Physica Medica 84 (2021) 168–177

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

Multi-input deep learning approach for Cardiovascular Disease diagnosis using Myocardial Perfusion Imaging and clinical data

Ioannis D. Apostolopoulos [a,*], Dimitris I. Apostolopoulos [b], Trifon I. Spyridonidis [b], Nikolaos D. Papathanasiou [b], George S. Panayiotakis [a]

[a] *Department of Medical Physics, School of Medicine, University of Patras, GR 265-00 Patras, Greece*
[b] *University Hospital of Patras, Department of Nuclear Medicine, GR 265-00 Patras, Greece*

ARTICLE INFO

Keywords:
Deep Learning
Machine Learning
Coronary Artery Disease
Medical imaging

ABSTRACT

Purpose: Accurate detection and treatment of Coronary Artery Disease is mainly based on invasive Coronary Angiography, which could be avoided provided that a robust, non-invasive detection methodology emerged. Despite the progress of computational systems, this remains a challenging issue. The present research investigates Machine Learning and Deep Learning methods in competing with the medical experts' diagnostic yield. Although the highly accurate detection of Coronary Artery Disease, even from the experts, is presently implausible, developing Artificial Intelligence models to compete with the human eye and expertise is the first step towards a state-of-the-art Computer-Aided Diagnostic system.
Methods: A set of 566 patient samples is analysed. The dataset contains Polar Maps derived from scintigraphic Myocardial Perfusion Imaging studies, clinical data, and Coronary Angiography results. The latter is considered as reference standard. For the classification of the medical images, the InceptionV3 Convolutional Neural Network is employed, while, for the categorical and continuous features, Neural Networks and Random Forest classifier are proposed.
Results: The research suggests that an optimal strategy competing with the medical expert's accuracy involves a hybrid multi-input network composed of InceptionV3 and a Random Forest. This method matches the expert's accuracy, which is 79.15% in the particular dataset.
Conclusion: Image classification using deep learning methods can cooperate with clinical data classification methods to enhance the robustness of the predicting model, aiming to compete with the medical expert's ability to identify Coronary Artery Disease subjects, from a large scale patient dataset.

Introduction

Cardiovascular Diseases (CVD) remain fatal diseases across the world [1], whilst the early, automatic, non-invasive, reliable detection and prognosis is still an open issue.

Myocardial Perfusion Imaging (MPI) with Single-Photon Emission Computed Tomography (SPECT) is an established modality for the identification of significant Coronary Artery Disease (CAD), the risk assessment of major adverse cardiac event risks, and the evaluation of myocardial viability [2–4]. The presence of artefacts from tissue attenuation of radioactivity may cause distortions in the quality of the acquired MPI images and intricate medical diagnosis. To overcome this, SPECT prone imaging, external radioactive sources providing transmission maps for attenuation correction (AC), and various software techniques have been proposed [5]. The advent of hybrid SPECT/CT technology enabled the use of Computed Tomography (CT) images as transmission maps for the AC of SPECT data. Despite software and technological advances, MPI still performs sub-optimally [2]. Medical experts still evaluate MPI findings together with the available clinical information [6]. This leads to a subjective interpretation of MPI, since clinical information [7] entered is complex, including patients' demographic characteristics, type of symptoms, CAD predisposing factors, concomitant diseases, etc. All these factors impose different weights of importance on final diagnostic interpretation.

Coronary Angiography remains the gold standard modality for confirmation or exclusion of CAD. However, it is invasive and may pose unnecessary risks in a substantial portion of patients.

Machine Learning (ML) is a subset of Artificial Intelligence, with

[*] Corresponding author.
E-mail address: ece7216@upnet.gr (I.D. Apostolopoulos).

https://doi.org/10.1016/j.ejmp.2021.04.011
Received 9 November 2020; Received in revised form 2 April 2021; Accepted 12 April 2021

integrated methods, algorithms, and statistical models intending to make predictions based on existing data. The main objective is to discover patterns from the available data called "training" data [8]. The pattern discovery can be performed by relying on specific factors, which have either been provided by the model's designer, or mined automatically by a computer model [9]. ML may use the extracted knowledge to make predictions, classify a new instance, or suggest an underlying connection between seemingly unrelated data.

For over a decade, the results were promising but failed to constitute a reliable method due to data insufficiency. Large scale datasets are usually desirable to employ ML in full strength [10].

Recently, the emergence of large-scale imaging datasets catalysed the evolution of a field called Deep Learning (DL). DL alludes to a variety of ML methods usually focused on images. Also called Convolutional Neural Networks (CNN) [11], DL automatically extracts millions of features from images. Moreover, by the aid of ML algorithms, those features can be classified according to their significance related to a specific task.

There has been much research regarding the non-invasive diagnosis of CAD. In the past decade, several ML models [12–16] were proposed to analyse and classify CAD datasets, such as the Z-Alizadeh Sani dataset [12] and the Cleveland Heart Disease Dataset [13]. These methods relied solely on categorical and continuous variables to ensure the maximal achieved accuracy. Despite their excellent results, these studies did not utilise image data to enhance the decision-making process. DL has recently been employed to aid in the classification or segmentation task, yielding promising results in datasets of limited size [17–22]. Whereas DL is widely investigated in medical imaging tasks, minimal research has been conducted using MPI SPECT images.

Most of the related research is limited by four vital issues. The first relates to the amount of data used. The effective application of ML and DL techniques usually requires large-scale datasets. Only the Cleveland and the Z-Alizadeh Sani datasets contain adequate amount of data for applying ML methods. The second issue is the quality of data incorporated into the publically available datasets. For example, the Z-Alizadeh Sani dataset contains a few healthy instances compared to the diseased ones. The third issue is the handling of typical angina, which is the most deciding factor for diagnosis. In everyday clinical practice, patients with typical angina are considered to suffer from CAD. Both the Cleveland and the Z-Alizadeh Sani datasets contain a significant numberφangina instances, making the decision algorithms' task easier. The final issue is that the used datasets are incomplete. Several diagnostic tests are absent, such as MPI SPECT, which is a reliable diagnostic test, with a reported accuracy between 65% and 75% [23,24].

A relatively large-scale patient dataset to investigate DL methods' effectiveness in analysing both image data and clinical outcomes for CAD prediction is used in the present retrospective study. The study is motivated by the recent results of our group [25], which revealed the ability of CNNs to identify significant CAD by use of MPI polar maps. The well-known CNN architecture, called Inception [26], is applied to classify the MPI images into normal and abnormal. The Random Forest and the Neural Network classifiers are selected for the classification of the clinical-condition data. The experiments' outcome highlights the hybrid method's efficiency, integrating both image and clinical data into a robust multi-input model.

Methods

Dataset of the study and data preprocessing

Description

The study used patient data recorded in the laboratory of the author's Institution from 16/2/2018 to 28/02/2020. Over this period, 2036 consecutive patients underwent gated-SPECT MPI with ^{99m}Tc-tetrofosmin. Two-hybrid SPECT/CT gamma-camera systems (Varicam, Hawkeye and Infinia, Hawkey-4, GE Healthcare) were used for MPI imaging. Computed tomography-based AC in both stress and rest images was applied in all cases. Five hundred and eighty-six patients (28.8%) were subsequently subjected to invasive coronary angiography (ICA) within 60 days from MPI for further evaluation. Twenty patients were excluded from the dataset due to inconclusive MPI results or missing ICA reports. The final dataset of the present study includes 566 samples. The Ethical Committee approved the study of our Institution. The nature of the survey waives the requirement to obtain patients' informed consent. The clinical characteristics of the dataset are presented in Table 1. The reader should note that no missing values are reported for this particular study. The presented clinical characteristics are collected by the medical staff based on specific guidelines [27].

