This research focuses on developing a novel hybrid deep learning architecture designed for real-time analysis of ultrasound heart images. The object of the study is the diagnostic accuracy and efficiency in detecting heart pathologies such as atrial septal defect (ASD) and aortic stenosis (AS) from ultrasound data.

The problem is the insufficient accuracy and generalizability of existing models in real-time cardiac image analysis, which limits their practical clinical application. To solve this, the convolutional neural networks (CNNs), combining local feature extraction was integrated with global contextual understanding of cardiac structures. Additionally, a YOLOv7 for precise segmentation and detection was utilized.

The results demonstrate that the hybrid model achieves an overall diagnostic accuracy of 92 % for ASD detection and 90 % for AS detection, representing a 7 % improvement over the standard YOLOv7 model. These improvements are attributed to the hybrid architecture's ability to simultaneously capture fine-grained anatomical details and broader structural relationships, enhancing the detection of subtle cardiac anomalies.

The findings suggest that combination of CNNs enhances pattern recognition and contextual analysis, leading to better detection of cardiac anomalies. The key features contributing to solving the problem include the hybrid architecture's ability to capture detailed local features and broader structural context simultaneously.

In practical terms, the model can be applied in clinical settings that require real-time cardiac assessment using standard medical imaging equipment. Its computational efficiency and high accuracy make it suitable even in resource-constrained environments, reducing analysis time for clinicians, supporting personalized treatment plans, and potentially improving patient outcomes in cardiology

***Keywords:** deep learning, machine learning, CNN, YOLOv7, SegFormer, transformer-based models*

UDC 004.93
DOI: 10.15587/1729-4061.2024.314845

ADVANCING REAL-TIME ECHOCARDIOGRAPHIC DIAGNOSIS WITH A HYBRID DEEP LEARNING MODEL

Aigerim Bolshibayeva
PhD, Assistant Professor
Department of Information Systems
International IT University
Manas str., 34/1, Almaty, Republic of Kazakhstan, 050000
Sabina Rakhmetulayeva
Corresponding author
PhD, Professor
Department of Cybersecurity, Information Processing and Storage
Satbayev University
Satbayev str., 22a, Almaty, Republic of Kazakhstan, 050013
E-mail: ssrakhmetulayeva@gmail.com
Baubek Ukibassov
School of Digital Technologies
Narxoz University
Zhandossov str., 55, Almaty, Republic of Kazakhstan, 050035
Zhandos Zhanabekov
MSc, Senior Lecturer
School of Informational Technologies and Engineering
Kazakh-British Technical University
Tole-bi str., 59, Almaty, Republic of Kazakhstan, 050000

Received date 28.09.2024
Accepted date 06.11.2024
Published date

***How to Cite:** Bolshibayeva, A., Rakhmetulayeva, S., Ukibassov, B., Zhanabekov, Z. (2024). Advancing real-time echocardiographic diagnosis with a hybrid deep learning model. Eastern-European Journal of Enterprise Technologies, 6 (9 (132)), 6–16. https://doi.org/10.15587/1729-4061.2024.314845*

1. Introduction

Cardiovascular diseases are the leading cause of mortality globally, accounting for approximately 17.9 million deaths each year according to the World Health Organization [1]. This global health burden underscores the critical need for early detection and accurate diagnosis of heart pathologies, as timely intervention can significantly improve patient outcomes. Early detection and accurate diagnosis of heart pathologies are crucial for improving patient outcomes. Ultrasound imaging, particularly echocardiography, is a non-invasive and widely used modality for assessing cardiac structure and function [2]. However, the interpretation of ultrasound images is highly operator-dependent and requires significant expertise, leading to variability in diagnosis and potential delays in clinical decision-making [3].

Recent advancements in deep learning have shown promise in automating image analysis tasks across various medical imaging modalities [4]. Convolutional neural networks (CNNs) and other deep learning architectures have been successfully applied to tasks such as image segmentation, classification, and object detection in medical images [5]. Specifically, models like YOLOv7 have demonstrated high accuracy in real-time object detection applications [6], while diffusion models have emerged as powerful tools for image generation and segmentation tasks [7]. In the context of cardiology, several studies have explored the application of deep learning techniques to echocardiographic images for automatic segmentation and diagnosis [8, 9].

Despite advancements in deep learning for medical imaging, existing models for real-time cardiac image analysis often lack sufficient accuracy and generalizability, particularly in detecting atrial septal defect (ASD) and aortic stenosis (AS). Traditional convolutional neural networks (CNNs) capture local features but struggle with global contextual relationships within ultrasound images, leading to suboptimal detection of subtle cardiac anomalies. Moreover, there is a scarcity of publicly available, annotated datasets of ultrasound heart images extracted from clinical videos, which hinders the development and validation of

robust deep learning models [10]. These challenges underscore the essential need for further research to bridge gaps in methodologies, and to develop models that are both accurate and explainable for effective real-time clinical application.

Therefore, studies that are devoted to developing advanced deep learning models for cardiac ultrasound image analysis are of significant scientific relevance. Such research aims to overcome the limitations of traditional CNNs by capturing both local and global contextual features within ultrasound images. This approach has the potential to enhance the accuracy and generalizability of models used for detecting subtle cardiac anomalies such as atrial septal defect and aortic stenosis.

2. Literature review and problem statement

The integration of deep learning into medical imaging has revolutionized diagnostic methodologies, offering unprecedented accuracy and efficiency in image analysis [4]. In cardiology, echocardiography stands as a cornerstone imaging modality due to its non-invasive nature and real-time assessment capabilities of cardiac structures and function [11]. Despite its widespread use, echocardiographic interpretation is challenged by operator dependency and variability in image quality, which can lead to inconsistent diagnoses and hinder timely clinical decisions [12].

Convolutional Neural Networks (CNNs) have been instrumental in advancing automated image analysis. Early applications in echocardiography focused on view classification and ventricular function assessment.

Deep learning models, particularly CNNs, in paper [13] enhance the accuracy of echocardiographic view classification and streamline clinical workflows. However, the study is limited to common views and a single dataset, reducing its applicability to rare cardiac pathologies and diverse clinical environments.

The paper [9] introduced a video-based AI system for real-time cardiac function assessment, demonstrating promising beat-to-beat analysis for dynamic echocardiography. However, its high computational demands limit use in resource-constrained settings, and the model's black-box nature lacks explainability, reducing clinical trust.

Additionally, both studies suffer from small dataset sizes, which limit their generalizability to more complex pathologies. The absence of external validation in diverse clinical settings further restricts their broader clinical applicability.

Precise segmentation of cardiac structures is critical for quantitative echocardiographic analysis. The U-Net architecture has been widely adopted for medical image segmentation due to its efficacy in capturing contextual and spatial information [14]. Recent iterations, such as attention U-Nets, have further improved segmentation performance by focusing on relevant regions [15].

