

Modality-agnostic Transfer Learning for Heterogeneous Cardiac Image Classification

Maxat Kabdullin ^{1,*}, Azat Kabdullin ¹, Lyazat Naizabayeva ², Vladimir Siládi ³,
and Assiya M. Issayeva ⁴

¹ Department of Information Systems, Satbayev University, Almaty, Kazakhstan

² Department of Information Systems, International Information Technology University (IITU), Almaty, Kazakhstan

³ Department of Computer Science, Matej Bel University, Banská Bystrica, Slovakia

⁴ Center for Innovation and Education, National Hospital of the Medical Center of the President's Affairs
Administration of the Republic of Kazakhstan, Almaty, Kazakhstan

Email: m.kabdullin@satbayev.university (M.K.); a.kabdullin@satbayev.university (A.K.);
l.naizabayeva@iitu.edu.kz (L.N.); vladimir.siladi@umb.sk (V.S.); issayeva_a@snh.kz (A.M.I.)

*Corresponding author

Abstract—Recent advances in artificial intelligence have improved automated cardiovascular image analysis, but clinical use remains *constrained* by small annotated datasets, heterogeneous imaging sources, and limited generalizability across acquisition protocols. This study evaluates a modality-agnostic transfer learning pipeline for binary classification of echocardiographic and cardiac magnetic resonance images under data-limited conditions. Images were standardized using resizing, normalization, noise reduction, and data augmentation. A Visual Geometry Group 16 (VGG16) network pretrained on ImageNet was adapted with global average pooling, a dense classification block, dropout regularization, and a sigmoid output layer. Stratified 5-fold cross-validation was performed on the internal dataset. External evaluation used the Automated Cardiac Diagnosis Challenge Segmentation (ACDC) for magnetic resonance imaging and Cardiac Acquisitions for Multi-structure Ultrasound (CAMUS) dataset for an ejection fraction-derived echocardiographic assessment. The pipeline achieved an internal F1-Score of approximately 0.88 and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) of approximately 0.91. External F1-Scores were approximately 0.83 on ACDC and 0.80 on CAMUS, reflecting domain shift. Gradient-weighted class activation mapping indicated attention to anatomically meaningful cardiac regions. The results support this reproducible proof-of-concept approach, while multicenter validation remains necessary before clinical application.

Keywords—cardiac image classification, heterogeneous medical imaging, modality-agnostic learning, echocardiography, cardiac Magnetic Resonance Imaging (MRI), transfer learning, external validation, Gradient-weighted Class Activation Mapping (Grad-CAM)

I. INTRODUCTION

In recent years, deep learning methods have become key tools in biomedical image analysis, showing strong

performance in classification, segmentation, and automated feature extraction, as reported in recent reviews and analytical studies on medical artificial intelligence [1–3]. In cardiology, Convolutional Neural Networks (CNNs) are widely used for the analysis of echocardiographic and cardiac magnetic resonance images, supporting automated visual assessment and reducing inter-observer variability [4]. Despite the progress achieved, the practical deployment of such solutions remains constrained by several technical factors.

One of the key limitations is the scarcity of annotated medical data. Unlike general computer vision datasets, cardiac image acquisition and annotation require substantial resources and expert involvement, which limits the scalability of deep learning approaches [5, 6]. Under small-sample conditions, training models from scratch leads to overfitting and unstable performance, thereby reducing the generalizability of the resulting models [7]. Accordingly, transfer learning and fine-tuning of pretrained models are widely adopted in contemporary studies [8].

Another important challenge is the heterogeneity of cardiac imaging data. Echocardiographic and magnetic resonance images differ substantially in their physical principles, spatial resolution, and noise characteristics [9]. In addition, recent studies indicate that strong performance on internal datasets does not guarantee robust model behavior when applied to external data obtained under different imaging protocols [5]. Therefore, developing robust and reproducible deep learning frameworks for cardiac image analysis under limited-data conditions remains an important research challenge.

The main contributions of this study can be summarized as follows:

- Development of a reproducible modality-agnostic preprocessing and classification pipeline for heterogeneous cardiac images, including

echocardiographic and cardiac Magnetic Resonance Imaging (MRI) data.

- Evaluation of transfer learning under limited-data conditions using a pretrained Visual Geometry Group 16 (VGG16) convolutional neural network and stratified 5-fold cross-validation.
- External evaluation of the trained model using independent public Automated Cardiac Diagnosis Challenge (ACDC) and the Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) datasets to assess robustness and generalization under domain shift.
- Clarification of the practical limitations of a shared modality-agnostic pipeline compared with explicit multimodal feature fusion approaches.
- Integration of Gradient-Weighted Class Activation Mapping (Grad-CAM)-based interpretability analysis for qualitative assessment of anatomically relevant model decision regions.

This work presents a technically validated proof-of-concept approach for the automatic binary classification of cardiac images (normal vs pathology) in the context of limited-data availability. A VGG16 convolutional neural network is used as the base architecture and is adapted for the analysis of echocardiographic and cardiac MRI images through the application of transfer learning. Instead of complex multimodal fusion schemes, a modality-agnostic approach is employed, in which a single model is trained on a heterogeneous set of images following unified preprocessing. Model performance is evaluated using stratified k-fold cross-validation, while generalization capability is additionally assessed on independent public datasets, namely ACDC and CAMUS.

It should be emphasized that, in this study, heterogeneous cardiac image analysis refers to the processing of different cardiac imaging sources within a shared computational pipeline. The proposed framework does not perform explicit multimodal feature fusion, late fusion, or joint modality-specific representation learning. Instead, this study evaluates whether a unified modality-agnostic transfer learning strategy can provide stable and reproducible binary classification performance across echocardiographic and cardiac MRI images under limited-data conditions.

II. LITERATURE REVIEW

Cardiovascular disease diagnosis is inherently based on multimodal imaging, including echocardiography, magnetic resonance imaging, and scintigraphy. These modalities differ substantially in their underlying physical principles and image characteristics [9]. Transparent reporting and validation standards are also important for artificial intelligence studies in medical imaging, particularly when heterogeneous datasets and external validation are involved [10]. The integration of heterogeneous data is considered a promising direction; however, most multimodal architectures require complex feature fusion mechanisms [11] and large volumes of data, which limits their applicability in small-sample settings [12].

Another important challenge discussed in the literature is model generalizability. Recent studies indicate that strong performance on internal datasets does not necessarily guarantee reliable model behavior on external data due to differences in acquisition protocols and imaging equipment [5]. Accordingly, the need for rigorous evaluation protocols, including cross-validation and external validation, is emphasized [13].

