scientific reports

Check for updates

OPEN

A hybrid segmentation and classification CAD framework for automated myocardial infarction prediction from MRI images

Mugahed A. Al-antari[1,6]✉, Riyadh M. Al-Tam[2], Aymen M. Al-Hejri[2], Zaid Al-Huda[3], Soojeong Lee[4], Özal Yıldırım[5] & Yeong Hyeon Gu[1,6]✉

Early diagnosis of myocardial infarction (MI) is critical for preserving cardiac function and improving patient outcomes through timely intervention. This study proposes an annovaitive computer-aided diagnosis (CAD) system for the simultaneous segmentation and classification of MI using MRI images. The system is evaluated under two primary approaches: a serial approach, where segmentation is first applied to extract image patches for subsequent classification, and a parallel approach, where segmentation and classification are performed concurrently using full MRI images. The multi-class segmentation model identifies four key heart regions: left ventricular cavity (LV), normal myocardium (Myo), myocardial infarction (MI), and persistent microvascular obstruction (MVO). The classification stage employs three AI-based strategies: a single deep learning model, feature-based fusion of multiple AI models, and a hybrid ensemble model incorporating the Vision Transformer (ViT). Both segmentation and classification models are trained and validated on the EMIDEC MRI dataset using five-fold cross-validation. The adopted ResU-Net achieves high F1-scores for segmentation: 91.12% (LV), 88.39% (Myo), 80.08% (MI), and 68.01% (MVO). For classification, the hybrid CNN-ViT model in the parallel approach demonstrates superior performance, achieving 98.15% accuracy and a 98.63% F1-score. These findings highlight the potential of the proposed CAD system for real-world clinical applications, offering a robust tool to assist healthcare professionals in accurate MI diagnosis, improved treatment planning, and enhanced patient care.

Keywords Myocardial infarction, Heart diseases, Ensemble fusion learning, Computer-Aided diagnosis, Segmentation, Classification, Visual explainable saliency maps

Cardiovascular disease (CVD) is a leading cause of global mortality, accounting for approximately 17.3 million deaths annually, with projections estimating an increase to 23.6 million by 2030[1]. Among CVD-related deaths, myocardial infarction (MI) and strokes contribute to 85% of fatalities[2]. MI occurs due to blockage in coronary arteries, leading to irreversible myocardial damage if not treated promptly. Early detection and precise assessment of myocardial functionality are critical for improving patient outcomes and guiding clinical interventions. Cardiac magnetic resonance imaging (MRI) plays a key role in post-MI evaluations, providing high-resolution insights into myocardial thickness, ejection fraction (EF), and infarcted tissue regions[3]. Delayed enhancement-MRI (DE-MRI) is particularly effective in visualizing infarcted myocardium, aiding physicians in diagnosis and treatment planning. A critical aspect of cardiac assessment is the identification of key myocardial structures, including myocardial infarction (MI), left ventricular cavity (LV), normal myocardium (Myo), and persistent microvascular obstruction (MVO). The detection of MVO is particularly significant as it is associated with poor prognosis, adverse left ventricular remodeling, and an increased risk of major adverse cardiovascular events (MACE). The accurate segmentation and classification of these regions are essential for risk stratification,

[1]Department of Artificial Intelligence and Data Science, College of AI Convergence, Daeyang AI Center, Sejong University, Seoul 05006, Korea. [2]School of Computational Sciences, Swami Ramanand Teerth Marathwada University, Nanded 431606, Maharashtra, India. [3]Stirling College, Chengdu University, Chengdu 610106, P. R. China. [4]Department of Computer Engineering, College of AI Convergence, Daeyang AI Center, Sejong University, Seoul 05006, Korea. [5]Department of Software Engineering, Technology Faculty, Fırat University, Elazığ, Turkey. [6]These authors contributed equally: Mugahed A. Al-antari and Yeong Hyeon Gu. ✉email: en.mualshz@sejong.ac.kr; yhgu@sejong.ac.kr

therapeutic decision-making, and personalized treatment strategies. Physicians investigate abnormal regions, as shown in Fig. 1, but these areas are relatively small compared to the entire image, posing challenges for both segmentation and classification tasks. Moreover, manual segmentation of myocardial regions is time-intensive and prone to variability, necessitating automated solutions for efficiency and accuracy. Medical image analysis has predominantly relied on Convolutional Neural Networks (CNNs), which excel at extracting high-resolution spatial features[4]. Models like U-Net[5] and its varients have been applied to segment various structures in medical imaging, including cancerous regions in breast, brain, and skin cancers. However, the intrinsic locality of convolutional operations limits the effectiveness of CNNs when analyzing structures with large inter-patient variations, such as differences in texture, shape, and size[6]. This limitation can hinder model accuracy and reliability, particularly in complex tasks like myocardial infarction detection. To overcome these challenges, recent research has integrated Vision Transformers (ViTs) with CNNs, combining the spatial feature extraction strengths of CNNs with the global context modeling capabilities of Transformers[7]. While Swin Transformers and other hybrid models have also been explored in medical image analysis, ViT offers a unique advantage in its ability to model global dependencies in the image, which is particularly important for tasks like MI detection, where spatial relationships across large regions of the myocardium need to be considered[8]. The ability of ViT to capture long-range dependencies enables it to effectively capture structural relationships and anomalies such as those seen in infarcted myocardium, making it well-suited for MI classification. Hybrid architectures such as ViT-based ResNet50 and ensemble-based ViT have shown promise in tasks like breast cancer detection[9–11] and cardiovascular disease recognition. In the realm of cardiac imaging, numerous deep learning techniques have demonstrated potential for automating cardiac MRI analysis. However, many existing studies struggle with accurately segmenting and classifying myocardial infarction (MI) and microvascular obstruction (MVO) due to the small and complex nature of the target regions[12,13]. Traditional single-model approaches[5] often face challenges

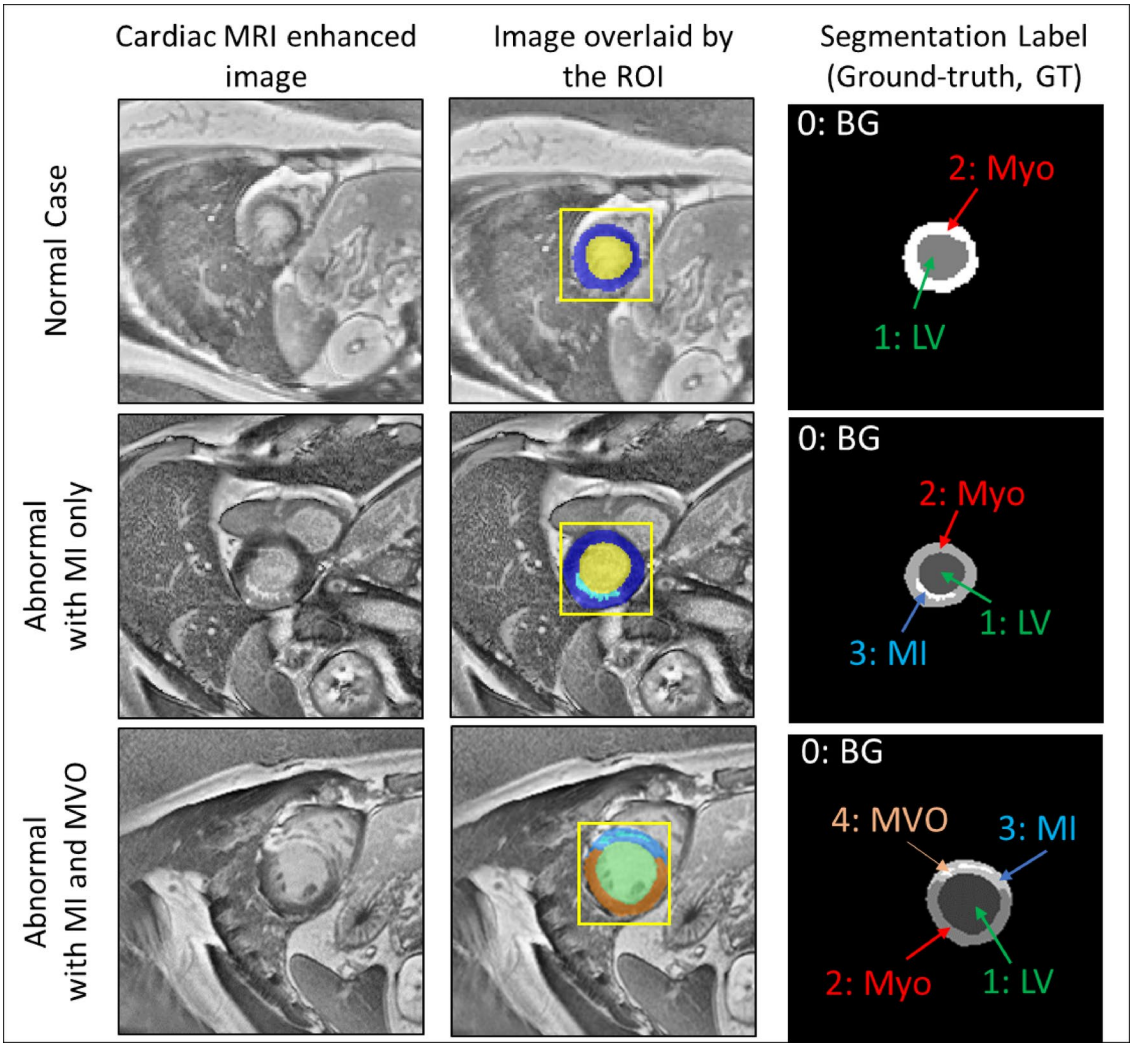

Fig. 1. Samples of heart MRI images with normal and abnormal cases. The normal image has only three segmentation labels which are background (BG), myocardium (Myo), and left ventricular (LV). The abnormal images in the second and third rows have abnormality of myocardial infarction (MI) and/or persistent microvascular obstruction (MVO).

in precise region boundary prediction and patient classification, while some CAD systems developed to segment the left or right ventricles alone[14–17], they often lack the robustness necessary for comprehensive MI detection. Moreover, a critical limitation of many existing systems is the lack of explainability and interpretability[18–20], which restricts their clinical applicability.To address these challenges, this study proposes a novel hybrid fully automatic computer-aided diagnosis (CAD) system that integrates both segmentation and classification stages. This hybrid system aims to enhance myocardial infarction detection by leveraging the complementary strengths of ResU-Net, CNNs, and ViT, through an end-to-end framework. A key innovation of this work is the use of a parallel approach, where segmentation and classification stages are executed simultaneously, enhancing both accuracy and computational efficiency. The major contributions of this research study are summarized as follows,

1. A novel hybrid fully automatic CAD system is proposed to predict cardiac Myocardial infarction (MI) in two consecutive segmentation and classification stages. The benefit of using the prior segmentation process is to extract the most suspicious ROIs enabling the deep learning classifiers to extract more powerful features. Based on the segmentation stage, we can define the various regions' boundaries that are important for heart attack prevention: Myocardial infarction (MI), left ventricular cavity (LV), normal myocardial (Myo), and persistent microvascular obstruction (MVO).
2. The classification stage is built based on the concept of hybridization of feature-based fuse learning and the power of the vision transformer concept. The ensemble or fusion learning is used to fuse the high-level deep features from multiple deep learning models or backbone networks, while the ViT is used to predict the patient-level class pathology.
3. The evaluation process is conducted in two separate approaches: Serial approach (1), and Parallel approach (2). For the first serial approach, two stages of segmentation and classification are consecutively built and verified. Thus, an area of interest (ROI) acting as an input to the classification stage is represented by cropping the entire segmented regions bounded by the outside bounds from the matching original image as shown in the second column of Fig. 1. In the second parallel approach, the prior segmentation process is conducted in parallel with the classification stage. Here, the whole input cardiac MR image is used in parallel scenarios for segmentation and classification stages simultaneously.
4. Both approaches employ three classification scenarios to construct the classification stage: (1) individual deep learning models, (2) feature-based concatenation or fusion of multiple AI models, and (3) a hybrid scenario that combines ensemble models with the Vision Transformer (ViT).
5. A validation and verification (V&V) ablation study is conducted to investigate the reliability of the proposed framework against two expert physicians using an unseen validation dataset.
6. Using Grad-CAM, the deep learning models generate explainable visual saliency maps (heat maps) that highlight the influential regions of the input images. These heat maps provide insights into the decision-making process of the models and enhance their interpretability. By understanding and trusting the model's predictions, users can easily identify potential errors or biases.

The rest of this paper is organized as follows. A review of recent literature study is presented in Sect. 2. Technical details of the proposed AI framework are presented in Sect. 3. The results of the experimental study are reported in Sect. 4. Section 5 discusses the most important findings. Finally, Sect. 6 concludes the important findings of this work.

Related works

Individual deep learning AI models

Deep learning has advanced medical image analysis, particularly in segmentation and detection for cardiovascular diagnosis using MRI. Many cardiac pathology segmentation methods still rely on traditional encoder-decoder networks[1], such as the widely used U-Net[2]. Deep learning models have proven to be effective in segmenting the myocardium (Myo) and left ventricle (LV) from late gadolinium enhancement (LGE) short-axis MRI images, making it a widely used technique for detecting pathologies like myocardial infarction (MI)[1]. Zabihollahy et al. presented a technique for segmenting the left ventricle (LV) and scar from late gadolinium enhancement (LGE) MRI images using a cascaded multi-planar U-Net[3]. Zakarya et al. introduced a fully convolutional neural network (FCN) for the segmentation of the left ventricle (LV) cavity and myocardium from late gadolinium enhancement (LGE) short-axis MRI images[21]. Brahim et al. presented a 3D network for the segmentation of myocardial infarction in late gadolinium enhancement (LGE) MRI[22]. Their approach involved an initial segmentation of the left ventricle (LV) and myocardium, followed by the fusion of a 3D U-Net architecture with a shape-prior-based framework to accurately segment the pathological tissues. Li et al. introduced a U-Net-inspired network, comprising two subnetworks: the multi-modal complementary information exploration network (MCIE-Net) and a lesion refinement network, to extract lesion features exclusively[23]. Chen et al. developed a CNN-based automatic model for myocardial infarction segmentation from short-axis delayed enhancement cardiac MRI (DE-MRI)[24]. Their approach involved two CNN networks, one for myocardium segmentation and another for MI segmentation. Heidenreich et al. presented a self-configuring CNN model with a U-Net architecture for the segmentation of the LV, myocardium, and infracted region in LGE MRI[25]. Moreover, several deep learning-based models have been proposed in the MICCAI 2020 challenge to segment MI using the EMIDEC dataset[2]. Zhang et al. introduced a cascaded convolutional neural network for the automatic segmentation of myocardial infarction from LGE cardiac MRI[26]. The network comprises a 2D U-Net for initial intra-slice segmentation and a 3D U-Net for extracting volumetric spatial information. Yang and Wang introduced a hybrid U-net network for simultaneous segmentation of the background, LV, left myocardium, MI, and no-reflow regions from LGE MRI[27]. The hybrid U-net enhances feature selection with squeeze-and-excitation in the encoder and selective kernels in

the decoder. Zhou et al. integrated attention for myocardium, MI, and no-reflow segmentation[28]. Huellebrand et al. compared a hybrid CNN and mixture model with two U-Net segmentation networks, one based on the EMIDEC challenge dataset and the other based on different training data[29]. Huellebrand et al. developed a deep learning framework for myocardial infarction (MI) detection. It involves LV segmentation, radiomics feature extraction, and classification based on previously extracted features and provided clinical information[30]. Shi et al. proposed a classification model based on 3D CNN and Random Forest for the classification of myocardial infarction on MRI images using clinical physiological data[31]. The two-stage model extracted spatial features from the images, which were then classified using RF based on encoded image features and physiological data. The results of RF with CNN outperformed the results of RF, CNN, and RF with CNN, which were 87%, 91%, and 95%, respectively. To detect and separate cardiac or anatomical structures in MRI scans, Poudel et al. introduced a recurrent complete convolutional network[18]. The method proposed by Isensee et al. utilizes a series of U-Net structures to segment the cardiac and combines segmentation and illness classification into a fully automated processing pipeline[32]. Saito et al. proposed a CNN model for classifying images of heart disease[19]. The model was trained twice, initially using a custom CNN model and then further trained using pre-trained networks. The CNN model achieved the highest accuracy rate of 96.17% and required the shortest training time of 61 min and 21 s. To automatically segregate myocardial infarction from delayed enhancement cardiac MRI, the authors in[20] proposed a cascaded convolutional neural network. They utilized a 2D U-Net to focus on intra-slice information for initial segmentation, followed by a 3D U-Net to leverage volumetric spatial information for a more refined segmentation. The model was tested on the dataset provided by the MICCAI EMIDEC challenge, and the results demonstrated average Dice scores of 0.8786, 0.7124, and 0.7851 for myocardial, infarction, and no-reflow respectively.In[33], Lourenco et al. proposed the use of deep-learning neural networks to automatically predict myocardial disease based on patient clinical information and DE-CMR. They suggested using DE-CMR to check for the presence and extent of late gadolinium enhancement (LGE). The model they suggested achieved an 85% accuracy in classification when submitted for classification. Furthermore, by including DE-CMR information, which includes metadata segmentation, the accuracy increased to over 95%. Brahim et al.[34] proposed ICPIU-Net segments LV myocardium, MI, and MVO tissues from LGE-MR images using cascaded subnets and topological constraints. It outperforms deep learning methods in the EMIDEC challenge, closely matching expert annotations.

