Journal of Nuclear Cardiology (xxxx) xxx, xxx

ORIGINAL RESEARCH

Explainable deep-learning-based ischemia detection using hybrid O-15 H_2O perfusion positron emission tomography and computed tomography imaging with clinical data

Jarmo Teuho, PhD [1,2,3,*], Jussi Schultz, MD, PhD [2,3], Riku Klén, PhD [2,3], Luis Eduardo Juarez-Orozco, MD, PhD [4,5], Juhani Knuuti, MD, PhD [2,3], Antti Saraste, MD, PhD [3,6], Naoaki Ono, PhD [1,7], Shigehiko Kanaya, PhD [1,7]

[1] Data Science Center, Nara Institute of Science and Technology, Nara, Japan
[2] Turku PET Centre, University of Turku, Turku, Finland
[3] Turku PET Centre, Turku University Hospital, Turku, Finland
[4] Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands
[5] Department of Cardiology, Meader Medical Center, Amersfoort, the Netherlands
[6] Heart Centre, Turku University Hospital and University of Turku, Turku, Finland
[7] Department of Science and Technology, Nara Institute of Science and Technology, Nara, Japan

[*] Corresponding author. Turku PET Centre c/o Turku University Hospital, Kiinamyllynkatu 4-8, 20520 Turku, Finland.
E-mail address: jatateu@utu.fi (Jarmo Teuho).

Abstract

Background: We developed an explainable deep-learning (DL)-based classifier to identify flow-limiting coronary artery disease (CAD) by O-15 H_2O perfusion positron emission tomography computed tomography (PET/CT) and coronary CT angiography (CTA) imaging. The classifier uses polar map images with numerical data and visualizes data findings.

Methods: A DLmodel was implemented and evaluated on 138 individuals, consisting of a combined image—and data-based classifier considering 35 clinical, CTA, and PET variables. Data from invasive coronary angiography were used as reference. Performance was evaluated with clinical classification using accuracy (ACC), area under the receiver operating characteristic curve (AUC), F1 score (F1S), sensitivity (SEN), specificity (SPE), precision (PRE), net benefit, and Cohen's Kappa. Statistical testing was conducted using McNemar's test.

Results: The DL model had a median ACC = 0.8478, AUC = 0.8481, F1S = 0.8293, SEN = 0.8500, SPE = 0.8846, and PRE = 0.8500. Improved detection of true-positive and false-negative cases, increased net benefit in thresholds up to 34%, and comparable Cohen's kappa was seen, reaching similar performance to clinical reading. Statistical testing revealed no significant differences between DL model and clinical reading.

Conclusions: The combined DL model is a feasible and an effective method in detection of CAD, allowing to highlight important data findings individually in interpretable manner.

Keywords: CAD, PET/CT, Myocardial perfusion, Deep learning, Explainability

Received 8 February 2024; Accepted 25 May 2024; Available online XXX

https://doi.org/10.1016/j.nuclcard.2024.101889

ABBREVIATIONS

ACC	accuracy
AUC	area under the receiver operating characteristic curve
CAD	coronary artery disease
CTA	coronary computed tomography
DL	deep learning
F1S	F1 score
GRAD-CAM	gradient-based class activation mapping
ICA	invasive coronary angiography
SHAP	shapely values
sMBF	stress myocardial blood

INTRODUCTION

Deep learning (DL) can introduce breakthrough applications in nuclear cardiology, for which one of the foremost is classification of clinical subjects [1,2]. DL methods for automated classification can bring benefits in diagnostic accuracy, decision-making, and clinical throughput by speeding up the diagnosis [3]. There is still unexploited potential in applying DL in evaluation of coronary artery disease (CAD) [4], conducted by myocardial perfusion imaging (MPI) with O-15H_2O hybrid positron emission tomography computed tomography (PET/CT).

Diagnostic work-up of CAD includes analyzing data from imaging tests and clinical risk factors, such as age, gender, diabetes, dyslipidemia, and hypertension [5–7]. Hybrid PET/CT imaging enables combined analysis of myocardial perfusion and detection of coronary stenosis with coronary CT angiography (CTA) [8]. Coronary CTA has high sensitivity enabling exclusion of obstructive CAD, whereas its specificity is lower [8,9]. This complementary information is combined in the diagnostic decision chain, and any DL approaches should operate on a similar manner. DL approaches can highlight useful target variables from this rich source of information and highlight regions in the images with subjects at risk to be further explored.

The use of current DL approaches in clinical decision-making is limited by lack of transparency or operating in a "black-box" fashion [10,11]. Ability to visualize the DL decision process and to indicate the priority of the input parameters is essential for clinical applications [12]. Interpretable models can result in increased trust from practitioners, can indicate new insights, and can assist the model design, training, and applicability [13]. This is advantageous in complex models using images and data, with more interconnections available. Thus, an intuitive explanation on which variables the model considered significant on an individual basis is desired.

There is a limited number of studies on the implementation of DL for the identification of myocardial ischemia in PET MPI [14]. Combined approaches using clinical features and image data from single-photon emission computed tomography (SPECT) MPI [10,15] or PET MPI [16] have been studied for diagnosis of CAD or event prediction. We have previously investigated detection of CAD using O-15 H_2O PET and only polar map images [17,18]. However, no studies using hybrid DL-based classifiers for the detection of myocardial ischemia, which operate on a combination of polar map images, CTA, quantitative stress myocardial blood flow (sMBF) values, and clinical data currently exist. Despite the vast amount of research performed, explainable DL models using both images and clinical variables have not been extensively explored. Our motivation is to build upon these prior findings to fill the gap on the implementation of hybrid DL approaches for the identification of myocardial ischemia in O-15 H_2O PET MPI and design an approach that improves the model explainability and interpretability.

This is achieved by combining image and data classification in a single, explainable DL pipeline for the detection of flow-limiting CAD, similar to diagnostic reading. Visual representation of blood flow as polar map images, quantitative sMBF values, data on coronary atherosclerosis, and anatomical coronary stenosis based on CTA as well as clinical characteristics are used in combination. To advance beyond the "black-box" approach and to understand which variables are given more weight by the model, we implement explainability methods for visualization and justification of the classification results on an individual basis. These include highlighting of perfusion defects on the polar maps as well as ranking of imaging and clinical variables, which affected positively or negatively to the classification result.

Our proposed method uses a combination of different sources of data from combined coronary CTA and O-15 H_2O PET myocardial perfusion imaging. The method adds on to the growing structure of application of DL in hybrid cardiovascular PET/CT imaging by using a combination of several variables including the following: a) features extracted from the polar maps indicating visual data, b) quantitative results from the PET perfusion examination as sMBF values, c) anatomical information from coronary CTA, as well as d) clinical patient-specific risk variables. An important aspect of our work is to use CTA

data in combination with polar map images and sMBF information, which has not been done previously. Finally, explainability methods have not been leveraged to the degree presented in this study for hybrid O-15 H_2O PET/CT imaging.

MATERIALS AND METHODS

Study population

The study was conducted in a prospective cohort of 138 individuals (73 men) with suspected obstructive CAD who had undergone coronary CTA and O-15 H_2O PET/CT perfusion study during adenosine stress at Turku PET Centre at the Turku University Hospital in Finland. All patients had invasive coronary angiography (ICA). The population characteristics with details from the image acquisition and analysis are given in detail in Refs. [8,18–20]. Details of the cohort with short description of the clinical reading are given in the following.

The cohort contains in total of 56 patients with significant obstructive CAD, with 36 individuals in the training set and 20 in the separate hold-out (test) set. Significant obstructive CAD is defined based on ICA, with hemodynamic significance confirmed by intracoronary fractional flow reserve (FFR) measuring <0.80 when appropriate. This information was used to define the reference CAD labels that were compared vs the clinical reading and the DL model predictions.

The human readers analyzing the CTA and PET were blinded to the results of ICA and FFR analysis, which was performed by an expert blinded to the CTA and PET results. For the clinical reading, the human readers had the CTA and PET data at their disposal and would weigh the final decision based on both data types and their concordance. In comparison, the DL model had all data available on a single setting, basing its decision on all the variables together.