MPI and ICA interpretation and reporting

MPI interpretation and reporting was carried out prospectively and independently by three experienced Nuclear Medicine physicians (DJA, NP and TS of the authors). The interpreters inspected the complete set of tomographic slices and the polar maps obtained from both AC and non-attenuation corrected (NAC) studies. Their final report has also considered all pertinent clinical information, such as detailed patient's history, symptoms, CAD predisposing factors, previous tests results, baseline ECG and ECG changes during stress, etc. MPI reports were considered positive for CAD if they described a reversible tracer defect of any size or a fixed defect not normalised by AC implementation, accompanied by normal/near normal wall motion or thickening.

Finally, the actual ground truth was obtained by invasive Coronary Angiography. Invasive Coronary Angiography was considered positive for flow-limiting CAD if lesions causing >50% lumen stenosis of the left main artery or >70% stenosis of the major coronary artery branches were identified. Fractional flow reserve (FFR) measurements were undertaken in case of intermediate lesions (causing 50%-70% stenosis). In such instances, an FFR value <0.8 was considered positive for flow-limiting CAD. The chosen thresholds are commonly used for similar purposes by the medical community [28,29].

In summary, there were two possible classes of each instance. The first class corresponded to 'healthy', and the second class corresponded to 'CAD'.

Table 1
Patient characteristics.

Clinical characteristics	Frequency
No	566
Age (mean ± sd)	66.07 ± 10.55 years
Sex (male/female)	79.55%/11.45%
History of CAD	33.03%
Previous myocardial infarction	17.84%
Previous revascularisation (PCI/CABG)*	21.37%
CAD predisposing factors	
Previous stroke	0.01%
Hypertension	76.85%
Dyslipidemia	64.48%
Smoking	40.28%
Diabetes	24.91%
Peripheral arteriopathy	0.01%
End-stage renal failure	0.05%
Family history of premature CAD	15.54%
Abnormal baseline ECG**	31.80%
Symptoms	
Asymptomatic	43.63%
Atypical chest pain	16.25%
Angina-like symptoms	13.25%
Incident of chest pain (with the subsequent biochemical exclusion of acute coronary syndrome)	12.36%
Dyspnea on exertion	15.90%

*PCI: percutaneous coronary intervention; CABG: coronary artery by-pass grafting; **ECG: electrocardiogram.

Image data

Images obtained by the two acquisition devices were presented in Dicom format. They were processed by the same commercial software to provide an aggregation of the dicom slices into one RGB Polar map image of tiff type (for each stress, rest, AC and NAC data set) pixel resolution 1400x1050 pixels and a bit depth of 24. AC and NAC polar maps in stress and rest were extracted from the SPECT studies in TIFF format. AC and NAC images of stress and rest were rescaled to 150x150 to reduce the computational cost. The four polar maps corresponding to each case were concatenated into one image. In Fig. 1, an overview of the image dataset creation is illustrated. Each final image contains 4 Polar maps, symmetrically located in a 390x390 black - backgrounded image, with 96 dpi and 24-bit depth. The compression format was tiff.

Clinical data

Auxilliary information was provided for each patient case. This data was characterised by 23 attributes without any missing values. Both categorical and numeric values were included. These attributes were summarised in Table 2.

The attribute values were transformed in order to be processed by the algorithms. For the Neural Network, categorical values were translated to 0 and 1, except for the actual label (“CAD”/“HEALTHY”), which was one-hot encoded [30]. The age was divided by 100 to fit the interval [0,1], and the BMI was divided by 50. Finally, the MPI ischemia score was divided by 3. In this way, all numeric attributes were transformed to continuous values belonging to [0,1]. This normalized the dataset and produced smooth training. For example, the values after transforming a 65-year-old patient with BMI of 30, and an MPI ischemia score of 3, would be 0.65, 0.6, and 1. For the categorical inputs of the Neural Network, each attribute’s potential value was a new attribute. For example, Gender = male and Gender = female were two different attributes for the Neural Network. If an instance was male, then the attribute “gender-male” was 1, whereas the attribute “gender-female” was 0. In this way, the NN completely ignored attributes with zero values (since there were no missing data, the NN did not confuse it with attribute values when zero).

For the Random Forest, categorical values and numeric values were retained, as shown in Table 2. In both algorithms, the Scaled MPI interpretation by medical experts was translated from a four-point scale to categorical (“CAD”/“Healthy”): 0 and 1 were addressed as “Healthy” and 2–3 as “CAD”.

Table 2
Clinical Data Attributes.

Attribute Name	Value Range
Gender	Male/Female
Age	Numeric
BMI	Numeric
History of known CAD	Yes/No
Previous myocardial infarction	Yes/No
Previous revascularisation PCI	Yes/No
Previous revascularisation CABG	Yes/No
Previous stroke	Yes/No
Diabetes	Yes/No
Smoking	Yes/No
Hypertension	Yes/No
Dislipidemia	Yes/No
Peripheral arteriopathy	Yes/No
End-stage renal failure	Yes/No
Family History of premature CAD	Yes/No
Previous ETT	Normal/Abnormal
Asymptomatic	Yes/No
Atypical chest pain	Yes/No
Angina-like	Yes/No
Dyspnea on Exertion	Yes/No
Incident of chest pain	Yes/No
Baseline ECG	Normal/Abnormal
Scaled MPI interpretation by medical experts	CAD/Healthy

Abbreviations: BMI: Body Mass Index; ETT: exercise treadmill test; PCI, CABG, ECG: as explained in the previous table.

Deep learning and machine learning models

InceptionV3

InceptionV3 is a 42-layer deep learning network, derived from the first version of Inception. Szegedy et al. [26] developed this updated deep learning network and was first named GoogLeNet. Compared to the first version, the third version had minor and major modifications, which are not analytically mentioned in the present study. The third version was the most broadly used due to its efficiency and advances over the older versions.

