Rom J Morphol Embryol 2025, 66(4):*in press* ISSN (print) 1220–0522, ISSN (online) 2066–8279 doi: 10.47162/RJME.66.4.y

ORIGINAL PAPER

RJME
Romanian Journal of Morphology & Embryology
http://www.rjme.ro/

Reliable deep learning for coronary artery disease detection: a patient-level, statistically validated MRI study

CHRISTIANA RALUCA DĂNCIULESCU[1)], CONSTANTIN RENATO IVĂNESCU[2)], DANIEL-ROBERT STĂNESCU[1)], ANDREI-FLORENTIN BĂIAŞU[1)], DRAGOŞ OVIDIU ALEXANDRU[3)], MIRCEA-SEBASTIAN ŞERBĂNESCU[3)]

[1)]*Doctoral School, University of Medicine and Pharmacy of Craiova, Romania*

[2)]*Department of Computers and Information Technology, University of Craiova, Romania*

[3)]*Department of Medical Informatics and Biostatistics, University of Medicine and Pharmacy of Craiova, Romania*

Abstract

Background: Accurate detection of coronary artery disease (CAD) from cardiac magnetic resonance (CMR) imaging can support earlier diagnosis and streamlined clinical decision-making. *Objective*: This study evaluated the performance and statistical robustness of two deep learning architectures – DenseNet121 and ResNet50 – for automated CAD classification using multiparametric CMR imaging. *Methods*: Images were preprocessed using a valid pipeline and partitioned strictly at the patient level. Model performance was quantified through average accuracy, area under the receiver operating characteristic (ROC) curve (AUC–ROC), precision recall, while distributional assumptions were assessed using Shapiro–Wilk tests and variance homogeneity was explored with Brown–Forsythe test. *Results*: ResNet50 demonstrated the strongest performance, achieving an average accuracy of 90.43%, AUC–ROC of 0.862, and area under the precision recall curve (PR–AUC) of 0.891. DenseNet121 showed lower accuracy (81.72%). Statistical analysis revealed non-normal performance distributions and significant variance differences between models. *Conclusions*: The findings indicate that ResNet50 offers a reliable and statistically validated solution for CAD detection from CMR imaging. The combined use of realistic preprocessing and comprehensive inferential testing supports the generation of reproducible and clinically meaningful performance estimates.

Keywords: coronary artery disease, cardiac MRI, deep learning, patient-level evaluation, statistical validation.

Introduction

Coronary artery disease (CAD) remains the most prevalent form of cardiovascular disease and a leading cause of morbidity and mortality worldwide. It is characterized by the narrowing of blockage of the coronary arteries, typically due to the buildup of atherosclerotic plaques, which restricts blood supply to the myocardium and can result in ischemia, myocardial infarction, or sudden death [1, 2]. Early and accurate diagnosis is essential, since therapeutic interventions such as percutaneous coronary intervention, coronary artery bypass grafting, or aggressive medical management can significantly improve patient outcomes if applied in time.

Conventional diagnosis of CAD relies on a combination of clinical assessment, biochemical markers, and imaging modalities. Among these, invasive coronary angiography continues to be regarded as the "gold standard" for confirming the presence and severity of arterial stenosis [3]. However, due to its invasive nature, risks, and costs, it is often preceded by non-invasive imaging methods. Computed tomography coronary angiography (CTCA) offers excellent anatomical details, while stress echocardiography and nuclear perfusion imaging provide functional information regarding myocardial ischemia [3]. In recent years, cardiac magnetic resonance (CMR) imaging has gained recognition as a powerful multiparametric tool, capable of combining high-resolution morphological visualization with functional and tissue characterization. CMR techniques, including late gadolinium enhancement (LGE), T2-weighted imaging (T2WI), perfusion sequences, and cine steady-state free precession (SSFP), allow clinicians to assess viability, edema, perfusion deficits, and wall motion abnormalities in a comprehensive manner [4].

Despite these advances, interpretation of CMR is complex and time-sensitive, requiring considerable expertise. Observer variability, limited availability of trained cardiologists, and the increasing volume of imaging data represent important challenges for clinical practice. In this context, artificial intelligence (AI) has emerged as a transformative approach in cardiovascular imaging. AI, particularly machine learning (ML) and deep learning (DL), enables automated extraction of complex patterns from high-dimensional medical images, offering the potential to improve diagnostic accuracy, reduce interpretation time, and standardize workflows [5].