The study was approved by an institutional review committee of the Hospital District of Southwest Finland and was conducted according to the guidelines of the Declaration of Helsinki. All individuals gave informed consent.

Polar map and clinical data preprocessing

The data consisted of polar maps and numerical data. Polar maps represent the distribution of quantitative myocardial blood flow during adenosine stress (sMBF) in the left ventricular myocardium. They contain information for quantitative and visual PET reading. The numerical CTA and clinical data contain information from patient-specific risk factors and potential coronary stenosis, which could indicate the presence of flow-limiting CAD.

A database of individuals was constructed, containing polar map images prepared identically to our previous study [18], with 35 numerical and categorical variables. Clinical variables were selected based on known clinical association with the prevalence of obstructive CAD. These consisted of a) purely clinical variables, b) PET results including the modeled sMBF value (mL/g/min) in 14 polar map segments excluding segments 2, 3, 17 due to variability [21], and c) CTA results describing the degree of stenosis in the main arteries and branches as well as the coronary calcium score [22]. A reference label for each individual was determined based on the ICA data (1 = ischemic, 0 = non-ischemic).

The number of missing variables is given in Supplemental Table 1, with a technical description of preprocessing the variables is given in Supplemental Methods. All variables are summarized in Table 1.

Continuous variables are presented as median with interquartile range. Categorical variables are expressed as counts and corresponding percentages. Statistical testing was conducted between the training and test datasets using the Wilcoxon rank-sum test for continuous variables and the chi-squared test for categorical ones, with P value < 0.05 denoting statistical significance.

Deep-learning pipeline description

The DL model is a custom approach, a result of careful curation and experimentation of different models, their combinations, and designs. Several approaches, mentioned in Ref. [17], and their combinations using ensembles were experimented, but none of them proved to be superior to our image-only approach in Ref. [18]. Building on our previous research, we designed an entirely new, explainable fusion approach for this specific purpose, which we refer specifically as the DL+data model.

The DL pipeline is shown in Figure 1, containing a) an image-based classifier for visualization of detected perfusion defects and image feature extraction and b) a data-based classifier, with a dual-input model concatenating the image features from input a) and the clinical data.

Image data are fed to the model in two complementary formats. The first are the sMBF values given directly to the data-based classifier, and the second are the features extracted from the polar map images by the image-based classifier. The features are low-level textural representations gained from the deep layers of the model and can be considered complementary to the sMBF

Table 1. List of all imaging and clinical variables. Continuous variables are presented as median with interquartile range in parenthesis. Categorical variables are expressed as counts and corresponding percentages in parentheses.

	All Patients (N = 138)	Training (N = 92)	Test (N = 46)	P Value
Clinical variables				
Age	62 (10.0)	63 (10.5)	61 (7.0)	0.391
Sex (M/F)	73/65	53/39	20/26	0.117
BMI	26.3 (4.7)	26.3 (3.8)	26.5 (5.5)	0.956
Family history of CAD	70 (50.7)	48 (52.2)	22 (47.8)	0.806
Angina[a]	85 (61.6)	56 (60.9)	29 (63.0)	0.835
Diabetes	114 (82.6)	77 (83.7)	37 (80.4)	0.827
Smoking	83 (60.1)	55 (59.8)	28 (60.9)	0.863
Dyslipidemia	33 (23.9)	18 (19.6)	15 (32.6)	0.197
Hypertension	55 (39.9)	38 (41.3)	17 (37.0)	0.798
PET sMBF[b]				
Segment 1	3.09 (1.45)	2.75 (1.46)	3.21 (1.44)	0.094
Segment 4	3.15 (1.52)	3.10 (1.64)	3.19 (1.24)	0.766
Segment 5	3.18 (1.58)	2.95 (1.52)	3.36 (1.32)	0.209
Segment 6	3.18 (1.43)	2.93 (1.48)	3.31 (1.14)	0.088
Segment 7	3.20 (1.70)	3.01 (1.57)	3.52 (1.57)	**0.011**
Segment 8	2.85 (1.60)	2.66 (1.56)	3.13 (1.53)	0.061
Segment 9	2.78 (1.51)	2.63 (1.62)	2.98 (1.16)	0.407
Segment 10	3.23 (1.72)	3.29 (1.77)	3.17 (1.42)	0.871
Segment 11	3.60 (1.99)	3.52 (2.07)	3.67 (1.78)	0.671
Segment 12	3.40 (1.63)	3.22 (1.69)	3.63 (1.38)	0.074
Segment 13	3.35 (1.79)	3.03 (1.81)	3.90 (1.58)	**0.016**
Segment 14	3.06 (1.90)	2.87 (2.03)	3.25 (1.51)	0.104
Segment 15	3.31 (1.78)	3.34 (1.81)	3.26 (1.63)	0.701
Segment 16	3.75 (1.85)	3.47 (2.20)	3.87 (1.30)	0.255
CTA findings[c]				
LM	56 (40.6)	42 (45.7)	14 (30.4)	0.196
LADA	93 (67.4)	66 (71.7)	27 (58.7)	0.149
LADB	77 (55.8)	55 (59.8)	22 (47.8)	**0.047**
LADC	40 (29.0)	32 (34.9)	8 (17.4)	0.417
LCXA	54 (39.1)	40 (43.5)	14 (30.4)	0.482
LCXB	40 (29.0)	32 (34.8)	8 (17.4)	0.306
RCAA	68 (49.3)	50 (54.3)	18 (39.1)	0.485
RCAB	51 (37.0)	34 (37.0)	17 (37.0)	0.479
RCAC	39 (28.3)	29 (31.5)	10 (21.7)	0.827
D1	39 (28.3)	29 (31.5)	10 (21.7)	**0.047**
RPD	15 (10.9)	14 (15.2)	29 (63.0)	0.063
Coronary calcium score	106 (406.3)	106.5 (477.5)	88 (314.8)	0.350

P value < 0.05 signifies statistical significance and is given in bold.

BMI, body mass index; *CAD*, coronary artery disease; *CTA*, coronary computed tomography; *PET*, positron emission tomography; *sMBF*, stress myocardial blood flow; *LM*, left main artery; *LADA, LADB, LADC*, the proximal, middle, and distal left anterior descending coronary artery; *M*, male; *F*, female; *LCXA, LCXB*, the proximal and middle left circumflex artery; *RCAA, RCAB, RCAC*, the proximal, middle and distal right coronary artery; *D1*, the first diagonal branch; *RPD*, the right posterior descending branch.

[a] Typical or atypical angina.

[b] Segments 2, 3, and 17 are excluded from the analysis.

[c] Grade 2–5.

values, guiding the visual interpretation of the polar maps by the DL+data model.

To visualize the regions that the model considers ischemic, we implemented gradient-based class activation mapping (GRAD-CAM) [23]. GRAD-CAM has been used in medical and non-medical contexts to visualize which spatial patterns in the image are considered important.

The data-based classifier concatenates the extracted-image features and clinical data. The first input layer receives extracted-image-based features from the flatten layer of the image classifier model. The second input layer receives directly the preprocessed clinical data as a data-frame. Both are concatenated and passed to a neural network classifier. The classifier uses both data types in combination to predict the final probability for ischemia in the individual, a value between 0 and 1.