Ensemble deep learning AI models

Ensemble learning is a machine learning approach that combines multiple base models or weak learners to enhance prediction accuracy and decision-making[35,36]. Ensemble learning improves performance by combining multiple models, reducing individual limitations, and enhancing prediction accuracy. In[37], Elmannai et al. proposed a stacking ensemble model based on CNNs. Three pre-trained models were used on two MI datasets for ECG heartbeat, namely MIT-BIH and PTB. The ensemble model collected the output probabilities of the CNN models, which were then classified by four machine learning algorithms. The random forest classifier achieved the highest accuracy for both datasets, with 99.8% for MIT-BIH and 99.7% for PTB. In[9], Al-Hejri et al. proposed a novel framework, called ETECADx, for breast cancer identification based on the feature ensemble concept. Their approach included the latest advancements in transfer learning, ensemble learning, and the transformer encoder model. The ensemble learning model performed better than the individual AI models with an accuracy of 97.16%. Additionally, the suggested ETECADx model surpassed the ensemble model with an accuracy of 99.58%. For COVID-19 prediction, Ukwuoma et al.[35,36] presented two consecutive research studies based on the ensemble and ViT concepts achieving promising evaluation results. This encourages us to apply such an ensemble technique for CVD classification. Al-Haidri et al.[38] presented a A deep learning framework using Mask-RCNN automatically segments lumbar vertebrae from MR images, detects six anatomical landmarks, and calculates vertebral heights. It accurately assesses vertebral deformities with high precision, significantly streamlining the manual measurement process. Moravvej et al.[39] presented an automatic myocarditis classification model using deep reinforcement learning and population-based algorithms to address imbalanced CMR data. Wang et al.[40] used radiomic analysis of fused multi-parametric CMRI sequences for myocardial infarction detection. The T1 + sBTFE-weighted fusion method achieved the AUC of 0.97%.

Transformer-based AI models

The Vision Transformer (ViT) excels in vision tasks like segmentation and classification. While used in AI research, its role in myocardial infarction classification remains underexplored. Explainable AI (XAI) enhances medical imaging by improving prediction accuracy[35,36]. However, their lack of interpretability can hinder their adoption in critical healthcare scenarios. Fortunately, XAI methods such as saliency maps, gradient-based techniques, and attention mechanisms can help visualize and explain the features or regions in an image that contribute the most to the model's decision[41–43]. This interpretability allows clinicians to understand the reasoning behind the model's predictions, thus increasing trust and facilitating decision-making. Moreover, XAI techniques can assist in the detection and localization of abnormalities in medical images. By highlighting the regions of interest or indicating areas that influence the model's decision, XAI methods help radiologists and clinicians focus their attention on potential abnormalities or subtle findings that may have been overlooked. Ultimately, this can enhance diagnostic accuracy and improve patient outcomes. Additionally, XAI techniques offer valuable tools for enhancing interpretability, detecting abnormalities, identifying biases, assessing quality, and educating professionals in the context of medical imaging. By providing transparent and interpretable insights, XAI can greatly improve the trust, effectiveness, and responsible use of AI models in medical diagnosis and decision-making.

Material and methods

The proposed fully automated hybrid CAD system, illustrated in Fig. 2, comprises myocardial infarction (MI) MRI dataset collection, data pre-processing, AI-based prediction, and final outcome generation. The AI model follows three key stages: auto-segmentation, deep feature extraction, and classification, supporting both serial and parallel approaches. In the serial approach, segmentation is performed first, extracting the ROI as input for feature extraction and classification, whereas in the parallel approach, segmentation and classification are executed simultaneously using pre-processed full MRI images. Segmentation plays a crucial role in delineating cardiac structures, including myocardial infarction (MI), left ventricular cavity (LV), normal myocardium (Myo), and microvascular obstruction (MVO), as shown in Fig. 1, enabling physicians to assess abnormal regions such as MI and MVO. The classification stage determines whether the heart condition is normal or abnormal, with abnormalities identified in scans containing MI and/or MVO. As presented in Fig. 3, the CAD system is designed following a structured end-to-end roadmap to provide intelligent MI segmentation, classification, and visually interpretable saliency maps. The comprehensive outputs include segmentation boundaries of cardiac regions, patient-level classification (normal or abnormal), and heat maps highlighting key areas influencing AI model decisions.

MRI dataset description

For this study, the 2020 EMIDEC MICCAI challenge benchmark dataset is used including MRI scans for 150 patients[2]. The images were scanned using Siemens 1.5T and 3T MRI scanners with 5–10 short-axis slices that cover the whole left ventricular myocardium from the base to the apex. This dataset was selected carefully due to its comprehensive segmentation boundaries for four heart regions (i.e., MI, LV, Myo, and MVO) and its expert-defined pathological classifications (normal vs. abnormal). This allows for an in-depth investigation using both segmentation and classification stages across the serial and parallel approaches. The training set includes 67 pathological cases and 33 normal cases, while the testing set includes 33 pathological cases and 17 normal cases. Each case has various short-axis 5 to 10 slices (MRI images) for the LV area from base to apex. Table 1 shows the data distribution in terms of patients and frames (images) levels over training, validation, and testing sets. Figure 1 shows examples of MRI images with their contour delineations of four cardiac regions: MI, LV, Myo, and MVO. As mentioned in the EMIDIC challenge event, the critical boundaries of all ROIs are created by biophysicists with over than 15 years of work experience in the medical MRI imaging domain. The patients who have MI and sometimes MVO are considered as abnormal cases, otherwise, the patients are classified as normal cases.

Data pre-processing

Contrast-limited adaptive histogram equalization (CLAHE) is applied to enhance and adjust the image contrast prior to any deep learning process[44]. Such pre-processing is important and assists the AI models in generating more powerful deep features. As we have applied the CLAHE technique for the former medical research works[9,10,44,45], it shows its capability for improving the contrast of medical images like chest X-rays,

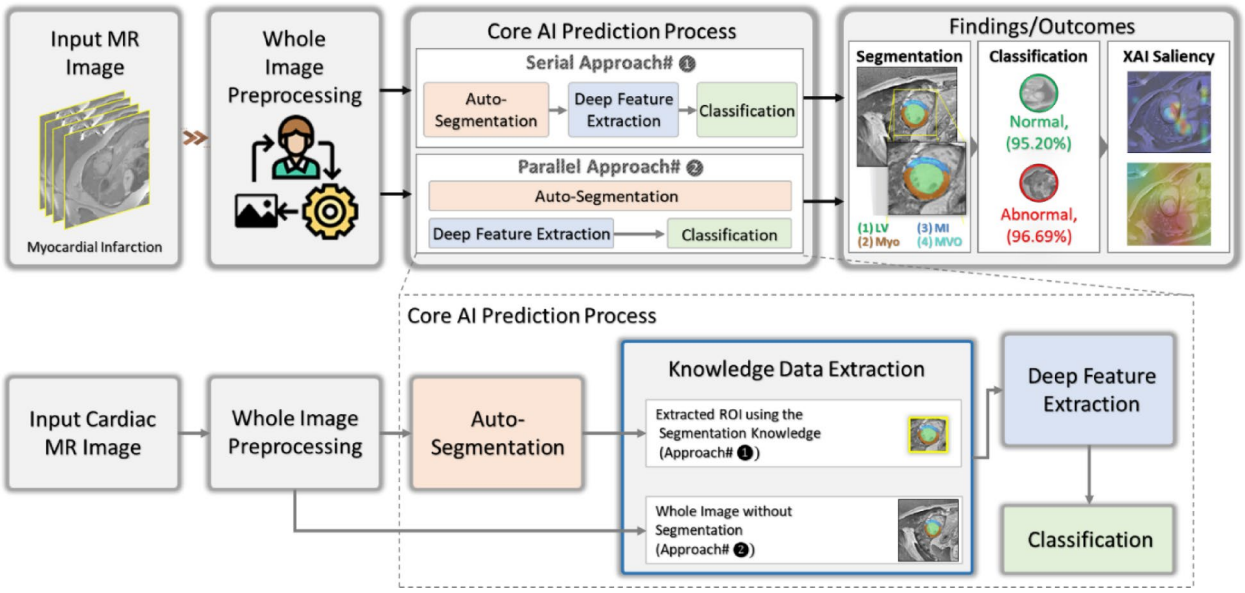

Fig. 2. The proposed hybrid fully automatic CAD system for myocardial infarction (MI) prediction. The CAD system is evaluated under serial and parallel approaches. The serial approach extracts heart region boundaries for ROI-based classification, while the parallel approach performs segmentation and classification simultaneously for visual inspection and prediction. The comprehensive approach provides segmentation boundaries, pathological classification, and visual heat maps, offering an end-to-end MI-based CAD prediction system.

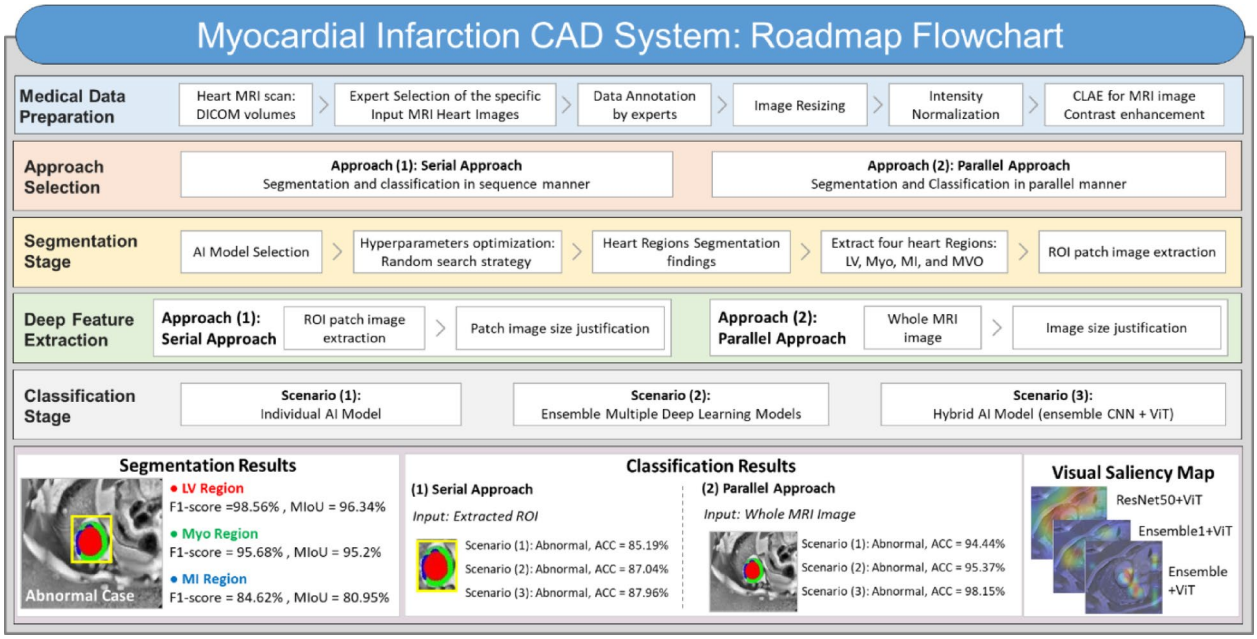

Fig. 3. Flowchart illustrating the proposed methodology for serial and parallel approaches. The serial approach extracts ROI from segmented regions for classification, while the parallel approach uses entire MRI images. Classification scenarios are executed separately for both approaches.

Pathological Condition	Training images (75%)	Validation images (10%)	Testing images (15%)	Total
Normal (50 Patients)	176	23	35	234
Abnormal (100 Patients)	355	46	73	474
Total	531	69	108	708

Table 1. EMIDIC MRI data distribution for normal and abnormal cases.

mammograms, MRI and CT scans. Instead of performing histogram equalization to the entire image at once, the images were separated into small blocks and histogram equalization was applied to each block individually and the equalized blocks were combined to generate the final enhanced image. This approach employs a clipping algorithm to limit the maximum contrast enhancement that may be applied to each block. Moreover, the image normalization is applied using bi-cubic interpolation to bring all pixel values in a unique pixel range of [0, 255][9,10,44,46]. Figure 4 shows an example of pre-processing MRI images for normal and abnormal cases corresponding with their histograms.

Data preparation for training, validation, and testing

As presented in Table 1, the MRI dataset is randomly split into 75% for training, 10% for validation, and 15% for testing. This splitting is conducted at the patient level to ensure that all frames or images that come from the same patients are solely used for training or testing purposes[44]. The training-validation sets are used to train or learn the AI models in both the segmentation and classification stages. Whereas, the unseen testing sets are used to assess the capability of the AI models for correctly predicting their pathological conditions. During the segmentation stage, the segmentation models are trained and tested to predict the different ventricular region boundaries including MI, LV, Myo, and MVO. The final patient-level prediction pathology condition (normal or abnormal) is predicted using the classification stage based on the extracted high-level deep features. Once the dataset is split, the training set is augmented to fulfill the requirements of the deep learning models including huge datasets for learning performance improvement. The benefits of data augmentation are to increase the training data size and diversity, improve the AI model generalization, mitigate overfitting, and improve the evaluation prediction performance[10,35,36,44,47]. For this purpose, the training images are rotated by 0º, 15º, 345º, 355º. Then, the rotated images are horizontally, vertically, and mirror flipped to generate at the end 6,372 augmented MRI cardiac images. All AI segmentation and classification models are trained and evaluated using the same data splitting and augmentation.