AI has become increasingly central across diverse areas of medical imaging, supporting tasks that range from automated cancer grading to structural-functional tissue assessment. The integration of DL architectures for Gleason grading in prostate cancer, including transfer-learning approaches built on general-purpose networks and validated against expert pathologists, illustrates the capacity of neural models to capture subtle morphological patterns with high diagnostic agreement [6, 7]. Similar advances have been reported in breast tumor classification from mammography, the quantification of stromal–tumoral interactions in prostate adenocarcinoma, and the structural characterization of basal cell carcinoma architecture [8–11]. Beyond oncology, DL methods have also shown strong potential in the automatic segmentation of liver lesions in ultrasound imaging, demonstrating the adaptability of AI across modalities [12]. Collectively, these contributions highlight how DL can reliably extract clinically relevant features from complex biomedical images, reinforcing the motivation of the present study to apply similarly rigorous, patient-level and statistically validated AI methodology for CAD classification from CMR imaging. The convergence of evidence across these domains supports the broader view that robust AI pipelines – whether in oncology, dermatology, hepatology, or cardiology – can enhance diagnostic precision when grounded in transparent and reproducible evaluation practices.

Over the past decade, significant progress has been made in applying AI to CAD detection and risk stratification. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models have been applied to classify perfusion deficits, quantify ventricular function, and identify coronary lesions from imaging modalities such as CTCA and CMR [13]. AI-based systems have demonstrated performance comparable to expert cardiologists in certain tasks, while also offering reproducibility and scalability. In addition, explainable AI methods are increasingly being developed to ensure clinical trust by providing transparency in the decision-making process [14].

The availability of large, annotated imaging datasets is central to the development and validation of these models. A notable example is the CAD multiparametric CMR dataset. The dataset's balanced design, multimodal nature, and robust labeling using invasive angiography as ground truth make it particularly well-suited for ML research. Several studies have already explored this dataset to propose and benchmark novel AI approaches for CAD detection. For instance, hybrid architectures combining CNNs with traditional classifiers such as Random Forest have reported strong performance [15]. Similarly, clustering-enhanced DL models have been applied to achieve high diagnostic accuracy [16]. These works illustrate both the promise and the versatility of AI when applied to multiparametric CMR.

Aim

In this manuscript, we carried out a comparative evaluation of two DL models for the classification of CAD *versus* non-CAD. Prior to model train, all images were preprocessed through resizing, normalization, and augmentation (flips, rotations, and intensity adjustments) to increase variability and reduce overfitting. Both models were initialized with weights pre-trained on ImageNet and subsequently fine-tuned on the dataset to adapt to the specific feature of CMR imaging. Performance evaluation was carried out using 10-fold cross-validation to ensure generalizability. A comprehensive statistical benchmarking process was performed to ensure that the comparative results are not only numerically robust but also statistically meaningful.

Materials and Methods

The dataset

The dataset used in this study consists of 63 648 multi-parametric CMR images, which collectively represent one of the largest publicly available collections of this type [17]. Within this dataset, 26 104 images correspond to subjects with a confirmed diagnosis of CAD, while the remaining 37 544 images originate from individuals without evidence of cardiovascular pathology, who served as the control group. Examples of both groups are presented in Figure 1. The CAD status in the patient cohort was established using invasive coronary angiography, widely accepted as the clinical "gold standard" for this condition.

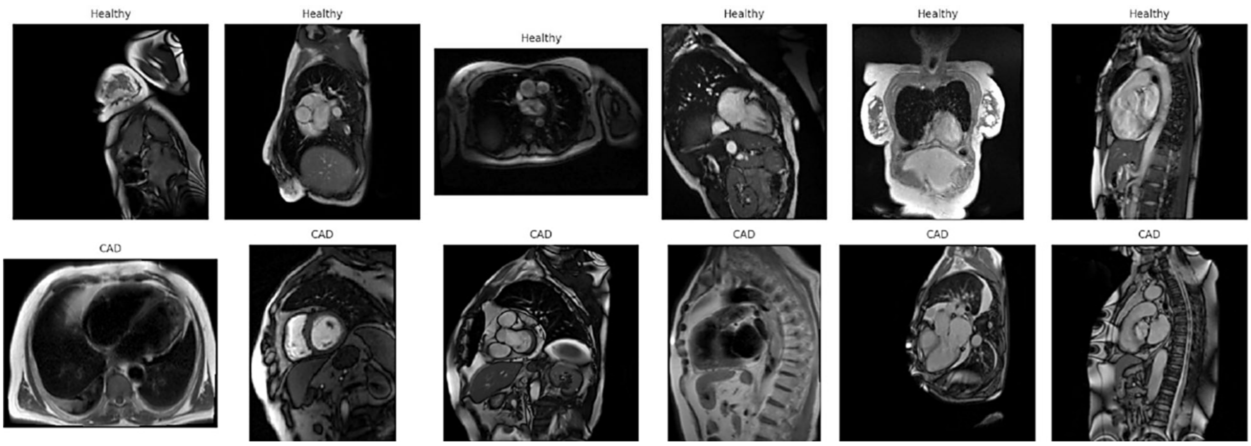

Figure 1 – ***Representative CMR imaging slices from the study cohort, illustrating both healthy subjects and patients diagnosed with CAD. The images display a range of anatomical planes and contrast characteristics, highlighting the visual variability relevant for downstream analysis and model evaluation. CAD: Coronary artery disease; CMR: Cardiac magnetic resonance.***

