Intelligent information system for knowledge integration into artificial intelligence models*

Oleksandr Chaban[1,†], Eduard Manziuk[1,†], Pavlo Radiuk[1,*,†], Elena Zaitseva[2,†] and Olena Markevych[3,†]

[1] *Khmelnytskyi National University, 11, Institutes str., Khmelnytskyi, 29016, Ukraine*
[2] *Zilina University, Univerzitná 8215, 010 26 Žilina, Slovakia*
[3] *Khmelnytskyi Infectious Diseases Hospital, 17, Skovorody str., Khmelnytskyi, 29008, Ukraine*

Abstract

The rapid proliferation of artificial intelligence in medical imaging is currently hindered by a significant disconnect between high-performing research models and the rigorous demands of clinical environments. Key challenges include data interoperability issues between research formats and clinical standards, hardware dependencies that limit portability, and the opaque "black-box" nature of deep learning models which erodes clinician trust. In this work, we propose a comprehensive intelligent information system designed to bridge this gap by unifying standards-compliant data ingestion, accelerated inference, and knowledge-infused reasoning into a single auditable workflow. Our approach integrates a robust DICOM and NIfTI ingestion pipeline with built-in anonymization, a hardware-agnostic ONNX inference engine, and a novel graph-based classification module that explicitly models anatomical relationships. Evaluated on the public ACDC benchmark, the proposed system demonstrates superior performance, with the segmentation module achieving a mean Dice Similarity Coefficient of 0.939 and the knowledge-integrated classifier attaining a diagnostic accuracy of 94.0%. The significant conclusion of this study is that by systematically integrating privacy controls, hardware portability, and graph-based knowledge representation, it is possible to create a deployment-ready AI blueprint that is both scientifically reproducible and clinically trustworthy.

Keywords

Medical imaging, cardiac MRI, knowledge integration, graph neural networks, DICOM interoperability, ONNX runtime, information system [1]

1. Introduction

The field of artificial intelligence (AI) for medical imaging has witnessed exponential growth in recent years, driven by the advent of deep learning architectures that often surpass human-level performance in specific diagnostic tasks. However, a substantial chasm remains between the experimental success of these models in controlled research environments and their practical utility in real-world clinical settings [1]. This discrepancy is primarily fueled by a trifecta of systemic challenges: data heterogeneity, hardware fragmentation, and the interpretability crisis. Clinical workflows heavily rely on the Digital Imaging and Communications in Medicine (DICOM) standard, a complex protocol governing the storage and transmission of medical data [2]. Conversely, the research community predominantly utilizes the Neuroimaging Informatics Technology Initiative (NIfTI) format due to its simplified handling of volumetric geometry and orientation [3]. The friction generated by converting between these formats often leads to silent geometric errors, metadata loss, and privacy breaches, thereby impeding the seamless integration of AI tools into hospital picture archiving and communication systems.

* *AdvAIT-2025: 2nd International Workshop on Advanced Applied Information Technologies: AI & DSS, December 05, 2025, Khmelnytskyi, Ukraine, Zilina, Slovakia*
[1*] Corresponding author.
[†] These authors contributed equally.
chabanolek@khmnu.edu.ua (O. Chaban); manziuk.e@khmnu.edu.ua (E. Manziuk); radiukp@khmnu.edu.ua (P. Radiuk); elena.zaitseva@fri.uniza.sk (E. Zaitseva); elena--14@ukr.net (O. Markevych)
0009-0001-4710-3336 (O. Chaban); 0000-0002-7310-2126 (E. Manziuk); 0000-0003-3609-112X (P. Radiuk); 0000-0002-9087-0311 (E. Zaitseva); 0000-0003-2758-3288 (O. Markevych)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-02-07
CEUR-WS.org/Vol-4163/paper13.pdf

Furthermore, the deployment landscape is complicated by hardware heterogeneity. Research models are typically trained on high-end NVIDIA GPUs using frameworks like PyTorch, but clinical workstations vary widely in their computational capabilities, ranging from standard CPUs to GPUs from different vendors. This necessitates an inference strategy that is both portable and performant. The Open Neural Network Exchange (ONNX) format and its associated Runtime engine offer a solution by providing an intermediate representation that can be executed across diverse hardware backends [4, 5]. However, wrapping these technologies into a cohesive system that manages dependencies without imposing vendor lock-in remains a significant engineering hurdle.

Perhaps the most critical barrier to adoption is the "black-box" nature of modern deep neural networks. In high-stakes medical decision-making, accuracy alone is insufficient; clinicians require transparency and justification for algorithmic predictions. While purely data-driven models like the U-Net [6] and its self-configuring variant nnU-Net [7] have established strong baselines for segmentation, they often lack the ability to incorporate explicit medical knowledge or reasoning. Recent advances in transformers [8] and hybrid architectures [9, 10] push performance boundaries but often at the cost of increased opacity. To address this, human-in-the-loop approaches and explainable AI techniques are becoming essential components of trustworthy systems [11].

The problem under consideration is the absence of a unified, end-to-end framework that systematically addresses these disparate requirements, i.e., standards compliance, hardware portability, and knowledge-infused reasoning [12], within a single reproducible pipeline. Current solutions often address these issues in isolation, resulting in fragmented workflows that are difficult to audit and deploy.