AI-based segmentation stage

To achieve the segmentation outcomes for the proposed framework, we adopt and use two AI-based segmentation models known as U-Net[5] and ResU-Net[48]. These models are selected due to their well-reputation in the AI-based

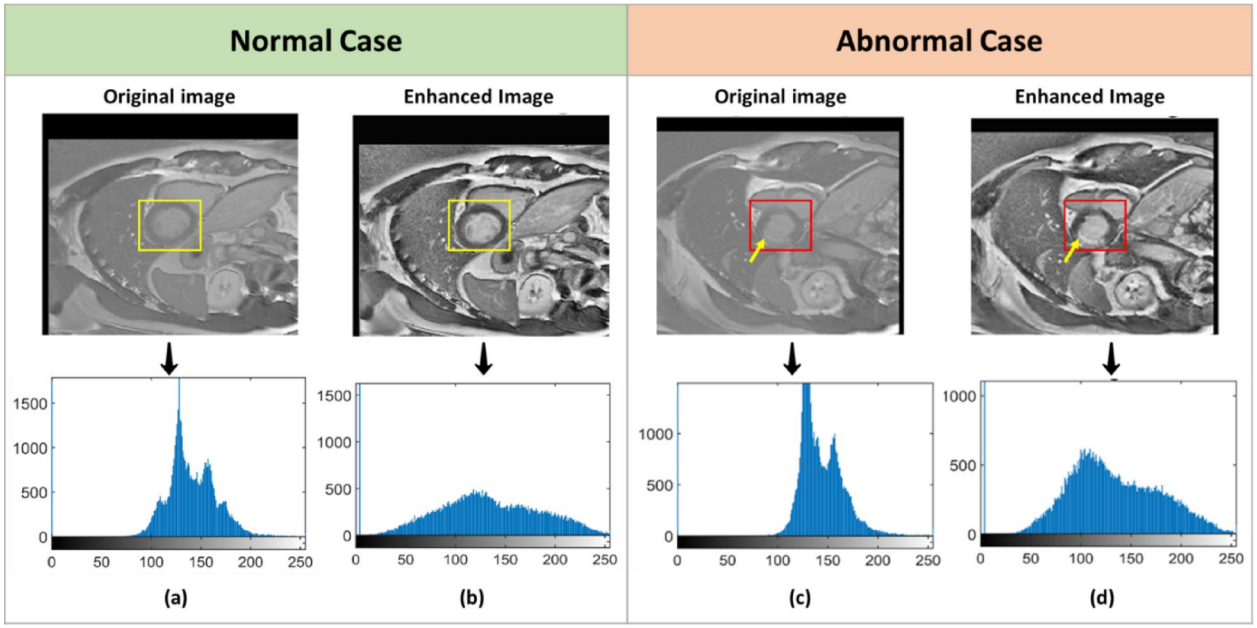

Fig. 4. Examples of pre-processed MRI cardiac images for normal and abnormal cases. (**a**) and (**c**) display the original images, while (**b**) and (**d**) illustrate the pre-processed images using the CLAHE technique. The histogram of each image is demonstrated to showcase the distribution achieved through histogram equalization, which enhances the quality of the image. The yellow and red boxes indicate the regions of interest (ROIs) for the LV regions, and the small arrows indicate the localization of abnormalities in the abnormal case.

medical research domain for segmenting various abnormal pathological regions such as breast cancer[44], brain tumors[49,50], and skin lesions[47]. Indeed, the purpose of various region segmentation of MI, MVO, LV, and Myo in the ventricular regions is to accurately identify and delineate the areas of damaged or ischemic (lack of blood supply) myocardial tissue in the ventricles of the heart[51]. Myocardial infarction, commonly known as a heart attack, occurs when the blood flow to a part of the heart muscle is blocked, leading to tissue damage or cell death. Accurate segmentation allows for quantifying the extent and severity of myocardial infarction. By measuring the volume or percentage of the infarcted tissue, clinicians can objectively evaluate the size and impact of the heart attack. This information helps in diagnosing the condition, determining the prognosis, and guiding the treatment plan[52]. Segmenting the infarcted regions helps cardiologists and healthcare professionals in planning appropriate interventions and therapies. By knowing the location and extent of the damaged tissue, they can identify the optimal treatment strategies, such as revascularization techniques or surgical interventions. Periodic segmentation of infarcted regions enables the monitoring of disease progression over time. By comparing the segmentations from different time points, clinicians can assess changes in the size or distribution of the infarcted tissue, which aids in evaluating the effectiveness of treatment and guiding further management decisions[3,51,52].

AI-based deep feature extraction and classification stages

For the stages of deep feature extraction and classification over both serial and parallel approaches, we designed three different scenarios which are (1) individual deep learning models, (2) feature-based concatenation or fusion of multiple AI models, and (3) a hybrid scenario that combines ensemble models with the Vision Transformer (ViT). Each scenario is launched twice for serial and parallel approaches. For the serial approach, the input data is represented by the extracted ROI image patch based on the segmentation knowledge data extraction. Whereas the entire cardiac MRI images are used as an input of all three scenarios in the parallel approach.

Individual deep learning structure

For the first scenario (individual deep learning models), we select and employ five AI-based backbone networks for deep feature extraction and classification. These models are InceptionResNetV2, VGG16, ResNet50, ResNet50-V2, and Xception. All these models are trained using the same training settings and dataset on the same execution and learning environment[9,10]. This is to achieve a fair comparison among their prediction performances. The backbone convolutional and pooling layers are used to extract the deep features, while the dense layers are used to predict the patient-level pathological condition to be normal or abnormal. Each AI model is trained and evaluated separately using the same split datasets. Figure 5 shows the conceptual diagram of the first scenario for feature extraction and classification where each AI model is trained and tested individually using the same training and testing settings for fair comparison. The proposed AI models are fine-tuned using consistent configuration settings to ensure a fair comparison of their performance, minimizing any potential biases. For the models based on deep learning architectures such as ResNet50, VGG16, Xception, InceptionResNetV2, and ResNet50-V2, which were pretrained on the ImageNet dataset, the classification layers are removed during this process. These models consist of three main blocks in their classification layers: a conventional dense layer with

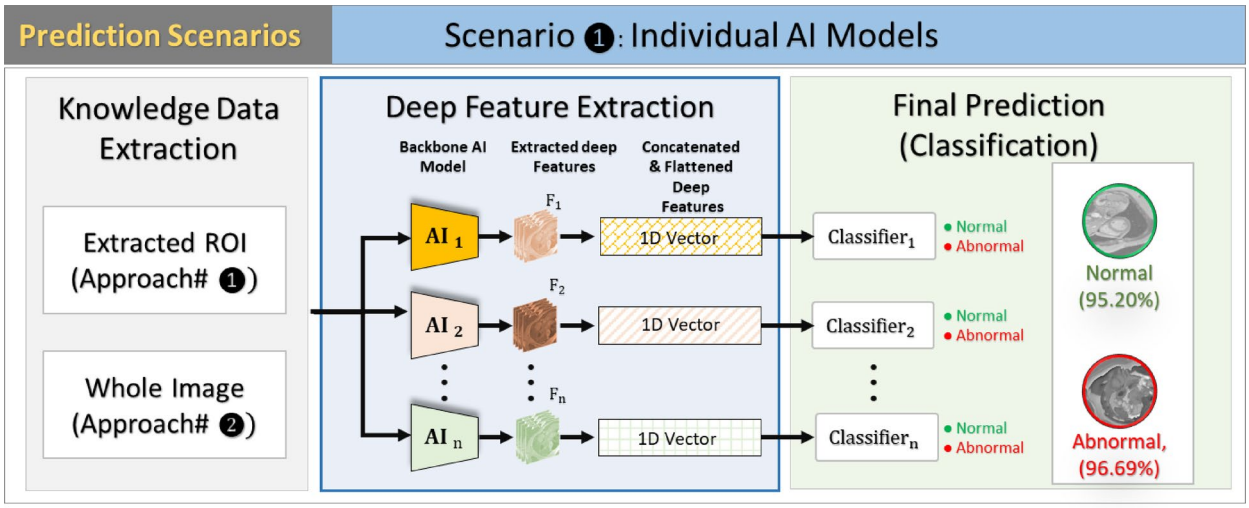

Fig. 5. Scenario 1: Deep feature extraction and classification using individual AI models. Five AI models are used to predict the final pathological condition for each case. The backbone convolutional and pooling layers are used to extract the deep features, while the dense layers on the top are used for the prediction purpose.

1,024 neurons, followed by batch normalization, and a dropout layer with a 50% dropout rate. Leveraging the transfer learning principle, the fine-tuning focuses on specific layers in each model. For example, in VGG16, only the layers from index 17 to the end are made trainable, while in ResNet50, layers from index 123 to the end are trainable. Similarly, for ResNet50-V2, training is focused on layers from index 672 to the end, and in DenseNet201, the trainable layers begin at index 481. In Xception, the trainable layers start from index 106, and in InceptionResNetV2, they begin from index 702.

Fusing multiple deep learning structures

For the second scenario, multiple AI models are used to extract the deep features individually, while the extracted features are fused or ensembled together to improve the learning process and then the prediction performance. Ensemble or fusion of deep features from multiple AI models can often lead to improved classification accuracy. The ensemble can leverage the strengths of individual models and compensate for their weaknesses, as each model captures different aspects or representations of the data. By combining their predictions or features, the ensemble achieves a more robust and accurate classification performance[35,36]. The fusion techniques also enhance the generalization capability of AI models, allowing them to effectively handle variations and complexities present in the data, thereby reducing the risk of overfitting[9,10]. Moreover, the ensemble approach strengthens the classification system against noise or outliers in the data. It fosters model diversity and exploration, resulting in a richer representation of the underlying patterns and improving the chances of capturing complex relationships within the data. Individual AI models may exhibit inherent biases due to their training data or architectures. Ensemble or fusion techniques help mitigate such biases by aggregating features or predictions from multiple models, providing a more balanced and unbiased view of the data. Figure 5 shows the technical diagram of the prediction scenario 2. To build the backbone network for scenario 2, we select the best AI models that investigated during the experimental study of the individual AI models (scenario 1). However, when combining the individual AI models in the ensemble scenario, the classification layers of each model are removed and flattened after training. A new classification layer is then added with 1,024 neurons, batch normalization, and a 50% dropout rate. This ensures that the ensemble benefits from a unified and optimized approach for classification.

Hybrid deep learning structure: ensemble AI structures with ViT

For the third feature extraction scenario, we combine the ensemble AI models, presented in Fig. 6, with the vision transformer (ViT) in a sequential manner to enhance the accuracy of our predictions. Such a hybridization structure is designed and built in our previous published research work[9,10,35,36]. It shows an impressive AI structure to predict various diseases such as breast cancer[9,10] and COVID-19[35,36]. Combining the ensemble deep features from various AI models with the ViT for classification purposes can improve representation learning, enhance model robustness, improve generalization, complementary feature extraction, and improve classification accuracy. Indeed, the ViTs have demonstrated strong performance in image classification tasks by leveraging self-attention mechanisms to capture global and long-range dependencies within images[35,36,53]. By combining the ensemble or fused deep features with ViT, the model can benefit from the complementary strengths of both approaches. The ensemble provides diverse representations learned by different models, while ViT captures fine-grained details and global context, leading to improved representation learning. The self-attention mechanism in ViT enables the model to focus on relevant regions and ignore noisy or irrelevant features, leading to improved resilience and more accurate classifications. For the proposed hybrid ensemble-

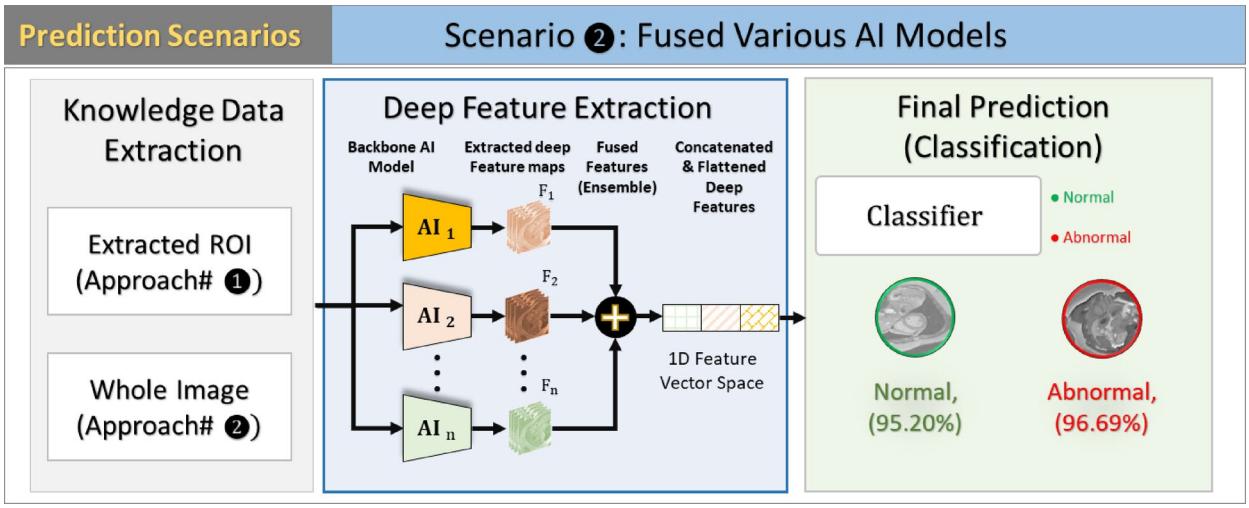

Fig. 6. Scenario 2: Deep feature extraction and classification using an ensemble of multiple AI models. In this scenario, various AI models are run concurrently to enhance the deep features, resulting in improved overall prediction accuracy. The top feature maps from each AI model are adjusted to a uniform size and then combined using a feature averaging aggregation strategy.

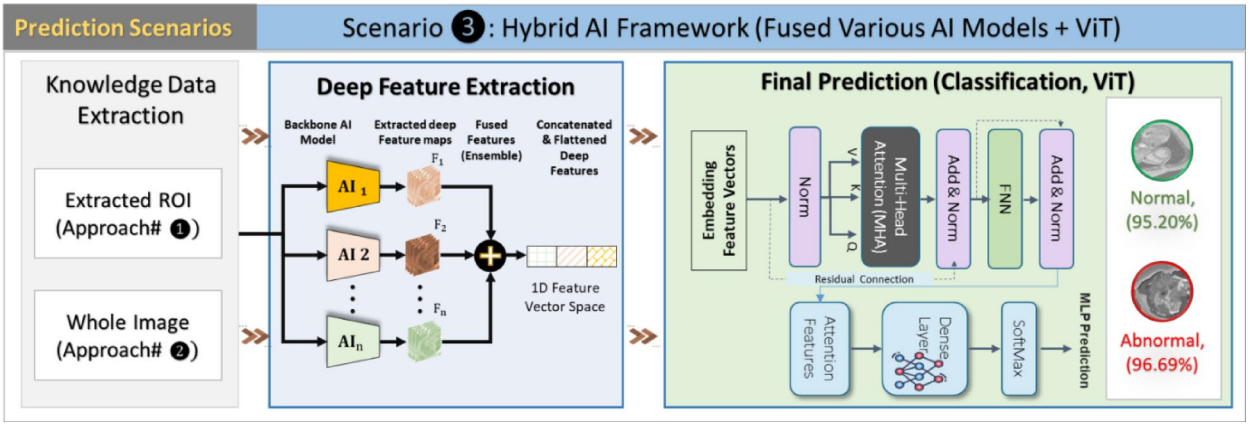

Fig. 7. Scenario 3: The proposed hybrid prediction framework integrates multiple AI networks, which are then combined with a Vision Transformer (ViT) as the classifier head to enhance prediction performance.

based Vision Transformer (ViT) model, the classification layers follow a similar configuration of Scenario 1. Ensemble methods extract features from multiple models, each with its own strengths and biases. ViT, on the other hand, focuses on self-attention and global context. Combining these features allows the model to leverage the strengths of both approaches, capturing both local and global information, leading to more informative and discriminative representations[53]. Such an ensemble with ViT combination allows the proposed framework to generalize well to unseen examples and handle variations, leading to better classification performance in real-world scenarios. The ensemble captures a diverse set of features, and ViT provides a powerful mechanism for learning high-level representations. This combination can effectively capture intricate patterns and relationships within the data, leading to better classification performance and higher accuracy. The hybrid concept between the ensemble AI models with the ViT is depicted in Fig. 7.