The imaging protocol involved the acquisition of the four distinct CMR sequences, each contributing complementary diagnostic information. These sequences were: LGE, which provides tissue characterization and highlights area of fibrosis; perfusion imaging, used to detect ischemia, T2WI, which is sensitive to edema and inflammatory processes; and SSFP, the most commonly used cine sequence for assessing cardiac morphology and function. For each sequence, images were obtained in both long-axis and short-axis orientations of the heart, thereby capturing a comprehensive anatomical and functional overview.

The long-axis orientation comprised two-chamber, three-chamber, and four-chamber views. For each of these, one slice was acquired from two different angular perspectives, effectively yielding six long-axis images per sequence. The short-axis orientation was even more detailed: it included 10 slices covering the entire heart, from the basal to the apical regions, with each slice again recorded from two separate angular viewpoints. This design resulted in 20 short-axis images per sequence.

In total, each subject contributed 13 slices per sequence (three long-axis + 10 short-axis), which each slice captured at two angles, generating 52 images per patient across the four sequences. The dataset includes 1244 individuals, of whom 502 were patients with angiographically confirmed CAD and 722 were health participants.

DenseNet121

DenseNet121 is a deep CNN that belongs to the family of densely connected convolutional networks (DenseNets) proposed by Huang *et al.* (2018) [18]. The central concept of DenseNet is that each layer receives as input not only the feature maps from the immediately preceding layer, but also from all earlier layers in the same dense block. This pattern of connectivity encourages feature reuse, strengthens the flow of gradients, and helps to alleviate the vanishing-gradient problem [18, 19].

DenseNet121 consists of four dense blocks separated by transition layers that reduce spatial dimensions *via* convolution and pooling. The full model has 121 layers, including convolutional, pooling, and fully connected components. Each convolutional operation is followed by batch normalization and rectified linear unit (ReLU) activation, while the classifier head uses global average pooling and a fully connected layer with softmax for prediction [20].

Compared to networks of similar depth, such as Residual Network (ResNet) or VGG, DenseNet121 is more parameter-efficient because of the concatenation-based feature propagation. This reduces the number of redundant parameters and leads to improved generalization, particularly when training on datasets that are not extremely large [21]. Compact design also lowers computational costs relative to networks of equivalent accuracy.

These characteristics have made DenseNet121 a popular backbone for medical imaging tasks, including radiology, pathology, and cardiology. Its ability to capture subtle spatial cues and to integrate multi-scale information across layers has proven effective for detecting fine-grained structural anomalies in clinical datasets [22, 23].

The defining feature of DenseNet121 is that each layer inside a dense block receives as input all feature maps generated by the previous layers and then produces a fixed number of new feature maps (the growth rate).

Let an input image be *x^((0))∈R^(H_0×W_0×C_0)*, where *H_0* is the height, *W_0* is the width, and *C_0* channels [*e.g.*, *C_0*=3 for red, green, blue (RGB) images].

Dense block

For the l-th layer ($l = 1, \ldots, L$) inside a block, the input is the concatenation of all feature maps produced so far:

$$u_l = [x_0, x_1, \ldots, x_{l-1}], \quad (1)$$

and the output is obtained by a composite function:

$$x_l = H_l(u_l), \quad (2)$$

Where H_l denotes batch normalization, ReLU, a 1×1 convolution (bottleneck), followed by batch normalization, ReLU, and a 3×3 convolution.

Each layer contributes k new channels (growth rate, typically k=32), giving

$$C_{out} = C_{in} + kL. \quad (3)$$

Transition layer

Between dense blocks, a transition layer reduces feature map size and channel number:

$$T(v) = AvgPool_{2\times2}\big(Conv_{1\times1}(v)\big), \quad (4)$$

with channel compression factor $\theta \epsilon (0, 1]$ (DenseNet121 uses θ=0.5).

Network configuration

DenseNet121 uses four dense blocks with layer counts:

$$(L_1, L_2, L_3, L_4) = (6, 12, 24, 16).$$

This yields progressively richer feature representations, while spatial dimensions shrink through pooling and transitions.

Classifier

After the final block, global average pooling generates a fixed-length feature vector:

$$h = GAP(x), \quad (5)$$

which is mapped to class probabilities by a fully connected layer and softmax:

$$p = softmax(Wh + b). \quad (6)$$

ResNet50

ResNet50 is a deep CNN introduce as part of ResNet family by He *et al.* (2016) [24]. The hallmark of this architecture is the use of residual connections, or “skip connections”, which allow the output of a layer to bypass intermediate layers and be added directly to later feature maps. This design addresses the vanishing-gradient problem and enables the successful training of very deep networks, something that was challenging for conventional feed-forward CNNs [24].