InceptionV3 was employed by transferring its architecture [31] and

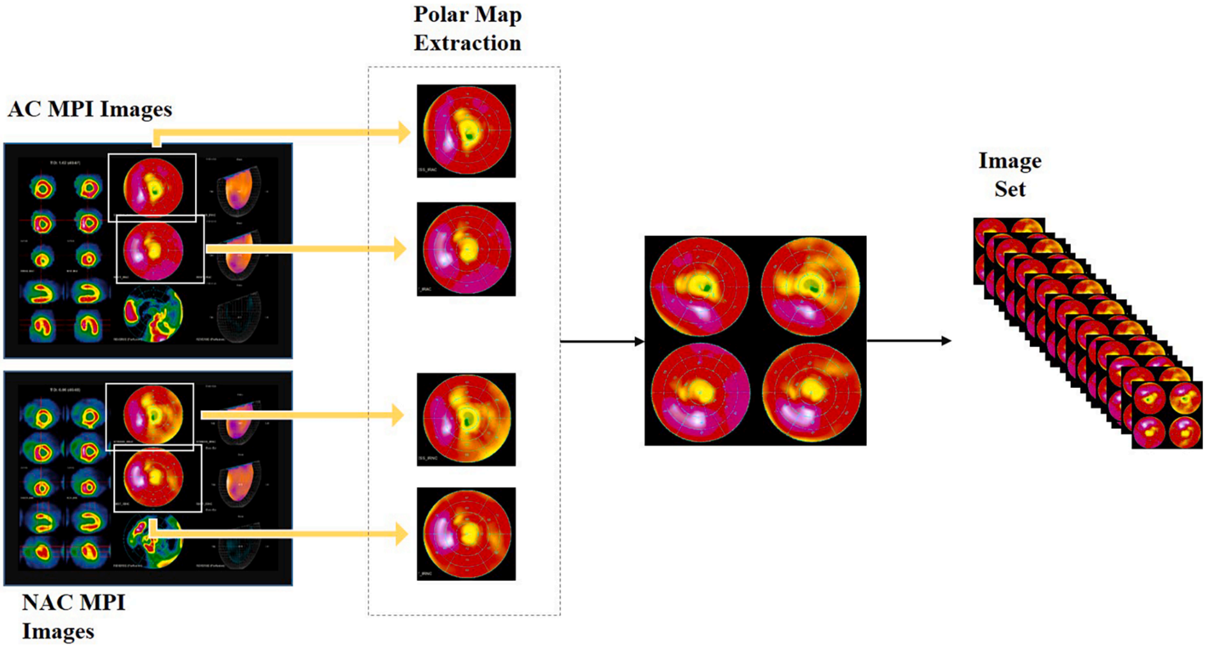

Fig. 1. Creation of the Polar Map dataset. AC and NAC polar maps are cropped and concatenated into one 300 × 300-pixel image in TIFF format.

some of the learned weights defined by simulations on the initial image dataset used to train the network. The dataset used for the initial training has been the 10-million ImageNet Large Scale Visual Recognition Competition (ILSVRC) image set.

For the experiments of this study, the model was fine-tuned. During this process, extensive experiments were conducted to define the optimal parameters for training. The best-performing combinations of parameters are illustrated in Table 3.

Neural Network and Random Forest

This study's Neural Network consisted of three hidden layers of 200, 100, and 30 nodes activated by the Rectified Linear Unit. For training optimisation, the Stochastic Gradient Descent was used. The calculation of the losses was according to the categorical cross-entropy function.

For the Random Forest, the framework was built with 300 estimators without applying depth restrictions or sample size restrictions to perform a split. Bootstrap samples were used when creating the inner tree-estimators.

Experiment setup

The experiments were performed with a single GPU setup (NVIDIA GeForce RTX 2060 Super), using Python programming language, with Tensorflow providing the necessary tools for deep learning. The rest of the computational aspects included 16 GB RAM and Intel® Core™ i5-9400F Processor.

Six classification schemes were performed to validate machine learning and deep learning methods using the mentioned networks. These schemes are illustrated in Fig. 2. The six experiments were categorised into two groups, namely single-input and multi-input. For the single-input strategy, the training was performed using either image-only data (processed by InceptionV3) or clinical-only data (processed by either Random Forest or Neural Network).

For the multi-input strategy, InceptionV3 was combined with the Random Forest or the Neural Network classifiers. This process was performed sequentially. Firstly, InceptionV3 was trained and evaluated using the image data. The model then predicted each image's class (following a 10-fold cross-validation training scheme, Fig. 3). These predictions were concatenated with the rest of the clinical data and supplied to the classifiers for the final predictions. Hence, the classifier received the output of the final layer of the CNN and the clinical data separately. In this case, the CNN acted as a helping hand to the classifier and provided its predictions based solely on the image data and not clinical conditions. Then the classifier used clinical data and CNN's output prediction to classify each instance as either healthy or diseased.

By applying such sequential multi-input strategy, the dense network of InceptionV3 did not utilise any clinical data to make decisions, which may have hampered its efficiency.

Therefore, an InceptionV3 with a multi-input densely connected classifier has also been developed (named InceptionV3_multi). In the latter network, the InceptionV3 architecture was used to extract features and classify the images using clinical data. This was achieved by densely concatenating the extracted image features and the clinical data, as shown in Fig. 2. This process is different from sequential-type. This operation involved two Neural Networks. The first to be responsible for the classification of the clinical data without using any image-related data. InceptionV3 architecture was used to extract image features. The image features and the first Neural Network outputs were supplied to a second Neural Network (the final classifier), responsible for making the final prediction/classification. In this way, one network was focused on image data and the other network on image data. A third network combined the predictions and decided the final class.

In Fig. 3, the 10-fold cross-validation training scheme is illustrated.

The reported metrics, which include True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), Accuracy, Sensitivity, Specificity, F1-Score, and AUC Score, were calculated for each test set (hidden set) during the 10-fold cross-validation. The complete dataset of 567 samples was split for training and testing (90% training-10% testing) for Fold 1. After the training, the experimental ML and DL models predicted the classes of the test set and evaluation metrics for Fold 1 were recorded. The complete dataset is split again (90% training-10% testing) but using an alternative subset for testing (Fold 2). After the training, the accuracy (and other metrics) of Fold 2 were recorded. This operation was repeated until each image is selected exactly once for testing (total 10 times). All the reported metrics were average values over the 10 folds, excluding the TP,TN,FP, and FN, calculated by adding the TP,TN,FP, and FN obtained from each iteration.

An observation was made that each Polar Map's precise position in the 4-in-1 image varies among the 566 images. This ±5-pixel variety could cause the CNN model to base its prediction on the Polar Map position rather than the colour distribution. For this reason, slight data augmentation was applied to each image.

These augmentations included small shifts in any direction. For particular experiments, heavy data augmentations were unnecessary due to the standard image acquisition protocol used to acquire and reconstruct the MPI images. Hence, there is no texture, colour, or brightness variety that could disorient the AI model's learning ability or cause either over-fitting or under-fitting. The Polar Map images included specific Regions of Interest at relatively fixed positions (e.g. upper left). More robust geometric transformations were, therefore, not preferable. The augmentations were applied to every image of the training set, while the original image was not used to train the model.

Table 3
Parameters of InceptionV3 (fine-tuned).

Parameter	Value
Trainable Layers starting from the last layer of the network	18 layers
Global Pooling	Average
Batch normalisation	After each convolution and before the activation (InceptionV3 default). Also, before the activations of the Dense Network at the top
Optimiser	Stochastic Gradient Descent (SGD)
Learning rate	0.01
Number of Densely Connected Layers (excluding the output layer)	1
Number of Nodes for each Dense Layer	2500
Activation Function	Rectified Linear Unit [32]
Activation Function of the output layer	Softmax (when used as a single input classifier)
Loss function	Categorical cross-entropy
Batch size	32
Epochs	30

Results

Medical experts' diagnostic yield

According to ICA results, 318 out of 566 patients (56.2%) had flow-limiting CAD. Using ICA findings as a reference standard, sensitivity, specificity, and accuracy, medical experts' positive and negative predictive value in diagnosing significant CAD from MPI studies was 89.17%, 71.20%, 79.15%, 70.10% and 90.20%, respectively.

Single input (either image or clinical data) model performance

The results of the experiments are presented in this section. In Table 4, the classification metrics for each case are presented. The reader should note that InceptionV3 processed only the MPI images, while the Neural Network and the Random Forest classifiers processed only clinical information.