Experimental setting

The hyperparameters of AI models have been optimized and selected using the random-search strategy as in our previous work[54,55]. For the segmentation stage, the AI models are trained for 50 epochs and a mini-batch size of 4 to accommodate GPU memory limitations. All cardiac MRI images are resized to a standardized dimension of 128×128 pixels. To ensure fair parameters training and improve the prediction performance, a normalization process is applied to adjust the image pixel values into the range of [0, 255]. Gradient Descent (SGD) optimizer is used for training the models using the weighted cross-entropy loss function with a learning rate of 1e-4. Besides the early stopping strategy, the learning rate is reduced by 10 for each 5 epochs. Given that the pixel representation of abnormal MI and MVO tissues is lower than that of normal regions, we have employed weighted

cross-entropy loss functions with the following class weights: 0.10 (BG), 0.12 (LV), 0.21 (Myo), 0.30 (MI), and 0.32 (MVO). This approach increases the contribution of minority classes, thereby balancing the model and mitigating bias or fitting issues during training. For classification stage, the same training environmental settings are used to train and evaluate the AI models used in this study. A mini-batch size of 4, a training duration of 100 epochs with Adam optimizer, and weighted cross-entropy loss function are empirically considered to train and validate the proposed AI prediction CAD system. The proposed ensemble AI model incorporates several key hyperparameters and training protocols to optimize its performance. Specifically, the model utilizes a patch size of 16×16 pixels, allowing it to strike a balance between capturing fine-grained details and long-range dependencies across the image. To enhance its ability to capture complex relationships within the data, the model features 12 attention heads in the multi-head self-attention layers, which enable it to focus on both local and global contextual information effectively. Additionally, the model is structured with 12 transformer layers, providing the necessary depth to model intricate interactions and extract high-level features. To maintain a balance between expressiveness and computational efficiency, the embedding dimension is set to 768. The learning rate is initialized at 1e-4, optimized using the SGD optimizer, ensuring stable convergence.

Performance evaluation strategy

The most popular metrics for quantitatively evaluating the efficacy of the AI classification framework are accuracy, specificity, sensitivity or recall, F1-score, precision, and area under the ROC curve (AUC)[9,35,36,44]. The mathematical definitions of these metrics are presented here,

$$\text{Accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Specificity (SPE)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2}$$

$$\text{Sensitivity (SEN)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F1} - \text{Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{4}$$

$$\text{Precision (Pre)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{MIoU} = \frac{1}{\eta_{\text{class}}} \frac{\sum_{\text{i}} \text{n}_{\text{ii}}}{\text{t}_{\text{i}} + \sum_{\text{j}} (\text{n}_{\text{ji}} - \text{n}_{\text{ii}})} \tag{6}$$

The confusion matrix is used to derive the parameters of true positive (TP), false positive (FP), false negative (FN), and true negative (TN). These metrics evaluated the segmentation and classification performance at the pixel-wise level and image-wise level, respectively. Meanwhile, visual XAI heat maps are generated using Grad-CAM to investigate visually the powerful prediction of the AI models. The MIoU method is employed to determine the degree of similarity between the designated mask and the binary segmentation outcome where η_{class} indicates the total number of classes, n_{ji} indicates the number of pixels that are classified to be in class j while they are originally from class *i*, and $t_i = \sum_i n_{ii}$ presents the total number of pixels in class *i*.

Execution environment

An MSI PC with the following specifications is used to experiment: Intel Core I 7 11th Generation (11800 H), 32 GB RAM with 2 TB SSD Nvme and RTX 3080 (16 GB) Graphics Card. To conduct the entire experimental study, we use Python 3.10 running on Windows 11 along with the Keras and TensorFlow backend libraries.

Experimental results

Both segmentation and classification results are recorded as an average overall five cross-validation tests. The cross-validation is used to help in robust model evaluation, overfitting detection, hyperparameter tuning, and handling various data challenges, ultimately leading to more reliable and generalizable models.

AI-based segmentation results

Based on the inputs from the entire MRI images, U-Net and ResU-Net are adopted and used to segment the various pathological normal (LV and Myo) and abnormal (MI and MVO) regions. These two models are selected based on the recommendations of the random serach hyperparameter optimization strategy, as explained in our previous works[54,55].Using the same learning process on the environmental execution settings, both AI models are fine-tuned and evaluated. To avoid any overfitting, the AI segmentation models are trained on the same training environmental settings and training portion of the dataset. The training convergence rates in terms of F1-score, MIoU, and loss function are depicted in Fig. 8, showing normal training convergence rates without any overfitting. Using the testing set, we evaluate the segmentation performance, with the weighted assessment results presented in Table 2. All normal and abnormal testing metrics are calculated as average weighted evaluation results. While MI and MVO classes appear only in abnormal images, LV and Myo classes are present in both normal and abnormal cases. When extracting small fragments of MI and MVO anomalies, ResU-Net outperforms U-Net in segmentation accuracy. As shown in Table 2, ResU-Net achieves superior average segmentation results, with ACC (88.48%), SEN (85.24%), Pre (85.46%), F1-score (85.35%), and MIoU (84.23%).

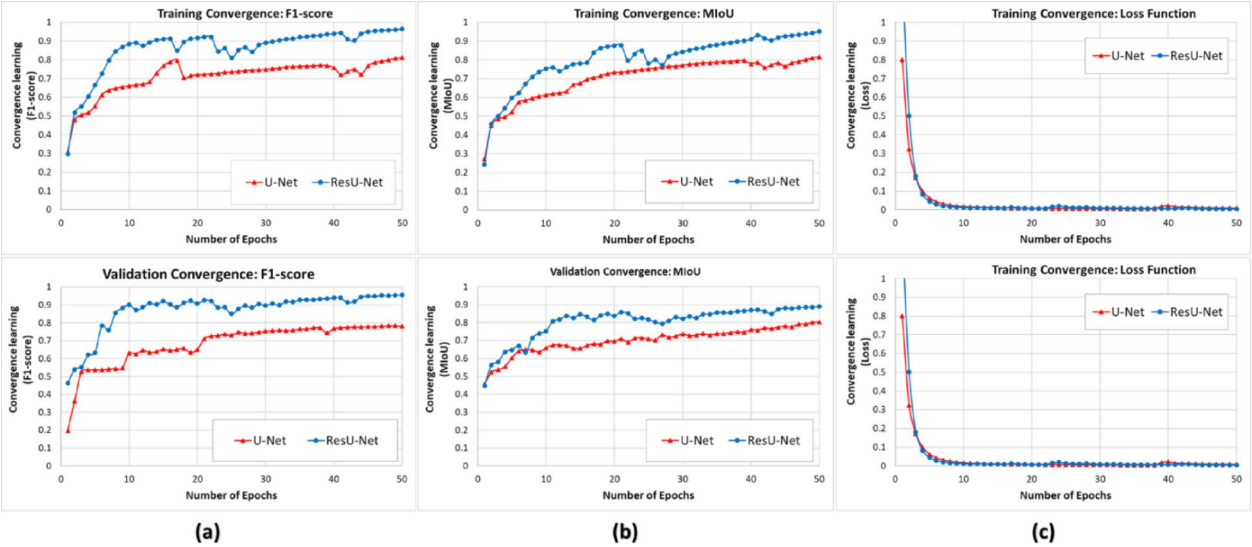

Fig. 8. The convergence rates of both U-Net and ResU-Net segmentation models during the training and validation process for Fold 3 are assessed based on (**a**) F1-score, (**b**) MIoU, and (**c**) loss functions. The convergence curves clearly demonstrate that the AI models began with lower values and progressively improved, ultimately reaching optimal learning. This indicates that the AI models do not encounter overfitting issues.

Tissue Type	Classes	U-Net					ResU-Net				
		ACC	SEN	Pre	F1-score	MIoU	ACC	SEN	Pre	F1-score	MIoU
Normal Regions	BG	98.74	97.25	98.02	97.63	95.05	99.78	99.04	99.23	99.13	98.59
	LV	89.23	86.45	88.03	87.23	85.05	95.89	89.52	92.73	91.12	91.86
	Myo	86.72	83.76	82.58	83.17	80.49	91.28	88.67	88.12	88.39	87.84
Abnormal Regions	MI	78.24	75.53	74.87	75.21	73.25	85.15	81.21	78.98	80.08	77.99
	MVO	63.02	61.92	63.47	62.69	58.63	70.31	67.78	68.25	68.01	64.89
Avg.	All	83.19	80.98	81.39	81.19	78.49	88.48	85.24	85.46	85.35	84.23

Table 2. Average evaluation segmentation performance (%) over five cross-validation per class for U-Net and ResU-Net using the testing MRI datasets.

Compared to U-Net, ResU-Net enhances segmentation performance by 5.29% (ACC), 4.26% (SEN), 4.07% (Pre), 4.16% (F1-score), and 5.74% (MIoU). ResU-Net al.so improves the segmentation of normal LV and Myo regions with gains of 5.61% in ACC, 4.56% in F1-score, and 7.08% in MIoU. Similarly, segmentation of abnormal MI and MVO regions is enhanced by 7.1% (ACC), 5.1% (F1-score), and 5.5% (MIoU). These improvements are critical, as they enhance the visual details of each pathology (LV, Myo, MI, and MVO), refining ROI extraction for the classification stage, particularly for approach (1).

To compare the U-Net and ResU-Net segmentation models, Fig. 9 showcases instances of normal and abnormal visual segmentation. It is evident that both segmentation methods effectively extract the small regions affected by cardiac attack, particularly the MVO regions. However, this poses a challenge for AI models due to the limited number of pixels in anomalous regions. Consequently, the F1-score for MI and MVO was lower, with respective scores of 75.21% and 62.69% (U-Net) and 80.08% and 68.01% (ResU-Net). That means the segmentation performance for the abnormal classes is recorded to lower than the normal regions. To preserve more abnormal pixels, the classification performance could be enhanced. As depicted in.

Figure 8, the ROI (Region of Interest) can be cropped by utilizing the outer boundaries of the segmented cardiac regions, which encompass the LV, Myo, MI, and MVO. The lower number of the MI and MVO segmented pixels negatively affects the classification performance recording lower prediction scores and vice versa.

AI-based classification results

As mentioned above, the classification results for both approaches are recorded over five cross-validation tests. Showing the training convergence for both approaches and all scenarios would be inconvenient for readers and make the article messy. However, Fig. 10 provides an example of the learning convergence of all classification models, particularly for Approach 2 with Fold 3. It has been proven that the convergence rates, in terms of accuracy and loss function, demonstrate that the learning process for the AI models is conducted without any overfitting in all three scenarios.

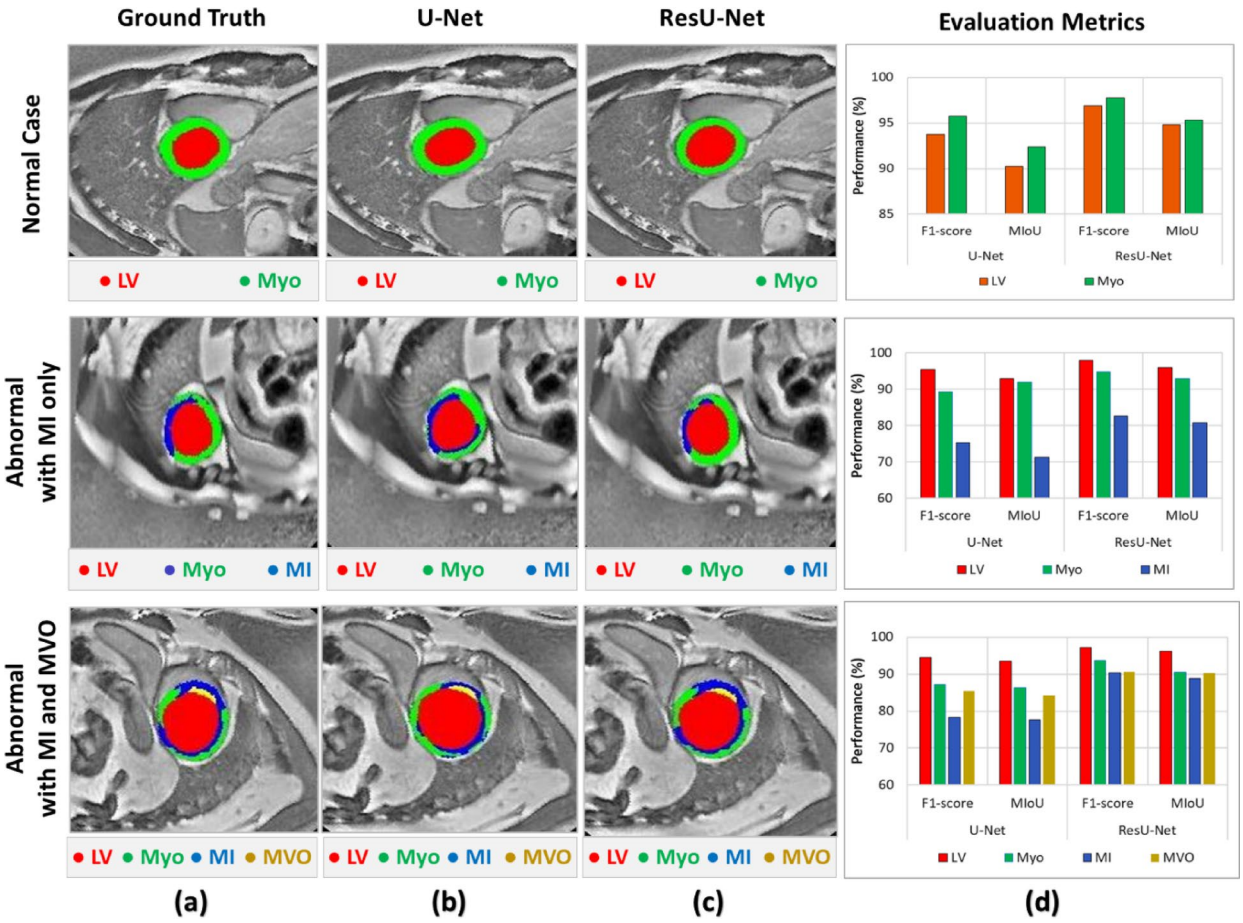

Fig. 9. Visual examples of the segmentation evaluation performance using three cases: normal (1st row), abnormal with MI (2nd row), and abnormal with MI and MVO (3rd row). (**a**) represents the original MRI image superimposed by the ground-truth (GT) of the segmented regions (LV, Myo, MI, and MVO). (**b**) and (**c**) represent the prediction outcomes by U-Net and ResU-Net, respectively. (**d**) depicts the segmentation evaluation metrics for each region in terms of F1-score and MIoU. The quantitative and qualitative evaluation results show the superiority of ResU-Net, providing better details of each segmented region.

Approach (1): serial approach

For the serial approach, the segmentation process is conducted first to extract the outer ROI that surrounds the whole area of LV and other sub-regions of Myo, MI, and MVO. The extracted whole image patch (ROI) is passed directly into the feature extraction and classification scenarios. In this section, the classification results of each scenario are presented separately.

Scenario 1: individual deep learning structure To perform the classification task, five state-of-the-art deep learning models are selected and trained separately using the same training and testing environment, settings, and dataset. Table 3 demonstrates the classification performance for each model.

The ResNet-50 produces the greatest results, with an F1-score of 85.45% and overall accuracy, sensitivity, and specificity of 85.19%, 86.26%, and 85.19% respectively. With an F1-score and AUC of 83.09%, the VGG16 records the second-best classification performance. The Xception AI model yields the best third performance, with an F1-score of 80.08% and an overall accuracy of 79.63%. With the help of InceptionResNetV2 and ResNet50-V2, the fourth and fifth performances are produced, respectively.

Scenario 2: fusing multiple deep learning structures Based on the individual evaluation results displayed in Table 3, we select the top three models to develop the suggested fused deep learning AI models. The following are two ensemble learning models made up of two and three deep learning structures. The first ensemble learning model (Ensemble 1) includes the top two AI models, ResNet50 and VGG16. The second ensemble AI model (Ensemble 2) includes ResNet50, VGG16, and Xception. The same execution environmental settings are used to separately train and evaluate each ensemble learning model. The classification outcomes for both ensemble learning models are shown in Table 4.