The ResNet50 variant contains 50 layers organized into an initial convolutional and max-pooling stage followed by four residual stages made up of bottlenecks blocks. Each bottleneck block consists of three convolutional layers: a 1×1 convolution that reduces dimensionality, a 3×3 convolutional for spatial feature extraction, and a final 1×1 convolution that restores dimensions. The identity mapping through skip connections ensures that the transformation learns only the “residual” portion of the mapping, which stabilizes training and improves optimization efficiency.

Compared to earlier architectures such as VGG-16 or VGG-19, ResNet50 achieves higher accuracy while using fewer parameters, thanks to its bottleneck design. Its ability to train deeper networks without performance degradation marked a breakthrough in computer vision, and the model achieved top performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC 2015) [24].

ResNet50 has since become a widely adopted backbone in both general computer vision and medical imaging applications. Its robust residual framework makes it particularly effective for extracting discriminative features in tasks such as disease classification, lesion detection, and image segmentation. Numerous studies in radiology, ophthalmology, cardiology, and digital pathology have leveraged ResNet50 as a standard benchmark for transfer learning and fine-tuning on domain-specific datasets [25].

From a mathematical point of view, let us consider the same input image described in formula (1).

Residual block (bottleneck design)

Each bottleneck block consists of three convolutions: a 1×1 convolution for dimensionality reduction, a 3×3 convolution for spatial feature extraction, and a final 1×1 convolution for dimensionality restoration. Given an input tensor, z, the block computes:

$$F(z) = W_3 \cdot \sigma(BN(W_2 \cdot \sigma(BN\left(W_1 \cdot \sigma\left(BN(z)\right)\right) \quad (7)$$

where W_1, W_2, W_3 denote convolutional kernels (1×1, 3×3,1×1), BN is batch normalization, σ is ReLU, and $\cdot$ denotes convolution. The output of the residual block is then:

$$y = F(z) + S(z), \quad (8)$$

where $S(z)$ is the shortcut connection (identity if dimensions match, or a 1×1 projection if they differ).

After the final residual block group, a global average pooling layer compresses the spatial dimensions of the feature maps into single feature vector expressed in formula (6). The vector is afterwards forwarded to a fully connected classification layer, which maps each learned representation to the corresponding target class. The layer outputs a set of logits which are further normalized by the softmax function to obtain the final class probabilities.

Results

To ensure a fair and reproductible evaluation framework, we adopted a well-defined benchmarking protocol. A rigorous evaluation procedure is essential, as inadequate statistical design may lead to misleading conclusions. Since both DenseNet121 and ResNet50 use stochastic training procedures (due to random weight initialization, mini-batch sampling, and dropout), single-run evaluations are not reliable. Therefore, to ensure robustness and reproducibility, we have executed 50 independent runs for each architecture. A statistical power analysis was carried out to determine a suitable experimental sample size. We targeted a two-tailed hypothesis test with statistical power ≥95% and a type I error rate α=0.05. A sample that is too large would unnecessarily increase computational cost with minimal benefit, whereas a small sample might reduce the precision and credibility of the findings. The chosen configuration provided an optimal compromise between computational efficiency and inference stability.

We used 10-fold cross-validation to ensure generalization reliability. For each of the 50 runs, the model was trained and evaluated across the 10 folds, and the average classification accuracy (ACA) over the test partitions was computed:

$$ACA = \frac{\text{Number of correctly classified MRI images}}{\text{Total number of MRI images in the test set}} \quad (9)$$

In addition to ACA, we calculated the standard deviation (SD) to evaluate score dispersion across runs, thereby assessing model stability. Since medical datasets typically exhibit class imbalance, we also employed area under the receiver operating characteristic (ROC) curve (AUC–ROC) and area under the precision-recall curve (PR–AUC) as complementary measures of discriminative performance, particularly under skewed class distributions.

To statistically compare the performance of DenseNet121 and ResNet50, we first assessed the normality of accuracy distributions using the Shapiro–Wilk test, given its sensitivity for small to medium sample sizes. For variance homogeneity, we applied Brown–Forsythe test, as heteroscedasticity may bias parametric tests. Based on the assumptions' outcomes, we used either an independent samples *t*-test (when normality and equal variance were satisfied) or the Mann–Whitney *U*-test (when these assumptions were violated). A summary of the evaluation metrics for both architectures is presented in Table 1.