To visualize which variables the model considers important in the final decision, we calculated the Shapely values (SHAP values, [24]). SHAP values have been proven to facilitate the explanation of highly non-linear models, allowing to

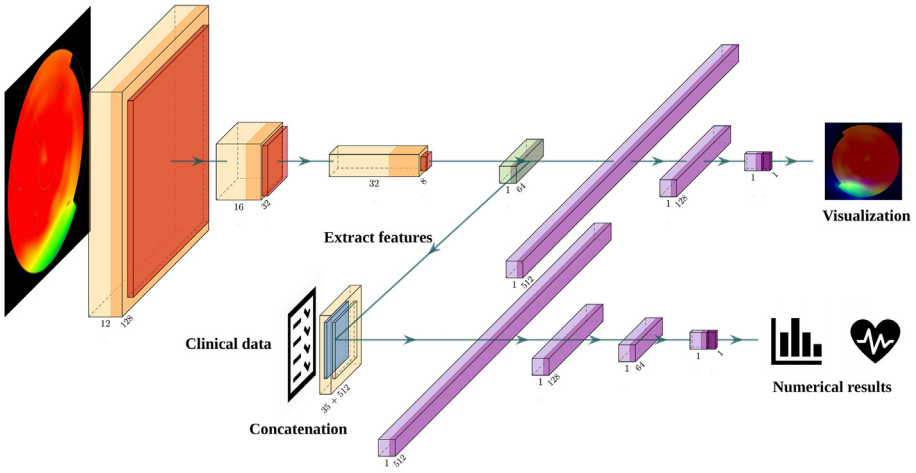

Figure 1. A description of the hybrid DL pipeline. The image classification model (upper model) takes only the JPEG polar map as an input and performs gradient-based class activation mapping (GRAD-CAM) visualizations. The polar map data are processed by convolutional layers until the image features are extracted before the first flatten layer. The features are given as an input to the dual-input model (lower model), which takes the raw tabulated clinical data (35 variables) as a secondary input. Both the image features (512 variables) and the tabulated data are concatenated, thereafter a neural network performs the final classification and outputs the numerical results consisting of a single value as a probability between 0 and 1, in addition to giving also Shapely values (SHAP). The max pool layers (red) contain a 2 × 2 window. The number of filters for each convolutional layer increases gradually as: 12, 16, and 32. The kernel sizes for all layers are 3 × 3, whereas strides were set as 2 × 2. The image classifier model dense layer sizes are 512 and 128. The data classifier model contains dense layers of 512, 128, and 64 in size, with 0.5 drop-out in between. Abbreviation: *DL*, deep learning.

break down the impact of input features on prediction [6]. They have been recently used as a tool for interpreting DL models and have been shown also to follow human intuition in the interpretation of model predictions. SHAP values are calculated for clinical data and image features and are visualized as an easily interpretable bar plot to explain which variables and features the model considered important.

The SHAP values represent how strongly each of the given feature (e.g., clinical variable) contributed to increasing or decreasing the probability of detecting ischemia in a specific patient. Consequently, this allows any outside reader to interpret what features a) the model considered important when making the final prediction and b) which affected the model judgement either positively or negatively. The final outputs from the pipeline are the following: 1) visualization of the GRAD-CAM maps highlighting perfusion defects on the polar map images, divided to each coronary artery segment, 2) SHAP values visualized as bar plots from the clinical data and the image features to indicate which variables contributed most significantly to the final decision of the classifier, and 3) a probability value based on image and clinical data to indicate whether the subject is ischemic or not.

A detailed technical description from the development of the DL processing pipeline, with the hyperparameters used, is given in Supplementary Methods. The final selection of the hyperparameters were decided on the training performance during k-fold cross-validation. The set of hyperparameters with the highest area under the receiver operating characteristic curve (AUC) was chosen and fixed for the final model, an optimization approach used in Ref. [25]. Thus, the hyperparameter selection was conducted on experimental basis, and no automatic optimization libraries were applied.

Training, validation, and testing procedures

Figure 2 shows an overview of training and testing procedures, where we follow the recommended approaches for evaluation of DL algorithms [26].

The dataset was divided with a single split of the entire data to separate training and testing sets. The training set consisted of 92 individuals. A separate hold-out (test) set containing 46 subjects was used for model testing only and calculation of the final results, the same split of data as in as in Ref. [18] was applied to make our results comparable.

For training, we applied a 10-fold repeated (N = 10 repeats) cross-validation using the dataset of 92 individuals. Training data were used to fix the model hyperparameters and evaluate the training performance. Repeated cross-validation was selected to ensure consistent performance and to minimize noise between different cross-validation sets. The training results of the model

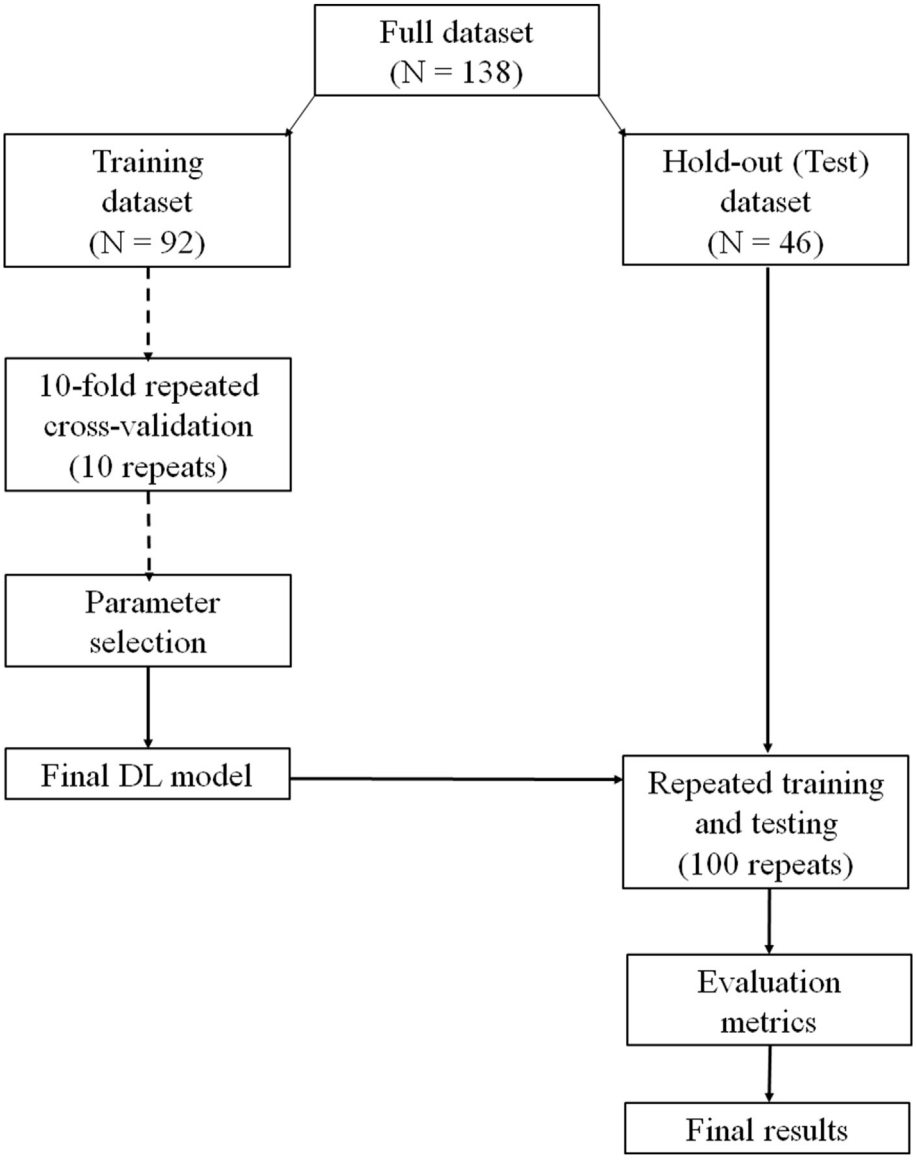

Figure 2. A workflow diagram of the training and testing performed to the DL pipeline. A single split of the data to training (92 cases) and test (46) datasets was performed. K-fold cross-validation was performed to train the model using the training set, with adjustment of the model hyperparameters. The final model performance evaluation was applied to test dataset only by repeated training and testing for 100 runs in total. Abbreviation: *DL*, deep learning.

from the 10-fold repeated cross-validation are given in Supplementary Table 2.

For model testing, we applied a separate hold-out (test) set of 46 individuals. The prediction performance and stability of the final DL+data model was evaluated with this set, which remained unseen during model training. The final testing was performed similarly to Ref. [18]. The evaluation included repeated training followed by hold-out testing with 100 repeats to estimate the variation of the model due to random number initialization.

The predicted probability values with the ICA reference labels during training and testing were saved to evaluate the model performance and stability.

Data analysis and visualization

All data visualization and analysis were performed with MATLAB v2020b (MathWorks Inc., US). The visualization of polar map data, GRAD-CAM visualization maps, and SHAP values were performed using the data processing libraries in Python 3.8.10.