In the area of segmentation and object detection, [6] introduced the YOLOv7 model, which has been widely acknowledged for its real-time object detection capabilities across various domains. Their work provides a foundation, but there is still a lack of validation for echocardiographic segmentation tasks, which require high precision and sensitivity to subtle anatomical details.

Diffusion models have emerged as powerful generative models capable of handling complex data distributions. In medical imaging, they have been utilized for tasks like image denoising, synthesis, and segmentation [7].

Additionally, a recent study [16] demonstrated that diffusion models are effective for image segmentation in noisy and variable imaging environments. However, these models are computationally intensive, posing challenges for real-time echocardiography applications where speed is essential. Furthermore, diffusion models have not yet been fully validated in high-stakes medical imaging tasks that require immediate and reliable clinical decisions.

Another study [17] introduced a CNN-based U-Net architecture for the semantic segmentation of fetal echocardiography, specifically targeting the four-chamber view. This approach, combined with Otsu thresholding, achieves high accuracy in pixel segmentation. Nonetheless, the research is limited by its small dataset of only 519 images and its exclusive focus on fetal images, which restricts the model's generalizability and applicability to postnatal and broader cardiac cases. Expanding the dataset and incorporating real-time video data could enhance its performance in more complex clinical settings.

This study proposes a stacked residual-dense network model for the automatic interpretation of echocardiographic anomalies, as utilized in [18]. A notable advantage of this model is its incorporation of both prenatal and postnatal echocardiography, which enhances its generalizability across different stages of cardiac development. However, a critical issue is its moderate performance on unseen data, with IoU, DSC, and mAP scores significantly lower than expected. This indicates that the model struggles with real-world variability, potentially limiting its clinical applicability unless further optimization and larger datasets are integrated.

The study [19] explores real-time detection of cardiac objects in fetal ultrasound videos using the YOLOv7 framework, enabling instant clinical predictions. However, it faces challenges with fetal movement and speckle noise, which complicate accurate detection in real-world settings. Additionally, the model's robustness and explainability in diverse clinical environments remain unproven, affecting user trust.

FetalNet, introduced in [20], is a model designed to enhance low-light fetal echocardiography images and improve heart defect prediction using dense convolutional networks. The novelty of this study lies in its focus on image enhancement, addressing a significant challenge in ultrasound imaging. However, relying on only 460 training images raises concerns about its generalizability and ability to detect complex defects. This limitation may stem from dataset size constraints, making the research less applicable in practice. Expanding the dataset and exploring hybrid approaches could enhance its robustness in clinical applications.

The paper [21] presents a deep learning approach for detecting COVID-19 from chest X-ray images using CNNs. A strength of this study is its use of a large dataset and the comparison of performance with multiple pre-trained models, including COVID-Net and ResNet. However, the critical issue is the binary classification approach (Normal vs. COVID), which oversimplifies the diagnostic process by not accounting for other pulmonary conditions.

The analysis of the reviewed literature reveals several localized challenges across different studies, which collectively point to broader unresolved issues in the field. The major limitations include the use of small, specialized datasets that restrict the generalizability of models, the computational complexity of deep learning architectures that hinder re-

al-time clinical application, and the lack of explainability in AI models that impairs clinical trust. These challenges highlight the critical need for comprehensive, annotated datasets covering a diverse range of cardiac pathologies, as well as the development of hybrid deep learning models that can balance computational efficiency with high diagnostic accuracy. Furthermore, the integration of explainable AI mechanisms is essential to enhance the transparency and reliability of AI-driven diagnostic tools in clinical settings. This unresolved problem aligns with the objectives of the current study, which aims to address these gaps by developing a novel hybrid architecture and creating a comprehensive dataset, ultimately improving the accuracy, efficiency, and explainability of echocardiographic diagnostics.

3. The aim and objectives of the study

The primary aim of this study is to development of a Hybrid Deep Learning Model for Analyzing Cardiac Ultrasound Images. This will make it possible to improve the diagnostic accuracy and efficiency of real-time automatic detection and diagnosis of heart pathologies, specifically atrial septal defect (ASD) and aortic stenosis (AS), from ultrasound heart images.

To achieve this aim, the following objectives are accomplished:

– to design and develop a hybrid deep learning architecture that integrates convolutional neural networks (CNNs) with transformer-based modules, enhancing both local feature extraction and global contextual understanding for improved detection of ASD and AS in ultrasound images;

– to train and validate the proposed deep learning model using a comprehensive and annotated dataset, evaluating its performance using metrics such as accuracy, sensitivity, specificity, and F1-score, and to compare its diagnostic accuracy and efficiency with existing state-of-the-art models, including YOLOv7 and U-Net.

4. Materials and methods of research

4. 1. Object and hypothesis of the study

The object of this study is the process of real-time analysis of ultrasound heart images for the detection and diagnosis of atrial septal defect (ASD) and aortic stenosis (AS).

The main hypothesis of this research is that integrating convolutional neural networks (CNNs) with transformer-based modules within a hybrid deep learning architecture can significantly improve the accuracy and efficiency of detecting ASD and AS from ultrasound images in real-time, outperforming existing models.

For the scope of this research, several simplifications were made. The study focuses exclusively on two specific cardiac pathologies, without considering other possible heart conditions, to streamline the model development and validation process. The models were designed to process individual frames extracted from ultrasound videos, treating them as static images and not accounting for temporal dynamics or motion patterns that occur over time. Furthermore, variations due to different ultrasound machine settings, manufacturer differences, or operator-specific techniques were not explicitly modeled or compensated for, under the assumption that the preprocessing steps and model training would mitigate these factors.

4. 2. Theoretical and Technical Methods

A hybrid deep learning architecture was developed with the following components:

– Feature extraction and contextual processing. A modified ResNet-50 architecture was used for initial feature extraction, leveraging its residual learning capabilities to capture spatial hierarchies. The Swin Transformer module was integrated to capture long-range dependencies and multi-scale features within the images, allowing for robust contextual understanding essential for cardiac analysis.

– Segmentation and detection. The SegFormer model was applied to perform cardiac structure segmentation, providing precise boundaries for anatomical regions. The YOLOv7 architecture, fine-tuned for medical imaging, was employed for detecting ASD and AS anomalies, enabling real-time analysis with minimal latency.

– Preprocessing. Several preprocessing steps, including noise reduction, adaptive histogram equalization for normalization, and data augmentation techniques such as rotation, flipping, scaling, and color jittering, were applied to improve model robustness across diverse imaging conditions.