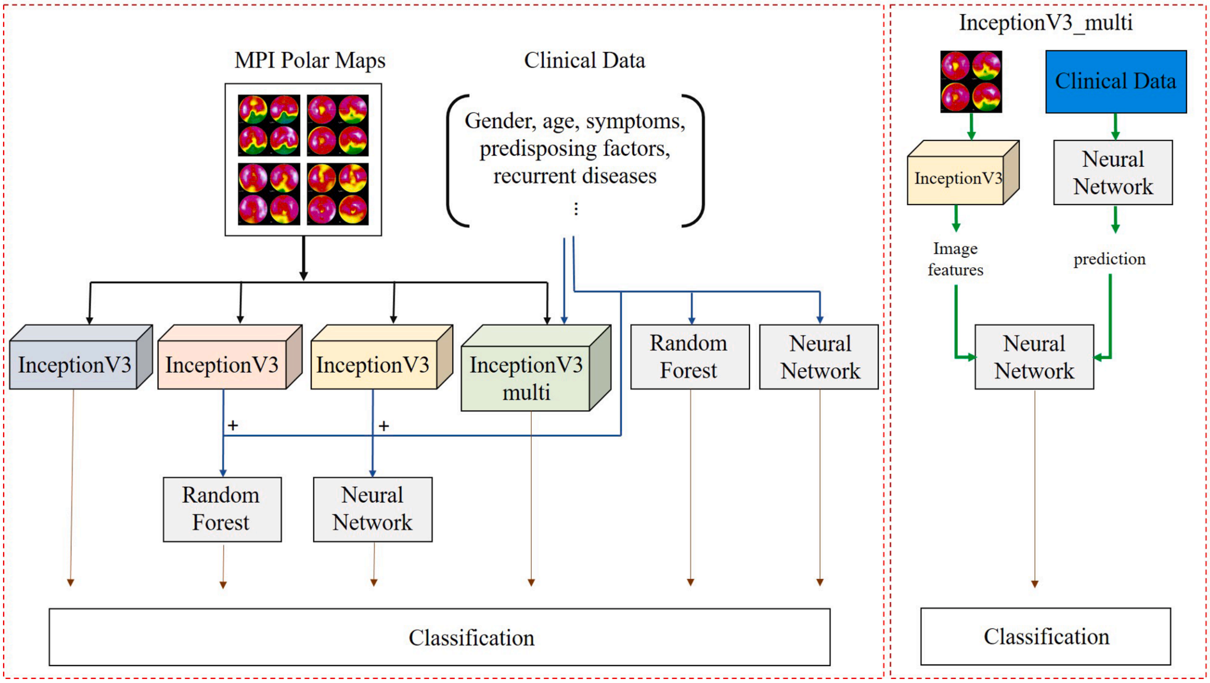

Fig. 2. Overview of the networks used for the experiments.

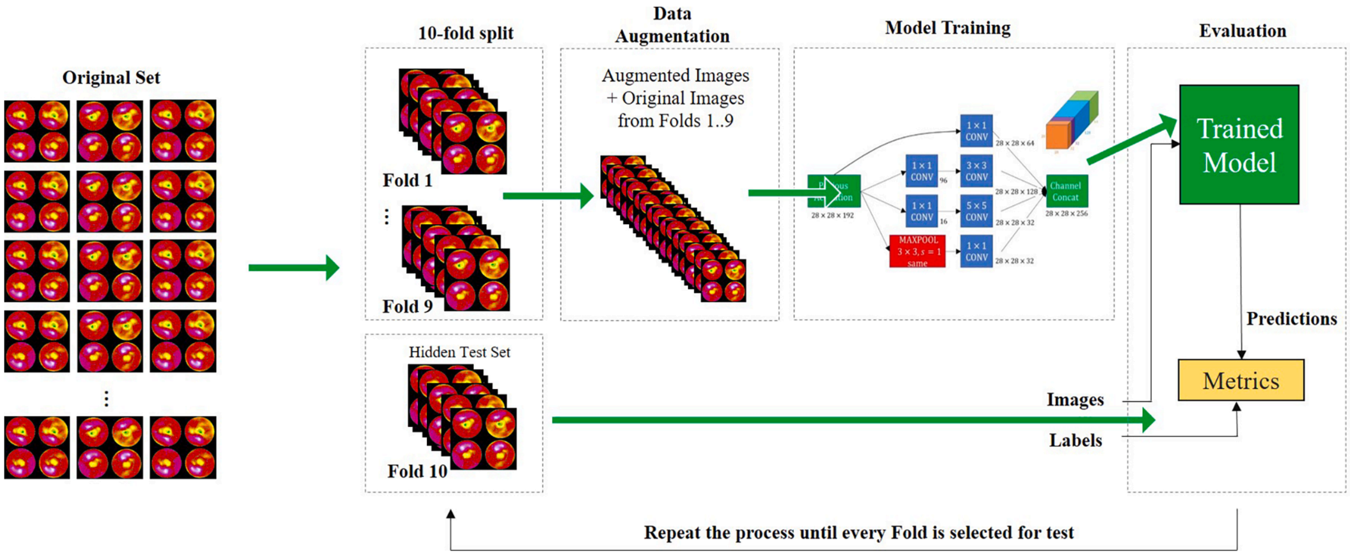

Fig. 3. The process of 10-fold cross-validation.

Random Forest classifier outperformed both the Neural Network and the CNN in identifying CAD, achieving an overall classification accuracy of 75.79%, a sensitivity of 74.07%, a specificity of 77.08%, and an AUC score of 70.22%.

The single-input strategy using image data (which is performed via the InceptionV3 classifier) yielded sub-optimal results and, especially, a poor sensitivity score of 64.15%. In Fig. 4, training accuracy checkpoints of the CNN are illustrated.

For the feature maps produced by the trained CNN to be tested, the outputs of selected layers were monitored by supplying a random image from the network's dataset. The CNN predicts the input image label according to the learned weights obtained from the training process. These weights altered the feature maps produced by each convolutional and Max-Pooling layer during processing. Selected feature maps are illustrated in Fig. 5. The feature maps produced by deeper convolutional layers were more abstract and reflect complex features extracted from the initial image as expected. These features do not necessarily represent decisive factors related to the patient's risk of suffering from CAD. The densely connected layers performed the classification of these features at the top of the CNN, wherein some features may be discarded.

Multi-Input model performance

The classification performance of the models designed to handle multiple types of data demonstrates this method's effectiveness. The complete results are presented in Table 5.

The overall accuracy of 78.43% was achieved for the multi-input dataset obtained by the sequential scheme of InceptionV3 followed by a Random Forest classifier. This method achieved a sensitivity of 77.36% and specificity of 79.25%, outperforming the rest of the models and

Table 4
Classification metrics of the single-input model experiments. Metrics assigned with * are computed for each fold and then summarised. Metrics assigned with ** are averaged for the 10 folds, and the standard deviation is reported. The AUC score is an established ML measure to estimate the prediction model's overall performance across every possible classification threshold.

Metric/Model	InceptionV3 (image only)	Neural Network (clinical data only)	Random Forest (clinical data only)
True Positives*	81	139	180
True Negatives*	282	246	249
False Positives*	41	77	74
False Negative*	162	104	63
Accuracy**	64.14 ± 4.49%	68.02 ± 4.92%	75.79 ± 4.00%
Sensitivity**	33 ± 4.00%	57.20 ± 5.37%	74.07 ± 4.00%
Specificity**	87.30 ± 4.36%	76.16 ± 4.72%	77.08 ± 3.94%
F1 Score**	44.38 ± 3.49	60.57 ± 1.60	72.43 ± 3.78
AUC score**	70.22 ± 6.77	69.04 ± 4.15	71.09 ± 2.57

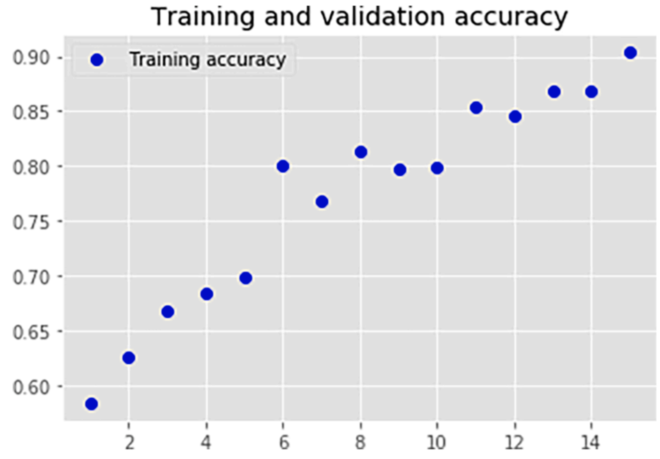

Fig. 4. Training accuracy by training-epoch for the best fold.

methods. The accuracy of ~75% obtained by the single Random Forest classifier (Table 4) was improved when the predictions of the InceptionV3 were supplied to the training dataset (InceptionV3 + Random Forest). This fact yielded an important conclusion. Despite the poor performance of the image classification, the predictions of the InceptionV3 model were useful to the Random Forest classifier. Those predictions enriched the clinical data information, resulting in improved classification accuracy by 3%.