Table 1 – ***Performance metrics for DenseNet121 and ResNet50***

Metric	DenseNet121	ResNet50
ACA	81.72%	90.43%
SD	0.012	0.0046
AUC	0.898	0.962
PR–AUC	0.8121	0.8919

ACA: Average classification accuracy; AUC: Area under the receiver operating characteristic (ROC) curve; PR–AUC: Area under the precision-recall curve; SD: Standard deviation.

Based on the results summarized in Table 1, there is a clear performance difference between the two convolutional architectures. ResNet50 achieved the highest ACA, reaching 90.43%, compared to 81.72% obtained by DenseNet121, which indicates that ResNet50 extracted more discriminative cardiac features from MRI slices. The SD is considerably lower for ResNet50 (0.0046), than for DenseNet121 (0.012), showing that ResNet50 provides more stable and consistent results across multiple runs. In terms of discriminative capability, ResNet50 again outperformed DenseNet121, with AUC of 0.96 *versus* 0.89, demonstrating superior ability to separate patients with CAD from health controls. The estimated PR–AUC values further confirm this trend, as ResNet50 obtained 0.8919, compared to 0.8121 for DenseNet121, which is particularly relevant for handling the class imbalance present in medical datasets. Overall, ResNet50 showed both higher accuracy and greater robustness than DenseNet121 on this dataset.

The Shapiro–Wilk test was used to assess whether the ACA values followed a normal distribution. The results shown in Table 2 indicate that neither DenseNet121 (*p*-level: <0.001), nor ResNet50 (*p*-level: 0.033) satisfied the normality assumption at a significance level of α=0.05. Furthermore, the Brown–Forsythe test was conducted to evaluate the homogeneity of variances. The outcome (p-level: 0.02) demonstrated a statistically significant difference between variances, indicating heteroscedasticity between the two sets of results. Consequently, a non-parametric statistical test such as the Mann–Whitney U-test is more appropriate for comparing the two models.

Table 2 – ***Performance metrics for DenseNet121 and ResNet50***

Model	Shapiro–Wilk test		Brown–Forsythe test	
	Shapiro–Wilk W	p-level	Brown–Forsythe (1, df)	p-level
DenseNet121	0.749	0.000	5.55	0.02
ResNet50	0.949	0.033		

After applying the Mann–Whitney U-test to compare the ACA of the two results, we observed a statistically significance difference between the two models (p-level: <0.0001). This confirms that the superior accuracy achieved by ResNet50 is not due to random variation but reflects a genuine improvement in classification performance over DenseNet121. Consequently, we reject the null hypothesis of equal means and conclude that ResNet50 demonstrates significantly better predictive capability on this dataset.

Figures 2–4 show a comparative analysis of the ACA distributions. The boxplot analysis shows that ResNet50 achieves a higher median accuracy with a narrower dispersion, indicating a more stable learning behavior across folds. In contrast, DenseNet121 exhibits a wider spread and several low performing outliers, suggesting reduced robustness.

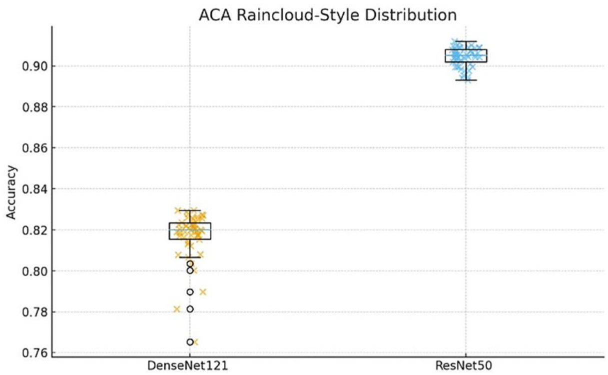

Figure 2 – ***Box-and-whiskers distribution of classification accuracy for DenseNet121 and ResNet50. ACA: Average classification accuracy.***

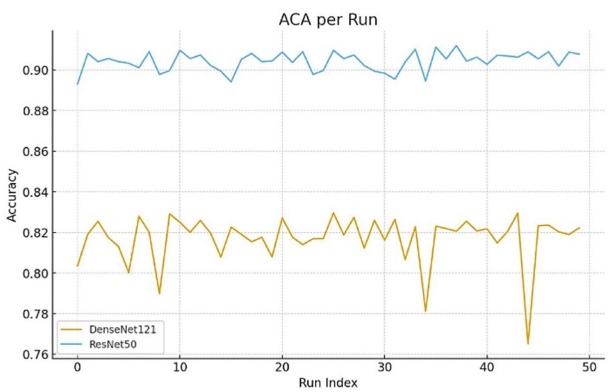

Figure 3 – ***Accuracy across 50 independent runs for DenseNet121 and ResNet50: ACA per run.***