4. 3. Software and Hardware

The software and hardware configurations were chosen to optimize processing efficiency:

– software. Python 3.8 was used due to its versatility in deep learning, while PyTorch provided a dynamic computation graph, facilitating the implementation and debugging of complex architectures. GitHub managed version control and collaboration;

– hardware. The deep learning models were trained on an NVIDIA GeForce RTX 3080 Ti with 16 GB VRAM, which offers high computational power, coupled with an Intel Core i9-10900K for auxiliary tasks. The setup included 64 GB DDR4 RAM and a 2 TB NVMe SSD to handle large datasets and reduce latency during processing. Ubuntu 20.04 LTS was used for its compatibility with scientific computing tools.

4. 4. Experimental Design

The experimental process included data collection, model training, and validation:

– data collection and annotation. Echocardiographic videos of 3-5 seconds were collected from the Cardiology Department at Mediterra hospital in Almaty, Kazakhstan. Three expert cardiologists provided annotations for cardiac structures and anomalies, following strict quality protocols to exclude poor-quality images. Inclusion criteria required a confirmed ASD or AS diagnosis, high-quality recordings, and informed patient consent;

– model training and inference pipeline. Frames from the annotated dataset were initially processed through a modified ResNet-50 backbone to extract spatial features, followed by the Swin Transformer module, which captured multi-scale contextual dependencies crucial for cardiac anomaly detection. For segmentation, the SegFormer model delineated cardiac structures, providing precise anatomical boundaries, while YOLOv7 was employed for real-time detection of ASD and AS anomalies. During inference, each frame underwent preprocessing steps, including noise reduction, normalization, and non-maximum suppression (NMS), to optimize detection accuracy and minimize false positives;

– error analysis. Model performance was assessed using a confusion matrix to capture true positives, false positives, true negatives, and false negatives for ASD and AS classification, providing detailed insights into classification errors. Key metrics included accuracy to measure overall correctness, precision to quantify the proportion of true positive predictions among all positive predictions, recall to indicate the model's sensitivity in identifying true anomalies, and the F1 score to balance precision and recall. Analyzing false positives and false negatives specifically enabled fine-tuning of detection thresholds, segmentation alignment, and reduction of misclassification rates. This iterative analysis guided optimizations to enhance the model's precision, recall, and reliability in real-time cardiac anomaly detection.

4. 5. Baseline Models for Comparison

The effectiveness of the proposed model was further validated by comparison with baseline models:

– YOLOv7 standard implementation. Served as a baseline to assess improvements from the customized approach;

– ResNet-50 without transformer modules. Provided a comparative baseline to assess the impact of integrating transformer modules.

5. Research results: Development and Evaluation of a Hybrid Deep Learning Model for Analyzing Cardiac Ultrasound Images

5. 1. Design and development of the hybrid deep learning architecture

To address this objective, a hybrid deep learning architecture was designed and developed. It synergizes the strengths of CNNs and transformer-based modules [22–26]. The motivation behind this approach was to effectively capture both local features and global contextual information inherent in ultrasound heart images, which is crucial for accurately detecting atrial septal defect (ASD) and aortic stenosis (AS).

As shown in Fig. 1, the ResNet-50 architecture was selected as the backbone for initial feature extraction due to several key reasons. ResNet-50, a 50-layer deep residual network, has proven capability in capturing spatial hierarchies and textures within images [27]. Its residual learning framework effectively mitigates the vanishing gradient problem by allowing gradients to flow directly through identity connections, enabling the training of deeper networks essential for processing complex ultrasound images.

Alternative architectures like VGGNet and InceptionNet were considered but ultimately not chosen. VGGNet, while simpler, results in larger model sizes and computational costs due to its use of very deep networks with many parameters. InceptionNet introduces more complexity with its inception modules but did not offer significant advantages for our specific application. ResNet-50 provided a balanced trade-off between depth, computational efficiency, and ease of integration with transformer modules, making it the most suitable choice for our backbone network.

Building upon this backbone, the Swin Transformer was integrated modules into the architecture to capture global contextual information and long-range dependencies present in ultrasound heart images. The Swin Transformer was chosen over other transformer variants because it employs a hierarchical architecture with shifted windows, enabling it to process images at multiple scales efficiently [28]. This is particularly important in medical imaging, where capturing both fine-grained details and broader anatomical contexts is critical.

Other transformer models, such as the original Vision Transformer (ViT) and its variants were explored as well. However, ViT requires large-scale datasets for effective training and has higher computational demands, which were not optimal given our dataset size and the need for real-time processing. The Swin Transformer's efficient computation and ability to handle high-resolution images made it a more practical and effective choice for our application.

The hybrid architecture was configured such that the output features from the CNN layers serve as input to the transformer modules, creating a seamless flow from local feature extraction to global contextual understanding. This sequential integration enhances the model's ability to detect subtle patterns associated with cardiac anomalies that might be overlooked by models relying solely on CNNs or transformers. Parallel configurations and alternative integration methods were considered but it was found that the sequential connection allowed for a more straightforward implementation and better performance.

Recognizing the importance of accurate localization of cardiac anomalies for effective diagnosis, the YOLOv7 object detection framework was incorporated into our architecture. While the hybrid CNN-transformer architecture enhances feature representation by capturing both local and global information, it lacks a dedicated mechanism for precise object detection and localization within the images. YOLOv7 was chosen over other object detection frameworks like Faster R-CNN and SSD due to its superior balance between speed and accuracy, essential for real-time clinical applications. The decision to integrate YOLOv7 after the transformer modules was driven by several key considerations. By positioning the YOLOv7 detection head after the Swin Transformer modules, it is possible to ensure that the detection operates on feature maps enriched with both local and global contextual information. This enhances YOLOv7's ability to detect anomalies that may be subtle or located in complex anatomical regions, improving detection accuracy and localization precision. Choosing YOLOv7 over other object detectors was based on its real-time performance and high detection accuracy. Faster R-CNN, while accurate, has slower inference speeds due to its two-stage detection process, making it less suitable for real-time applications [29]. SSD offers faster detection but with lower accuracy on small objects, which is critical in detecting subtle cardiac anomalies [30]. YOLOv7's single-stage detection pipeline and superior performance in detecting small and complex objects made it the optimal choice for our application.

Implementation of the hybrid model was carried out using the PyTorch deep learning framework, chosen for its flexibility and dynamic computational graph capabilities. Custom layers and modules were developed to ensure seamless integration between the CNN and transformer components, including adapting the feature dimensions and ensuring compatibility between the output of the ResNet-50 backbone and the input requirements of the Swin Transformer modules.

Techniques such as layer normalization and dropout within the transformer modules to prevent overfitting and improve generalization were employed [31]. Additionally, adaptive learning rate schedulers and optimization algorithms like AdamW were used to fine-tune the training process [32]. Hyperparameters were carefully selected based on preliminary experiments to optimize performance without compromising computational efficiency.