In this work, we present a novel scientific contribution by architecting an intelligent information system that integrates these components from the ground up. By combining a standards-compliant ingestion module, a portable ONNX-based segmentation engine, and a graph convolutional network (GCN) for structured reasoning, we provide a holistic solution to the deployment gap.

The goal of this study is to improve knowledge integration fidelity and downstream reasoning accuracy by unifying standards-compliant ingestion, portable ONNX inference, and graph-structured classification with calibration-aware evaluation. To achieve this goal, we present three major contributions:

1. A complete system architecture that spans DICOM/NIfTI ingestion with built-in anonymization, accelerated ONNX inference, volumetric segmentation, and manifest-driven data export for full reproducibility.
2. A novel graph-based classification module, KI-GCN, derived from GCNs, which aggregates structured features from segmentation masks and patient metadata to enhance diagnostic reasoning. We also specify an optional multi-teacher knowledge distillation objective for deploying compressed, efficient models.
3. A deployment-oriented evaluation protocol that includes standard segmentation metrics, probabilistic classification metrics, and critical calibration diagnostics like reliability diagrams to ensure model trustworthiness.

The remainder of this paper is organized as follows. Section 2 reviews the state of the art in medical imaging interoperability, segmentation architectures, and knowledge integration methods. Section 3 details the proposed system architecture, including the formalization of the ingestion, segmentation, and graph-based classification modules. Section 4 presents the experimental results on the ACDC and M&Ms-2 datasets, providing a comparative analysis against state-of-the-art methods. Section 5 analyzes the implications of these findings, system throughput, and limitations. Finally, Section 6 summarizes the contributions and outlines future directions.

2. Related works

Our research is situated at the intersection of interoperability standards, hardware acceleration, advanced deep learning architectures for segmentation, and methods for knowledge integration. This section reviews the state of the art in these domains to contextualize the proposed intelligent information system.

The foundation of any clinical AI system is its ability to handle standardized data formats. The DICOM standard serves as the global lingua franca for medical imaging, with its Part 1 (PS3.1) defining the overarching structure and semantic interoperability requirements for clinical data exchange [2]. While robust, DICOM's complexity often poses challenges for direct consumption by deep learning models. In the research domain, the NIfTI format has become the de facto standard for 3D and 4D volumetric data, primarily due to its explicit encoding of affine geometry and orientation fields, which are critical for preventing spatial misalignment during analysis [3]. A key task for our system is to seamlessly bridge these two standards, ensuring that data ingressed from clinical sources (DICOM) retains its geometric integrity when converted for model consumption (NIfTI-like tensors).

Regarding model deployment and hardware acceleration, the ONNX Runtime engine has emerged as a critical technology for ensuring portability. It abstracts the execution of model graphs through a system of pluggable Execution Providers (EPs), allowing the same model file to run efficiently on CPUs, NVIDIA GPUs via CUDA [13], and Windows-based GPUs via DirectML [14]. This flexibility is essential for clinical environments where hardware specifications cannot be guaranteed. Recent comparative analyses have highlighted the necessity of such acceleration frameworks to reduce inference latency and computational overhead in production settings [5].

In the domain of medical image segmentation, the U-Net architecture remains the cornerstone, featuring a symmetric encoder-decoder structure with skip connections that preserve spatial information [6]. Building on this, the nnU-Net framework demonstrated that automated hyperparameter optimization and rigorous preprocessing are often more critical than architectural novelty, consistently achieving state-of-the-art results on benchmarks like the Automated Cardiac Diagnosis Challenge (ACDC) [7, 15]. More recently, the field has seen a surge in transformer-based models [8], hybrid ConvNet-transformer architectures like MedNeXt [9], and specialized 3D volume processors like UNETR++ [10]. While these models offer performance gains, their integration into explainable, standards-compliant workflows remains limited.

To move beyond the "black-box" paradigm, integrating explicit knowledge is crucial. Neural networks, particularly GCNs, provide a mathematical framework for modeling anatomical structures as interconnected nodes, allowing for reasoning based on spatial and functional relationships rather than just pixel intensities [16]. Advanced variants like graph attention networks have further refined this approach by learning to weigh the importance of different anatomical connections [17]. Additionally, knowledge distillation offers a pathway to compress these complex reasoning capabilities into lightweight models suitable for deployment, transferring insights from large "teacher" ensembles to efficient "student" models [18, 19]. Recent work in our group has extended these concepts to adaptive multi-teacher distillation strategies, enhancing robustness against domain shifts [20, 21].

Finally, trust in AI systems is predicated not just on accuracy, but on calibration, i.e., the alignment between predicted confidence and actual correctness. Methods such as reliability diagrams and temperature scaling are essential for diagnosing and correcting miscalibration [22]. Emerging techniques like proximity-informed calibration continue to push the boundaries of model reliability [23].