Classification accuracy metrics

The prediction performance was evaluated using commonly applied classification evaluation metrics. After validation of the model performance and tuning of the hyperparameters using k-fold cross-validation, a comparison of the model prediction performance to a clinical observer and our

previously published image-only DL model in Ref. [18] was performed using the hold-out (test) dataset.

All the metrics were calculated in comparison of the reference labels from ICA. The probability value given by the DL+data model was converted to a binary value using a threshold of 0.5, where probability values greater than the threshold were given a value of 1. This binary label was then compared to the predictions of the clinical observer and reference CAD label.

The metrics were accuracy (ACC), area under the receiver operating characteristic curve (AUC), F1 score (F1S), sensitivity (SEN), specificity (SPE) and precision (PRE). F1 score provides a measure of the trade-off between false-positive and false-negative results in the test population. Furthermore, PRE in this context does not refer to the variability of the measurement but indicates which proportion of positive identifications were actually correct. In addition, we calculated the number of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs). All metrics are reported as median with the interquartile range over the different runs for the DL+data model. The equations for calculating the metrics are given in Supplementary Methods.

Net benefit

We calculated the net benefit [27] to compare the DL+data model with the clinical reading in the test set. Net benefit is a relatively new decision analytic technique, which allows to compare two models, markers or tests in regard whether they would lead to better clinical outcomes on average among suitable patients, and whether either of the techniques would be better than a default strategy of treating all patients or none. By definition, an optimal technique would perform better than alternative approaches, across a wide range of threshold probabilities [27]. In the analysis, both the DL+data model and clinical reading were assessed vs the CAD label based on ICA plus FFR data as the ground truth. The net benefit was visualized in a decision curve over threshold probabilities from 0% to 100%.

Agreement with the reference CAD data

The agreement between the ICA results compared to the DL+data model, and clinical reading was investigated by calculation of Cohen's Kappa coefficient κ for each of the 100 runs performed on the test set. We report the median value and the interquartile range of κ between the DL+data model and the clinical reading against the ICA labels.

Statistical testing

Finally, statistical testing between the clinical reading and the DL+data model was performed using the McNemar's test [28–30] with Edward's continuity correction [31]. The test was conducted for each of the 100 runs of the test data, with significance level $\alpha = 0.05$. Due to multiple comparisons performed, Bonferroni correction [32] was implemented, with the corrected P value of <0.0005 denoting statistical significance.

RESULTS

Figure 3 shows the box plots from the 100 runs of the DL+data model over the test data in comparison to the clinical classification. Comparable performance to the clinical classification can be seen.

Table 2 contains the classification accuracy metrics reported as median with interquartile range in parenthesis, over 100 runs of the test data. The results from the clinical reading and image-only DL model are presented for comparison. The DL+data model and the clinical reading had small differences in accuracy and achieved similar AUC and F1 score. Sensitivity was higher with the DL+data model, whereas clinical reading achieved better specificity and precision.

Table 3 contains the amount of TP, TN, FP, and FN cases classified by the DL+data model, with median and interquartile range of the 100 runs of the test data with the results of clinical reading and the image-only DL model. The DL+data model had improved performance over TP and FN cases, with difference of two subjects with TN and FP cases to the clinical reading.

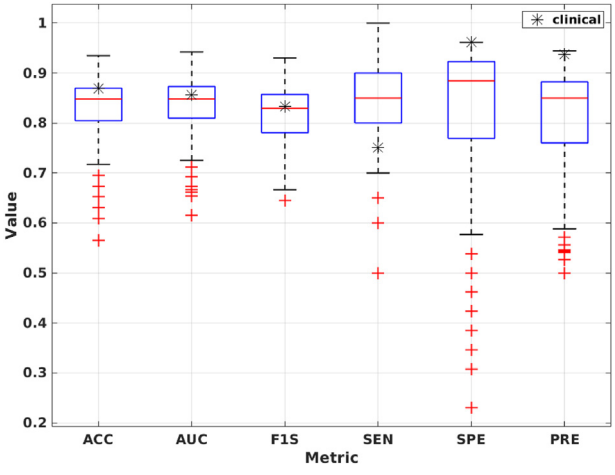

Figure 3. A box plot of prediction performance with the DL+data model with 100 runs over the test data, showing a comparison to the clinical observer to visualize the distribution of performance metrics over different runs. Abbreviation: *DL*, deep learning.

Table 2. Classification accuracy metrics of the DL+data model prediction performance. Metrics are reported as median and interquartile range (IQR) in parenthesis over 100 runs of the data in comparison to the results with the clinical reading and image-only DL model (DL) reported in Ref. [18] with the hold-out (test) dataset.

Method	ACC	AUC	F1S	SEN	SPE	PRE
DL+data	0.848 (0.065)	0.848 (0.064)	0.829 (0.077)	0.850 (0.100)	0.885 (0.154)	0.850 (0.122)
DL	0.826 (0.065)	0.806 (0.065)	0.765 (0.101)	0.650 (0.100)	0.962 (0.000)	0.929 (0.010)
Clinical	0.870	0.856	0.833	0.750	0.962	0.938

ACC, accuracy; *AUC*, area under the receiver operating characteristic curve; *F1S*, F1 score; *SEN*, sensitivity; *SPE*, specificity; *PRE*, precision; *DL*, deep learning.

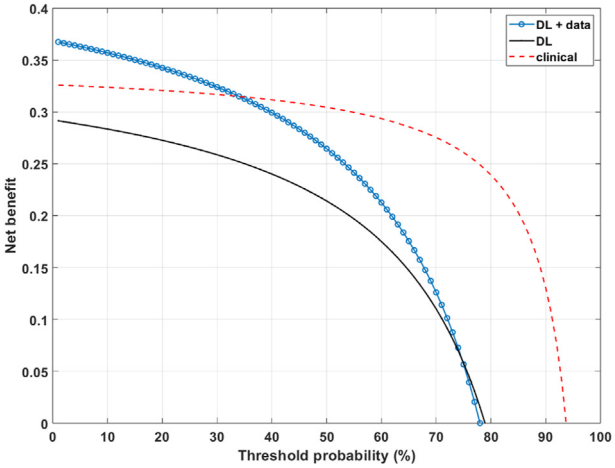

Figure 4. Net benefit of the developed DL model (DL+data) and the clinical observer (clinical) over decision thresholds of 0% to 100%. The net benefit of our previous image-only DL approach (DL) using data from Ref. [18] is shown in comparison. The net benefit is visualized as mean over the 100 runs of the test data. Only positive values in the range of 0 to 1 for the net benefit are shown. Abbreviation: *DL*, deep learning.

In Figure 4, net benefit is shown. Using both clinical and imaging data resulted in improved net benefit up to threshold probability of 34% with the DL+data model, showing improved performance over the previous image-only DL model.

We found a good agreement $\kappa = 0.692$ (0.129) between the DL+data model and the ICA data in classifying the individuals as ischemic, which was comparable to the clinical reading ($\kappa = 0.728$).

According to the McNemar's test, there was a non-significant difference (corrected *P* value > 0.0005) between the proportion of errors in the test set between the DL+data model and the clinical reading, in each of the 100 runs of the data. A single, worst-performing run had the lowest *P*-value closest to the corrected significance threshold (run number 9, *P* value $= 0.007963$). Majority of the runs (92 runs out of 100) had a *P*-value that was considerably higher even than the uncorrected significance threshold ($P > 0.05$).

In Figure 5, we show the classification results with SHAP and GRAD-CAM highlighting the ischemic regions in the polar maps for one ischemic case. Noteworthy is the complementary nature of the GRAD-CAM and SHAP highlighting the defect and individual risk variables from CTA and clinical data. The DL+data model highlighted image features which weigh the decision towards a CAD-positive case (Supplementary Figure 1).

In Figure 6, we show the SHAP values from image features for one severely ischemic case. Image features are highlighted according to the extent of the perfusion defect in the polar maps and are visualized by GRAD-CAM. Additional variables are highlighted in the CTA and clinical data.

In Figure 7, we show a case example in which the model classified as a TP, although the polar map was showing high sMBF. No defects were detected by GRAD-CAM or in the image features, but classification was made based on the CTA stenosis degree and patient-specific clinical variables, showing the complementary advantage of our approach.