Model interpretability is also of substantial importance. In medical and engineering applications, so-called “black box” models raise justified concerns regarding the reliability and reproducibility of their decisions [14, 15]. Explainable Artificial Intelligence (XAI) methods, such as Grad-CAM, enable the analysis of spatial regions within images that have the greatest influence on model predictions and are widely used in contemporary studies [16].

The objective of this study is to evaluate the robustness, reproducibility, and generalization capability of a modality-agnostic transfer learning framework for heterogeneous cardiac image classification under limited-data conditions. The results demonstrate that, with appropriate experimental design, even a relatively simple CNN architecture can serve as a reliable technical foundation for the further development of intelligent information systems for cardiac image analysis.

III. MATERIALS AND METHODS

A. Dataset Description and Preprocessing

1) Dataset sources and binary labeling

This study used a combined dataset of biomedical cardiac images obtained from the National Hospital of the Medical Center of the President’s Affairs Administration of the Republic of Kazakhstan, comprising both echocardiographic and cardiac MRI data. All images were anonymized and annotated according to a binary classification scheme: “normal” (absence of detected cardiac pathology) and “pathological” (presence of a diagnosed cardiovascular disease). The annotations were performed by a certified cardiologist based on the corresponding clinical diagnoses. This approach is consistent with a typical formulation of automated diagnostic tasks under limited-data conditions and is widely adopted in proof-of-concept studies [1, 5, 17].

The internal dataset was limited in size, comprising several hundred images, which reflects the practical challenges of medical data acquisition and annotation and is characteristic of many applied studies in medical computer vision [18]. Accordingly, particular attention was paid to the proper design of the experimental protocol and to the model validation procedures. To achieve stable and representative performance estimates under small-sample conditions, stratified k-fold cross-validation ($k = 5$) was employed, ensuring class balance within each fold and allowing the entire available dataset to be utilized for both training and testing [19].

2) External datasets and label construction

In addition to internal cross-validation, external validation was performed using two publicly available benchmark datasets that are widely used in research to evaluate algorithms for cardiac image analysis. The first external dataset was ACDC, which contains 150 cardiac cine-MRI studies from 150 patients, categorized into five clinical classes, including normal cases and various forms of cardiac pathology [20]. The ACDC dataset was used exclusively for independent evaluation of the model on MRI data and was not employed during the training stage.

For the binary classification protocol used in this study, the original ACDC diagnostic categories were mapped into two classes. The normal category was assigned to the “normal” class, whereas all disease-related categories, including dilated cardiomyopathy, hypertrophic cardiomyopathy, myocardial infarction, and abnormal right ventricle cases, were merged into the “pathological” class, in accordance with previously proposed binary diagnostic approaches based on this dataset [21]. This conversion was adopted to match the screening-oriented binary classification objective of the present proof-of-concept study. It should be noted that this binary formulation does not aim to distinguish between specific cardiac pathologies and may obscure clinically meaningful differences between disease subtypes. Therefore, multiclass classification of cardiac pathologies is considered an important direction for future work.

The second external source was the CAMUS dataset, a large-scale echocardiographic dataset comprising two- and

four-chamber cardiac views from 500 patients with detailed annotations of cardiac structure contours [22]. Although CAMUS was originally designed for segmentation tasks, it has been widely used in research for the analysis of cardiac functional status and for evaluating the generalization capability of ultrasound image processing algorithms [23]. In this study, the CAMUS dataset was used as an external echocardiographic benchmark to assess the robustness and transferability of the proposed framework without additional fine-tuning.

Although CAMUS was originally developed for cardiac structure segmentation and functional assessment rather than for direct binary disease classification, it provides expert manual annotations and clinical functional information, including Left Ventricular Ejection Fraction (LVEF). For the supplementary binary evaluation in this study, CAMUS cases were grouped according to LVEF-based functional status: patients with LVEF lower than 45% were considered as the pathological-risk group, whereas patients with LVEF higher than 55% were considered as the normal-function group. Cases with intermediate LVEF values were not used for binary label assignment to avoid ambiguity. Therefore, the CAMUS-based results should be interpreted as an LVEF-derived external echocardiographic assessment rather than as a direct disease-subtype classification experiment.

The dataset composition is summarized in Table I, and the overall datasets and validation protocol are illustrated in Fig. 1.

TABLE I. SUMMARY OF DATASETS USED IN THE STUDY

Dataset	Modality	Data Type	Patients	Images	Normal	Pathology	Usage
Internal (Echo)	Echo	2D (ES/ED)	200	515	240	275	Train + Val
Internal (MRI)	MRI	Cine MRI (SA)	45	555	275	280	Train + Val
ACDC	MRI	Cine MRI (SA)	150	~1500	30	120	External Test
CAMUS	Echo	2D (2CH/4CH)	500	1000	LVEF > 55%	LVEF < 45%	External Echo Assessment

Note: For CAMUS, the Patients and Images columns describe the full publicly available dataset, whereas the Normal and Pathology columns indicate the LVEF-based criteria used for binary grouping in the supplementary echocardiographic assessment. Cases with intermediate LVEF values were excluded from binary label assignment to avoid ambiguous functional-status classification.

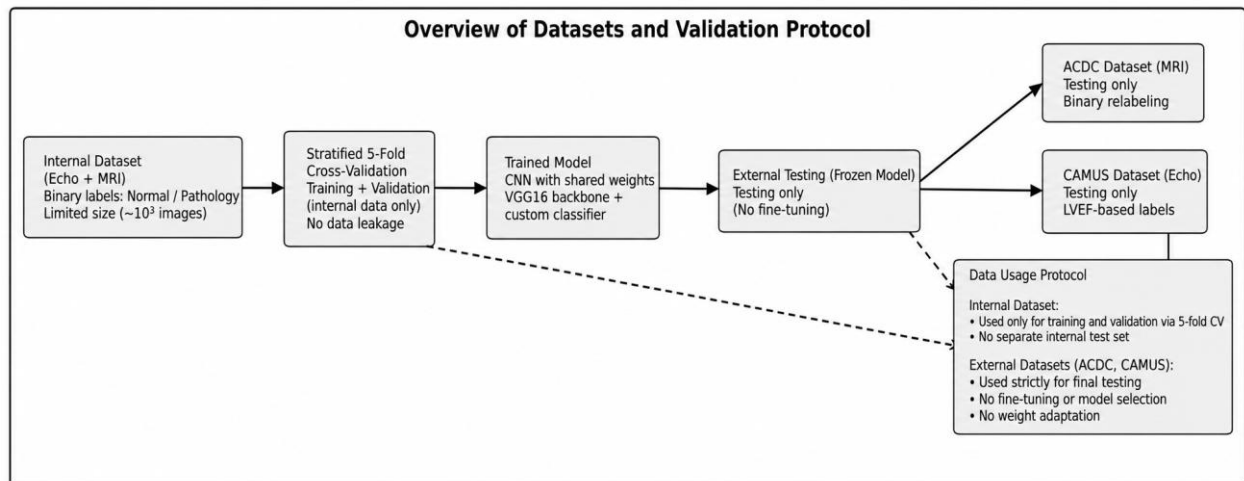


Fig. 1. Schematic overview of the datasets and validation protocol used in this study.