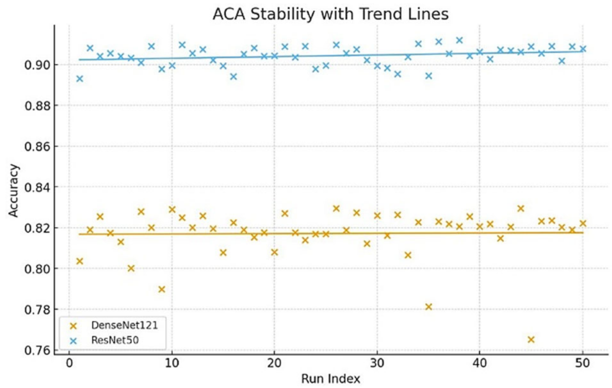

Figure 4 – ***Accuracy across 50 independent runs for DenseNet121 and ResNet50: ACA stability with trend lines.***

Discussions

Our study provides a rigorous evaluation of CAD detection from multiparametric CMR imaging using two CNNs DenseNet121 and ResNet50. Under strict patient-level separation, ResNet50 achieved an average accuracy of 90.43%, a performance that is lower than the near-perfect accuracies (99%) which have been occasionally reported in previous publications using the same dataset [15, 16]. Such discrepancies, however, must be interpreted with care considering methodological considerations. Multiple studies show that exceptionally high accuracies in medical imaging are frequently associated with data leakage systematically inflates predictive performances in data mining workflows [25]. Kapoor & Narayanan [26] documented that leakage affects hundreds of published papers what regard machine-learning, contributing to widespread reproducibility issues. Rosenblatt *et al.* [27] further showed in neuroimaging that subject-level non-independence can increase apparent accuracy compared to separated evaluations. Tello *et al.* [28] demonstrated that conventional random cross-validation can overestimate performance by over 10 percentage points relative to subject-wise tests. Apicella *et al.* [29] reported same effects when MRI data from the same individual appear simultaneously in training and testing sets. Taken altogether, these studies indicate that accuracies that are close to 99% on complex diagnostic task are more consistent with procedural artefacts than with genuine model discrimination. Consequently, the 90.43% average accuracy reported here, obtained under rigorous patient-wise separation – is likely to reflect a more realistic estimate of clinically relevant performance.

Another key distinction between our study and the others relates to statistical rigor. Prior work on this dataset typically reports only descriptive metrics and provides a limited insight into distributional properties, variance structure, or the significance of the observed differences. In contrast, the present study incorporates a complete statistical validation pipeline: the Shapiro–Wilk test was used to evaluate normality, the Brown–Forsythe test to examine homogeneity of variances, and the Mann–Whitney U-test for inference when normality assumptions were not met. Statistical validation offers a more transparent interpretation of performance variation and supports statistically grounded comparisons across model configurations.

Differences between studies appear regarding the preprocessing choices also. While some authors rely on aggressive filtering, manual curation, or extensive augmentation to optimize performance, our approach intentionally preserves the heterogeneity of the dataset, thereby better approximating real-world clinical conditions. Although this limits maximum achievable accuracy, the resulting evaluation in more conservative, robust, and generalizable.

These considerations demonstrate that methodological rigor, in data partitioning and statistical validation, is essential for obtaining reliable performance estimates in medical imaging. The present study shows that when patient-level independence is strictly enforced and appropriate inferential tests are applied, performance stabilizes at realistic levels, even if this leads to lower headline accuracy compared with previous reports. Rather than aiming for nominal maxima, our findings emphasize the importance of robustness, transparency, and generalizability as the key prerequisites for clinically meaningful deployment of DL models in CMR imaging analysis.

Conclusions

This study provides a statistically robust and rigorous evaluation of two DL architectures for CAD detection from multiparametric CMR imaging. Using strict patient-level segregation and comprehensive inferential testing, the ResNet50 network reached an average accuracy of 90.43%, AUC of 0.962, and PR–AUC of 0.8919, substantially outperforming DenseNet121 (average accuracy 81.72%, AUC 0.898). The statistical tests, Shapiro–Wilk and Brown–Forsythe, further confirmed the statistically significant differences between the two models, validating the robustness of the observed performance gap. The study highlights that realistic performance estimates arrive only when patient-level independence is enforced and statistical assumptions are rigorously tested, contrasting with earlier publications reporting near-perfect accuracies that are inconsistent with established evidence on data leakage and overestimation. Overall, the results underscore that methodological transparency, statistical robustness and valid data preprocessing are essential prerequisites for trustworthy and clinically translatable AI solutions in CMR imaging.

Conflict of interests

The authors declare no conflict of interests.

Funding

This research received no external funding.