Table 3. The number of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) with median and interquartile range (IQR) of the DL+data model on 100 runs of the data in addition to the results with the clinical reading and the image-only DL model (DL) reported in Ref. [18] with the hold-out (test) dataset.

Method	TP	TN	FP	FN
DL+data	17 (2)	23 (4)	3 (4)	3 (2)
DL	13 (2)	25 (0)	1 (0)	7 (2)
Clinical	15	25	1	5

TP, true positive; *TN*, true negative; *FP*, false positive; *FN*, false negative; *DL*, deep learning.

DISCUSSION

We introduced a DL-based approach using polar maps, segmental sMBF values, coronary CTA findings, and clinical modifiers of the likelihood of obstructive CAD. The approach includes explainability methods to visualize perfusion defects and highlighting important clinical features. The model achieved a good performance compared to the clinical reading. Even greater benefit is the ability to highlight possible risk variables and perfusion defects on individual

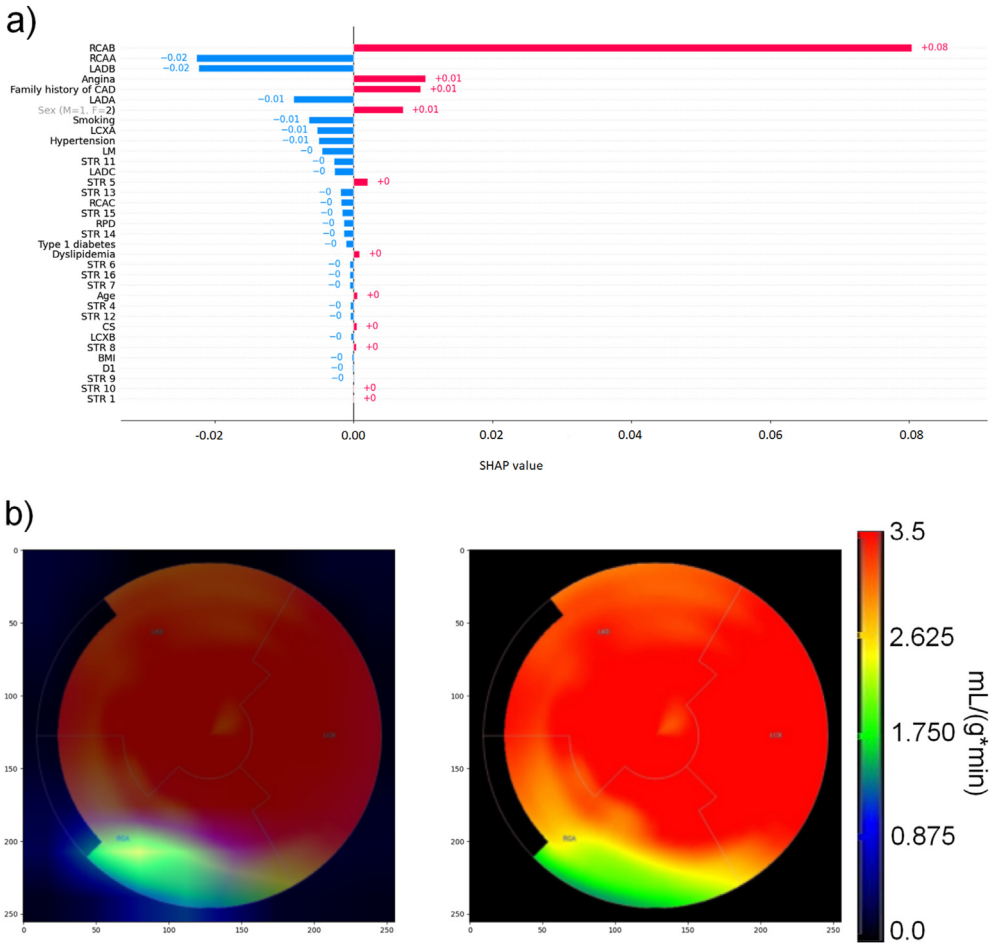

Figure 5. Detection examples with a TP ischemic case showing (A) SHAP values from CTA, PET, and clinical data represented as a bar plot and (B) GRAD-CAM detection overlaid on polar map (left polar map). Noteworthy is that the SHAP values of the coronary CTA data (RCAB) and GRAD-CAM visualization maps in polar maps match each other. The probability of CAD as predicted by the model was 0.643. SHAP acronyms are identical to Table 1, with exception of STR1 to STR16, which describe the segmental sMBF values and CS which indicates coronary calcium score. Abbreviations: *CTA*, coronary computed tomography; *PET*, positron emission tomography; *GRAD-CAM*, gradient-based class activation mapping; *SHAP*, Shapely values; *TP*, true positive.

basis, giving extended information compared to majority of current DL-based classifiers.

Our approach proposes potential solutions on how to address the following, timely issues in application of DL in PET MPI [33]. First, the approach allows to incorporate multi-parametric information to a single, explainable pipeline. Second, both imaging and other clinical information are considered together. Third, explainability methods for both images, and clinical variables are used to understand what the model is basing its decisions.

The hybrid DL+data approach showed improved performance and stability over our previous image-only approach [18] (Table 2). Increased ACC, AUC, F1S, and SEN were seen (Table 2). The number of TP results increased from 13 to 17 and that of FN results decreased from 7 to 3 (Table 3). Increased net benefit was also seen (Figure 4). Thus, the hybrid DL+data approach showed improved performance over the image-only approach, with only a slight increase in the number of FP results. Furthermore, the performance of DL+data model was superior to a data-only model (Supplementary Methods, Supplementary Figures 3 and 4).

The DL+data model compared to clinical reading showed similar ACC, AUC, and F1S, but the model showed improved SEN (Figure 3, Table 2). Clinical interpretations of ACC, AUC, and F1S fall close to the median predictions, and are within the interquartile range, indicating comparable performance (Figure 3). However, SPE and PRE of the model could be improved (Figure 3).

An improved net benefit of decision thresholds up to 34% over the clinical reading was seen with superior performance compared to the image-only model (Figure 4). The increased net benefit of the model indicates that it could be useful to a