Statistical significance among the two best-performing methods

To compare the statistical significance of the results between InceptionV3 + Random Forest (sequential) and Random Forest, which were the best-performing strategies, 25 times 10-fold cross-validation was performed. Each classifier was trained-tested using 10-fold cross-validation ten times, and the mean accuracy and F1-Score from each 10-fold cross-validation were recorded. For each repetition of the 10-fold cross-validation, the reader should note that the training-test sets' data distribution remained the same. A T-test was performed to evaluate the significance of the results for the mean accuracies and F1 scores. For the particular test, the alpha was set to 0.05. The results are illustrated in Table 6.

The low p-values indicate that there is sufficient evidence that the results are statistically significant.

Comparisons with the medical expert's diagnostic yield

The Deep Learning method competed with the diagnostic performance of the medical experts on the dataset. The human cognitive process's overall accuracy reached 79.15%, which is approximately 1% better than the automatic model's accuracy (78.43%). The experts and the model had a 12% difference regarding sensitivity and specificity, with the experts achieving the optimal result. There is an 8% difference between the experts and the model in terms of specificity, with the model coming first at 79.25%. The results are summarised in Table 7. A graphical comparison of the diagnostic metrics of various classification strategies used in the present study is presented in Fig. 6.

In Fig. 6, a summary of the classification metrics for each method is illustrated.

A case-to-case comparison between the sequential InceptionV3 Random Forest classifier and the expert's prediction is provided in Table 8. This comparison revealed the agreement rates between the automatic model and human expertise. As Table 8 suggests, there was an 86.04% overall agreement.

Also, Cohen's Kappa coefficient was calculated using Table 9, which provided a comparison between experts' and InceptionV3 Random Forest diagnostic yields. Cohen's Kappa coefficient is 72.24.

The errors of the automatic model's errors were very close to the doctors' errors, which enhanced the deep learning method's significance, as it is an indicator of its transparency. It also indicated that the model did not make random predictions, but it was most probably based on doctors' same factors. Other proof for the above reasoning is the fact that 83.57% of the common forecasts were correct. Finally, out of 79 disagreements in the forecasts, in 41 cases the correct forecast was that of the expert, while in 38 cases the correct decision was that of the model.

The agreement rate is defined by comparing the doctors' diagnosis with the diagnosis made by the model. This study's ground truth is not the experts' diagnosis but the Coronary Angiography outcome**.** Hence, the model was trained using Coronary Angiography results as ground truth and not the experts' diagnosis, which is why the agreement rating may be high, but the actual accuracy may be lower.

Robustness on acquisition device variation

To evaluate the robustness of the proposed method on image acquisition device variation, InceptionV3 + Random Forest was trained using images from the first device and tested using images captured by the alternative device. The results are reported in Table 10. Training and testing had been performed ten times, and the results are the mean scores accompanied by the standard deviation.

Images derived from the first device (438 images) had been used as the training set, where 207 images refer to diseased instances and 261 to healthy. For the second device, 36 diseased and 62 healthy instances constituted the testing image data of 98 size. As the images derived from the second device were very few to train a deep network, the process could not be performed vice-versa.

Although training and testing images constituted imbalanced datasets, the results suggest that InceptionV3 + Random Forest was slightly sensitive to device variation. However, more research is required in the future to test the validity of the outcomes.

A T-test was performed to validate the statistical significance of the results. The high p-values (0.4734 for the Accuracy and 0.4061 for the F1 score) suggest no evidence that the results are statistically insignificant. The results are presented in Table 11.

Comparison with state-of-the-art CNNs

InceptionV3 was also compared with other state-of-the-art CNNs. The reader should note that the comparison was held using image data and clinical data, following the CNN + Random Forest strategy. The

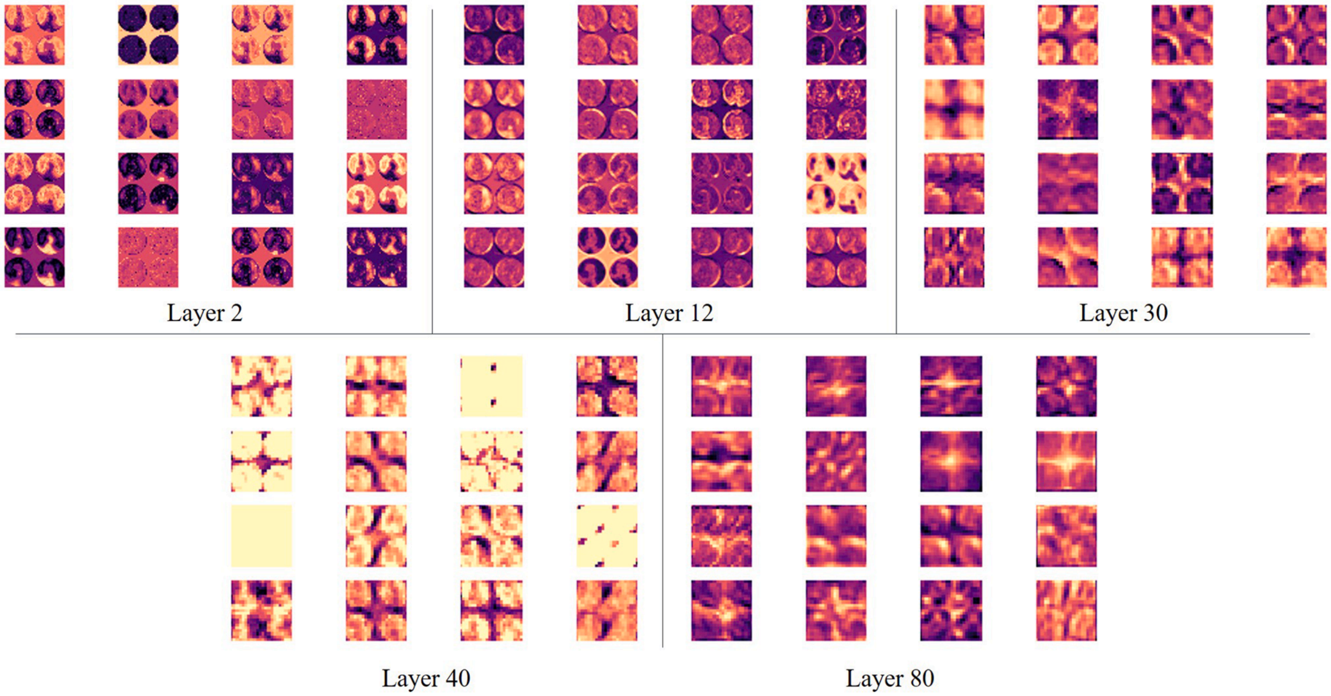

Fig. 5. Feature maps produced by the 2nd, 12th, 30th, 40th, and 80th layer of the InceptionV3 CNN. The colour depends on the mapping chosen to plot the feature maps.