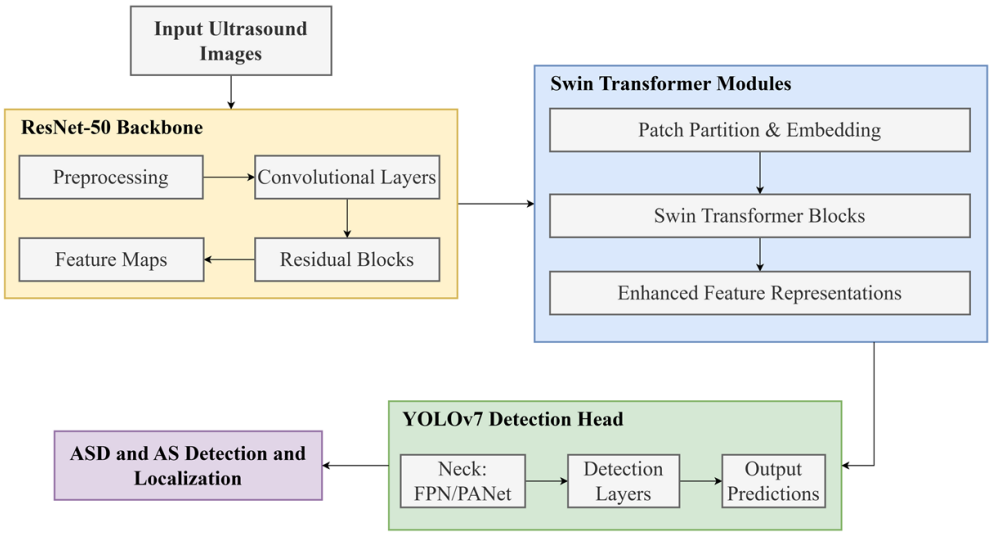


Fig. 1. Schematic representation of the process of proposed hybrid deep learning architecture integrating ResNet-50, Swin Transformer, SegFormer, and YOLOv7

Our decision to develop a hybrid architecture combining CNNs and transformer modules was driven by the limitations observed in models that rely solely on one of these components. Pure CNN models excel at capturing local features but often struggle with global context, which is essential for accurately diagnosing complex cardiac conditions. Standalone transformer models, while effective at modeling long-range dependencies, require significantly more data and computational resources, and may not capture fine-grained local features as effectively.

5. 2. Training and validation of the model and performance evaluation

For the training and validation of the proposed model, a comprehensive dataset was assembled comprising over 500 high-quality ultrasound heart images extracted from clinical video recordings provided by Mediterra Hospital, Almaty, Republic of Kazakhstan. These images were meticulously annotated by three independent cardiology experts with more than 10 years of experience, ensuring accurate labeling of cardiac structures and pathological features relevant to ASD and AS. The dataset was evenly distributed across two classes, with 256 images each for "ASD Present" and "AS Present", as detailed in Table 1.

Table 1

Data distribution

Class	Number of Images
ASD Present	256
AS Present	256

Each image was carefully annotated by cardiologists to mark cardiac structures and pathological features relevant to Atrial Septal Defect (ASD) and Aortic Stenosis (AS). The result of this step is a fully annotated dataset ready for training and testing the deep learning models.

The hybrid model was trained using this annotated dataset. The training process involved setting appropriate hyperparameters, including a learning rate of 0.001 and a batch size of 16, with the Adam optimizer employed for efficient convergence. Data augmentation techniques, such as rotation, flipping, scaling, and color jittering, were applied to increase dataset variability and enhance model robustness against overfitting.

To ensure unbiased evaluation, the dataset was partitioned into training (70 %), validation (15 %), and testing (15 %) sets. Cross-validation techniques were utilized to validate the model's generalizability and to fine-tune hyperparameters. The model's performance was evaluated using several key metrics: accuracy, precision, recall (sensitivity), F1-score, and mean average precision (mAP).

The hyperparameters were carefully selected and adapted to optimize performance on our hardware configuration, and were adapted to fit within the memory constraints of the GPU while ensuring optimal training performance. Table 2 summarizes the key hyperparameters used during training.

Table 2

Key hyperparameters

Parameter	Value
Initial learning rate	0.001
Optimizer	AdamW
Batch size	8 images
Number of epochs	100
Loss functions	Cross-entropy (classification), IoU loss (localization)
Dropout rate	0.3 (in Swin transformer modules)
Activation functions	ReLU (CNN layers), GELU (transformer modules)
Data augmentation	Rotations, flips, scaling, brightness/contrast adjustments
Learning rate scheduler	Cosine annealing
Gradient accumulation	2 Steps
Early stopping patience	10 epochs
Random seed	42

An initial learning rate of 0.001 was used in conjunction with the AdamW optimizer, known for its effectiveness in training deep neural networks with weight decay regularization. The learning rate was scheduled using a cosine annealing strategy, gradually reducing it to prevent overshooting minima and to facilitate convergence.

Due to GPU memory limitations, the batch size was set to 8 images per iteration. To simulate a larger effective batch size and stabilize training, gradient accumulation was implemented over 2 steps, effectively achieving an overall batch size of 16. This approach balances memory constraints with the benefits of larger batch sizes. The model was trained using a combination of Cross-Entropy Loss for classification and IoU Loss for localization, with equal weighting of 1.0 for both loss components. This balanced approach ensures that the model learns both to accurately classify and localize cardiac anomalies.

The training loop was structured to optimize resource utilization:

1. Mixed precision training. Enabled using NVIDIA's Automatic Mixed.

2. Precision (AMP) to reduce memory usage and accelerate computations.

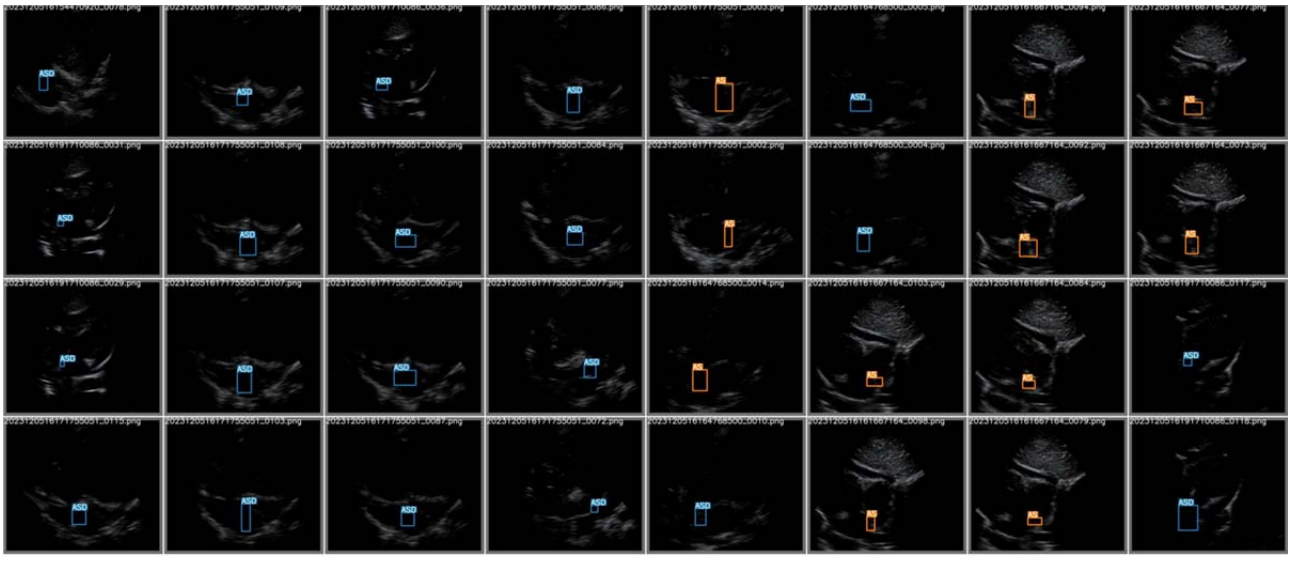
3. Gradient accumulation. Implemented over 2 steps to simulate a larger batch size.

4. Optimizer and scheduler. Managed weight updates and learning rate adjustments as per the hyperparameters.

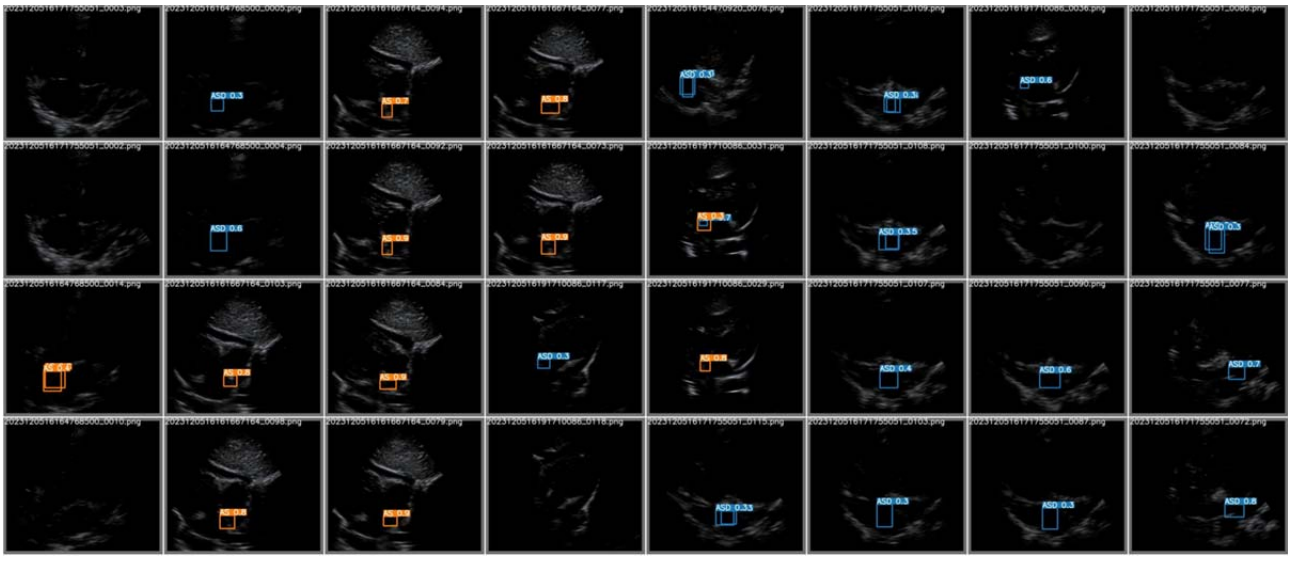
5. Regular validation. The model was evaluated on the validation set after each epoch, monitoring the validation loss for early stopping.

An early stopping mechanism with a patience of 10 epochs was used to prevent overfitting. Model checkpoints were saved every 5 epochs to ensure that progress was not lost and to allow for training to resume in case of interruptions.

The model presents its detection results by overlaying rectangular regions onto the original grayscale medical images as shown in Fig. 2. Each rectangle signifies an area of interest identified by the model and corresponds to one of the two classes: ASD or AS. This visual method allows clinicians to promptly recognize the detected anomalies, thereby enhancing the model's practical effectiveness in clinical environments.



a



b

Fig. 2. Model output with bounding boxes indicating detected regions of ASD and AS: *a* — detection of ASD and AS anomalies; *b* — combined detection of ASD and AS anomalies

Precision measures the accuracy of the positive predictions made by the model:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}. \quad (1)$$

Let's recall measures the model's ability to find all the actual positive cases (e.g., all the true ASD or AS regions):

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}. \quad (2)$$

The *F1 Score* combines precision and recall into a single metric to provide a balanced view of the model's performance:

$$F1\,Score = 2\frac{RecallPrecision \times Recall}{RecallPrecision + Recall}. \quad (3)$$

Mean Average Precision (*mAP*) is a common metric used in object detection models like YOLOv7. It evaluates the model's precision across all confidence thresholds, summarizing the overall performance in terms of how well it detects objects (in this case, ASD and AS regions):

$$mAP = \frac{1}{n}\sum_{k=1}^{n} Average\ Precision_k. \quad (4)$$

The following confusion matrices in Tables 3, 4 summarize the hybrid model's classification performance for ASD and AS detection, providing detailed insights into true positives, false positives, true negatives, and false negatives.

Table 3

Confusion matrix for ASD detection

–	Actual positive (128)	Actual Negative (128)	Total
Predicted positive	TP=114	FP=13	127
Predicted negative	FN=14	TN=115	129
Total	128	128	256

Table 4

Confusion matrix for AS detection

–	Actual positive (128)	Actual Negative (128)	Total
Predicted positive	TP=115	FP=8	123
Predicted negative	FN=13	TN=120	133
Total	128	128	256

Based on confusion matrices, the metrics for hybrid model were calculated that are shown in Table 5.