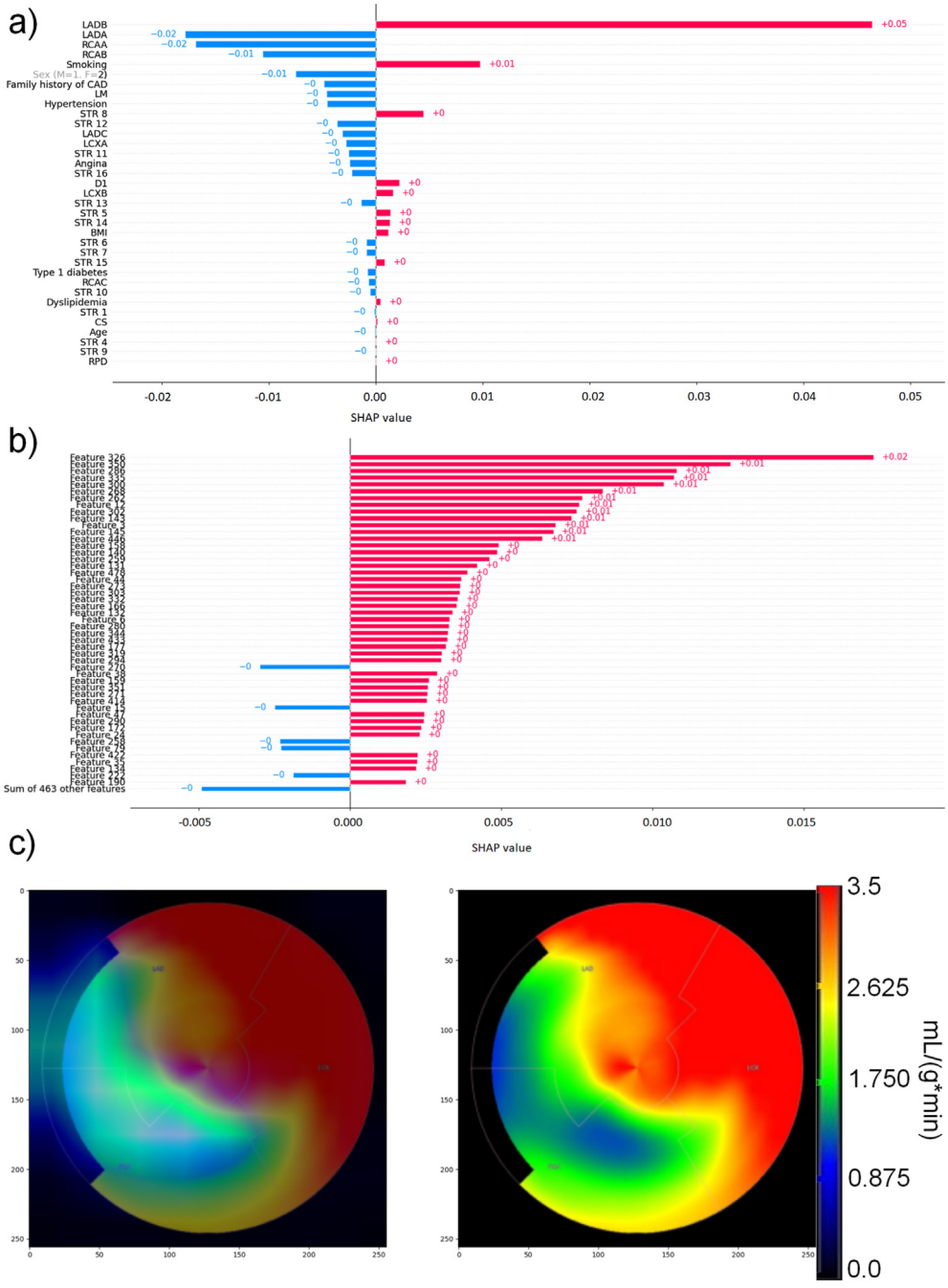

Figure 6. Example case of SHAP values from (A) clinical data, (B) image features, and (C) GRAD-CAM detection (left polar map with overlay) with a severely ischemic case. Image features in (B) are highlighted according to the degree of the perfusion defect seen in the polar map in addition to clinical variables in (A), as well as the location of the defect by GRAD-CAM in (C). The final probability for CAD-positive predicted by the DL+data model was 0.858. SHAP acronyms are identical to Table 1, with exception of STR1 to STR16, which describe the segmental sMBF values and CS which indicates coronary calcium score. Abbreviations: *CAD*, coronary artery disease; *DL*, deep learning; *GRAD-CAM*, gradient-based class activation mapping; *SHAP*, Shapely values.

subset of preferences when applied in conjunction with clinical reading. For example, additional clinical analysis could be conducted in those deemed positive by the DL+data model.

Notably, the amount of FN results with the use of DL+data model was 3 vs. 5 based on clinical reading. In individual and epidemiological standpoint, the cost of FN results is higher than that of FP results [6], indicating that the fusion model is preferable to our previous, image-based approach, which resulted in 7 FN cases. An example of the classification results is shown in Supplementary Figure 2.

Slightly lower Kohen's kappa with the DL+data model compared to clinical classification ($\kappa = 0.692$ (0.129) vs $\kappa = 0.728$) was found. There

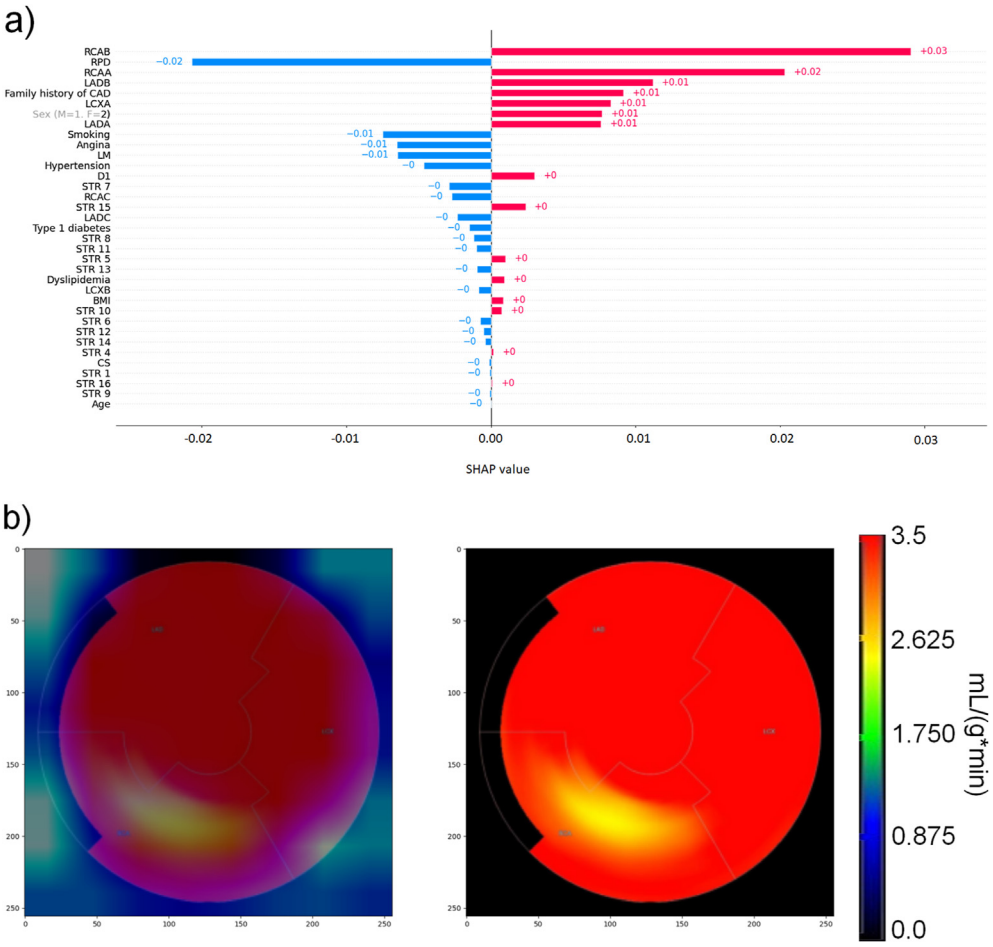

Figure 7. A case example which the DL+data model classified as TP, although no reduction of sMBF is shown in the polar map image, with (A) SHAP values from clinical data and (B) GRAD-CAM detection. GRAD-CAM (left polar map) has not detected any regions with clear defects in conjunction with high sMBF shown in the original polar map (right polar map). However, the SHAP values show the DL model considering the CTA stenosis degree in major vessels with additional, patient-specific risk factors in making the final classification decision (red bars). These included CTA stenosis degree in major arteries, or individual risk factors such as family history of CAD, angina, or diabetes. Noteworthy is that the final probability for CAD-positive given by the DL+data model of 0.574 (range 0 to 1) shows that the case can be considered borderline. SHAP acronyms are identical to Table 1, with exception of STR1 to STR16, which describe the segmental sMBF values and CS which indicates coronary calcium score. Abbreviations: *CAD*, coronary artery disease; *CTA*, coronary computed tomography; *DL*, deep learning; *PET*, positron emission tomography; *GRAD-CAM*, gradient-based class activation mapping; *SHAP*, Shapely values; *TP*, true positive.

was also a non-significant difference in proportion of errors in the test set between the model and clinical reading according to the McNemar's test. This shows that addition of the imaging and clinical data could effectively improve the classification performance to the level comparable to clinical reading, considering all metrics. The model does have a slight advantage in having additional 9 clinical variables to use, whereas the human reader is operating on CTA and PET data.

Special attention was placed to the explainability and visualization of the model decision process. This offered several advantages to highlight why certain cases are deemed CAD-positive (Figures 5–7), especially when investigating TP cases with high sMBF. Adding complementary data from coronary CTA and clinical variables allowed detection of CAD even in some cases with preserved sMBF. An additional advantage of visualization is highlighting the regions with reduced sMBF and clinical variables individually, allowing to justify the classifier decisions (Figures 5-6).