Both ensemble models produced predictions with similar outcomes recording an overall accuracy of 87.04%. When comparing SEN and SPE, there is a tiny difference that becomes apparent since Ensemble 1 performs better in terms of predicting normal cases, while the second model performs better in terms of predicting

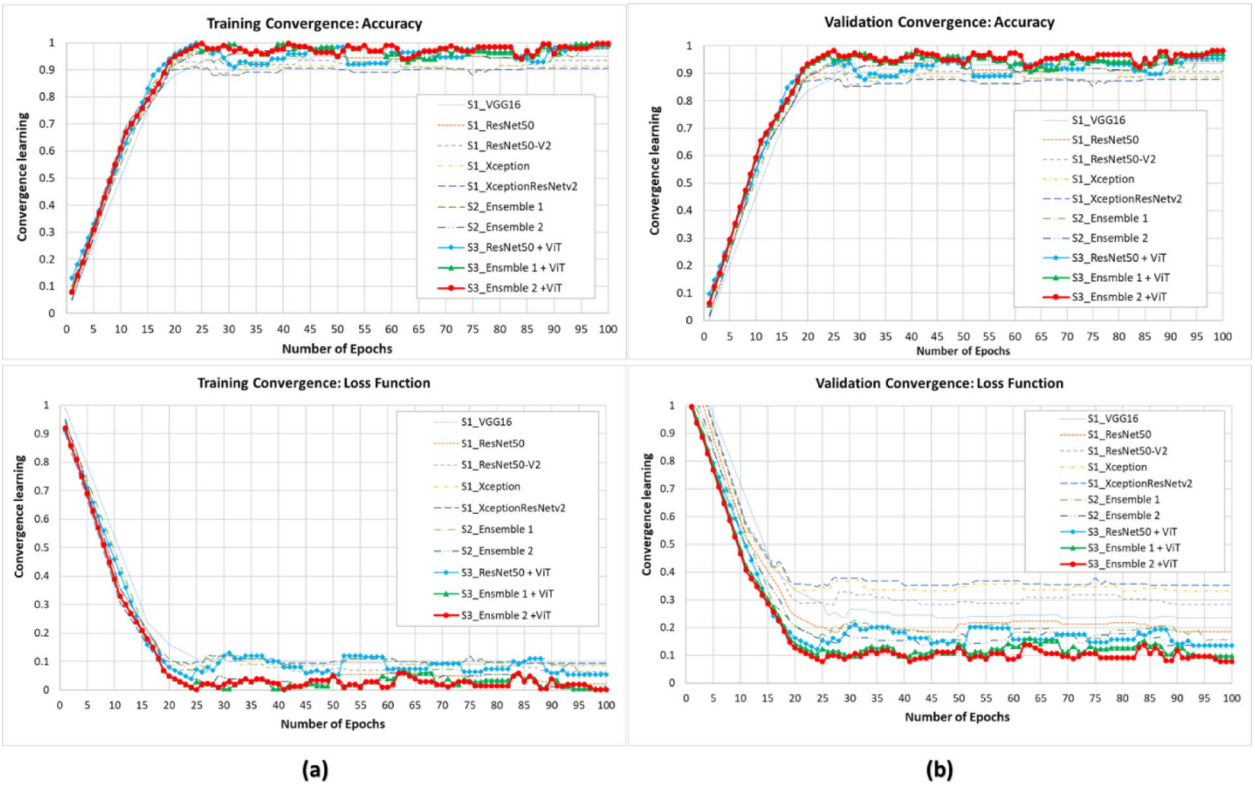

Fig. 10. The learning convergence rates of all AI classification models used for Approach (2): Scenario 1(S1), Scenario 2(S2), and Scenario 3(S3). These results are recorded in terms of accuracy and loss function values over 100 epochs during the (**a**) training and (**b**) validation process for Fold 3. The curves indicate the AI models do not encounter overfitting issues.

AI Model	ACC	SEN	SPE	F1-score	AUC
ResNet50	**85.19**	**85.19**	**86.26**	**85.45**	**85.32**
VGG16	85.19	83.09	83.09	83.09	83.09
Xception	79.63	79.63	81.42	80.08	79.73
InceptionResNetV2	72.22	72.22	73.14	72.58	69.78
ResNet50-V2	70.37	70.37	68.24	67.96	60.98

Table 3. Approach (1) - Scenario 1: average classification evaluation performance (%) over five cross-validation tests of each individual model using the testing sets.

AI Model	ACC	SEN	SPE	F1-score	AUC
(1) Ensemble 1 Model: (ResNet50 and VGG16)	87.04	84.93	**91.43**	89.86	**88.18**
(2) Ensemble 2 Model: (ResNet50, VGG16, and Xception)	**87.04**	**87.67**	85.71	**90.14**	86.69

Table 4. Approach (1) - Scenario 2: average classification evaluation results (%) over five cross-validation tests of the ensemble learning models: ensemble 1 (ResNet50 + VGG16) and ensemble 2 (ResNet50 + VGG16 + Xception).

abnormal cases. The performance of ensemble models is compressed as seen by the confusion matrices in Fig. 11. Eleven abnormal cases that were incorrectly categorized are listed in Ensemble 1, whereas 9 cases are recorded for Ensemble 2. In contrast, Ensemble 1 outperforms Ensemble 2 in circumstances where there are just 3 normal cases versus 5 normal cases that are misclassified. As a conclusion, both AI ensemble learning achieved encouraged evaluation results and achieved better performance than the individual AI models. The

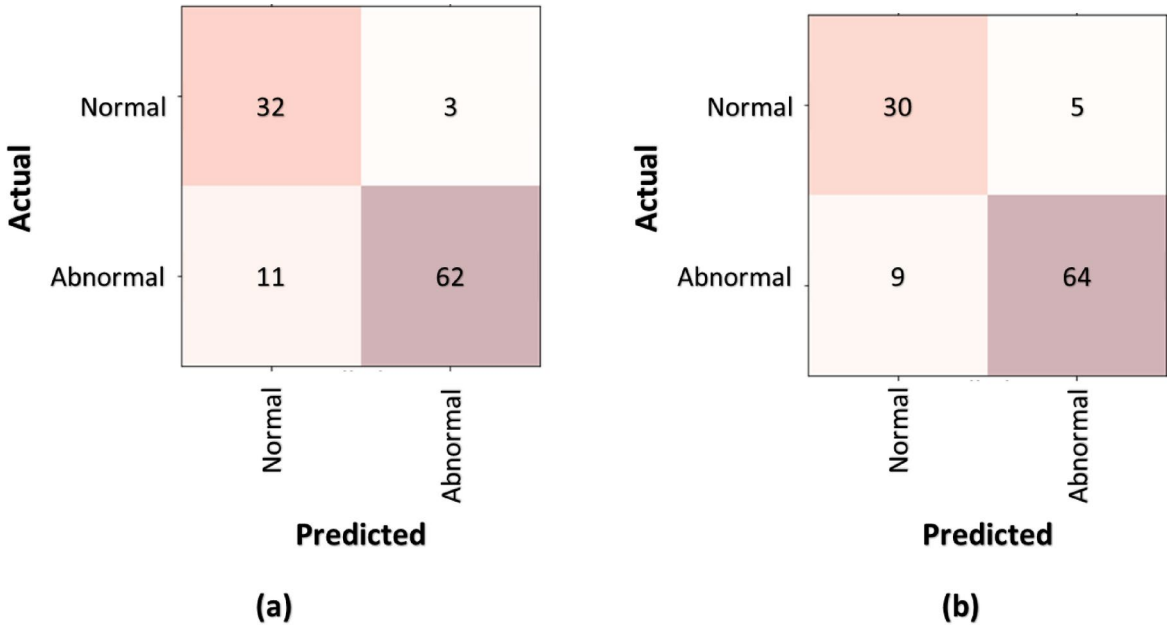

Fig. 11. Approach (1) - Scenario 2: Classification confusion matrix of the ensemble learning models for Fold 3: (**a**) Ensemble model 1 (ResNet50 and VGG16) and (**b**) Ensemble model 2 (ResNet50, VGG16, and Xception).

AI Model	ACC	SEN	SPE	F1-score	AUC
(1) ResNet50 + ViT	86.11	86.30	85.71	89.36	86.01
(2) Ensemble 1 + ViT	**87.96**	84.93	**94.29**	**90.51**	89.61
(3) Ensemble 2 + ViT	87.04	**87.67**	85.71	90.14	**86.69**

Table 5. Approach (1) - Scenario 3: average classification evaluation results (%) over five cross-validation tests of the proposed hybrid model (CNN + ViT).

overall accuracy of the ensemble models is improved by 1.85% and 16.67% compared with the best (ResNet50) and worst (ResNet50-V2) individual AI models.

Scenario 3: hybrid deep learning structure As shown in Fig. 7, we propose employing a hybrid scenario to improve classification performance by combining ensemble deep learning CNN architectures with the newly developed ViT prediction technique. The CNN backbone networks are utilized to prepare and extract deep learning features, which are subsequently classified into two pathological cases (normal and abnormal) using ViT. Three different hybridization structures of CNN and ViT are created and investigated. The deep features are generated based on the capabilities of ResNet50 and classified using ViT's head classifier in the first structure, which combines the optimal CNN architecture of ResNet50 and ViT. Ensemble 1 (i.e., ResNet50 and VGG16) is employed as a deep feature extractor in the second structure, while Ensemble 2 (i.e., ResNet50, VGG16, and Xception) is utilized in the third hybrid. The evaluation prediction result is summarized in Table 5.

The ViT assists the ensemble backbone in marginally enhancing the categorization results based on the gathered data. The hybrid ResNet50 and ViT exhibit an improvement rate of 0.92% AUC and 0.69% accuracy compared with the individual ResNet50. Similar enhancements in recording accuracy rate (0.92%), F1-score rate (0.65%), and AUC rate (1.43%) are achieved in Ensemble 1's performance with ViT. In contrast, the ViT fails to support the Ensemble 2 prediction performance where the prediction results remain with the same behaviour. in Fig. 12 shows the corresponding confusion matrices of this scenario.

Approach (2): parallel approach

For the parallel approach, both segmentation and classification stages are launched in parallel to predict the abnormality boundaries (LV, Mayo, MI, and MVO) and the pathology condition (Normal Vs vs. abnormal). Both stages have the same input of the entire cardiac MR images as shown in Fig. 2. To investigate the prediction performance in terms of the use of the entire cardiac MR image, we reinvestigate the three scenarios presented in Sect. 3.5. In the following sub-section, the overall prediction performance is demonstrated for each scenario separately.

Scenario 1: individual deep learning structure Table 6 shows the overall classification performance of the individual AI models when the entire cardiac MR images are used instead of the extracted ROIs. When compar-

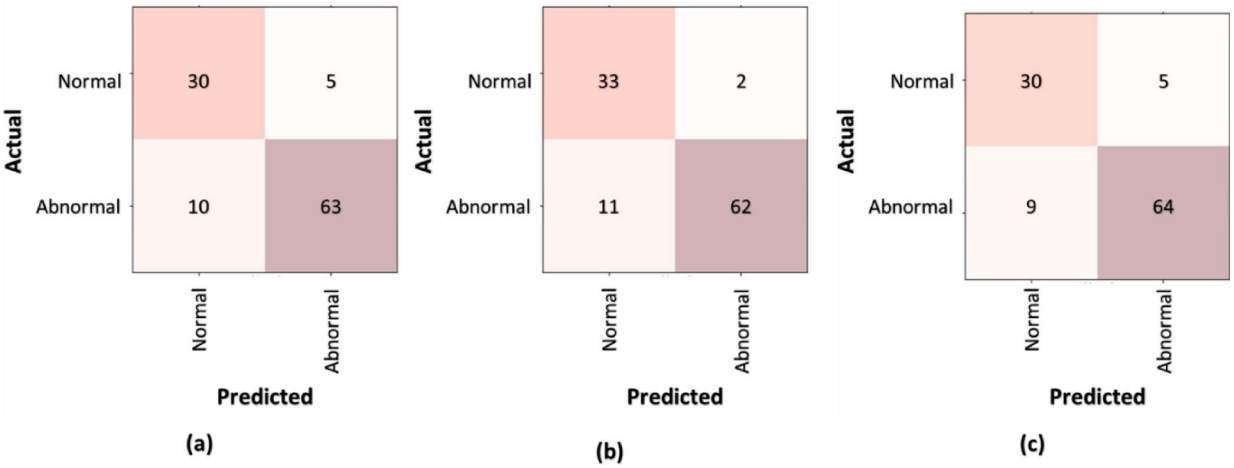

Fig. 12. Approach (1) - Scenario 3: Classification performance in term of the derived confusion matrices of the proposed hybrid ensemble CNN and the ViT for the Fold 3: (**a**) ResNte50 + ViT, (**b**) Ensemble 1 + ViT, and (**c**) Ensemble 2 + ViT.

AI Model	ACC	SEN	SPE	F1-score	AUC
ResNet50	**94.44**	**94.44**	**94.43**	**94.40**	**92.92**
VGG16	93.52	93.52	93.72	93.38	90.74
ResNet50-V2	90.74	90.74	90.72	90.58	87.95
Xception	88.89	88.89	88.80	88.70	85.83
InceptionResNetV2	87.96	87.96	88.07	88.01	86.63

Table 6. Approach (2) - Scenario 1: average classification evaluation performance (%) over five cross-validation tests of each individual AI model over the testing dataset.

ing Scenario 1 for both approaches, Scenario 1 for approach 2 yields more precise results for the prediction evaluation. This indicates that the entire cardiac images have more robust feature sets than the extracted ROIs, leading to improved classification accuracy. The decision to classify cardiac myocardial infarction using the entire medical image or an extracted ROI depends on multiple factors. The whole image approach allows the AI model to gather contextual data and potentially important features from adjacent tissues. This can be beneficial for understanding the general heart structure and identifying small trends. However, including unnecessary areas in the image may introduce noise, decrease model performance, increase computational complexity, and require more resources. On the other hand, the extracted ROI approach focuses on a specific ROI, such as the area surrounding the heart, which can help eliminate noise and irrelevant data. This strategy may result in a more precise and effective categorization procedure. However, in some cases, removing contextual information from other areas of the image may lead to the loss of important features. Additionally, since mistakes at this stage can impact the overall classification process, the accuracy of ROI extraction is crucial. In general, we can conclude that the whole image approach may be practical when a large training dataset is available, as in this study. However, the ROIs may be more useful for smaller datasets. Therefore, the most effective strategy depends on various factors, including the dataset size, the type of features being captured, and the computational resources available. To achieve the highest prediction performance, the proposed AI models for both the whole picture and ROI techniques need to be iteratively tested and refined and this is what we investigate in our study.

Scenario 2: fusing multiple deep learning structures Similarly, to scenario 2 in Approach (2), we select the top three classification performance models to build the proposed ensemble deep learning model. Two ensemble learning models are designed and called Ensemble 1 (ResNet50 and VGG16) and Ensemble 2 (ResNet50, VGG16, and ResNet50-V2). The second ensemble AI model (Ensemble 2) includes ResNet50, VGG16, and ResNet50-V2. Both ensemble models are trained and tested using the same execution environmental settings and the same data distribution. The classification performance of both ensemble learning is reported in Table 7.

The Ensemble 1 seems to achieve similar classification results of individual ResNet50 with ACC of 94.44%. On the other hand, Ensemble 2 clearly improves the classification results achieving overall classification ACC, SNE, SPE, F1-score, and AUC of 95.37%, 98.63%, 88.57%, 96.64%, and 94.60, respectively. Figure 13 shows the classification performance in terms of confusion matrices of this scenario. The proposed Ensemble 2 model could decrease the FN and FP cases into one and four misclassified cases as shown in Fig. 12, respectively.