Table 5
Classification metrics of the multi-input model experiments. Metrics assigned with * are computed for each fold and then summarised. Metrics assigned with ** are averaged for the 10 folds, and the standard deviation is reported.

Metric/ Model	InceptionV3 + Neural Network (nested)	InceptionV3 + Neural Network (sequential)	InceptionV3 + Random Forest (sequential)
True Positives*	118	162	188
True Negatives*	276	252	256
False Positives*	47	71	67
False Negative*	125	81	55
Accuracy**	69.60 ± 4.61%	73.14 ± 4.20%	78.44 ± 3.71%
Sensitivity**	48.55 ± 4.29%	66.66 ± 3.98%	77.36 ± 3.87%
Specificity**	85.44 ± 5.44%	78.01 ± 4.14%	79.25 ± 5.12%
F1 Score**	57.84 ± 2.98	68.07 ± 2.54	75.50 ± 2.89
AUC score**	78.09 ± 9.26	80.54 ± 7.62	79.26 ± 3.66

Table 6
T-test results for 25 times 10-fold cross-validation between InceptionV3 + Random Forest (sequential) and Random Forest.

Type	Mean	Variance
InceptionV3 + Random Forest Accuracy	78.37	3.37
Random Forest Accuracy	75.48	2.25
p-value	<0.01	
InceptionV3 + Random Forest F1 Score	75.62	2.64
Random Forest F1 Score	72.65	2.62
p-value	<0.01	

Table 7
Comparison with the diagnostic yield of medical experts.

Metric/Model	Medical Experts	InceptionV3 + Random Forest (sequential)
True Positives	218	188
True Negatives	230	256
False Positives	93	67
False Negative	25	55
Accuracy	79.15%	78.44 ± 3.71%
Sensitivity	89.17%	77.36 ± 3.87%
Specificity	71.20%	79.25 ± 5.12%

results in terms of accuracy, sensitivity, specificity, and F1 score are presented in Table 12. 10-fold cross-validation is used for every evaluation. All networks shared the same hyperparameters with InceptionV3, except for the trainable layers. 40% of the total convolutional layers are made trainable for all networks, starting from the layers close to the Average Pooling layer.

VGG19 also produced acceptable results, although slightly worse than InceptionV3. Several state-of-the-art networks could also be evaluated in future research. Additionally, several fine-tuning experiments could further optimise other CNNs, aiming to improve the results.

Discussion

This research aimed to develop Artificial Intelligence diagnostic models to detect Coronary Artery Disease using the image and clinical data. The Deep Learning approach was applied to automatically identify abnormal MPI studies, based on polar map images. However, these images were not sufficient enough for accurate diagnosis for an accurate diagnosis due to image acquisition and processing technology's sub-optimal performance. Some inherent discrepancies between MPI's functional information and the anatomical findings revealed by invasive Coronary Angiography; these inconsistencies often lead to confusing results. Hence, several clinical parameters must be considered for the final clinical diagnosis.

Deep Learning has become an established and valuable medical image classification approach and segmentation on many medical imaging domains [33]. Deep Learning has even succeeded in discovering complex and abstract decisive features from medical images and features that the human eye cannot detect [34,35]. The probability that

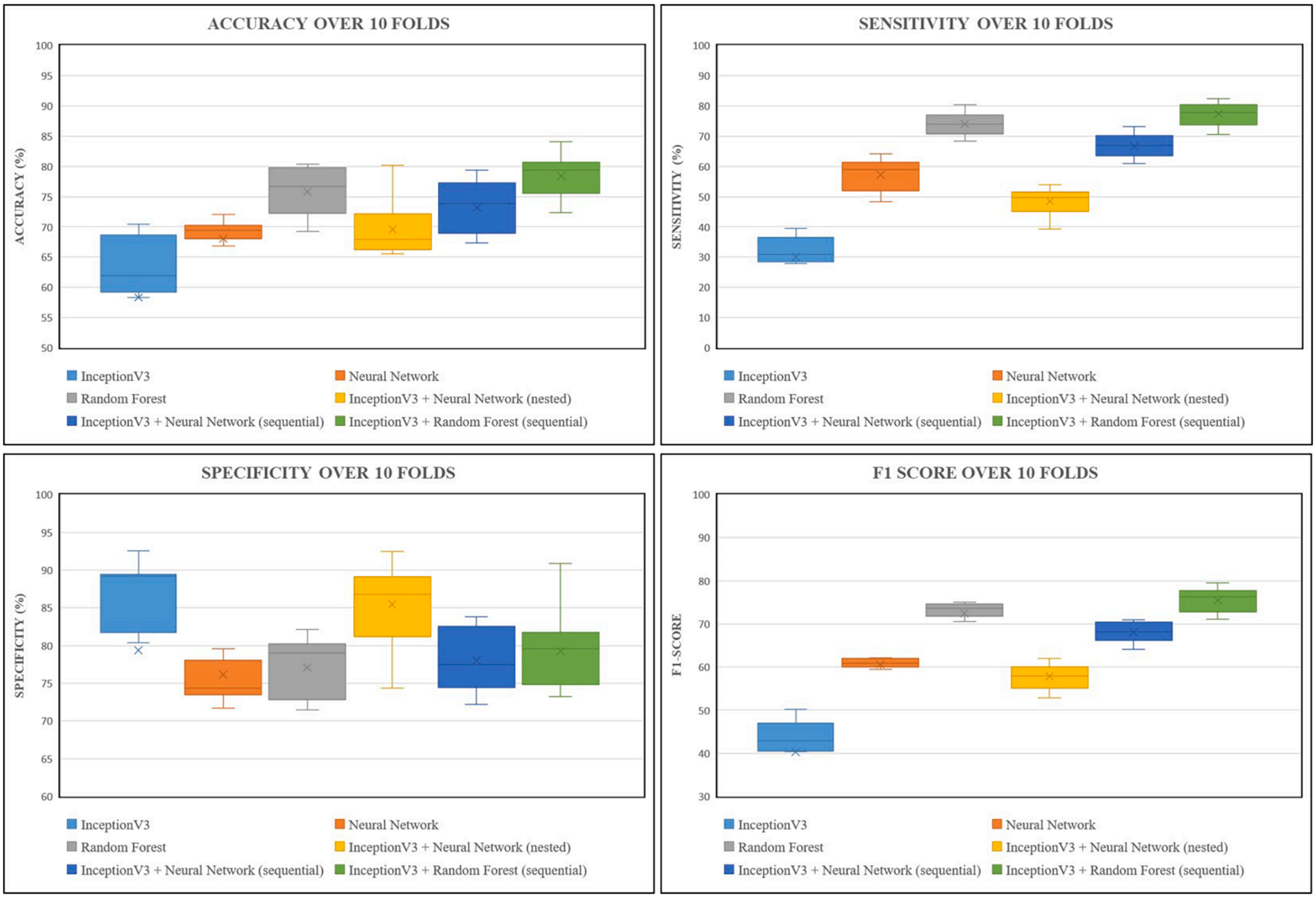

Fig. 6. Summary of the performance of the various classification strategies followed in the present study.

Table 8
Case-to-case comparison between the sequential InceptionV3 Random Forest classifier and the experts' prediction.