References

[1] Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR, Delling FN, Djousse L, Elkind MSV, Ferguson JF, Fornage M, Jordan LC, Khan SS, Kissela BM, Knutson KL, Kwan TW, Lackland DT, Lewis TT, Lichtman JH, Longenecker CT, Loop MS, Lutsey PL, Martin SS, Matsushita K, Moran AE, Mussolino ME, O'Flaherty M, Pandey A, Perak AM, Rosamond WD, Roth GA, Sampson UKA, Satou GM, Schroeder EB, Shah SH, Spartano NL, Stokes A, Tirschwell DL, Tsao CW, Turakhia MP, VanWagner LB, Wilkins JT, Wong SS, Virani SS; American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics – 2019 Update: a Report from the American Heart Association. Circulation, 2019, 139(10):e56–e528. https://doi.org/10.1161/CIR.0000000000000659 PMID: 30700139

[2] Dey D, Slomka PJ, Leeson P, Comaniciu D, Shrestha S, Sengupta PP, Marwick TH. Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. J Am Coll Cardiol, 2019, 73(11):1317–1335. https://doi.org/10.1016/j.jacc.2018.12.054 PMID: 30898208 PMCID: PMC6474254

[3] Greenland P, Alpert JS, Beller GA, Benjamin EJ, Budoff MJ, Fayad ZA, Foster E, Hlatky MA, Hodgson JM, Kushner FG, Lauer MS, Shaw LJ, Smith SC Jr, Taylor AJ, Weintraub WS, Wenger NK, Jacobs AK, Smith SC Jr, Anderson JL, Albert N, Buller CE, Creager MA, Ettinger SM, Guyton RA, Halperin JL, Hochman JS, Kushner FG, Nishimura R, Ohman EM, Page RL, Stevenson WG, Tarkington LG, Yancy CW; American College of Cardiology Foundation; American Heart Association. 2010 ACCF/AHA Guideline for assessment of cardiovascular risk in asymptomatic adults: a Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol, 2010, 56(25):e50–e103. https://doi.org/10.1016/j.jacc.2010.09.001 PMID: 21144964

[4] Kramer CM, Barkhausen J, Bucciarelli-Ducci C, Flamm SD, Kim RJ, Nagel E. Standardized cardiovascular magnetic resonance imaging (CMR) protocols: 2020 update. J Cardiovasc Magn Reson, 2020, 22(1):17. https://doi.org/10.1186/s12968-020-00607-1 PMID: 32089132 PMCID: PMC7038611

[5] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal, 2017, 42:60–88. https://doi.org/10.1016/j.media.2017.07.005 PMID: 28778026

[6] Şerbănescu MS, Manea NC, Streba L, Belciug S, Pleşea IE, Pirici I, Bungărdean RM, Pleşea RM. Automated Gleason grading of prostate cancer using transfer learning from general-purpose deep-learning networks. Rom J Morphol Embryol, 2020, 61(1): 149–155. https://doi.org/10.47162/RJME.61.1.17 PMID: 32747906 PMCID: PMC7728132

[7] Şerbănescu MS, Oancea CN, Streba CT, Pleşea IE, Pirici D, Streba L, Pleşea RM. Agreement of two pre-trained deep-learning neural networks built with transfer learning with six pathologists on 6000 patches of prostate cancer from Gleason 2019 Challenge. Rom J Morphol Embryol, 2020, 61(2):513–519. https://doi.org/10.47162/RJME.61.2.21 PMID: 33544803 PMCID: PMC7864291

[8] Nica RE, Şerbănescu MS, Florescu LM, Camen GC, Streba CT, Gheonea IA. Deep learning: a promising method for histological class prediction of breast tumors in mammography. J Digit Imaging, 2021, 34(5):1190–1198. https://doi.org/10.1007/s10278-021-00508-4 PMID: 34505960 PMCID: PMC8554900

[9] Mitroi G, Pleşea RM, Pop OT, Ciovică DV, Şerbănescu MS, Alexandru DO, Stoiculescu A, Pleşea IE. Correlations between intratumoral interstitial fibrillary network and vascular network in Srigley patterns of prostate adenocarcinoma. Rom J Morphol Embryol, 2015, 56(4):1319–1328. PMID: 26743277

[10] Bungărdean RM, Şerbănescu MS, Colosi HA, Crişan M. High-frequency ultrasound: an essential non-invasive tool for the pre-therapeutic assessment of basal cell carcinoma. Rom J Morphol Embryol, 2021, 62(2):545–551. https://doi.org/10.47162/RJME.62.2.21 PMID: 35024743 PMCID: PMC8848273

[11] Şerbănescu MS, Bungărdean RM, Georgiu C, Crişan M. Nodular and micronodular basal cell carcinoma subtypes are different tumors based on their morphological architecture and their interaction with the surrounding stroma. Diagnostics (Basel), 2022, 12(7):1636. https://doi.org/10.3390/diagnostics12071636 PMID: 35885545 PMCID: PMC9323345