AI Model	ACC	SEN	SPE	F1-score	AUC
(1) Ensemble 1 Model: (ResNet50 and VGG16)	94.44	95.89	**91.14**	95.89	93.66
(2) Ensemble 2 Model: (ResNet50, VGG16, and ResNet50-V2)	**95.37**	**98.63**	88.57	**96.64**	**94.60**

Table 7. Approach (2) - Scenario 2: average classification evaluation results (%) over five cross-validation tests of the ensemble learning models: ensemble 1 (ResNet50 and VGG16), while ensemble 2 (ResNet50, VGG16, and ResNet50-V2).

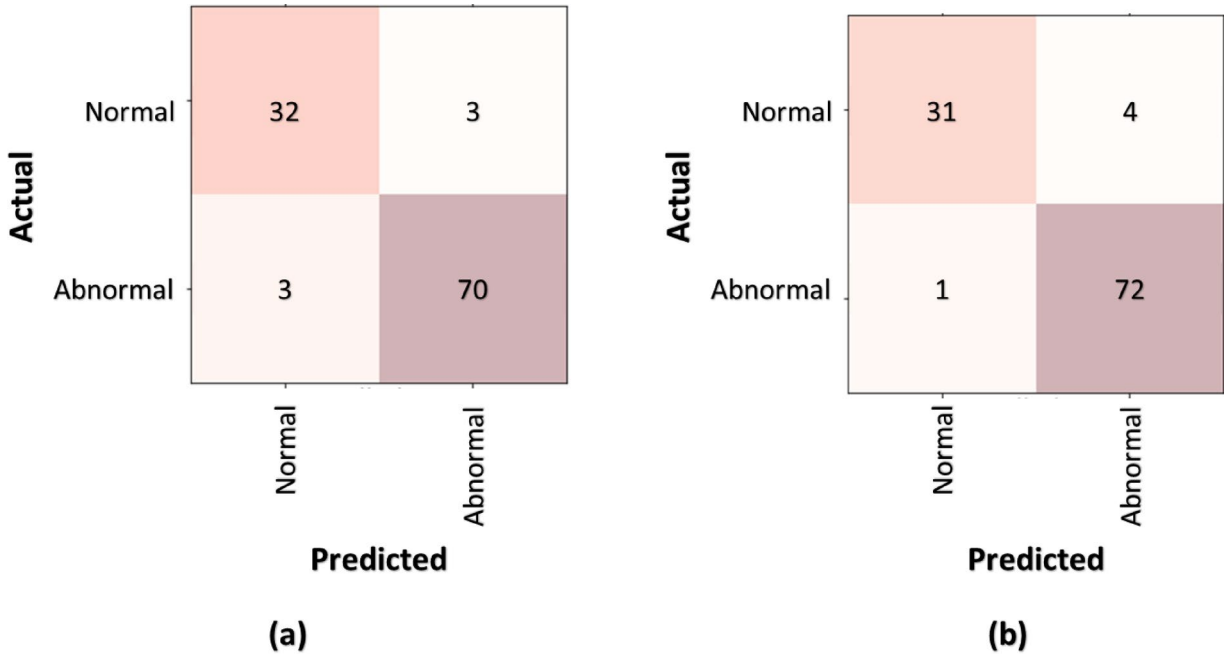

Fig. 13. Approach (2) - Scenario 2: Classification confusion matrix of the ensemble learning models for the Fold 3: (**a**) Ensemble model 1 (ResNet50 and VGG16) and (**b**) Ensemble model 2 (ResNet50, VGG16, and ResNet50-V2).

AI Model	ACC	SEN	SPE	F1-score	AUC
(1) ResNet50 + ViT	95.37	97.26	91.43	96.60	94.86
(2) Ensemble 1 + ViT	97.22	**98.63**	94.29	97.96	95.64
(3) Ensemble 2 + ViT	**98.15**	**98.63**	**97.14**	**98.63**	**97.13**

Table 8. Approach (2) - Scenario 3: average classification evaluation results (%) over five cross-validation tests of the proposed hybrid model between CNN and ViT.

Scenario 3: hybrid deep learning structure We repeat the experiments of the proposed hybridization technique for three different deep learning structures including both CNN backbone and ViT: (1) ResNet50 and ViT, (2) Ensemble 1 and ViT, and (3) Ensemble 2 and ViT. Table 8 records the evaluation results of all three hybrid deep learning structures. Table 8 shows the improvement of the classification performance due to the robustness of the hybridization technique of CNN and ViT. The best classification performance is achieved using Ensemble 2 as a backbone network where the achieved results are recorded to be 98.15% overall accuracy, 98.63% SEN, 97.14% SPE, 98.63% F1-score, and 97.13% AUC. In contrast to Approach (1), the proposed three hybridization models of CNN and ViT of Approach (2) achieve better classification results than the individual or ensemble models.

As shown in Fig. 14, the confusion matrix of the best hybrid model (Ensemble 1 + ViT) achieves the lowest rate of the misclassified normal and abnormal cases with only one case for both classes. Such impressive prediction performance could be an acceptable application for practical applications.

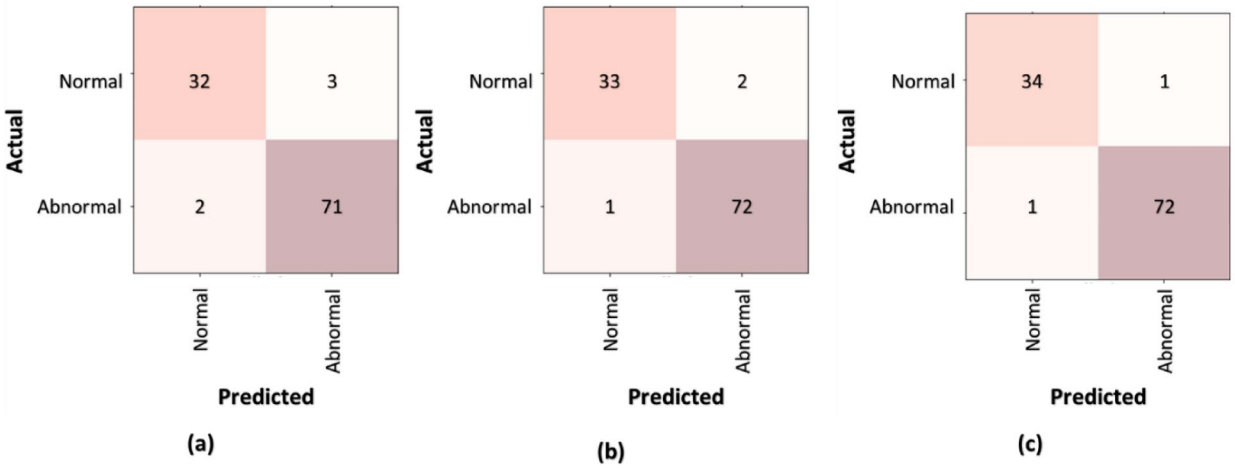

Fig. 14. Approach (2) - Scenario 3: Classification performance in terms of the derived confusion matrices of the proposed hybrid ensemble CNN and the ViT for the Fold 3: (**a**) ResNte50 + ViT, (**b**) Ensemble 1 + ViT, and (**c**) Ensemble 2 + ViT.

Discussion

The proposed CAD system is evaluated using serial and parallel approaches to determine the most effective method for heart disease prediction. Both approaches rely on a prior segmentation stage to extract heart regions. In the serial approach, segmentation boundaries are crucial for defining ROIs used as input patches for classification. In the parallel approach, these boundaries assist physicians in visually inspecting abnormal tissue growth, particularly for MI and MVO. This study carefully examines segmentation models to identify the most accurate method for extracting the boundaries of four heart regions. Notably, the ResU-Net model achieves superior segmentation performance, effectively distinguishing both normal and abnormal tissues, as summarized in Table 2. Specifically, for MI and MVO abnormalities, ResU-Net improves segmentation accuracy by 6.91% and 7.29%, and MIoU by 4.74% and 6.26%, respectively. As shown in Fig. 9, the accurate delineation of the myocardium's outer borders plays a crucial role in precise ROI extraction, as it encapsulates the LV, MI, and MVO regions. ResU-Net demonstrates strong segmentation performance, achieving 91.28% accuracy, an 88.39% F1-score, and 87.84% MIoU. Accurate boundary extraction enhances classification performance and aids in visually assessing abnormal tissue growth within normal myocardial tissue. For classification, three different scenarios utilizing CNN and ViT hybrid models are explored to identify the optimal approach for CAD-based heart disease prediction, particularly for MI and MVO abnormalities. The results indicate that the hybrid classifier combining CNN and ViT outperforms individual models and ensemble approaches. In the serial approach, the best hybrid classifier (Ensemble 1 + ViT) improves accuracy, F1-score, and AUC by 2.77%, 5.06%, and 4.29% compared to the individual ResNet50 model, and by 0.92%, 0.37%, and 1.43% compared to the best ensemble model (ResNet50 + VGG16). These findings highlight the effectiveness of the proposed CAD system in supporting accurate heart disease prediction and aiding clinical decision-making.

Approach (1) against approach (2): scenarios comparison

We compare the top classification results for the same Scenario of both approaches. To do this, we select the scenario with the best performance, regardless of how well the AI models align. For example, in Scenario 2, Ensemble 1 for Approach (1) and Ensemble 2 for Approach (2) are selected because they achieve the highest categorization outcomes. To compare Scenario 3 in both approaches, we also utilize the hybrid AI models of Ensemble 1 with ViT and Ensemble 2 with ViT. Such a comparison is conducted to pick up the accurate AI model for cardiac infraction classification. Figure 15 shows the comparison results among all scenarios of both approaches. We can conclude that Approach (2) performs well for each scenario and always produces better classification performance. In addition, Scenario 3 (Ensemble 2 with ViT) in Approach (2) achieves the best classification performance. Thus, the proposed hybrid Ensemble 2 with ViT (Scenario 3) model seems to be applicable for practical applications of myocardial infarction classification.

Execution computational cost

We compare the time computation of all AI models that are used to build the structures of the proposed three scenarios as reported in Table 9. We show the comparison results in terms of Approach (2) since it achieved the best classification performance than Approach (1) for all scenarios. This comparison is conducted in terms of trainable parameters, training time per epoch, inferencing time per image, the number of FPS, and the corresponding overall accuracy. The proposed hybrid model (Ensemble 2 + ViT, Scenario 3) could achieve the best classification accuracy of 98.15% even if it has heavy structures with 192 M trainable parameters. Meanwhile, it needs about 2.31 s to predict the pathology of the entire single cardiac MR image. The huge trainable parameters came due to the hybridization strategy that was used to build the proposed model of Scenario 3. The tradeoff between model complexity and prediction performance is always present. However,

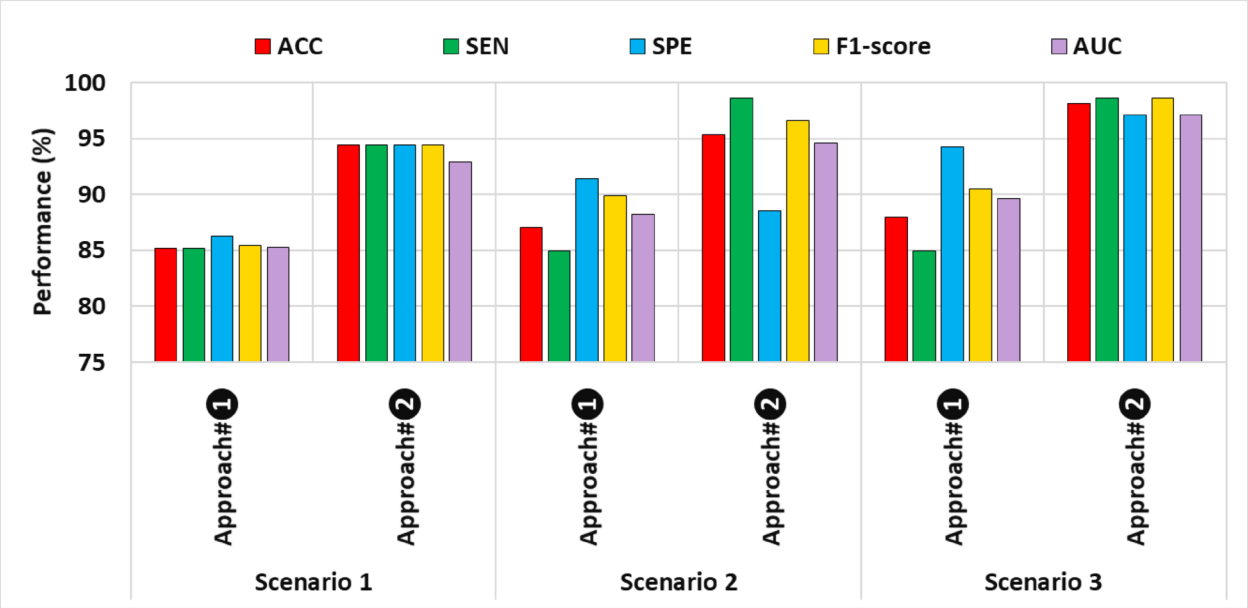

Fig. 15. Comparison average evaluation results between the performance of all scenarios in Approaches (1) and (2). For each Scenario, the best classification model is selected to conduct such comparison regardless the type of deep learning models.

Scenario	AI Model	No. of Trainable Parameters (Million)	Training Time/Epoch (msec)	Testing Time/Image (sec)	Frame Per Second (FPS)	ACC (%)
Scenario 1	ResNet50	50.79	24	0.64	1.54	94.44
	VGG16	20.89	28	0.56	1.79	93.52
	ResNet50-V2	51.64	25	0.74	1.35	90.74
	Xception	10.50	63	2.88	0.35	**88.89**
	InceptionResNetV2	18.48	308	1.49	0.67	87.96
Scenario 2	Ensemble 1	70.17	50	0.97	1.03	94.44
	Ensemble 2	76.21	67	1.26	0.79	**95.37**
Scenario 3	ResNet + ViT	88.31	57	1.53	0.65	95.37
	Ensemble 1 + ViT	89.29	191	2.29	0.44	97.22
	Ensemble 2 + ViT	89.96	192	2.31	0.43	**98.15**

Table 9. Approach (2): execution computational cost for all AI classification models over all scenarios.

practical implementation on edge devices or integration into clinical workflows depends on platform-specific requirements. In medical applications, higher accuracy is often prioritized, even if it results in a slight processing time to ensure comprehensive outcomes.

The standard absolute error is calculated for all models in terms of ACC, SEN, SPE, F1-score, and AUC to compare the error variations among the various AI models utilized in this study with the best classification performance of the suggested AI model of Scenario 3 (Ensemble 2 + ViT). The results of the standard error comparison for each model employed in Approach (2) are shown in Fig. 16. As shown in Fig. 16; Table 9, the hybrid model of Ensemble 1 + ViT achieves the closest classification performance recording standard errors of 0.93% ACC, 2.85% SPE, 0.67% F1-score, and 1.49% AUC. The individual InceptionResNetV2 model, on the other hand, produces the worst classification results, with standard errors of 10.19%, 10.67%, 9.07%, 10.62%, and 10.5%, respectively, for the ACC, SEN, SPE, F1-score, and AUC.