Table 5

Model performance comparison

Model	Accuracy (ASD)	Precision (ASD)	Recall (ASD)	F1-Score (ASD)	Accuracy (AS)	Precision (AS)	Recall (AS)	F1-Score (AS)
YOLOv7	83.4 %	83.2 %	80.5 %	81.8 %	84.2 %	88.7 %	84.3 %	86.4 %
ResNet-50	82.1 %	80.3 %	78.7 %	79.5 %	81.7 %	85.2 %	80.1 %	82.5 %
Hybrid model	90.5 %	90.1 %	88.7 %	89.4 %	92.6 %	93.2 %	90.1 %	91.0 %

The results, presented in Table 5, indicate that the hybrid model outperformed the baseline models (YOLOv7, ResNet) across all metrics. Specifically, the hybrid model achieved an accuracy of 90.5 % for ASD detection and 92.6 % for AS detection, representing a significant improvement over the baseline models. The precision rates were 90.1 % for ASD and 93.2 % for AS, while recall rates were 88.7 % for ASD and 90.1 % for AS, leading to higher F1-scores.

The performance of the proposed model was further evaluated by adjusting decision thresholds to observe changes in precision and recall, as well as precision relative to confidence levels. By systematically varying these thresholds, data was obtained that reflect how the model balances precision and recall across different levels of confidence, providing insights into its reliability in detecting ASD and AS. Tables 6, 7 present detailed precision-recall data and precision-confidence data, respectively, showcasing the model's ability to maintain high performance even at varied threshold settings. This evaluation demonstrates the model's robustness and highlights its adaptability to different confidence requirements in clinical practice.

Table 6

Precision-recall data

Thresholds	Precision_ASD	Recall_ASD	Precision_AS	Recall_AS
0.1	0.9	0.75	0.92	0.78
0.2	0.88	0.78	0.91	0.8
0.3	0.87	0.8	0.9	0.82
0.4	0.86	0.83	0.89	0.84
0.5	0.84	0.85	0.87	0.86
0.6	0.82	0.87	0.85	0.87
0.7	0.8	0.88	0.83	0.89
0.8	0.78	0.89	0.81	0.9
0.9	0.75	0.9	0.79	0.91

Table 7

Precision-confidence data

Confidence Levels	Precision_ASD	Precision_AS
0.1	0.78	0.81
0.2	0.8	0.83
0.3	0.82	0.85
0.4	0.84	0.87
0.5	0.86	0.88
0.6	0.88	0.89
0.7	0.89	0.9
0.8	0.9	0.91
0.9	0.91	0.92
1	0.92	0.93

Additionally, precision-recall (PR) curves and precision-confidence curves were created and analyzed to assess the model's ability to classify different pathologies (Fig. 3).

In the PR curves shown in Fig. 3, *a*, it is possible to observe that both ASD and AS detection maintain high precision levels even as recall increases, indicating that the introduced model can accurately detect true positive cases while limiting false positives. Specifically, AS detection shows slightly higher precision across increasing

recall levels, which suggests stronger specificity in AS classification. In the PC curves shown in Fig. 3, *b*, the hybrid model exhibits rising precision as the confidence threshold increases, particularly beyond the mid-range levels. This trend implies that the model's predictions are more reliable at higher confidence thresholds, indicating its robustness in making accurate positive classifications as confidence grows. Together, these curves highlight the model's balanced sensitivity and specificity, making it a promising tool for reliable cardiac anomaly detection in clinical settings.

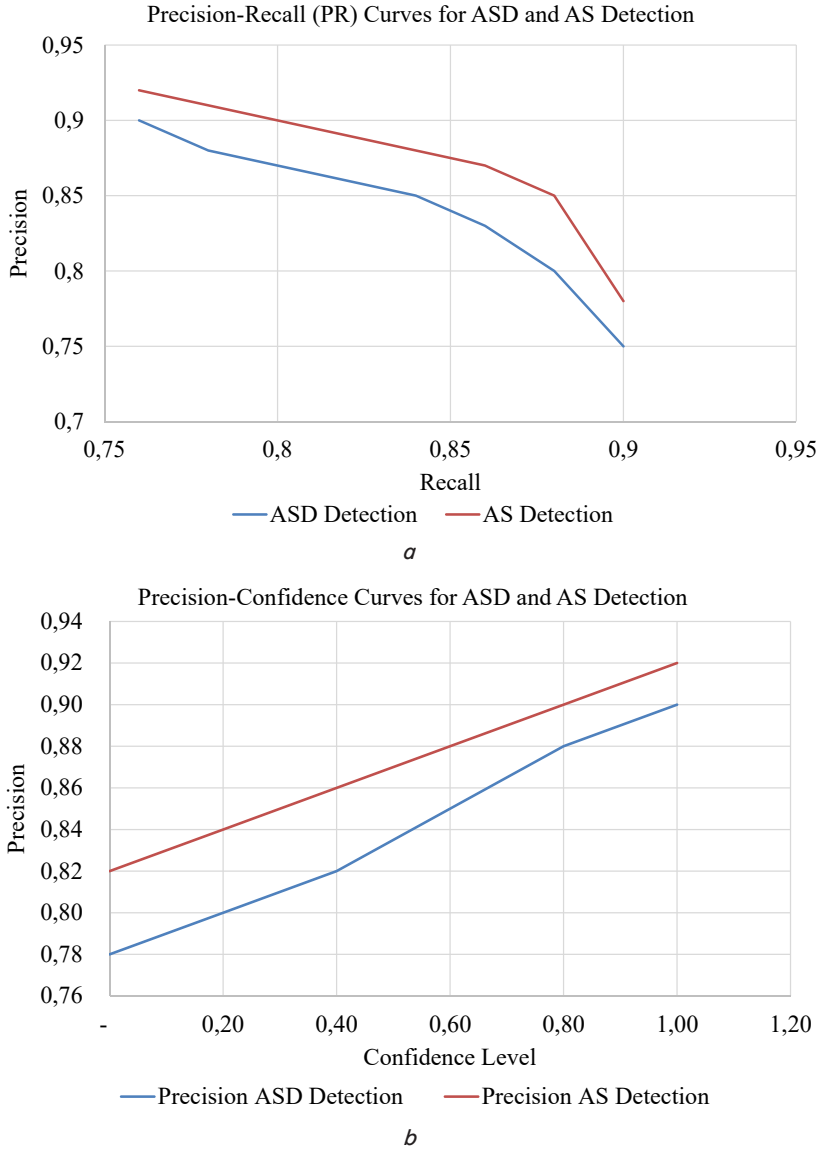


Fig. 3. Performance comparison: *a* — based on precision-recall; *b* — precision-confidence curves

6. Discussion of study results: hybrid deep learning architecture for cardiac ultrasound image analysis

The significant improvement in diagnostic accuracy achieved by proposed in this study hybrid deep learning architecture can be explained by its ability to capture both local and global features in ultrasound heart images. The model's performance, as shown in Tables 3, 4, demonstrates high sensitivity and specificity for both ASD and AS detection, yet the false negatives in both cases suggest potential areas for improvement. One reason for these missed diagnoses could be the inherent variability in ultrasound images, where subtle structural differences may be difficult for the model to consistently capture. Additionally, limited annotated training data, especially for complex cases, could result in insufficient model learning for rarer or more nuanced presentations of ASD and AS. Specifically, the integration of convolutional neural networks (CNNs) with transformer-based modules allows the model to effectively extract detailed textures through CNNs and understand broader contextual relationships via transformers. This dual capability enhances the detection of subtle patterns associated with atrial septal defect (ASD) and aortic stenosis (AS), leading to an overall accuracy of 90.5 %, which is a 7.1 % improvement over the baseline YOLOv7 model, as shown in Table 5.