It is important to note that external datasets were used exclusively at the evaluation stage, without any weight adaptation or retraining. ACDC was used for independent MRI-based external validation, whereas CAMUS was used for LVEF-derived external echocardiographic assessment. This protocol enables an objective assessment of the model's ability to generalize to data that differ in source, acquisition protocols, and imaging equipment characteristics, which is a critically important requirement in the development of algorithms for real-world clinical scenarios [5].

The ACDC and CAMUS datasets are publicly available from their respective repositories. The internal clinical dataset is not publicly available because of institutional and data-protection restrictions.

3) Unified preprocessing and augmentation

After dataset preparation, all images underwent a unified preprocessing procedure to ensure consistent model training and data comparability across imaging modalities. The preprocessing procedure included modality-specific noise reduction, pixel intensity normalization, spatial standardization, and data augmentation. Representative input images and the unified preprocessing pipeline are shown in Fig. 2.

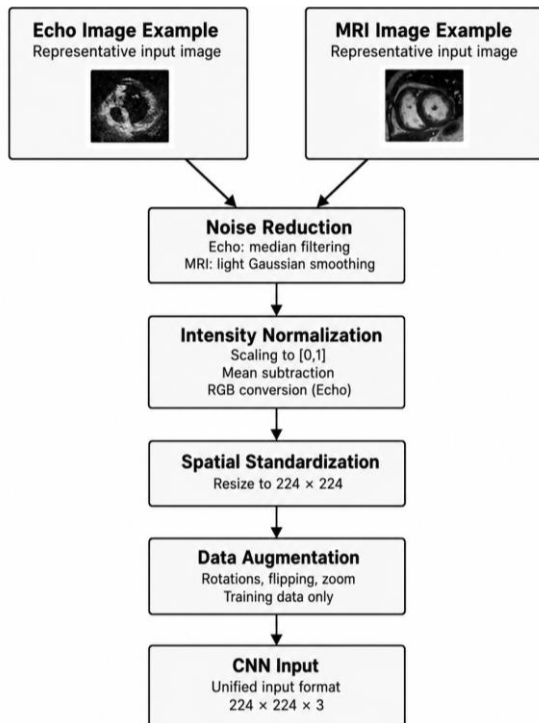


Fig. 2. Unified image preprocessing pipeline with representative input examples.

Spatial standardization was performed by resizing all images to a uniform spatial resolution. The original frames were resized to 224×224 pixels, which corresponds to the input requirements of widely used convolutional architectures such as VGG16 and Residual Network (ResNet) [24]. The use of a fixed spatial resolution simplifies the application of pretrained models and ensures

comparability of the extracted features across different image types.

Pixel intensity normalization was also performed. Pixel intensity values were scaled to the $[0, 1]$ range, after which the mean value computed from the training set was subtracted from the images. For models pretrained on the ImageNet dataset, standard channel-wise normalization was additionally applied (mean subtraction and division by the standard deviation for each channel), aligning the input data statistics with the distribution on which the original model was trained [24]. Since echocardiographic images are originally represented in grayscale, they were replicated into a three-channel Red-Green-Blue (RGB) format to ensure compatibility with architectures designed for color images.

Modality-specific noise was also addressed. Echocardiographic images are characterized by pronounced speckle noise, which can hinder the extraction of informative features [25]. To reduce this noise, median filtering with a 3×3 window was applied, as it can effectively suppress speckle noise while preserving anatomical boundaries. For MRI images, which exhibit a higher signal-to-noise ratio, only mild smoothing was applied when necessary, using a Gaussian filter with a small standard deviation to suppress high-frequency noise without compromising diagnostically relevant details.

To improve model robustness and reduce the risk of overfitting under limited-data conditions, data augmentation was applied during training. The use of data augmentation is a standard practice when training deep neural networks on small medical datasets and has been shown to be effective in improving model generalization [26, 27]. During training, the following transformations were randomly applied to the input images:

- horizontal flipping.
- random rotations within a $\pm 15^\circ$ range.
- scaling (zooming) within $\pm 10\%$.
- small spatial translations.
- brightness and contrast adjustments.
- addition of weak Gaussian noise or simulated speckle noise.
- random cropping followed by resizing to the original input dimensions.

The augmentation parameters were selected to ensure that the synthetically generated images remained clinically plausible and did not distort the anatomical structure of the heart. Data augmentation was applied exclusively to the training data and was not used during validation or testing, in accordance with commonly accepted recommendations for the proper evaluation of model performance [5].

B. Model Architecture and Training Procedure

The model architecture, classification head, transfer learning strategy, and training configuration are described together in this subsection to provide a coherent overview of the proposed learning pipeline.

1) Base architecture and classification head

VGG16 was selected as the base architecture for biomedical cardiac image classification because it is one

of the most extensively studied and widely used architectures in image analysis tasks [24]. Despite its relatively simple structure by current standards, VGG16 remains widely used in medical imaging because of its training stability, predictable behavior, and transferability of pretrained weights [24].

VGG16 consists of sequential convolutional blocks with small 3×3 filters followed by subsampling operations, enabling hierarchical feature extraction from local edges to high-level structural patterns [24]. The original model was trained on ImageNet, which contains more than 1.2 million images, making it a suitable starting point for limited-data tasks.

In this work, a pretrained VGG16 model was used, and the original fully connected classifier layers designed for the 1,000 ImageNet classes were removed. On top of the convolutional base, a custom classification block was added and adapted for the binary classification task (“normal” vs “pathological”).

Heterogeneous image processing was implemented through a simplified and reproducible pipeline consistent with the objectives of a proof-of-concept study. A single convolutional neural network with a shared architecture and shared weights was trained on a heterogeneous set of cardiac images from different modalities, namely echocardiography and magnetic resonance imaging. This design evaluated whether a common transfer learning pipeline could extract task-relevant features across different cardiac imaging domains without adding architectural complexity. Specialized multi-branch architectures, explicit feature-level fusion, decision-level fusion, or joint modality-specific representation learning were not employed in this study. Instead, a modality-agnostic learning approach was adopted, in which the model was trained to extract robust and discriminative features relevant to the binary classification task, regardless of the imaging source. Therefore, the proposed framework should be interpreted as a shared heterogeneous-image classification pipeline rather than as a full multimodal fusion architecture.