Examination of the statistics for performance significance

To demonstrate the significance of our proposed model (Scenario 3: Ensemble 2 + ViT), we conducted a machine learning statistical analysis using the robust non-parametric Friedman's statistical test, applying the method of multiple classifiers over multiple datasets[56]. We selected the four best classifiers presented in this study to assess the classification performance using six medical datasets, as shown in Table 10. To evaluate the statistical significance of our proposed model, we set the null hypothesis (H0) to state that the differences in results among different classifiers are not significant, while the alternative hypothesis (AH) is formulated to state that there are significant differences between the performance of the proposed model and other classifiers. As depicted

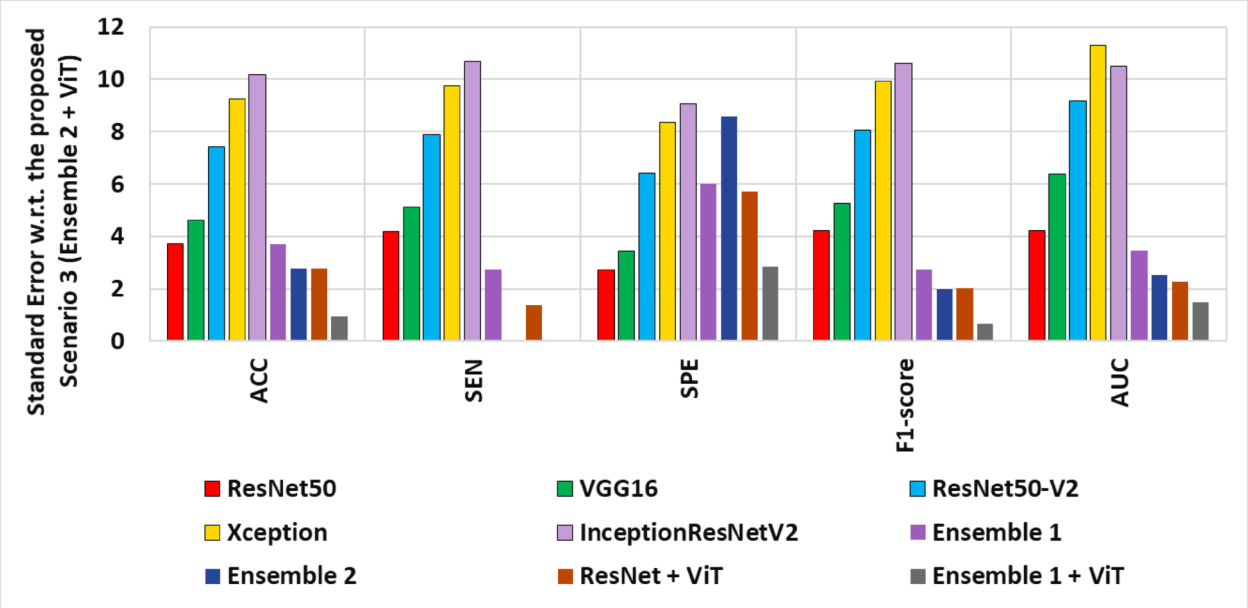

Fig. 16. Approach (2): The standard absolute error with respect to the best proposed model (Ensemble 2 + ViT) against all AI models used in this study. The best error is recorded when the proposed hybrid classification model (Ensemble 1 + ViT) is used. The InceptionResNetV2 records the lower classification results among all classification scenarios.

Dataset Description			Classification performance: F1-score (%)			
Dataset	Data size	Splitting Ratio (Train/Validation/Test)	Ensemble 2 + ViT	Ensemble 1 + ViT	Ensemble 1	Ensemble 2
Cardiac EMIDIC	708	75/10/15	**98.15**	97.06	93.98	94.44
BreastMNIST[57]	780	70/10/20	**96.84**	95.85	93.02	93.12
Chest X-ray[35]	15k	70/20/10	**97.35**	96.89	92.74	93.86
Chest X-ray[36]	14.4k	70/20/10	**97.63**	96.56	93.16	94.01
HEp-2 I3A Task-1[56]	13.5k	64/16/20	**98.05**	97.85	94.23	94.48
PneumoniaMNIST[57]	5.8 K	80/9/11	**96.20**	95.14	92.05	93.21

Table 10. The statistical evaluation results of approach (2) using Friedman's test using confidence level $\alpha = 0.05$. Friedman's test Conclusion: F (calculated) = 691.2, F (critical, $\alpha = 0.05$, df = 3) = 7.815, and p-values < 0.0001 $\Rightarrow$ H_0 is Rejected.

in Table 10, Friedman's test indicates significant differences with F (calculated) = 691.2 is much larger than F (critical, $\alpha = 0.05$, df = 3) = 7.815. Based on the chi-squared distribution with df = 3 and the calculated F statistic of 691.2, the p-value is extremely close to zero (in practice, it is effectively zero ($p < 0.0001$)). This implies that the null hypothesis is rejected, and the alternative hypothesis is accepted. Consequently, we can conclude that the proposed AI model (Scenario 3: Ensemble 2 + ViT) is significantly superior, achieving the highest prediction accuracy.

The impact of MR image Pre-processing

The best AI models from Scenario 3 are used in the training and evaluation phases of this comparison study twice: ResNet + ViT, Ensemble 1 + ViT, and Ensemble 2 + ViT. The original and processed cardiac MR images are used to train and support the AI candidates individually while keeping the same training environment and settings. Figure 17 compares the study's findings for overall accuracy and F1-score evaluation metrics, with and without pre-processing. With both approaches, the pre-processing phase improves the classification outcomes. The improvement rates are 1.85% and 5.63% for accuracy and 5.53% and 4.31% for F1-score for Approaches(1) and (2) respectively, when using the best model (Ensemble 1 + ViT).

Ablation study: validation and verification of the proposed model using unseen cardiac MRI images

The unseenset of 358 cardiac MRI images was used to validate the model and assess the generalization of the selected AI hyperparameters. These images are publicly posted as a testing sub-set of the 2020 EMIDEC MICCAI challenge event that was used to evaluate the participants. Unfortunately, till this time, neither the segmentation

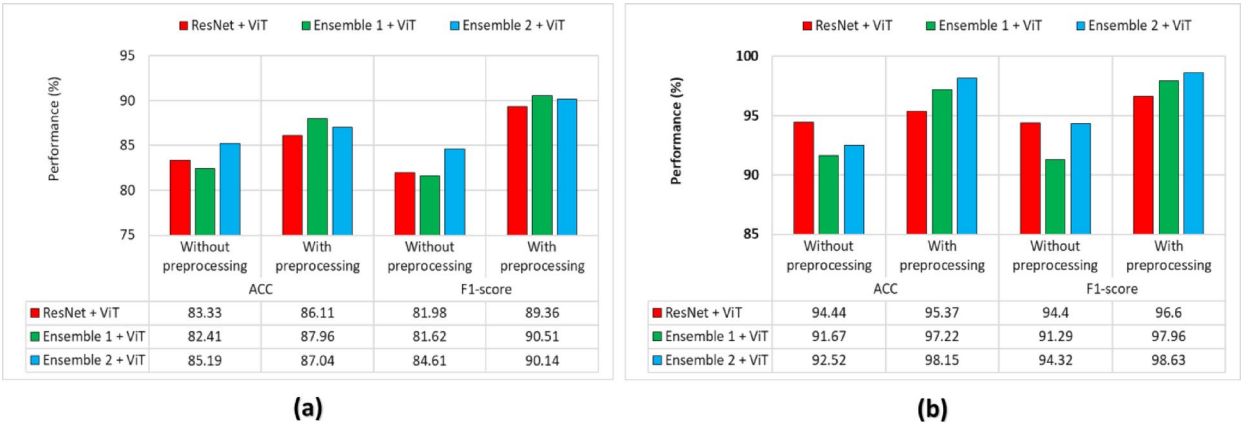

Fig. 17. The impact of pre-processing on the overall classification performance for Approaches (1) and (2) that are shown in (**a**) and (**b**), respectively. The AI models of the best Scenario 3 are selected to conduct this comparison over Fold 3: ResNet50 + ViT, Ensemble 1 + ViT, and Ensemble 2 + ViT. It is shown clearly the preprocessing stage supports the CAD system to improve the final prediction results.

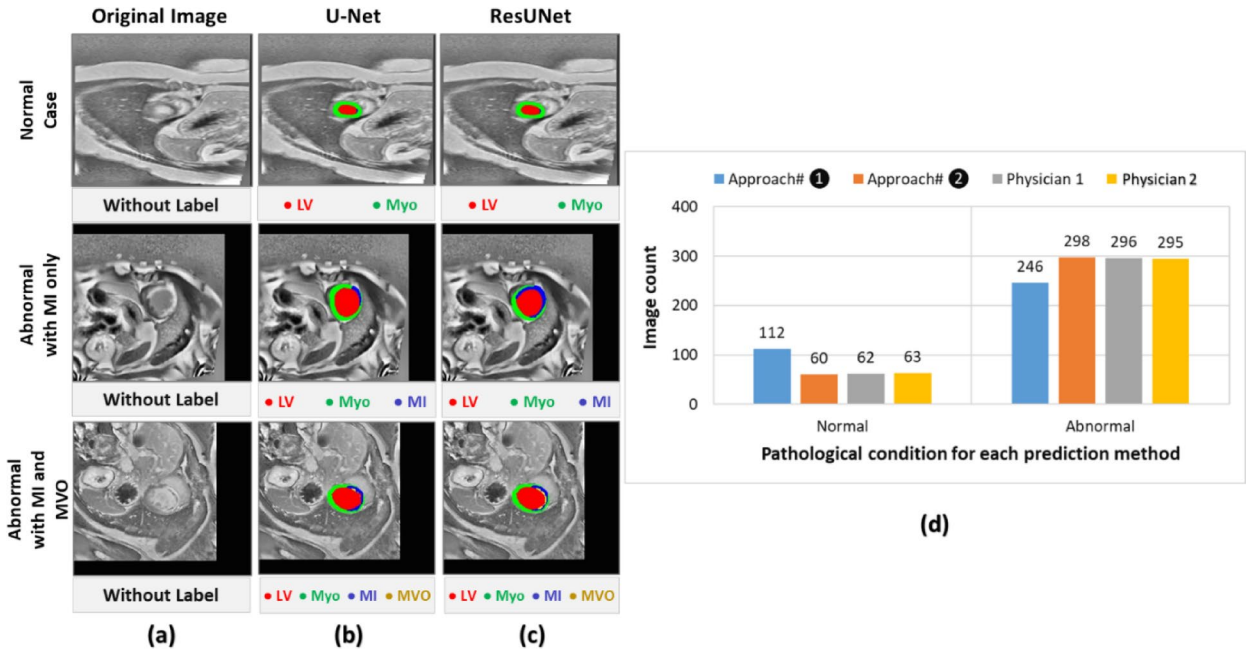

Fig. 18. Ablation study: Validation and verification evaluation results for unseen dataset. (**a**) shows the original image without segmentation label (GT), while (**b**) and (**c**) depict examples of the segmentation prediction via U-Net and ResU-Net, respectively. (**d**) shows the statistical analysis to count the normal and abnormal classification conditions for each method: best models for Approach (1) (Ensemble 1 + ViT), Approach (2) (Ensemble 2 + ViT) belongs with the Physician 1, and Physician 2.

(binary mask) nor the classification (pathology condition) labels are available online. As a result, we solely use them to verify the best-suggested model for segmentation (ResU-Net) and classification stages: Approach (1) (Ensemble 1 + ViT) and Approach (2) (Ensemble 2 + ViT). Simultaneously, we ask two medical professionals in parallel to examine and verify the results of the AI models, and their assessments closely aligned with the AI model predictions as shown in Fig. 18. All prediction methods agree to predict 23 normal and 209 abnormal cases similarly. However, the remaining cases are predicted differently based on the power of the prediction approach. The total number of predicted normal and abnormal cases for each method is reported in Fig. 18**(d)**. This suggests that the selected hyperparameters are well-suited for predicting other datasets for MI-related heart diseases.

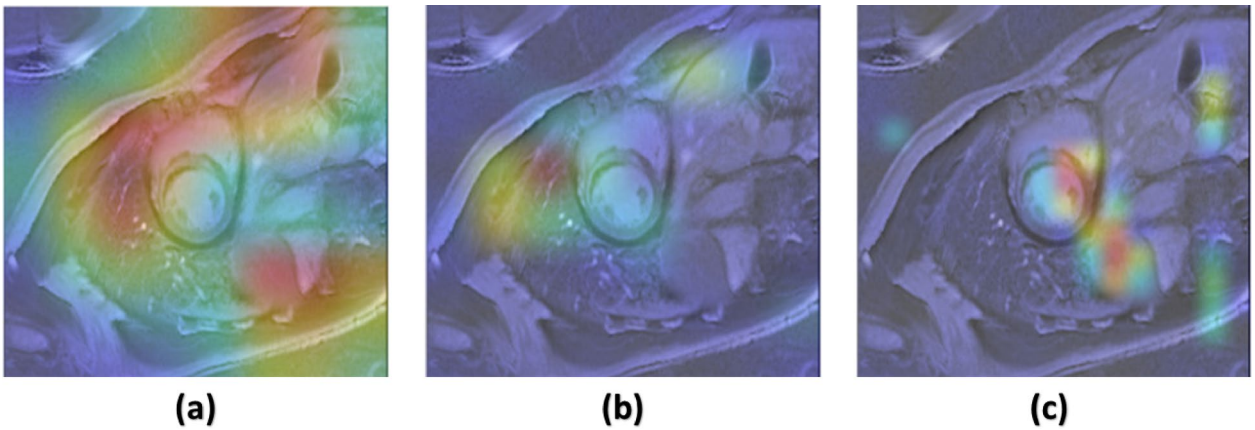

Fig. 19. Approach (2): Visual explainable saliency maps for abnormal case using (**a**) ResNet50 + ViT, (**b**) Ensemble 1 + ViT, and (**c**) Ensemble 2 + ViT. The saliency map provides a visual explanation to inspect the abnormality regions of the black box deep learning model.

Reference	Prior Segmentation	AI Model	ACC	SEN	SPE	F1-score
Sharma et al. (2021)[59]	✗	Multimodal Neural Network (M2N2)	62.0	72.70	41.20	-
Invantsits et al. (2021)[60]	✓	LV Segmentation + CNN-based Radiomics features	76.0	72.70	82.30	-
Girum et al. (2021)[16]	✓	Support vector machine (SVM)	82.0	78.79	88.24	-
Shi et al. (2021)[31]	✗	3D CNN and Random Forest (RF)	92.0	90.91	88.20	-
Lourenço et al. (2021)[33]	✗	CNN with fully connected layers	82	87.88	70.59	-
Moravvej et al. (2022)[39]	✗	deep reinforcement learning and population-based algorithms	88.86	86.63	90.1	85.1
Wang et al., (2024)[40]	✗	Radiomics analysis with image fusion	93.0	89.0	95.0	89.0
Hadamitzky et al. (2025)[61]	✓	3D ResNet-34 ensemble	Mean AUC = 75%			
The proposed CAD System (2025)	✓	**Approach (1): Ensemble 2 + ViT**	87.04	87.67	85.71	90.14
	✗	**Approach (2): Ensemble 2 + ViT**	**98.15**	**98.63**	**97.14**	**98.63**

Table 11. Evaluation results (%) are compared with recent AI methods in literature, emphasizing prediction performance after the segmentation stage.

Visual explainable saliency maps

The proposed AI framework is investigated against other AI models in terms of visual heat maps that explain how the AI black box predicts the pathological condition of each case considering the important regions of the entire MRI image. We test and record the XAI heat maps from each AI model (Approach (2)) based on the 2D deep features from the top layers after removing the classification dense layers as in our previous works[36,58]. Figure 19 shows an example of the derived saliency maps for each AI model when the entire image is used.

Comparison with existing works

Table 11 lists the recent AI research work for myocardial infraction segmentation using the same 2020 EMIDEC MICCAI challenge dataset. We compare the proposed approaches in terms of ACC, SEN, SPE, F1-score, and AUC. This comparison clearly shows the superiority of the proposed Scenario 3 AI hybrid model (Ensemble 1 + ViT) achieving the top prediction accuracy when Approach (2) is applied. In conclusion, our proposed AI model outperforms the other AI techniques achieving an overall accuracy of 98.15%, SEN of 98.63%, SPE of 97.14%, and F1-score of 98.63%. The classification performance of Approach (1) is comparable with other models in the literature where it achieves the ACC of 87.04%. The slight performance variation could be a result of data splitting or different training execution environments.