Type	Number	Percentage (%)
Total Agreements	487	86.04
Total Disagreements	79	13.96
Number of Correct Agreements	407	83.57
Number of Mistaken Agreements	80	16.43
Number of Disagreements where expert's decision was the correct	41	51.89
Number of Disagreements where the model's decision was the correct	38	48.11

Table 9
Comparison between experts' and InceptionV3 Random Forest diagnostic yields. The reported values are used to calculate Cohen's Kappa.

		Expert diagnosis	
		CAD	Healthy
InceptionV3 Random Forest diagnosis	CAD	249	17
	Healthy	62	238

Table 10
Accuracy of the InceptionV3 + Random Forest (sequential) model for the second image acquisition device's test set. The experiment was conducted ten times, and the mean and standard deviation results are reported.

Metric/Model	InceptionV3 + Random Forest (sequential)
Accuracy	76.53 ± 3.38%
Sensitivity	77.77 ± 5.75%
Specificity	75.80 ± 3.6%
F1 Score	70.88 ± 4.07
AUC Score	75.24 ± 3.66

Table 11
Statistical Significance Test for the Accuracy and the F1-score regarding the robustness to device variation experiment. The experiment was conducted ten times.

Type	Mean	Variance
InceptionV3 + Random Forest Accuracy	76.53	3.38
p-value	0.4734	
InceptionV3 + Random Forest F1 Score	70.88	4.07
p-value	0.4061	

important hidden underlying features representing biological signatures or other significant processes exist in MPI images was negligible in the current experiment. For this reason, the deep learning framework was not expected to perform surprisingly better than the human eye. This expectation was confirmed by the results. The reader should note that this work aimed not to perform a deep analysis on biomarker extraction from the specific MPI imaging data but to propose a hybrid image-clinical data classification method to handle the complex risk factors affecting CAD severity. This research demonstrated that the image process, combined with the available clinical information, can turn a sub-optimal classifier into a robust one, matching the medical experts' validity. The hybrid InceptionV3 – Random Forest framework, which initially learned to distinguish between normal and abnormal Polar

Table 12
Comparison with state-of-the-art CNNs.

Network	Accuracy	Sensitivity	Specificity	F1 score
InceptionV3 + Random Forest	78.44 ± 3.71	77.36 ± 3.87	79.25 ± 5.12	75.50 ± 2.89
VGG19 + Random Forest	77.56 ± 2.78	73.25 ± 4.05	80.80 ± 6.45	73.70 ± 3.16
ResNet V2 + Random Forest	65.37 ± 5.22	66.25 ± 3.55	64.70 ± 6.74	62.16 ± 3.45
MobileNet v2 + Random Forest	70.31 ± 2.41	72.83 ± 1.36	68.42 ± 2.08	67.81 ± 2.55

Maps images, associated the predicted classes with the rest of the clinical conditions. The rest of the evaluated similar methodologies (InceptionV3 + Neural Network and InceptionV3 with an embedded auxiliary input) performed weaker.

The interpretation of the generated feature maps is undeniably a matter of thorough ongoing research [36]. The behaviour of deep models as black boxes is directing the attention towards finding explanations making use of, but not limited to, visualisation methods. Deep models, however, involve many convolutional layers, which extract hundreds of feature maps and millions of unique features. Therefore, a complete and analytic investigation is almost impossible. Feature maps could be used as a confirmation that the designed framework is seeking patterns in the correct direction and is not learning irrelevant information through its training. In the feature maps sample (Fig. 5), the reader can notice that feature maps coming from the first layers of the network segment the polar maps into smaller regions using the colour variance. It is also worth noticing that areas of the image, such as the centre and the edges, which do not belong to the Region of Interest, are not generating any useful features. Those two observations support the assumption that the model is learning significant features and generally follow the human observation procedure. However, a great interest of origin and the meaning of the deeper-extracted features come from the network's last layers. Those features are supposed to be high-level and potentially reveal the important underlying information. Layers around layer 80 (as illustrated in Fig. 5) extract deep features. The confirmation or the denial of their importance is impossible to be defined optically and is a matter of further research.

The significant difference in distributing the false positives and the true negatives between the human experts and the deep learning model is explainable. Medical experts carry a heavy burden when submitting medical reports since they take full responsibility for their suggestions. A false positive decision is less than the cost of a false negative, which directs medical experts towards overestimating patients' risk. In contrast, the automatic diffusion model stabilises when it obtains the best accuracy, regardless of the distribution of False Positives and False Negatives.

Moreover, the agreement analysis between the deep learning method and the human expertise (86% agreement, 72.24 Cohen's Kappa) is a prime indication that the proposed methodology bases these predictions on actual data and does not do so by chance. Although deep networks suffer from inexplicability, the above comparison aids to that front.

In a previous paper by our research group [25], it was mentioned that larger-scale MPI datasets should be used to confirm the effectiveness of deep learning methods in MPI imaging analysis. In that work, a ~ 74% classification accuracy was achieved, using only MPI polar maps. The present study was based on a larger dataset, which is adequate to draw more firm conclusions. In this work, it has been demonstrated that the deep learning method can match the experts' performance using this larger sample. Hence, this study confirms that the previous results were indeed significant.

Finally, an interesting topic for future research would be manipulating the original Polar Map images towards applying specific data augmentation methods that generate realistic images and benefit the deep learning models.

Conclusions

Two major conclusions are drawn. Firstly, despite the moderate performance of the CNN in image characterisation, the predictions enhanced the Random Forest classifier's robustness, which associated the image-based predictions with the rest of the clinical data and achieved better results. Secondly, this hybrid method competed with the medical expert's diagnostic ability in Accuracy and Specificity. Therefore, enabling utilisation of this algorithm for medical assisting tools incorporated into computer-aided support systems.

Moreover, developing a computer-aided decision-making support system for CAD diagnosis using Machine Learning and Deep Learning methods is possible. In future research, an issue that must be addressed is the transparency of the deep learning model's decision-making mechanism, acting as a black box. Explainable and interpretable deep learning models and developed methods are vital so that medical experts can trust them.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The dataset is not publicly available due to ethical reasons.

Data ethics

The present study is retrospective

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Benjamin EJ, Muntner P, Bittencourt MS. Heart disease and stroke statistics-2019 update: a report from the American Heart Association. Circulation 2019;139: e56–528. https://doi.org/10.1161/cir.0000000000000659.

[2] Apostolopoulos DJ, Kaspiri A, Spyridonidis T, Patsouras N, Savvopoulos CA, Davlouros P, et al. Assessment of absolute Tc-99m tetrofosmin retention in the myocardium as an index of myocardial blood flow and coronary flow reserve by gated-SPECT/CT: a feasibility study. Ann Nucl Med 2015;29:588–602. https://doi.org/10.1007/s12149-015-0982-6.

[3] Willerson JT, Wellens HJJ, Cohn JN, Holmes DR, editors. Cardiovascular Medicine. London: Springer London; 2007. https://doi.org/10.1007/978-1-84628-715-2.

[4] van Dijk J, Mouden M, Ottervanger J, van Dalen J, Knollema S, Slump C, et al. value 99999999of attenuation correction in stress-only myocardial perfusion imaging using CZT-SPECT. J Nucl Cardiol 2017;24:395–401.

[5] Apostolopoulos DJ, Savvopoulos C. What is the benefit of CT-based attenuation correction in myocardial perfusion SPET? Hell J Nucl Med 2016;19:89–92. https://doi.org/10.1967/s0024499100360.