In this work, modality heterogeneity is interpreted in an engineering sense: different cardiac imaging sources are processed within a unified computational pipeline rather than through architectural feature fusion. This interpretation is consistent with the objectives and constraints of a proof-of-concept study.

To ensure correct joint data processing, images from both modalities underwent a unified preprocessing stage and were converted to a standardized input format (spatial resolution, number of channels, and normalization). As a result, echocardiographic and MRI images were treated as distinct input data domains processed by the same model. Such an approach enables the assessment of the ability of a standard CNN architecture to generalize across heterogeneous medical images without increasing architectural complexity and is widely adopted in studies focused on robustness and reproducibility under limited-data conditions [28–30]. An overview of the proposed modality-agnostic CNN architecture is illustrated in Fig. 3.

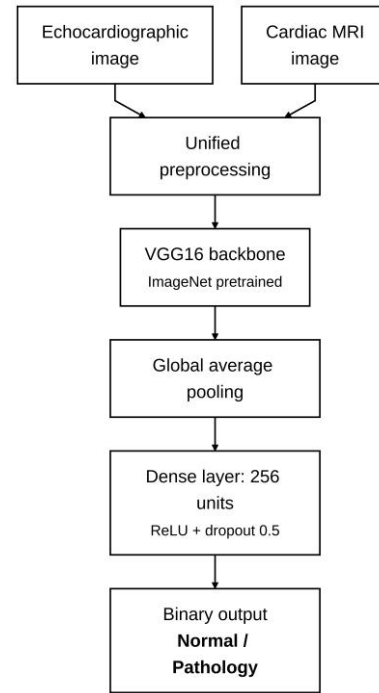


Fig. 3. Modality-agnostic CNN architecture for heterogeneous cardiac image classification.

To reduce the number of trainable parameters and mitigate overfitting, a Global Average Pooling (GAP) layer was used instead of a standard fully connected layer. GAP aggregates spatial features across each channel while preserving semantic information and reducing the dimensionality of the feature representation [31].

Following the GAP layer, the following components were added:

- a fully connected layer with 256 neurons and ReLU activation.
- a Dropout layer with a rate of 0.5 for regularization.
- a single-neuron output layer with a sigmoid activation function, producing the probability of an image belonging to the “pathological” class.

This classifier configuration is widely used in medical applications and provides a favorable balance between training stability and generalization performance in small-sample settings [1].

2) Transfer learning and fine-tuning protocol

Given the limited size of the internal dataset, training the model from scratch would likely be unstable and prone to overfitting. Therefore, this study adopted a transfer learning strategy using pretrained weights from the VGG16 convolutional backbone [24].

Model training was conducted in two stages:

a) Convolutional layer freezing phase

In the initial stage, all VGG16 convolutional layers were frozen, and training was restricted to the newly added classification block. This strategy allowed the high-level feature representations to be adapted to the new task without disrupting the universal filters learned from ImageNet.

b) Fine-tuning phase

After the validation metrics stabilized, the upper convolutional blocks of the model, namely the last one to two VGG16 blocks, were unfrozen, and the network was fine-tuned using a reduced learning rate. This approach enables careful adaptation of features to the specific characteristics of medical images without causing abrupt degradation of previously learned weights [24].

Such a staged training strategy is widely used in medical imaging tasks and is considered one of the most reliable approaches when working with small datasets [24]. The two-stage transfer learning strategy adopted in this study is schematically illustrated in Fig. 4.

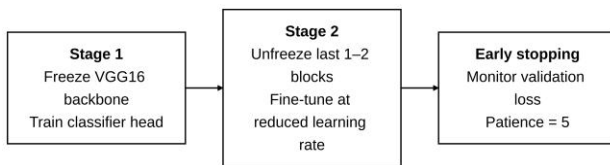


Fig. 4. Two-stage transfer learning strategy.

3) Training configuration and hyperparameters

Model training was performed using the TensorFlow/Keras framework. Binary cross-entropy was used as the loss function, as it is standard for binary classification tasks. Parameters were optimized using the Adam algorithm, which has demonstrated high stability and efficiency in training deep neural networks [32].

The main training hyperparameters were as follows:

- Initial learning rate: 1×10^{-4} (during classifier training).
- Learning rate during fine-tuning: 1×10^{-5} .
- Batch size: 32.
- Maximum number of epochs: 25.

To prevent overfitting, early stopping was applied based on validation loss dynamics. Training was automatically terminated when no improvement was observed over several consecutive epochs. In this study, early stopping was configured to monitor the validation loss ('val_loss') with a patience of 5 epochs, and the model weights corresponding to the lowest validation loss were restored for final evaluation.

C. Experimental Design and Evaluation Protocol

The experimental design combined internal cross-validation, external evaluation, complementary performance metrics, and confusion matrix analysis to assess model robustness under limited-data conditions.

1) Cross-validation procedure

Given the limited size of the internal dataset, model training and evaluation were performed using stratified 5-fold cross-validation. At each iteration, four folds were used for model training and one-fold was used for validation, while class proportions ("normal" vs "pathological") were preserved in each split. Final performance metrics were reported as the mean value and standard deviation across all five folds. This scheme helps reduce the dependence of results on a specific data split and provides more stable estimates of model performance under small-sample conditions [19].

2) External validation strategy

To assess the generalization capability of the trained model, external evaluation was performed using two publicly available datasets. The ACDC dataset was used for independent MRI-based external validation. The CAMUS dataset was used for LVEF-derived external echocardiographic assessment, based on functional grouping of cases according to left ventricular ejection fraction thresholds. Neither external dataset was involved in model training or fine-tuning, and both were used exclusively for evaluation. Such an evaluation protocol is consistent with recommendations for preventing data leakage and provides a more realistic assessment of model robustness under domain shift, which remains a major challenge in deep learning-based medical image analysis [33, 34].

3) Performance metrics

Given the binary nature of the task ("normal" vs "pathological") and the potential clinical significance of classification errors, a set of complementary performance metrics was used to evaluate model performance. The use of multiple metrics helps avoid an overly narrow interpretation of the results and provides a more comprehensive characterization of model behavior [35]. The selection of these metrics is also motivated by their widespread applicability in evaluating binary classifiers in engineering and information systems operating with noisy and limited-volume data.

Accuracy was used as the primary aggregate performance metric and was defined as the proportion of correctly classified images relative to the total number of samples. However, given the potential asymmetry in the cost of errors in medical tasks, accuracy alone is insufficient for a comprehensive evaluation of model performance [35].