[12] Mămuleanu M, Urhuţ CM, Săndulescu LD, Kamal C, Pătraşcu AM, Ionescu AG, Şerbănescu MS, Streba CT. Deep learning algorithms in the automatic segmentation of liver lesions in ultrasound investigations. Life (Basel), 2022, 12(11):1877. https://doi.org/10.3390/life12111877 PMID: 36431012 PMCID: PMC9695234

[13] Zreik M, van Hamersvelt RW, Khalili N, Wolterink JM, Voskuil M, Viergever MA, Leiner T, Isgum I. Deep learning analysis of coronary arteries in cardiac CT angiography for detection of patients requiring invasive coronary angiography. IEEE Trans Med Imaging, 2020, 39(5):1545–1557. https://doi.org/10.1109/TMI.2019.2953054 PMID: 31725371

[14] Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv, 2017, arXiv:1708.08296. https://doi.org/10.48550/arXiv.1708.08296

[15] Khozeimeh F, Sharifrazi D, Izadi NH, Joloudari JH, Shoeibi A, Alizadehsani R, Tartibi M, Hussain S, Sani ZA, Khodatars M, Sadeghi D, Khosravi A, Nahavandi S, Tan RS, Acharya UR, Islam SMS. RF–CNN–F: random forest with convolutional neural network features for coronary artery disease diagnosis based on cardiac magnetic resonance. Sci Rep, 2022, 12(1): 11178. https://doi.org/10.1038/s41598-022-15374-5 PMID: 35778476 PMCID: PMC9249743

[16] Joloudari JH, Saadatfar H, GhasemiGol M, Alizadehsani R, Sani ZA, Hasanzadeh F, Hassannataj E, Sharifrazi D, Mansor Z. FCM–DNN: diagnosing coronary artery disease by deep accuracy fuzzy C-means clustering model. Math Biosci Eng, 2022, 19(4):3609–3635. https://doi.org/10.3934/mbe.2022167 PMID: 35341267

[17] Sharifrazi D. CAD cardiac MRI dataset: an image dataset to detect CAD disease, very suitable for deep learning methods. Kaggle, 2021. https://www.kaggle.com/datasets/danialsharifrazi/cad-cardiac-mri-dataset

[18] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. arXiv, 2018, arXiv:1608.06993. https://doi.org/10.48550/arXiv.1608.06993

[19] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang L, Wang G, Cai J, Chen T. Recent advances in convolutional neural networks. arXiv, 2017, arXiv:1512.07108. https://doi.org/10.48550/arXiv.1512.07108

[20] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv, 2016, arXiv:1602.07360. https://doi.org/10.48550/arXiv.1602.07360

[21] Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev, 2020, 53(8):5455–5516. https://doi.org/10.1007/s10462-020-09825-6

[22] Rajpurkar R, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Lungren MP, Ng AY. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv, 2017, arXiv:1711.05225. https://doi.org/10.48550/arXiv.1711.05225

[23] Zhou T, Ruan S, Canu S. A review: deep learning for medical image segmentation using multi-modality fusion. Array, 2019, 3–4:100004. https://doi.org/10.1016/j.array.2019.100004

[24] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: ***. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016, 770–778. https://doi.org/10.1109/CVPR.2016.90

[25] Kaufman S, Rosset S, Perlich C. Leakage in data mining: formulation, detection, and avoidance. KDD ’11: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA (August 21–24, 2011), ACM Trans Knowl Discov Data, 2011, 6(4):556–563. https://doi.org/10.1145/2020408.2020496

[26] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y), 2023, 4(9): 100804. https://doi.org/10.1016/j.patter.2023.100804 PMID: 37720327 PMCID: PMC10499856

[27] Rosenblatt M, Tejavibulya L, Jiang R, Noble S, Scheinost D. Data leakage inflates prediction performance in connectome-based machine learning models. Nat Commun, 2024, 15(1): 1829. https://doi.org/10.1038/s41467-024-46150-w PMID: 38418819 PMCID: PMC10901797

[28] Tello A, Degeler V, Lazovik A. Too good to be true: accuracy overestimation in (re)current practices for Human Activity Recognition. In: ***. 2024 EEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), Biarritz, France, 11–15 March 2024, 511–517. https://doi.org/10.1109/PerComWorkshops59983.2024.10503465

[29] Apicella A, Isgrò F, Prevete R. *Don’t push the button!* Exploring data leakage risks in machine learning and transfer learning. Artif Intell Rev, 2025, 58(11):339. https://doi.org/10.1007/s10462-025-11326-3

Corresponding author
Daniel-Robert Stănescu, MD, PhD Student, Doctoral School, University of Medicine and Pharmacy of Craiova, 2 Petru Rareş Street, 200349 Craiova, Dolj County, Romania; Phone +40767–438 591, e-mail: daniel.stanescu@umfcv.ro

Received: December 3, 2025

Accepted: December 29, 2025