The explainability methods allowed us to note several interesting findings using feature ranking. Previously, integration of coronary CTA and clinical data has been shown to improve the detection of ischemia [34]. Our results expand this by confirming that both image-based and data-based (CTA, PET, clinical) features provide complementary information for CAD detection. Additionally, we found the features were ranked consistently in the following order: CTA, clinical, and PET variables (Supplementary Figure 5).

The clinical variables were ranked higher than PET variables, which may be caused by sMBF values being in the normal range in most of the segments. Only a small number of individuals showed significantly reduced perfusion below the sMBF $<$ 2.3 g/mL/min threshold [35]. Another explanation is that the visual features already contributed the information needed, and thus the quantitative values did not bring additional benefit. Furthermore, the coronary CTA variables indicating coronary stenosis were ranked the highest, whereas coronary calcium score was ranked the lowest. This is an interesting finding, as it might imply how the CTA stenosis variables could be co-linear with the coronary calcium score, which mainly translates to the presence of plaque burden. Whether this finding has further clinical significance should be investigated in more detail.

In addition to the clinical variables, the model receives low-level textural representations of the image. Currently, the image feature importance visualization is qualitative, showing significant trend with extreme cases (Figure 7). It is challenging to correlate these features in spatial terms, compared to the GRAD-CAM visualization maps. This is better understood based on what the image features represent: the most salient responses over the whole image while discarding the spatial distributions of the individual pixels [36]. Thus, the information contained in these features is a condensed textural representation of the visual patterns in the image.

Only a few DL approaches have focused on CAD detection in MPI, with majority existing for SPECT [10,37] and not for PET. One study [37] presented a fusion approach with SPECT polar maps and clinical data. Our approach extends this by allowing the model to learn from both polar maps and clinical data before final classification. Similarly to Ref. [10], the GRAD-CAM maps matched the perfusion defects (Figure 6). Furthermore, another important aspect in our work was combining coronary anatomy information from CTA, which was not used in the aforementioned studies.

Our study is currently limited to a feasibility study, as the number of participants (N = 138) is small. However, we noted a good stability with low variation in the k-fold cross-validation and repeated hold-out testing. Considering a relatively large proportion of patients with positive results in our cohort, a larger cohort would allow to include more individuals with no or only minimal CTA findings and low number of clinical risk variables. In our study, individuals typically showed elevated probability for ischemia by the model (30%-40%). However, the results with the k-fold cross-validation and the hold-out dataset show the feasibility of our approach. Naturally, adding additional data would help to optimize the model performance better. Finally, this could increase the statistical power of the McNemar's test as we did not detect significant differences after Bonferroni correction.

Thus, our future intent is to extend our evaluation to a large cohort of patients who have undergone hybrid O-15 H_2O PET/CT MPI, although constructing such a database is challenging. To develop the methodology to operate in a fully automated form, the variables could be extracted and processed automatically and passed to the model at the time of clinical reading.

Finally, the practice of modern medicine relies heavily on synthesis of data from multiple sources [38]. Thus, DL models operating on data synthesis from multiple sources, with increased transparency in model decision-making are needed.

NEW KNOWLEDGE GAINED

- A fusion DL model was developed for hybrid stress O-15 H_2O MPI-PET and CTA imaging. The model incorporates both polar map images and diagnostic data and shows good performance in detection of flow-limiting CAD.
- With increased input data, model explainability is needed to justify the model decision process in visualizing the defects and highlighting the clinical variables. These might reveal important diagnostic variables to be investigated further.
- Knowledge of the contribution of relevant variables on an individual basis offers a significant advantage, allowing to understand where the model bases its decisions, increasing the adaptability of DL approaches to clinical routine.

CLINICAL PERSPECTIVE

- A deep-learning approach allows to incorporate multiparametric information, such as multi-modality imaging data from coronary computed tomography, stress myocardial perfusion imaging positron emission tomography polar maps, and clinical risk variables for detecting flow-limiting coronary artery disease in an explainable pipeline.
- When combining imaging and diagnostic data to a deep-learning model, explainability methods are needed to understand the model decisions and might potentially reveal important diagnostic variables to be investigated.

- Knowledge of the contribution of relevant clinical and imaging variables with highlighting the ischemic regions in the polar maps on an individual basis offers a significant advantage over traditional deep-learning models, improving their adaptability to clinical routine.

CONCLUSIONS

A combined DL-based classifier using polar map images, coronary CTA, sMBF and diagnostic data in hybrid coronary CTA and O-15 H_2O PET/CT perfusion imaging is a feasible method for the detection of flow-limiting CAD, showing diagnostic performance up to the level of clinical reading.

This work expands from our prior efforts, showing the improved performance of the hybrid DL approach. One of the major strengths is the ability to highlight the culprit regions with the contribution of relevant clinical variables in an interpretable manner and on an individual basis.

FUNDING AND SUPPORT

Dr Jarmo Teuho is an International Research Fellow of the Japan Society for the Promotion of Science, supported by JSPS Grant Number P19748 (Postdoctoral Fellowships for Research in Japan (Standard)). In addition, Dr Teuho would also like to acknowledge the Academy of Finland mobility funding (Academy Decision Number 322019) for supporting this research by allowing to combine both research work and family life in Japan. This research was also funded by the Maire and Aimo Mäkinen Fund of the Finnish Cultural Foundation (Dr Riku Klén), Academy of Finland (Dr Juhani Knuuti, Academy Decision Number 351482) and by the JSPS KAKENHI grant 21KK0183 and 21K12111 (Dr Naoaki Ono), and Turku University Foundation.

DISCLOSURES

Jarmo Teuho, Jussi Schultz, Riku Klén, Luis Eduardo Juarez-Orozco, Juhani Knuuti, Naoaki Ono, and Shigehiko Kanaya have nothing to disclose. Antti Saraste discloses grants from the Academy of Finland, Finnish Foundation for cardiovascular research and fees for lectures or consultation from Abbott, Astra Zeneca, Bayer, Novartis, and Pfizer outside the submitted work.

DATA AVAILABILITY STATEMENT

The data presented in this study are available on a reasonable request from the corresponding author. The Python source codes, the trained weights, and parameters for the model generated in the current study will be made publicly available at the University of Turku GitLab portal of the corresponding author (https://gitlab.utu.fi/jarmo.teuho).

ACKNOWLEDGMENTS

Dr Jarmo Teuho is an International Research Fellow of the Japan Society for the Promotion of Science, supported by JSPS Grant Number P19748 (Postdoctoral Fellowships for Research in Japan (Standard)). In addition, Dr Teuho acknowledges the Academy of Finland mobility funding (Academy Decision Number 322019) for supporting this research by allowing to combine both research work and family life in Japan.

APPENDIX A. SUPPLEMENTARY DATA

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nuclcard.2024.101889.

REFERENCES

[1] Juarez-Orozco LE, Klén R, Niemi M, Ruijsink B, Daquarti G, Van Es R, et al. Artificial intelligence to improve risk prediction with nuclear cardiac studies. Curr Cardiol Rep 2022;24:1–10. https://doi.org/10.1007/S11886-022-01649-W.

[2] Slomka P. Future of nuclear cardiology is bright: promise of cardiac PET/CT and artificial intelligence. [Editorial] J Nucl Cardiol. 2022. https://doi.org/10.1007/s12350-022-02942-5.

[3] Adlung L, Cohen Y, Mor U, Elinav E. Machine learning in clinical decision making. Med 2021;2:642–65. https://doi.org/10.1016/J.MEDJ.2021.04.006.

[4] Juarez-Orozco LE, Martinez-Manzanera O, Storti AE, Knuuti J. Machine learning in the evaluation of myocardial ischemia through nuclear cardiology. Curr Cardiovasc Imaging Rep 2019;12. https://doi.org/10.1007/s12410-019-9480-x.

[5] Sciagrà R, Lubberink M, Hyafil F, Saraste A, Slart RHJA, Agostini D, et al. EANM procedural guidelines for PET/CT quantitative myocardial perfusion imaging. Eur J Nucl Med Mol Imaging 2020;48:1–30. https://doi.org/10.1007/s00259-020-05046-9.

[6] Moore A, Bell M. XGBoost, A novel explainable AI technique, in the prediction of myocardial infarction: a UK biobank cohort study. Clin Med Insights Cardiol 2022;16:1–6. https://doi.org/10.1177/11795468221133611.