Therefore, the following metrics were additionally computed:

- Precision (positive class precision)—the proportion of truly pathological images among all images classified by the model as "pathological".
- Recall (sensitivity)—the proportion of correctly identified pathological cases among all true pathological instances.
- F1-Score, which is particularly informative under limited-data conditions and is widely used as a key performance indicator in medical imaging studies [36].

In addition, the Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) was computed to analyze the model's ability to discriminate between classes across different decision thresholds. This metric reflects the probability that the model assigns a higher score to a pathological image than to a healthy one and is widely used for evaluating binary classifiers in biomedical applications [35].

4) Confusion matrix analysis

For a more detailed analysis of classification errors, a confusion matrix was constructed to summarize True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions. Analysis of the

confusion matrix allows identification of characteristic error types and assessment of which cases pose the greatest challenges for the model [35].

A confusion matrix was constructed for the internal cross-validation procedure, while external evaluation was summarized using aggregate performance metrics. Such analysis is particularly important in the context of medical applications, where false-negative errors may have more serious consequences than false-positive predictions [35].

IV. RESULTS AND DISCUSSION

Classification performance was evaluated on the internal dataset using stratified 5-fold cross-validation. For each iteration, the main performance metrics were computed, after which the results were averaged across all folds and reported as mean $\pm$ standard deviation. This approach enabled stable performance estimation under limited-data conditions and reduced the dependence of results on a specific data split.

Model optimization dynamics are shown in Figs. 5 and 6. The loss curves in Fig. 5 indicate stable convergence and support the use of early stopping, while the accuracy curves in Fig. 6 show consistent performance improvement across epochs without substantial divergence between training and validation.

The results indicate stable and balanced model performance. The small standard deviations of the performance metrics across folds indicate limited sensitivity to the choice of the training subset, which is particularly important when working with small medical datasets [19]. An F1-Score of approximately 0.88 indicates a favourable balance between sensitivity to pathological cases and false-positive predictions.

Recall slightly exceeds precision, indicating that the model is more oriented toward detecting pathological cases while allowing a moderate number of false positive classifications. In the context of cardiological diagnostics, such asymmetry is considered acceptable and often

preferable, as missing a pathology may have more serious consequences than referring a patient for additional examination [1]. The quantitative results obtained from the stratified 5-fold cross-validation on the internal dataset are summarized in Table II. The model demonstrates stable and balanced performance across all evaluated metrics, with relatively low standard deviations, indicating robustness with respect to different data splits. The substantial performance gap between the proposed VGG16-based framework and the dummy baselines confirms that the model learns meaningful discriminative representations rather than relying on class distribution bias.

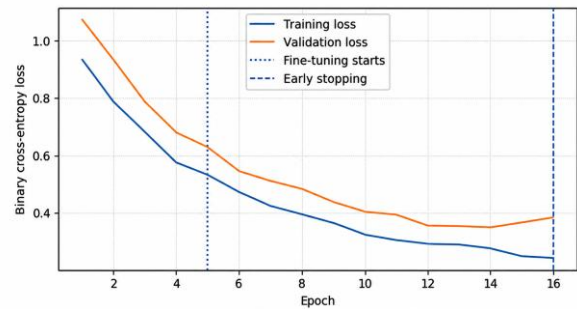


Fig. 5. Training and validation loss with fine-tuning and early stopping.

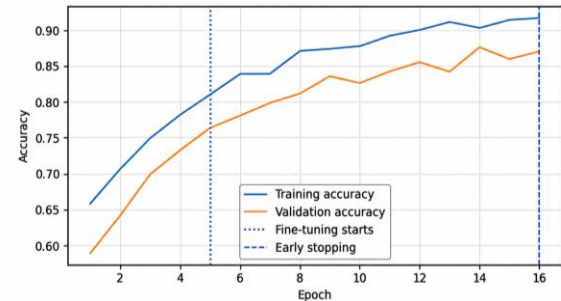


Fig. 6. Training and validation accuracy with fine-tuning and early stopping.

TABLE II. PERFORMANCE METRICS OF THE VGG16 MODEL ON THE INTERNAL DATASET OBTAINED VIA STRATIFIED 5-FOLD CROSS-VALIDATION (MEAN $\pm$ STANDARD DEVIATION)

Model	Accuracy	Precision	Recall	Specificity	F1-Score	ROC-AUC
Dummy (Most frequent)	0.52 $\pm$ 0.00	-	-	-	-	0.50 $\pm$ 0.01
Dummy (Stratified)	0.50 $\pm$ 0.02	-	-	-	-	0.50 $\pm$ 0.01
VGG16 (proposed)	0.89 $\pm$ 0.03	0.87 $\pm$ 0.04	0.90 $\pm$ 0.03	0.89 $\pm$ 0.03	0.88 $\pm$ 0.03	0.91 $\pm$ 0.02

For reference, baseline performance was estimated using dummy classifiers. The most-frequent baseline corresponds to predicting the majority class (pathology), yielding an accuracy of approximately 52%, consistent with the class distribution in the internal dataset. In addition, the proposed model achieved a specificity of 0.89, indicating reliable identification of normal cardiac cases alongside high sensitivity to pathology.

A. Confusion Matrix Analysis

To further analyze model behavior, a confusion matrix averaged across all cross-validation folds was constructed. The analysis showed that most classification errors

occurred in borderline cases, where visual differences between normal and pathological images were less pronounced.

The proportion of False-Positive (FP) predictions remains moderate, which is reflected in a precision value of approximately 0.87. False-Negative (FN) predictions account for a smaller proportion, confirming the high recall value. The confusion matrix obtained for internal cross-validation is shown in Fig. 7.

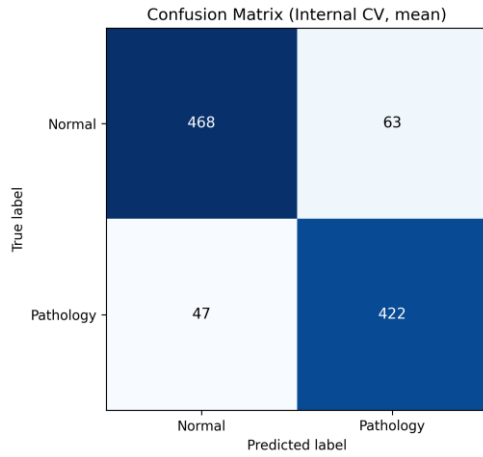


Fig. 7. Confusion matrix for internal cross-validation.