[6] Verweij N, Eppinga RN, Hagemeijer Y, van der Harst P. Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. Sci Rep 2017;7:1–9. https://doi.org/10.1038/s41598-017-03062-8.

[7] Moss AJ, Williams MC, Newby DE, Nicol ED. The updated NICE guidelines: cardiac ct as the first-line test for coronary artery disease. Curr Cardiovasc Imaging Rep 2017;10:15. https://doi.org/10.1007/s12410-017-9412-6.

[8] Kortesniemi M, Tsapaki V, Trianni A, Russo P, Maas A, Källman H-E, et al. The European Federation of Organisations for Medical Physics (EFOMP) White Paper: Big data and deep learning in medical imaging and in relation to medical physics profession. Physica Med 2018;56:90–3. https://doi.org/10.1016/j.ejmp.2018.11.005.

[9] Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature 2018;559:547–55. https://doi.org/10.1038/s41586-018-0337-2.

[10] Giger ML. Machine learning in medical imaging. J Am College Radiol 2018;15: 512–20. https://doi.org/10.1016/j.jacr.2017.12.028.

[11] LeCun Y, Kavukcuoglu K, Farabet C. Convolutional networks and applications in vision. Proceedings of 2010 IEEE International Symposium on Circuits and

Systems, Paris, France: IEEE; 2010, p. 253–6. https://doi.org/10.1109/ISCAS.2010.5537907.

[12] Alizadehsani R, Habibi J, Hosseini MJ, Mashayekhi H, Boghrati R, Ghandeharioun A, et al. A data mining approach for diagnosis of coronary artery disease. Comput Methods Programs Biomed 2013;111:52–61. https://doi.org/10.1016/j.cmpb.2013.03.004.

[13] Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid J-J, Sandhu S, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol 1989;64:304–10. https://doi.org/10.1016/0002-9149(89)90524-9.

[14] Babaoğlu I, Fındık O, Bayrak M. Effects of principle component analysis on assessment of coronary artery diseases using support vector machine. Expert Syst Appl 2010;37:2182–5. https://doi.org/10.1016/j.eswa.2009.07.055.

[15] Abdar M, Książek W, Acharya UR, Tan R-S, Makarenkov V, Pławiak P. A new machine learning technique for an accurate diagnosis of coronary artery disease. Comput Methods Programs Biomed 2019;179:104992. https://doi.org/10.1016/j.cmpb.2019.104992.

[16] Arabasadi Z, Alizadehsani R, Roshanzamir M, Moosaei H, Yarifard AA. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Comput Methods Programs Biomed 2017;141:19–26. https://doi.org/10.1016/j.cmpb.2017.01.004.

[17] Betancur J, Hu L-H, Commandeur F, Sharir T, Einstein AJ, Fish MB, et al. Deep learning analysis of upright-supine high-efficiency SPECT myocardial perfusion imaging for prediction of obstructive coronary artery disease: a multicenter study. J Nucl Med 2019;60:664–70. https://doi.org/10.2967/jnumed.118.213538.

[18] Spier N, Nekolla S, Rupprecht C, Mustafa M, Navab N, Baust M. Classification of polar maps from cardiac perfusion imaging with graph-convolutional neural networks. Sci Rep 2019;9:7569. https://doi.org/10.1038/s41598-019-43951-8.

[19] Sharma M, Acharya UR. A new method to identify coronary artery disease with ECG signals and time-Frequency concentrated antisymmetric biorthogonal wavelet filter bank. Pattern Recogn Lett 2019;125:235–40. https://doi.org/10.1016/j.patrec.2019.04.014.

[20] Tan JH, Hagiwara Y, Pang W, Lim I, Oh SL, Adam M, et al. application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. Comput Biol Med 2018;94:19–26. https://doi.org/10.1016/j.compbiomed.2017.12.023.

[21] Moradi S, Oghli MG, Alizadehasl A, Shiri I, Oveisi N, Oveisi M, et al. MFP-Unet: A novel deep learning based approach for left ventricle segmentation in echocardiography. Physica Med 2019;67:58–69. https://doi.org/10.1016/j.ejmp.2019.10.001.

[22] Butun E, Yildirim O, Talo M, Tan R-S, Rajendra AU. 1D-CADCapsNet: One dimensional deep capsule networks for coronary artery disease detection using ECG signals. Physica Med 2020;70:39–48. https://doi.org/10.1016/j.ejmp.2020.01.007.

[23] Biagini E, Shaw LJ, Poldermans D, Schinkel AF, Rizzello V, Elhendy A, et al. Accuracy of non-invasive techniques for diagnosis of coronary artery disease and prediction of cardiac events in patients with left bundle branch block: a meta-analysis. Eur J Nucl Med Mol Imaging 2006;33:1442–51. https://doi.org/10.1007/s00259-006-0156-9.

[24] Takx RA, Blomberg BA, Aidi HE, Habets J, de Jong PA, Nagel E, et al. Diagnostic accuracy of stress myocardial perfusion imaging compared to invasive coronary angiography with fractional flow reserve meta-analysis. Circulation: Cardiovascular Imaging 2015;8:e002666. https://doi.org/10.1161/circimaging.114.002666.

[25] Apostolopoulos ID, Papathanasiou ND, Spyridonidis T, Apostolopoulos DJ. Automatic characterisation of myocardial perfusion imaging polar maps employing deep learning and data augmentation. Hellenic J Nucl Med 2020;23:125–32. https://pubmed.ncbi.nlm.nih.gov/32716403/.

[26] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 2818–26. https://doi.org/10.1109/CVPR.2016.308.

[27] Knuuti J, Wijns W, Saraste A, Capodanno D, Barbato E, Funck-Brentano C, et al. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes: the Task Force for the diagnosis and management of chronic coronary syndromes of the European Society of Cardiology (ESC). Eur Heart J 2020;41:407–77. https://doi.org/10.1093/eurheartj/ehz425.

[28] Hajjiri MM, Leavitt MB, Zheng H, Spooner AE, Fischman AJ, Gewirtz H. Comparison of positron emission tomography measurement of adenosine-stimulated absolute myocardial blood flow versus relative myocardial tracer content for physiological Assessment of coronary artery stenosis severity and location. JACC: Cardiovasc Imag 2009;2:751–8. https://doi.org/10.1016/j.jcmg.2009.04.004.

[29] Gautier M, Pepin M, Himbert D, Ducrocq G, Iung B, Dilly M-P, et al. Impact of coronary artery disease on indications for transcatheter aortic valve implantation and on procedural outcomes. EuroIntervention 2011;7:549–55. https://doi.org/10.4244/eijv7i5a90.

[30] Rodríguez P, Bautista MA, Gonzalez J, Escalera S. Beyond one-hot encoding: Lower dimensional target embedding. Image Vis Comput 2018;75:21–31. https://doi.org/10.1016/j.imavis.2018.04.004.

[31] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2009;22:1345–59. https://doi.org/10.1109/TKDE.2009.191.

[32] Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10), 2010, p. 807–14. http://www.cs.toronto.edu/~hinton/absps/reluICML.pdf.

[33] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical image analysis 2017;42:60–88.https://doi.org/10.1016/j.media.2017.07.005.

[34] Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilising transfer learning with convolutional neural networks. Phys Eng Sci Med 2020;43:635–40. https://doi.org/10.1007/s13246-020-00865-4.

[35] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559–67. https://doi.org/10.1038/s41591-018-0177-5.

[36] Avanzo M, Stancanello J, El Naqa I. Beyond imaging: the promise of radiomics. Physica Med 2017;38:122–39. https://doi.org/10.1016/j.ejmp.2017.05.071.