[7] Juarez-Orozco LE, Knol RJJ, Sanchez-Catasus CA, Martinez-Manzanera O, van der Zant FM, Knuuti J. Machine learning in the integration of simple variables for identifying patients with myocardial ischemia. J Nucl Cardiol 2020;27:147–55. https://doi.org/10.1007/s12350-018-1304-x.

[8] Kajander S, Joutsiniemi E, Saraste M, Pietilä M, Ukkonen H, Saraste A, et al. Cardiac positron emission tomography/computed tomography imaging accurately detects anatomically and functionally significant coronary artery disease. Circulation 2010;122:603–13. https://doi.org/10.1161/CIRCULATIONAHA.109.915009.

[9] Menke J, Kowalski J. Diagnostic accuracy and utility of coronary CT angiography with consideration of unevaluable results: a systematic review and multivariate

bayesian random-effects meta-analysis with intention to diagnose. Eur Radiol 2016;26:451–8.

[10] Otaki Y, Singh A, Kavanagh P, Miller RJH, Parekh T, Tamarappoo BK, et al. Clinical deployment of explainable artificial intelligence of SPECT for diagnosis of coronary artery disease. JACC Cardiovasc Imaging 2021. https://doi.org/10.1016/J.JCMG.2021.04.030.

[11] Papandrianos N, Papageorgiou E. Automatic diagnosis of coronary artery disease in SPECT myocardial perfusion imaging employing deep learning. Appl Sci 2022;11:1–14. https://doi.org/10.3390/app11146362.

[12] Nakajima K, Maruyama K. Nuclear cardiology data analyzed using machine learning. Ann Nucl Cardiol. 2022;8:80–5. https://doi.org/10.17996/anc.22-00164.

[13] Fan F-L, Xiong J, Li M, Wang G. On interpretability of artificial neural networks: a survey. IEEE Trans Radiat Plasma Med Sci 2021;5:741–60. https://doi.org/10.1109/trpms.2021.3066428.

[14] Yeung MW, Benjamins JW, Knol RJJ, van der Zant FM, Asselbergs FW, van der Harst P, et al. Multi-task deep learning of myocardial blood flow and cardiovascular risk traits from PET myocardial perfusion imaging. J Nucl Cardiol 2022;29:3300–10. https://doi.org/10.1007/s12350-022-02920-x.

[15] Pieszko K, Shanbhag AD, Singh A, Hauser MT, Miller RJH, Liang JX, et al. Time and event-specific deep learning for personalized risk assessment after cardiac perfusion imaging. npj Digit Med 2023;6:1–11. https://doi.org/10.1038/s41746-023-00806-x.

[16] Juarez-Orozco LE, Martinez-Manzanera O, van der Zant FM, Knol RJJ, Knuuti J. Deep learning in quantitative PET myocardial perfusion imaging: a study on cardiovascular event prediction. JACC Cardiovasc Imaging 2020;13:180–2. https://doi.org/10.1016/j.jcmg.2019.08.009.

[17] Teuho J, Schultz J, Klen R, Saraste A, Ono N, Kanaya S. Comparison of 12 machine learning methods for polar map classification in cardiac perfusion PET. In: 2021 IEEE Nucl Sci Symp Med Imaging Conf Rec. NSS/MIC 2021 28th Int Symp Room-Temperature Semicond Detect RTSD 2022 2021; 2021. p. 2021–3. https://doi.org/10.1109/NSS/MIC44867.2021.9875597.

[18] Teuho J, Schultz J, Klen R, Knuuti J, Saraste A, Ono N, et al. Classification of ischemia from myocardial polar maps in 15O–H2O cardiac perfusion imaging using a convolutional neural network. Sci Rep 2022;12:1–12. https://doi.org/10.1038/s41598-022-06604-x.

[19] Danad I, Uusitalo V, Kero T, Saraste A, Raijmakers PG, Lammertsma AA, et al. Quantitative assessment of myocardial perfusion in the detection of significant coronary artery disease: cutoff values and diagnostic accuracy of quantitative [15O]H2O PET imaging. J Am Coll Cardiol 2014;64:1464–75. https://doi.org/10.1016/J.JACC.2014.05.069.

[20] Stenström I, Maaniitty T, Uusitalo V, Pietilä M, Ukkonen H, Kajander S, et al. Frequency and angiographic characteristics of coronary microvascular dysfunction in stable angina: a hybrid imaging study. Eur Heart J Cardiovasc Imaging 2017;18:1206–13. https://doi.org/10.1093/ehjci/jex193.

[21] Nesterov SV, Han C, Mäki M, Kajander S, Naum AG, Helenius H, et al. Myocardial perfusion quantitation with 15O-labelled water PET: high reproducibility of the new cardiac analysis software (Carimas™). Eur J Nucl Med Mol Imaging 2009;36:1594–602. https://doi.org/10.1007/s00259-009-1143-8.

[22] Cury RC, Abbara S, Achenbach S, Agatston A, Berman DS, Budoff MJ, et al. CAD-RADS™ coronary artery disease - reporting and data system. An expert consensus document of the Society of Cardiovascular Computed Tomography (SCCT), the American College of Radiology (ACR), and the North American Society for Cardiovascular Imaging (NASCI). Endorsed by the American college of cardiology. J Cardiovasc Comput Tomogr 2016;10:269–81. https://doi.org/10.1016/J.JCCT.2016.04.005.

[23] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020;128:336–59. https://doi.org/10.1007/S11263-019-01228-7.

[24] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems; 2017. p. 4766–75.

[25] Tamarappoo BK, Lin A, Commandeur F, McElhinney PA, Cadet S, Goeller M, et al. Machine learning integration of circulating and imaging biomarkers for explainable patient-specific prediction of cardiac events: a prospective study. Atherosclerosis 2021;318:76–82. https://doi.org/10.1016/J.ATHEROSCLEROSIS.2020.11.008.

[26] Tohka J, van Gils M. Evaluation of machine learning algorithms for health and wellness applications: a tutorial. Comput Biol Med 2021;132:104324. https://doi.org/10.1016/J.COMPBIOMED.2021.104324.

[27] Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352. https://doi.org/10.1136/bmj.i6.

[28] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153–7. https://doi.org/10.1007/BF02295996.

[29] Salzberg SL. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min Knowl Discov 1997;1:317–28. https://doi.org/10.1023/A:1009752403260.

[30] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10:1895–923. https://doi.org/10.1162/089976698300017197.

[31] Edwards AL. Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. Psychometrika 1948;13:185–7. https://doi.org/10.1007/BF02289261.

[32] Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Seeber: Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze; 1936.

[33] Williams MC. Machine learning models for positron emission tomography myocardial perfusion imaging. J Nucl Cardiol 2024;32:101805. https://doi.org/10.1016/j.nuclcard.2024.101805.

[34] Benjamins JW, Yeung MW, Maaniitty T, Saraste A, Klén R, van der Harst P, et al. Improving patient identification for advanced cardiac imaging through machine learning-integration of clinical and coronary CT angiography data. Int J Cardiol 2021;335:130–6. https://doi.org/10.1016/J.IJCARD.2021.04.009.

[35] Danad I, Raijmakers PG, Driessen RS, Leipsic J, Raju R, Naoum C, et al. Comparison of coronary CT angiography, SPECT, PET, and hybrid imaging for diagnosis of ischemic heart disease determined by fractional flow reserve. JAMA Cardiol 2017;2:1100–7. https://doi.org/10.1001/JAMACARDIO.2017.2471.

[36] Feng J, Ni B, Tian Q, Yan S. Geometric ℓp-Norm feature pooling for image classification. In: Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit; 2011. p. 2697–704. https://doi.org/10.1109/CVPR.2011.5995370.

[37] Apostolopoulos ID, Apostolopoulos DI, Spyridonidis TI, Papathanasiou ND, Panayiotakis GS. Multi-input deep learning approach for cardiovascular disease diagnosis using myocardial perfusion imaging and clinical data. Phys Medica 2021;84:168–77. https://doi.org/10.1016/j.ejmp.2021.04.011.

[38] Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. npj Digit Med 2020;3. https://doi.org/10.1038/s41746-020-00341-z.