B. External Validation and LVEF-Derived Echocardiographic Assessment

To assess the model's generalization capability, independent external evaluation was performed using the ACDC and CAMUS datasets. ACDC was used as an MRI-based external validation dataset, whereas CAMUS was used as an LVEF-derived echocardiographic assessment dataset based on functional grouping according to left ventricular ejection fraction thresholds. Neither dataset was used during training, cross-validation, or fine-tuning. The resulting performance metrics are summarized in Table III.

TABLE III. PERFORMANCE METRICS OBTAINED DURING EXTERNAL VALIDATION AND LVEF-DERIVED ECHOCARDIOGRAPHIC ASSESSMENT

Dataset	Accuracy	F1-Score	ROC-AUC
ACDC (MRI external validation)	0.84	0.83	0.87
CAMUS (Echo, LVEF-derived assessment)	0.81	0.80	0.85

Compared with internal cross-validation, external evaluation showed a moderate decrease in performance. Accuracy decreased by approximately 5% on ACDC and by approximately 8% on CAMUS, while the F1-Score decreased by approximately 0.05–0.08. This degradation is expected and reflects the effect of domain shift caused by differences in dataset origin, acquisition protocols, imaging equipment, image quality and annotation criteria [33]. The lower performance on CAMUS compared with ACDC may also be explained by the different nature of the external labels. In ACDC, the binary grouping was derived from diagnostic categories, whereas in CAMUS the binary assessment was based on left ventricular ejection fraction thresholds rather than direct disease-subtype labels. In addition, echocardiographic images are typically more variable and noise-sensitive than cardiac magnetic resonance images, which may further contribute to the larger performance decrease on CAMUS. Therefore, the CAMUS-based results should be interpreted as a supplementary LVEF-derived echocardiographic assessment rather than as a direct disease classification experiment. Nevertheless, the

model-maintained performance above random chance on both external datasets, supporting the robustness of the proposed modality-agnostic pipeline as a proof-of-concept solution.

Given the limited data volume and the small number of cross-validation folds, rigorous statistical comparisons may produce unstable and difficult-to-interpret results. Therefore, within the scope of this proof-of-concept study, the emphasis was placed on the reproducibility and stability of performance metrics across cross-validation folds and external datasets, rather than on formal statistical ranking of models.

C. Model Interpretability via Grad-CAM

To assess the qualitative interpretability of the proposed model, Grad-CAM visualizations were generated for representative echocardiographic images from the CAMUS dataset. This analysis identified image regions that most strongly influenced the model's predictions.

Fig. 8 shows Grad-CAM activation maps overlaid on original echocardiographic inputs. The highlighted regions predominantly correspond to anatomically relevant cardiac structures, particularly the left ventricular area, suggesting that the model focuses on clinically meaningful features rather than background artifacts.

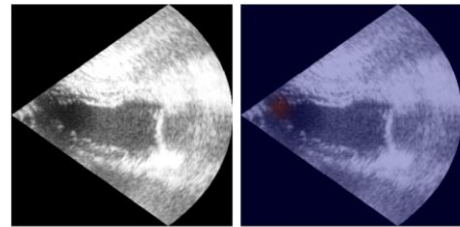


Fig. 8. Grad-CAM visualization for echocardiographic images from the CAMUS dataset. Left: original echocardiographic input image. Right: corresponding Grad-CAM heatmap overlaid on the input, highlighting regions that contributed most to the model's prediction.

D. Comparison with Alternative Architectures

In additional experiments, the baseline VGG16 architecture was compared with more modern and compact convolutional neural networks—50-layer Residual Network (ResNet50) and MobileNet version 2 (MobileNetV2)—trained under an identical experimental protocol on the same internal dataset. This comparison was not intended to identify the best-performing architecture in absolute terms, but rather to assess the robustness and reproducibility of the results under limited-data conditions. Consistent with earlier findings [37], an architecture of moderate complexity, when properly adapted to the task, demonstrates comparable performance while exhibiting better stability of metrics across cross-validation folds.

Table IV presents the quantitative comparison results. ResNet50 showed a slight advantage in terms of ROC-AUC; however, it also exhibited substantially higher variability of performance metrics across cross-validation folds. Such variability may be attributed to the higher

model complexity and, consequently, increased sensitivity to data partitioning under limited sample size conditions.

TABLE IV. COMPARISON OF DIFFERENT CNN ARCHITECTURES ON THE INTERNAL DATASET (MEAN $\pm$ STANDARD DEVIATION OVER 5-FOLD CROSS-VALIDATION)

Model	Accuracy	F1-Score	ROC-AUC
VGG16	0.89 ± 0.03	0.88 ± 0.03	0.91 ± 0.02
ResNet50	0.88 ± 0.05	0.87 ± 0.05	0.92 ± 0.03
MobileNetV2	0.88 ± 0.05	0.84 ± 0.04	0.88 ± 0.03

The differences between the evaluated architectures (VGG16, ResNet50, and MobileNetV2) were not subjected to formal statistical significance testing within this study (e.g., t-test or Wilcoxon test). Accordingly, the observed differences in performance metrics should be interpreted as qualitative trends rather than as statistically validated evidence of one architecture outperforming another. In contrast, MobileNetV2 has substantially fewer parameters and lower computational complexity; however, it demonstrates reduced accuracy and F1-Score compared with VGG16. This may indicate a limited ability of the model to extract stable, diagnostically relevant features in the considered problem setting.

In the context of medical applications and proof-of-concept studies, such stability is often a more important factor than marginal improvements in individual metrics, particularly under limited-data conditions.

Based on the results obtained, the VGG16 architecture was selected as the baseline model, providing a balanced trade-off between classification performance, training stability, and interpretability. Thus, the selection of VGG16 was driven not by the pursuit of maximum metric values, but by a focus on the reproducibility and reliability of the computational pipeline, in line with the objectives and constraints of the present study.

E. Discussion

The overall results indicate that transfer learning and convolutional neural networks can effectively address binary classification of cardiac images, even under limited-data availability. An important outcome of this study is not only the achievement of satisfactory performance metrics under internal cross-validation, but also the maintenance of acceptable performance on external datasets, which is rarely demonstrated in studies of this class [33, 34].

