Systematic Review and Meta-analysis," *JAMA Intern Med*, vol. 180, no. 10, pp. 1295–1304, Oct. 2020, doi: 10.1001/jamainternmed.2020.3050.

[4] S. Vaidyanathan, C. N. Patel, A. F. Scarsbrook, and F. U. Chowdhury, "FDG PET/CT in infection and inflammation—current and emerging clinical applications," *Clin Radiol*, vol. 70, no. 7, pp. 787–800, 2015, doi: https://doi.org/10.1016/j.crad.2015.03.010.

[5] O. Espitia *et al.*, "Specific features to differentiate Giant cell arteritis aortitis from aortic atheroma using FDG-PET/CT," *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-96923-2.

[6] R. H. J. A. Slart *et al.*, "FDG-PET/CT(A) imaging in large vessel vasculitis and polymyalgia rheumatica: joint procedural recommendation of the EANM, SNMMI, and the PET Interest Group (PIG), and endorsed by the ASNC," *Eur J Nucl Med Mol Imaging*, vol. 45, no. 7, pp. 1250–1269, 2018, doi: 10.1007/s00259-018-3973-8.

[7] M. Casali *et al.*, "State of the art of 18F-FDG PET/CT application in inflammation and infection: a guide for image acquisition and interpretation," *Clin Transl Imaging*, vol. 9, no. 4, pp. 299–339, 2021, doi: 10.1007/s40336-021-00445-w.

[8] L. Alic, W. J. Niessen, and J. F. Veenland, "Quantification of Heterogeneity as a Biomarker in Tumor Imaging: A Systematic Review," *PLoS One*, vol. 9, no. 10, pp. e110300-, Oct. 2014, [Online]. Available: https://doi.org/10.1371/journal.pone.0110300

[9] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—'how-to' guide and critical reflection," *Insights Imaging*, vol. 11, no. 1, p. 91, 2020, doi: 10.1186/s13244-020-00887-2.

[10] P. P. Kafouris, I. P. Koutagiar, A. T. Georgakopoulos, G. M. Spyrou, D. Visvikis, and C. D. Anagnostopoulos, "Fluorine-18 fluorodeoxyglucose positron emission tomography-based textural features for prediction of event prone carotid atherosclerotic plaques," *Journal of Nuclear Cardiology*, vol. 28, no. 5, pp. 1861–1871, 2021, doi: 10.1007/s12350-019-01943-1.

[11] L. Duff *et al.*, "A methodological framework for AI-assisted diagnosis of active aortitis using radiomic analysis of FDG PET–CT images: Initial analysis," *Journal of Nuclear Cardiology*, 2022, doi: 10.1007/s12350-022-02927-4.

[12] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable Deep Learning Methods in Medical Diagnosis: A Survey," *arXiv preprint arXiv:2205.04766*, 2022.

[13] Christoph Molnar, *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. 2022.

[14] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[15] Y. van Sleen *et al.*, "Angiopoietin-2/-1 ratios and MMP-3 levels as an early warning sign for the presence of giant cell arteritis in patients with polymyalgia rheumatica," *Arthritis Res Ther*, vol. 24, no. 1, p. 65, 2022, doi: 10.1186/s13075-022-02754-5.

[16] S. A. de Boer *et al.*, "Arterial stiffness is positively associated with 18F-fluorodeoxyglucose positron emission tomography-assessed subclinical vascular inflammation in people with early type 2 diabetes," *Diabetes Care*, vol. 39, no. 8, pp. 1440–1447, 2016, doi: 10.2337/dc16-0327.

[17] J. Goodfellow *et al.*, "Endothelium and inelastic arteries: an early marker of vascular dysfunction in non-insulin dependent diabetes," *Bmj*, vol. 312, no. 7033, pp. 744–745, 1996.

[18] B. D. Nielsen, L. C. Gormsen, I. T. Hansen, K. K. Keller, P. Therkildsen, and E.-M. Hauge, "Three days of high-dose glucocorticoid treatment attenuates large-vessel 18F-FDG uptake in large-vessel giant cell arteritis but with a limited impact on diagnostic accuracy," *Eur J Nucl Med Mol Imaging*, vol. 45, no. 7, pp. 1119–1128, 2018, doi: 10.1007/s00259-018-4021-4.

[19] A. Kaalep *et al.*, "EANM/EARL FDG-PET/CT accreditation - summary results from the first 200 accredited imaging systems," *Eur J Nucl Med Mol Imaging*, vol. 45, no. 3, pp. 412–422, 2018, doi: 10.1007/s00259-017-3853-7.

[20] G. D. van Praagh *et al.*, "Toward reliable uptake metrics in large vessel vasculitis studies," *Diagnostics*, vol. 11, no. 11, p. 1986, 2021.

[21] R. T. H. Leijenaar *et al.*, "The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis," *Sci Rep*, vol. 5, no. 1, p. 11075, 2015, doi: 10.1038/srep11075.

[22] "IBSI - Image Biomarker Standardisation Initiative," *(Available At Https://Theibsi.Github.Io)*, vol. 0, no. December, 2019, [Online]. Available: https://theibsi.github.io

[23] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Comput Stat Data Anal*, vol. 143, p. 106839, 2020, doi: https://doi.org/10.1016/j.csda.2019.106839.

[24] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. J. W. L. Aerts, "Machine Learning methods for Quantitative Radiomic Biomarkers," *Sci Rep*, vol. 5, no. 1, p. 13087, 2015, doi: 10.1038/srep13087.

[25] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[26] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 818–833.

[27] M. D. Stellingwerff *et al.*, "Different Scoring Methods of FDG PET/CT in Giant Cell Arteritis: Need for Standardization," *Medicine*, vol. 94, no. 37, 2015, [Online]. Available: https://journals.lww.com/md-journal/Fulltext/2015/09030/Different_Scoring_Methods_of_FDG_PET_CT_in_Giant.36.aspx

[28] N. Pipitone, A. Versari, and C. Salvarani, "Role of imaging studies in the diagnosis and follow-up of large-vessel vasculitis: an update," *Rheumatology*, vol. 47, no. 4, pp. 403–408, Apr. 2008, doi: 10.1093/rheumatology/kem379.

[29] J. ben Shimol, H. Amital, M. Lidar, L. Domachevsky, Y. Shoenfeld, and T. Davidson, "The utility of PET/CT in large vessel vasculitis," *Sci Rep*, vol. 10, no. 1, p. 17709, 2020, doi: 10.1038/s41598-020-73818-2.

[30] S. Reuzé *et al.*, "Radiomics in Nuclear Medicine Applied to Radiation Therapy: Methods, Pitfalls, and Challenges.," *Int J Radiat Oncol Biol Phys*, vol. 102, no. 4, pp. 1117–1142, Nov. 2018, doi: 10.1016/j.ijrobp.2018.05.022.

[31] B. Koçak, E. Ş. Durmaz, E. Ateş, and Ö. Kılıçkesmez, "Radiomics with artificial intelligence: a practical guide for beginners.," *Diagn Interv Radiol*, vol. 25, no. 6, pp. 485–495, Nov. 2019, doi: 10.5152/dir.2019.19321.

[32] L. M. Duff *et al.*, "An Automated Method for Artifical Intelligence Assisted Diagnosis of Active Aortitis Using Radiomic Analysis of FDG PET-CT Images," *Biomolecules*, vol. 13, no. 2, p. 343, 2023.

[33] S. S. F. Yip and H. J. W. L. Aerts, "Applications and limitations of radiomics," *Phys Med Biol*, vol. 61, no. 13, p. R150, 2016.