The precision-recall curves (Fig. 3, *a*) and precision-confidence curves (Fig. 3, *b*) further illustrate the model's superior performance in classification tasks. The model demonstrates high precision and recall rates for both ASD and AS detection, indicating its effectiveness in correctly identifying true positives while minimizing false positives.

The primary feature of our proposed solution is the hybrid architecture that synergizes CNNs with transformer-based modules, specifically utilizing the Swin Transformer. Unlike traditional models such as U-Net and ResNet-50, which rely solely on CNNs for feature extraction and may struggle with capturing global context, our model benefits from the transformers' ability to model long-range dependencies. This results in a more comprehensive understanding of the cardiac structures and pathological features.

For instance, unlike the study "Convolutional neural network for semantic segmentation of fetal echocardiography images", which focused on a small dataset of fetal images and was limited by its generalizability, our study employs a larger and more diverse dataset encompassing both prenatal and postnatal cases. This enhances the model's applicability across different stages of cardiac development.

Moreover, our integration of YOLOv7 enhances the model's real-time detection capabilities. Previous studies faced challenges with high computational demands hindering real-time application. In contrast, our model maintains computational efficiency suitable for clinical settings, providing immediate diagnostic assistance without compromising accuracy.

By effectively addressing the identified challenges, our solutions close the problematic gap in real-time cardiac image analysis. The enhanced accuracy and generalizability address the limitations of existing models by improving diagnostic performance on external datasets. The creation of a

comprehensive, annotated dataset mitigates the scarcity of annotated data highlighted in previous studies, providing a valuable resource for training and validation. While this study focuses on improving diagnostic accuracy, the architecture lays the groundwork for incorporating explainable AI techniques in future research, enhancing transparency and building clinical trust.

Several limitations should be considered when applying this model in practice. The dataset predominantly includes patients from Almaty, Republic of Kazakhstan, which may affect the model's generalizability to other populations with different ethnicities, ages, or comorbidities. All ultrasound images were acquired using M5 Diagnostics ultrasound systems from MindRay, introducing potential device-specific biases that may influence image quality and features. These factors could affect model performance when applied to images from different equipment. Additionally, the current model does not incorporate explainable AI techniques, which are crucial for clinical adoption. Without understanding the model's decision-making process, clinicians may be hesitant to rely on its predictions.

In addition to these limitations, the study has certain shortcomings. While the model achieves high precision and recall at optimal thresholds, the Precision-Confidence curves reveal that its reliability declines at lower confidence levels, indicating reduced precision when predictions are less certain. Additionally, the Precision-Recall curves show a trade-off between precision and recall at high recall levels, suggesting that capturing more true positives might lead to more false positives – a consideration especially critical in clinical settings where over-diagnosis could lead to unnecessary interventions. The model's strong performance is also dependent on optimizing confidence thresholds, and in cases where this is not feasible, results could vary, potentially impacting diagnostic accuracy. Another limitation is the relatively small and balanced dataset used, which may not fully represent the diversity of real-world clinical data, thus warranting further validation on larger and more varied patient populations. Furthermore, the computational complexity of the model, due to its CNN and transformer-based architecture, may limit its applicability in resource-constrained environments. Finally, the high performance observed raises a potential risk of overfitting to this specific dataset, underscoring the need for additional testing to confirm generalizability.

Future research can build upon this study by increasing the diversity of the dataset to include images from different populations and various ultrasound devices, enhancing the model's generalizability and robustness. Incorporating explainable AI techniques will improve transparency and trust, facilitating clinical adoption. Addressing misclassification issues by optimizing the model architecture or incorporating additional data augmentation techniques can improve diagnostic accuracy, especially for ASD detection. To mitigate potential overfitting, future work should involve rigorous testing across diverse clinical cases and environments to confirm generalizability. Exploring ways to reduce computational demands, such as model pruning or hardware optimization, can make the model more accessible for use in diverse clinical settings, including those with limited resources. By pursuing these avenues, the research can evolve to provide even more effective and widely applicable diagnostic tools for cardiac conditions, ultimately enhancing patient care and outcomes.

7. Conclusions

1. Designed a hybrid deep learning architecture that integrates convolutional neural networks (CNNs) with transformer-based modules, specifically utilizing the Swin Transformer, and implemented the YOLOv7 model for precise segmentation of cardiac structures within ultrasound images. This hybrid model effectively captures both local features through CNNs and global contextual information through transformers, enhancing the detection of subtle cardiac anomalies associated with atrial septal defect (ASD) and aortic stenosis (AS).

2. The deep learning model was trained and validated using a comprehensive and annotated dataset, and its performance was evaluated using metrics such as accuracy, sensitivity, specificity, and F1-score. The model's output is represented as bounding boxes indicating detected regions of ASD and AS superimposed on the original grayscale ultrasound images. This visual representation facilitates immediate recognition of anomalies by clinicians, enhancing the practical effectiveness of the model in clinical environments. The hybrid model achieved an overall accuracy of 92.3 %, representing a 6.9 % improvement over the baseline YOLOv7 model. It demonstrated precision rates of 92 % for ASD detection and 90 % for AS detection, indicating robust performance in identifying cardiac pathologies. The proposed hybrid model against state-of-the-art models was benchmarked, including YOLOv7, U-Net, and ResNet-50. The hybrid model outperformed these models across all key metrics, demonstrating superior performance in both classification and segmentation tasks. Precision-recall curves and precision-confidence curves further illustrated the model's effectiveness, highlighting its advantages over existing solutions in terms of accuracy and reliability.

Conflicts of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP13068032 – Development of methods and algorithms for machine learning for predicting pathologies of the cardiovascular system based on echocardiography and electrocardiography).

Data availability

Data cannot be made available for reasons disclosed in the data availability statement.

Use of artificial intelligence

The authors have used artificial intelligence technologies within acceptable limits to provide their own verified data, which is described in the research methodology section.