The obtained results are also consistent with recent studies emphasizing the importance of robust feature representation, reproducible classification protocols, and clinically meaningful interpretation in healthcare-oriented machine learning. Karimi et al. emphasized that feature relevance and representation quality are central to improving classification reliability, computational efficiency, and interpretability in large medical databases [38]. In contrast to explicit feature selection strategies, the present study relies on a pretrained convolutional network to perform automatic hierarchical feature extraction from cardiac images, which is more suitable for image-based classification tasks but remains sensitive to dataset size and domain shift. Kurdi

demonstrated the applicability of machine learning-based classification frameworks in healthcare monitoring systems, including approaches based on handcrafted image descriptors and supervised classifiers [39]. Compared with such pipelines, the proposed CNN-based framework reduces the need for manual feature engineering, although it requires careful validation under heterogeneous imaging conditions. Iftikhar et al. showed that hybrid machine learning approaches combining Support Vector Machine (SVM) with evolutionary optimization can support heart disease classification and risk factor identification using structured clinical data [40]. Unlike these tabular-data-oriented studies, the present work focuses specifically on cardiac image classification and evaluates a modality-agnostic transfer learning pipeline using internal cross-validation, external evaluation, and Grad-CAM-based interpretability analysis.

Overall, the external validation results indicate that generalization performance depends not only on the model architecture, but also on dataset origin, label construction, and modality-specific image characteristics. This supports the proof-of-concept interpretation of the study and highlights the need for further modality-specific validation before clinical application.

The present study should be interpreted as a proof-of-concept evaluation rather than as a finalized clinical diagnostic system. The main objective was not to achieve maximum classification performance, but to test whether a simple modality-agnostic preprocessing and transfer learning pipeline could provide stable and reproducible behavior under heterogeneous and data-limited conditions. For this reason, the experimental design emphasized stratified cross-validation, external evaluation, comparison with alternative architectures, and qualitative interpretability analysis rather than aggressive model optimization. This conservative strategy reduces the risk of overestimating model performance and provides a more realistic assessment of the technical feasibility of the proposed approach [17]. Further validation on larger multicentre datasets, including more detailed modality-specific and multiclass diagnostic analysis, is required before considering clinical application.

V. CONCLUSION AND FUTURE WORK

This study presents a modality-agnostic transfer learning pipeline for heterogeneous cardiac image classification under limited-data conditions. The proposed approach was designed as a proof-of-concept framework rather than as a finalized clinical diagnostic system. The results showed that a unified preprocessing and classification pipeline based on transfer learning can provide stable internal performance and maintain acceptable performance under external evaluation, despite moderate degradation caused by domain shift.

The main significance of the study lies in demonstrating that a relatively simple and reproducible pipeline can support binary cardiac image classification across heterogeneous imaging sources without explicit multimodal feature fusion. The external evaluation on independent cardiac magnetic resonance imaging and

echocardiographic datasets further highlighted both the potential and the limitations of this approach. Grad-CAM-based visual analysis showed that the model focused mainly on anatomically meaningful cardiac regions, supporting the qualitative interpretability of the learned representations.

Several limitations should be emphasized. The internal dataset was limited in size, the classification task was formulated as a binary normal-versus-pathological problem, and the CAMUS-based evaluation relied on left ventricular ejection fraction-derived grouping rather than direct disease-subtype labels. Therefore, the reported results should be interpreted as preliminary technical evidence of feasibility and reproducibility, not as evidence of clinical readiness.

Future work will focus on expanding the dataset, incorporating multicenter clinical data, performing more detailed modality-specific analysis, and moving toward multiclass classification and functional cardiac assessment. Additional validation using larger independent cohorts and more advanced architectures may further improve robustness and support the development of transparent intelligent decision-support systems for cardiac image analysis.

ETHICS APPROVAL

The study protocol was reviewed and approved by the Research Ethics Committee of the Association of Bioethics and Medical Law, Almaty, Kazakhstan (Protocol No. ER-66-01-26; review date: January 28, 2026). The study was classified as retrospective and non-interventional, and the Committee determined that it posed no greater than minimal risk and approved a waiver of re-consent. All biomedical images used in this study were anonymized before analysis and processed exclusively for research purposes in accordance with applicable data protection principles. No personally identifiable patient information was used in the study. The external ACDC and CAMUS datasets used in this study are publicly available anonymized research datasets.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Maxat Kabdullin conceptualized the study, designed the methodology, implemented the deep learning framework, performed the experiments, analysed the results, and wrote the manuscript; Vladimír Siládi contributed to the scientific supervision, methodological review, and manuscript editing; Lyazat Naizabayeva contributed to literature analysis, data preparation, and manuscript review; Azat Kabdullin participated in software implementation, validation, and experimental analysis; Assiya M. Issayeva contributed to clinical data collection, annotation, and medical interpretation of the results. All authors reviewed and approved the final version of the manuscript.

FUNDING

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant No. AP25794129, “Development of an Algorithm for Filtering and Preprocessing Biomedical Images for Cardiology Diagnostics”).

ACKNOWLEDGMENT

The authors thank the National Hospital of the Medical Center of the President’s Affairs Administration of the Republic of Kazakhstan for supporting the data collection process used in this study.

REFERENCES

- [1] E. J. Topol, “High-performance medicine: The convergence of human and artificial intelligence,” *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019. doi: 10.1038/s41591-018-0300-7
- [2] S. Bennett *et al.*, “Development and validation of echocardiography artificial intelligence models: A narrative review,” *Journal of Clinical Medicine*, vol. 14, no. 19, 7066, 2025. doi: 10.3390/jcm14197066
- [3] G. Sanjeevi *et al.*, “Deep learning supported echocardiogram analysis: A comprehensive review,” *Artificial Intelligence in Medicine*, vol. 151, 102866, 2024. doi: 10.1016/j.artmed.2024.102866
- [4] P. L. Myhre *et al.*, “Artificial intelligence-enhanced echocardiography in cardiovascular disease management,” *Nature Reviews Cardiology*, vol. 23, no. 3, pp. 164–182, 2025. doi: 10.1038/s41569-025-01197-0
- [5] M. Roberts *et al.*, “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans,” *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199–217, 2021. doi: 10.1038/s42256-021-00307-0
- [6] B. Recht *et al.*, “Do ImageNet classifiers generalize to ImageNet?” in *Proc. International Conference on Machine Learning*, 2019, pp. 5389–5400.
- [7] A. B. Khalifa *et al.*, “Deep transfer learning for classification of late gadolinium enhancement cardiac MRI images into myocardial infarction, myocarditis, and healthy classes,” *Diagnostics*, vol. 15, no. 2, 207, 2025. doi: 10.3390/diagnostics15020207
- [8] N. Srikijkasemwat, M. Villarreal, and A. Banerjee, “Multi-phase deep learning model for automated disease classification from cardiac cine MRI,” *J. R. Soc. Interface*, vol. 22, no. 231, 20250303, 2025. doi: 10.1098/rsif.2025.0303
- [9] L. Alwan *et al.*, “Current and evolving multimodality cardiac imaging in managing transthyretin amyloid cardiomyopathy,” *JACC: Cardiovascular Imaging*, vol. 17, no. 2, pp. 195–211, 2024. doi: 10.1016/j.jcmg.2023.10.010
- [10] A. S. Tejani *et al.*, “Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update,” *Radiology: Artificial Intelligence*, vol. 6, no. 4, e240300, 2024. doi: 10.1148/ryai.240300
- [11] S.-C. Huang *et al.*, “Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: A case-study in pulmonary embolism detection,” *Scientific Reports*, vol. 10, no. 1, 22147, 2020. doi: 10.1038/s41598-020-78888-w
- [12] C. F. Baumgartner *et al.*, “PHiSeg: Capturing uncertainty in medical image segmentation,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 119–127. doi: 10.1007/978-3-030-32245-8_14
- [13] B. Kocak *et al.*, “Evaluation metrics in medical imaging AI: Fundamentals, pitfalls, misapplications, and recommendations,” *European Journal of Radiology Artificial Intelligence*, vol. 3, 100030, 2025. doi: 10.1016/j.ejrai.2025.100030
- [14] D. Muhammad and M. Bendechache, “Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis,” *Computational and Structural Biotechnology Journal*, vol. 24, pp. 542–560, 2024. doi: 10.1016/j.csbj.2024.08.005