Work limitations

Classifying myocardial infarction (MI) from MRI images is a complex task that requires identifying and describing heart muscle damage caused by reduced blood flow. Although the desired categorization performance has been achieved, several limitations remain. First, the clarity and resolution of MRI images play a significant role in MI categorization accuracy. Low-resolution images or those affected by motion artifacts can make it difficult to accurately identify and characterize changes in cardiac tissue. Second, variations in patient anatomy, including differences in heart size, shape, and orientation, complicate the interpretation of MRI images, making it challenging to establish consistent classification criteria. Additionally, the timing of the imaging process can

also impact the MI categorization. Third, MI can manifest with varying degrees of tissue damage and patterns, such as sub-endocardial, sub-epicardial, and transmural involvement, which further complicates accurate classification from MRI scans. Fourth, data imbalance poses a critical challenge, as it can hinder proper AI model training and lead to major-class biases or overfitting. Lastly, the absence of an annotated textual dataset limits the proposed CAD system's ability to generate detailed medical reports, which would provide comprehensive textual explanations of abnormal findings.

Future work

A comprehensive classification CAD system including more evaluation results such as explainable textual interpretation is the future for the medical domain including MI classification[62]. Provide output of segmentation, classification, and XAI visual and textual explanation is our next step in the near future once the related dataset becomes available. Also, testing and verifying the proposed framework using 3D DICOM MRI volums is a future research point when the annotated dataset be available.

Conclusion

In this study, we introduce a novel CAD system for heart disease prediction that integrates both segmentation and classification stages through serial and parallel approaches. The segmentation stage accurately identifies key heart regions, including MI, LV, Myo, and MVO, while the classification stage determines heart conditions, distinguishing between normal and abnormal cases, especially when MI and/or MVO are present. This comprehensive approach offers segmentation boundaries, pathological classification, and visual heat maps, resulting in an end-to-end MI-based CAD prediction system. To enhance interpretability, visual explainable saliency maps are generated for each AI model, emphasizing the most relevant heart regions used for segmentation and classification decisions. Through the implementation of two distinct approaches with three classification scenarios, we conclude that segmentation plays a critical role in visually defining heart regions, supporting physicians in decision-making and detecting abnormal tissue growth in MI and MVO cases. Furthermore, the proposed hybrid classification model (Scenario 3: Ensemble 2 + ViT) in the second approach proves to be the most promising for practical and industrial applications. By selecting the highest accuracy segmentation and classification scenario, early MI prediction performance can be improved, ultimately reducing the mortality rate. However, the absence of a textual dataset limits the ability to generate textual interpretations alongside visual heat maps. Future work will focus on integrating advanced large language models (LLMs) for automated text report generation, visual question answering (VQA), and applying agentic AI for multi-tasking across various heart diseases.

Data availability

To achieve this study, a free public "EMIDEC Dataset" is used: (https://emidec.com, Accessed on March 08, 2025).

Code availability

The code is available here: https://github.com/AISSLab2025/Hybrid-CAD-Segmentation-Classification-of-Myocardial-Infarction-from-MRI.

Received: 24 December 2024; Accepted: 15 April 2025
Published online: 23 April 2025

References

1. Attallah, O. & Ragab, D. A. Auto-MyIn: automatic diagnosis of myocardial infarction via multiple GLCMs, CNNs, and SVMs. *Biomed. Signal Process. Control*. **80**, 104273 (2023).
2. Lalande, A. et al. Emidec: a database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac MRI. *Data*. **5**, 89 (2020).
3. Florian, M. et al. Interplay of obesity, ethanol, and contaminant mixture on clinical profiles of cardiovascular and metabolic diseases: evidence from an animal study. *Cardiovasc. Toxicol.* **22**, 558–578 (2022).
4. Motevalli, M. M., Sohrabi, M. K. & Yaghmaee, F. Aspect-based sentiment analysis: A dual-task learning architecture using imbalanced maximized-area under the curve proximate support vector machine and reinforcement learning. *Inf. Sci.* **689**, 121449 (2025).
5. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9*, Proceedings, Part III 18, 2015, pp. 234–241. (2015).
6. Chen, J. et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, (2021).
7. Khan, A. et al. A survey of the vision Transformers and their CNN-transformer based variants. *Artif. Intell. Rev.* **56**, 2917–2970 (2023).
8. Takahashi, S. et al. Comparison of vision Transformers and convolutional neural networks in medical image analysis: a systematic review. *J. Med. Syst.* **48**, 84 (2024).
9. Al-Hejri, A. M. et al. ETECADx: Ensemble Self-Attention Transformer Encoder for Breast Cancer Diagnosis Using Full-Field Digital X-ray Breast Images. *Diagnostics*. **13**, 89 (2022).
10. Al-Tam, R. M. et al. A Hybrid Workflow of Residual Convolutional Transformer Encoder for Breast Cancer Classification Using Digital X-ray Mammograms. *Biomedicines*. **10**, 2971 (2022).
11. Addo, D. et al. A hybrid lightweight breast cancer classification framework using the histopathological images. *Biocybernetics Biomedical Eng.* **44**, 31–54 (2024).
12. Kosmala, W., Sanders, P. & Marwick, T. H. Subclinical myocardial impairment in metabolic diseases. *JACC: Cardiovasc. Imaging*. **10**, 692–703 (2017).

13. Gupta, S., Ge, Y., Singh, A., Gräni, C. & Kwong, R. Y. *Multimodality Imaging Assess. Myocard. Fibros. Cardiovasc. Imaging*, **14**, 2457–2469, (2021).
14. Chang, Y. & Jung, C. Automatic cardiac MRI segmentation and permutation-invariant pathology classification using deep neural networks and point clouds, *Neurocomputing*, vol. 418, pp. 270–279, (2020).
15. Vesal, S., Maier, A. & Ravikumar, N. Fully automated 3d cardiac mri localisation and segmentation using deep neural networks. *J. Imaging*. **6**, 65 (2020).
16. Girum, K. B. et al. Automatic myocardial infarction evaluation from delayed-enhancement cardiac MRI using deep convolutional networks, in *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4*, Revised Selected Papers 11, 2021, pp. 378–384. (2020).
17. Bhan, A., Mangipudi, P. & Goyal, A. Deep Learning Approach for Automatic Segmentation and Functional Assessment of LV in Cardiac MRI. *Electronics*. **11**, 3594 (2022).
18. Poudel, R. P., Lamata, P. & Montana, G. Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation, in *Reconstruction, Segmentation, and Analysis of Medical Images: First International Workshops, RAMBO 2016 and HVSMR 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17*, Revised Selected Papers 1, 2017, pp. 83–94. (2016).
19. Saito, K., Zhao, Y. & Zhong, J. Heart diseases image classification based on convolutional neural network, in *International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, pp. 930–935. (2019).
20. Zhang, Y. Cascaded convolutional neural network for automatic myocardial infarction segmentation from delayed-enhancement cardiac MRI, in *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4*, Revised Selected Papers 11, 2021, pp. 328–333. (2020).
21. Shaaf, Z. F., Jamil, M. M. A. & Ambar, R. A convolutional neural network model to segment myocardial infarction from MRI images. *Int. J. Online Biomedical Eng.* **19**, (2023).
22. Brahim, K. et al. A 3D network based shape prior for automatic myocardial disease segmentation in delayed-enhancement MRI. *IRBM*. **42**, 424–434 (2021).
23. Li, W., Wang, L., Li, F., Qin, S. & Xiao, B. Myocardial pathology segmentation of multi-modal cardiac MR images with a simple but efficient Siamese U-shaped network. *Biomed. Signal Process. Control*. **71**, 103174 (2022).
24. Chen, Z. et al. Automatic deep learning-based myocardial infarction segmentation from delayed enhancement MRI. *Comput. Med. Imaging Graph*. **95**, 102014 (2022).
25. Heidenreich, J. F., Gassenmaier, T., Ankenbrand, M. J., Bley, T. A. & Wech, T. Self-configuring nnU-net pipeline enables fully automatic infarct segmentation in late enhancement MRI after myocardial infarction. *Eur. J. Radiol.* **141**, 109817 (2021).
26. Zhang, Z. et al. Multi-modality pathology segmentation framework: application to cardiac magnetic resonance images, in Myocardial Pathology Segmentation Combining Multi-Sequence Cardiac Magnetic Resonance Images: First Challenge, MyoPS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 1, pp. 37–48. (2020).
27. Yang, S. & Wang, X. A hybrid network for automatic myocardial infarction segmentation in delayed enhancement-mri, in *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4*, Revised Selected Papers 11, 2021, pp. 351–358. (2020).
28. Zhou, Y., Zhang, K., Luo, X., Wang, S. & Zhuang, X. Anatomy prior based u-net for pathology segmentation with attention, in *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 392–399. (2020).
29. Huellebrand, M. et al. Comparison of a hybrid mixture model and a cnn for the segmentation of myocardial pathologies in delayed enhancement MRI, in Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, Revised Selected Papers 11, 2021, pp. 319–327. (2020).
30. Huellebrand, M., Ivantsits, M., Tautz, L., Kelle, S. & Hennemuth, A. A collaborative approach for the development and application of machine learning solutions for CMR-based cardiac disease classification. *Front. Cardiovasc. Med.* **9**, 829512 (2022).
31. Shi, J., Chen, Z. & Couturier, R. Classification of pathological cases of myocardial infarction using convolutional neural network and random forest, in *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4*, Revised Selected Papers 11, 2021, pp. 406–413. (2020).
32. Isensee, F. et al. Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features, in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10–14*, Revised Selected Papers 8, 2018, pp. 120–129. (2017).
33. Lourenço, A. et al. Automatic myocardial disease prediction from delayed-enhancement cardiac MRI and clinical information, in *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4*, Revised Selected Papers 11, 2021, pp. 334–341. (2020).
34. Brahim, K. et al. An Improved 3D Deep Learning-Based Segmentation of Left Ventricular Myocardial Diseases from Delayed-Enhancement MRI with Inclusion and Classification Prior Information U-Net (ICPIU-Net), *Sensors*, vol. 22, p. 2022. (2084).
35. Ukwuoma, C. C. et al. A Hybrid Explainable Ensemble Transformer Encoder for Pneumonia Identification from Chest X-ray Images. *J. Adv. Res.*, (2022).
36. Ukwuoma, C. C. et al. Deep Learning Framework for Rapid and Accurate Respiratory COVID-19 Prediction Using Chest X-ray Images. *J. King Saud University-Computer Inform. Sci.*, 101596 (2023).
37. Elmannai, H. et al. Diagnosis Myocardial Infarction Based on Stacking Ensemble of Convolutional Neural Network. *Electronics*. **11**, 3976 (2022).
38. Al-Haidri, W. et al. Deep learning-assisted framework for automation of lumbar vertebral body segmentation, measurement, and deformity detection in MR images. *Biomed. Signal Process. Control*. **106**, 107770 (2025).
39. Moravvej, S. V. et al. RLMD-PA: a reinforcement learning-based myocarditis diagnosis combined with a population-based algorithm for pretraining weights. *Contrast Media & Molecular Imaging*. vol. p. 8733632, 2022. (2022).
40. Wang, D. et al. Multi-parametric assessment of cardiac magnetic resonance images to distinguish myocardial infarctions: A tensor-based radiomics feature. *J. X-Ray Sci. Technol.* **32**, 735–749 (2024).
41. D. W. A. David gunning, DARPA's explainable artificial intelligence program. *AI Mag.* **40**, p. 44, (2019).
42. Brunese, L., Mercaldo, F., Reginelli, A. & Santone, A. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. *Comput. Methods Programs Biomed.* **196**, 105608 (2020).
43. You, C., Zhou, Y., Zhao, R., Staib, L. & Duncan, J. S. Simcvd: simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Trans. Med. Imaging*, (2022).
44. Al-Antari, M. A., Al-Masni, M. A., Choi, M. T., Han, S. M. & Kim, T. S. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. *Int. J. Med. Informatics*. **117**, 44–54 (Sep 2018).
45. Al-Masni, M. A. et al. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Comput. Methods Programs Biomed.* **157**, 85–94 (2018).
46. Al-antari, P. R. M. A. et al. *An Automatic Recognition of Multi-class Skin Lesions Via Deep Learning Convolutional Neural Networks, Presented at the ISIC 2018* (Skin Lesion Analysis Towards Melanoma Detection, 2018).

47. Al-Masni, M. A., Al-Antari, M. A., Choi, M. T., Han, S. M. & Kim, T. S. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Comput. Methods Programs Biomed.* **162**, 221–231 (2018).
48. Han, C. & Shi, L. ML–ResNet: A novel network to detect and locate myocardial infarction using 12 leads ECG. *Computer methods and programs in biomedicine*, **185**, 105138 (2020).
49. Chinnam, S. K. R., Sistla, V. & Kolli, V. K. K. Multimodal attention-gated cascaded U-Net model for automatic brain tumor detection and segmentation. *Biomed. Signal Process. Control.* **78**, 103907 (2022).
50. Raza, R., Bajwa, U. I., Mehmood, Y., Anwar, M. W. & Jamal, M. H. dResU-Net: 3D deep residual U-Net based brain tumor segmentation from multimodal MRI. *Biomed. Signal Process. Control.* **79**, 103861 (2023).
51. Abdelhamed, M. K. & Meriaudeau, F. NesT UNet: pure transformer segmentation network with an application for automatic cardiac myocardial infarction evaluation, in *Medical Imaging 2023: Computer-Aided Diagnosis*, pp. 608–619. (2023).
52. Li, L. et al. MyoPS: A benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images. *Med. Image. Anal.* **87**, 102808 (2023).
53. Shamshad, F. et al. Transformers in Medical Imaging: A Survey. arXiv preprint arXiv:2201.09873, (2022).
54. Saied Salem, A. S., Elbadawy, R., Mukhlis, E., Bütün & Al-Antari, M. A. Artificial Intelligence Segmentation Framework for Identifying Significant Pathological Areas Causing Lumbar Spinal Stenosis, presented at the International conferences on Engineering and Natural Science, Republic of Korea (Jeju), (2024).
55. Salem, S. et al. A Novel AI-based Hybrid Ensemble Segmentation CAD System for Lumber Spine Stenosis pathological Regions Using MRI Axial Images, in *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–5. (2024).
56. Anaam, A., Al-antari, M. A. & Gofuku, A. A deep learning self-attention cross residual network with Info-WGANGP for mitotic cell identification in HEp-2 medical microscopic images. *Biomed. Signal Process. Control.* **86**, 105191 (2023).
57. Yang, J. et al. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Sci. Data*. **10**, 41 (2023).
58. Ukwuoma, C. C. et al. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. *Journal of Advanced Research*. **48**, 191–211 (2023).
59. Sharma, R., Eick, C. F. & Tsekos, N. V. Sm2n2: A stacked architecture for multimodal data and its application to myocardial infarction detection, in *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4*, Revised Selected Papers 11, 2021, pp. 342–350. (2020).
60. Ivantsits, M. et al. Deep-learning-based myocardial pathology detection, in *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4*, Revised Selected Papers 11, 2021, pp. 369–377. (2020).
61. Hadamitzky, M. et al. AI-based myocardial segmentation and cardiac disease classification using multi-sequence cardiac MRI. *J. Cardiovasc. Magn. Reson.* **27**, (2025).
62. Zhang, K. et al. BiomedGPT: A Unified and Generalist Biomedical Generative Pre-trained Transformer for Vision, Language, and Multimodal Tasks. arXiv preprint arXiv:2305.17100, (2023).

Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2024-00437191) supervised by the IITP(Institute for Information & Communications Technology Planning& Evaluation. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2023-00256517).

Author contributions

Mugahed A. Al-antari: Conceptualization; Methodology; Software; Data Curation; Formal Analysis; Writing - Original Draft; Writing - review & editing; Funding acquisition. Riyadh M. Al-Tam: Formal Analysis; Validation; Software. Aymen M. Al-Hejri: Formal Analysis; Validation; Software. Zaid Al-Huda: Visualization; Resources; Investigation; Software. Soojeong Lee: Resources; Validation. Özal Yıldırım: Resources; Validation; Formal Analysis; Yeong Hyeon Gu: Conceptualization; Writing - review & editing; Supervision; Funding acquisition.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.A.A.-a. or Y.H.G.