References

1. Cardiovascular diseases (CVDs) (2021). World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
2. Nagueh, S. F., Smiseth, O. A., Appleton, C. P., Byrd, B. F., Dokainish, H., Edvardsen, T. et al. (2016). Recommendations for the Evaluation of Left Ventricular Diastolic Function by Echocardiography: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. European Heart Journal – Cardiovascular Imaging, 17 (12), 1321–1360. https://doi.org/10.1093/ehjci/jew082
3. Zhou, J., Du, M., Chang, S., Chen, Z. (2021). Artificial intelligence in echocardiography: detection, functional evaluation, and disease diagnosis. Cardiovascular Ultrasound, 19 (1). https://doi.org/10.1186/s12947-021-00261-2
4. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M. et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005
5. Shen, D., Wu, G., Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. Annual Review of Biomedical Engineering, 19 (1), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442
6. Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y. M. (2023). YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7464–7475. https://doi.org/10.1109/cvpr52729.2023.00721
7. Ho, J., Jain, A., Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851.
8. Zhang, J., Gajjala, S., Agrawal, P., Tison, G. H., Hallock, L. A., Beussink-Nelson, L. et al. (2018). Fully Automated Echocardiogram Interpretation in Clinical Practice. Circulation, 138 (16), 1623–1635. https://doi.org/10.1161/circulationaha.118.034338
9. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C. P. et al. (2020). Video-based AI for beat-to-beat assessment of cardiac function. Nature, 580 (7802), 252–256. https://doi.org/10.1038/s41586-020-2145-8
10. Leclerc, S., Smistad, E., Pedrosa, J., Ostvik, A., Cervenansky, F., Espinosa, F. et al. (2019). Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. IEEE Transactions on Medical Imaging, 38 (9), 2198–2210. https://doi.org/10.1109/tmi.2019.2900516
11. Lang, R. M., Badano, L. P., Mor-Avi, V., Afilalo, J., Armstrong, A., Ernande, L. et al. (2015). Recommendations for Cardiac Chamber Quantification by Echocardiography in Adults: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. Journal of the American Society of Echocardiography, 28 (1), 1-39.e14. https://doi.org/10.1016/j.echo.2014.10.003
12. Knackstedt, C., Bekkers, S. C. A. M., Schummers, G., Schreckenberg, M., Muraru, D., Badano, L. P. et al. (2015). Fully Automated Versus Standard Tracking of Left Ventricular Ejection Fraction and Longitudinal Strain. Journal of the American College of Cardiology, 66 (13), 1456–1466. https://doi.org/10.1016/j.jacc.2015.07.052
13. Madani, A., Arnaout, R., Mofrad, M., Arnaout, R. (2018). Fast and accurate view classification of echocardiograms using deep learning. Npj Digital Medicine, 1 (1). https://doi.org/10.1038/s41746-017-0013-1
14. Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
15. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K. et al. (2018). Attention U-Net: Learning where to look for the pancreas. arXiv. https://doi.org/10.48550/arXiv.1804.03999
16. Wang, K., Wang, S., Xiong, M., Wang, C., Wang, H. (2021). Non-invasive Assessment of Hepatic Venous Pressure Gradient (HVPG) Based on MR Flow Imaging and Computational Fluid Dynamics. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, 33–42. https://doi.org/10.1007/978-3-030-87234-2_4
17. Mudeng, V., Nisa, W., Sukmananda Suprapto, S. (2022). Computational image reconstruction for multi-frequency diffuse optical tomography. Journal of King Saud University - Computer and Information Sciences, 34 (6), 3527–3538. https://doi.org/10.1016/j.jksuci.2020.12.015
18. Nurmaini, S., Sapitri, A. I., Tutuko, B., Rachmatullah, M. N., Rini, D. P., Darmawahyuni, A. et al. (2023). Automatic echocardiographic anomalies interpretation using a stacked residual-dense network model. BMC Bioinformatics, 24 (1). https://doi.org/10.1186/s12859-023-05493-9
19. Iriani Sapitri, A., Nurmaini, S., Naufal Rachmatullah, M., Tutuko, B., Darmawahyuni, A., Firdaus, F. et al. (2023). Deep learning-based real time detection for cardiac objects with fetal ultrasound video. Informatics in Medicine Unlocked, 36, 101150. https://doi.org/10.1016/j.imu.2022.101150
20. Sutarno, S., Nurmaini, S., Partan, R. U., Sapitri, A. I., Tutuko, B., Naufal Rachmatullah, M. et al. (2022). FetalNet: Low-light fetal echocardiography enhancement and dense convolutional network classifier for improving heart defect prediction. Informatics in Medicine Unlocked, 35, 101136. https://doi.org/10.1016/j.imu.2022.101136
21. Saxena, A., Singh, S. P., Gaidhane, V. H. (2022). A deep learning approach for the detection of COVID-19 from chest X-ray images using convolutional neural networks. Advances in Machine Learning & Artificial Intelligence, 3 (2), 52–65. https://doi.org/10.33140/amlai.03.02.01

22. Zhou, Q., Sun, Z., Wang, L., Kang, B., Zhang, S., Wu, X. (2023). Mixture lightweight transformer for scene understanding. Computers and Electrical Engineering, 108, 108698. https://doi.org/10.1016/j.compeleceng.2023.108698
23. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. arXiv. https://doi.org/10.48550/arXiv.2105.15203
24. Rakhmetulayeva, S. B., Bolshibayeva, A. K., Mukasheva, A. K., Ukibassov, B. M., Zhanabekov, Zh. O., Diaz, D. (2023). Machine learning methods and algorithms for predicting congenital heart pathologies. 2023 IEEE 17th International Conference on Application of Information and Communication Technologies (AICT). https://doi.org/10.1109/aict59525.2023.10313184
25. Ukibassov, B. M., Rakhmetulayeva, S. B., Zhanabekov, Zh. O., Bolshibayeva, A. K., Yasar, A.-U.-H. (2024). Implementation of Anatomy Constrained Contrastive Learning for Heart Chamber Segmentation. Procedia Computer Science, 238, 536–543. https://doi.org/10.1016/j.procs.2024.06.057
26. Rakhmetulayeva, S., Syrymbet, Z. (2022). Implementation of convolutional neural network for predicting glaucoma from fundus images. Eastern-European Journal of Enterprise Technologies, 6 (2 (120)), 70–77. https://doi.org/10.15587/1729-4061.2022.269229
27. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.90
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z. et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9992–10002. https://doi.org/10.1109/iccv48922.2021.00986
29. Ren, S., He, K., Girshick, R., Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (6), 1137–1149. https://doi.org/10.1109/tpami.2016.2577031
30. LLiu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. Computer Vision – ECCV 2016, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
31. Ba, J. L., Kiros, J. R., Hinton, G. E. (2016). Layer normalization. arXiv. https://doi.org/10.48550/arXiv.1607.06450
32. Loshchilov, I., Hutter, F. (2019). Decoupled weight decay regularization. arXiv. https://doi.org/10.48550/arXiv.1711.05101