- [15] E. H. Houssein *et al.*, “Explainable artificial intelligence for medical imaging systems using deep learning: A comprehensive review,” *Cluster Computing*, vol. 28, no. 7, 469, 2025. doi: 10.1007/s10586-025-05281-5
- [16] R. R. Selvaraju *et al.*, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020. doi: 10.1007/s11263-019-01228-7
- [17] C. J. Kelly *et al.*, “Key challenges for delivering clinical impact with artificial intelligence,” *BMC Medicine*, vol. 17, no. 1, 195, 2019. doi: 10.1186/s12916-019-1426-2
- [18] G. Maicas *et al.*, “Pre and post-hoc diagnosis and interpretation of malignancy from breast DCE-MRI,” *Medical Image Analysis*, vol. 58, 101562, 2019. doi: 10.1016/j.media.2019.101562
- [19] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” *BMC Bioinformatics*, vol. 7, no. 1, 91, 2006. doi: 10.1186/1471-2105-7-91
- [20] O. Bernard *et al.*, “Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?” *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018. doi: 10.1109/TMI.2018.2837502
- [21] J. M. Wolterink *et al.*, “Automatic segmentation and disease classification using cardiac cine MR images,” in *Proc. International Workshop on Statistical Atlases and Computational Models of the Heart*, vol. 10663, pp. 101–110, 2018. doi: 10.1007/978-3-319-75541-0_11
- [22] S. Leclerc *et al.*, “Deep learning for segmentation using an open large-scale dataset in 2D echocardiography,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2198–2210, 2019. doi: 10.1109/TMI.2019.2900516
- [23] D. Ouyang *et al.*, “Video-based AI for beat-to-beat assessment of cardiac function,” *Nature*, vol. 580, pp. 252–256, 2020. doi: 10.1038/s41586-020-2145-8
- [24] S. Kornblith, J. Shlens, and Q. V. Le, “Do better ImageNet models transfer better?” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2661–2671. doi: 10.1109/CVPR.2019.00277
- [25] F. J. Putz *et al.*, “Sonographic 3-D power doppler imaging enhances rapid assessment of morphologic and pathologic arteriovenous fistula variations,” *Ultrasound in Medicine and Biology*, vol. 47, no. 6, pp. 1484–1494, 2021. doi: 10.1016/j.ultrasmedbio.2021.02.014
- [26] J. Nalepa, M. Marcinkiewicz, and M. Kawulok, “Data augmentation for brain-tumor segmentation: A review,” *Frontiers in Computational Neuroscience*, vol. 13, 83, 2019. doi: 10.3389/fncom.2019.00083
- [27] L. M. Hein *et al.*, “Metrics reloaded: recommendations for image analysis validation,” *Nature Methods*, vol. 21, pp. 195–212, 2024. doi: 10.1038/s41592-023-02151-z
- [28] S.-C. Huang *et al.*, “Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines,” *NPJ Digital Medicine*, vol. 3, no. 1, 136, 2020. doi: 10.1038/s41746-020-00341-z
- [29] J. R. Zech *et al.*, “Variable generalization performance of a deep learning model to detect pneumonia,” *PLoS Medicine*, vol. 15, no. 11, e1002683, 2018. doi: 10.1371/journal.pmed.1002683
- [30] V. Gulshan *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016. doi: 10.1001/jama.2016.17216
- [31] B. Zhou *et al.*, “Learning deep features for discriminative localization,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929. doi: 10.1109/CVPR.2016.319
- [32] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv Preprint*, arXiv:1412.6980, 2015. doi: 10.48550/arXiv.1412.6980
- [33] D. F. Narro *et al.*, “Unsupervised characterization of temporal dataset shifts as an early indicator of AI performance variations,” *JMIR Medical Informatics*, vol. 13, no. 1, e78309, 2025. doi: 10.2196/78309
- [34] S. Arun *et al.*, “Systematic scoping review of external validation studies of AI pathology models for lung cancer diagnosis,” *NPJ Precision Oncology*, vol. 9, no. 1, 166, 2025. doi: 10.1038/s41698-025-00940-7
- [35] M. E. Klontzas *et al.*, “ESR Essentials: common performance metrics in AI–Practice recommendations by the European society of medical imaging informatics,” *European Radiology*, vol. 36, no. 2, pp. 1528–1540, 2025. doi: 10.1007/s00330-025-11890-w
- [36] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLOS One*, vol. 10, no. 3, e0118432, 2015. doi: 10.1371/journal.pone.0118432
- [37] A. Kabdullin, M. Kabdullin, and L. Naizabayeva, “Optimizing neural network performance to predict coronary heart disease,” in *Proc. IEEE International Conference on Smart Information Systems and Technologies (SIST)*, 2021, pp. 2656–2666. doi: 10.1109/SIST50301.2021.9465925
- [38] M. Karimi *et al.*, “Feature selection methods in big medical databases: A comprehensive survey,” *International Journal of Theoretical and Applied Computational Intelligence*, vol. 2025, pp. 181–209, 2025. doi: 10.65278/IJTACI.2025.21
- [39] S. Z. Kurdi, “Machine learning–based classification framework for human health care monitoring,” *International Journal of Theoretical & Applied Computational Intelligence*, vol. 2026, pp. 1–15, 2026. doi: 10.65278/IJTACI.2026.1
- [40] S. Ifthikhar *et al.*, “An evolution-based hybrid approach for heart diseases classification and associated risk factors identification,” *Biomedical Research*, vol. 28, no. 8, pp. 3451–3455, 2017.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](#)).