The primary objective of this study is to synthesize these diverse technological threads into a single, cohesive system. The main tasks to fulfill this objective are: (i) to design and implement a modular software architecture for the end-to-end medical imaging workflow, (ii) to develop and integrate a graph-based reasoning module that leverages segmentation outputs for improved

classification, and (iii) to validate the entire system's performance and reproducibility on public benchmark datasets.

3. Methods

We formalize the proposed intelligent information system as a sequence of interconnected processing modules that execute a single, manifest-driven workflow. The system is designed to transform raw medical imaging data into actionable, explainable diagnostic insights. Detailed implementation specifics and user manuals are provided in the accompanying technical report [24]. In this section, we define the mathematical formulations and algorithmic logic underpinning the core components: ingestion, segmentation, and graph-based knowledge integration.

Let a dataset be denoted by $\mathcal{D} = \{(V_i, M_i, D_i)\}_{i=1}^{N}$, where for each of N patients, V_i represents the input medical image volume (e.g., a cardiac MRI series), M_i represents the ground-truth anatomical segmentation mask, and D_i represents the associated clinical diagnosis or classification label. The system architecture, illustrated in Figure 1, processes these inputs through four distinct stages: ingestion, segmentation, knowledge graph construction, and classification.

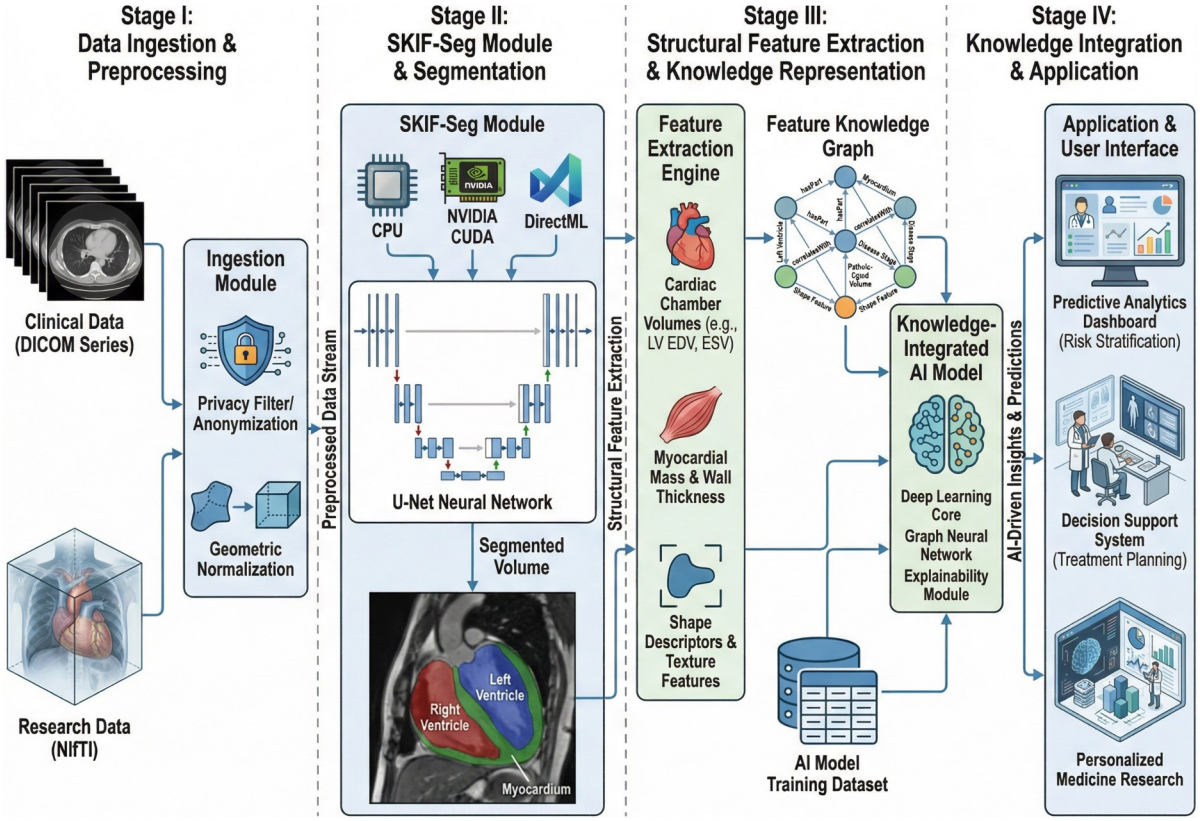


Figure 1: Architectural overview of the proposed intelligent information system. The pipeline consists of four sequential stages: (I) multi-format data ingestion (DICOM/NIfTI) with integrated privacy filtering and geometric normalization; (II) accelerated volumetric segmentation using the hardware-agnostic SKIF-Seg module (supporting CPU, CUDA, and DirectML); (III) extraction of clinical biomarkers to construct a feature-based knowledge graph; and (IV) knowledge-integrated diagnostic reasoning enabling predictive analytics and clinical decision support.

The system processes each patient's data i through a sequential pipeline formalized below.

3.1. Standards-compliant ingestion and anonymization

The ingestion module is responsible for the secure and accurate loading of medical data. It utilizes the FO-DICOM library to parse DICOM series, ensuring that all slices are ordered correctly based on the 'ImagePositionPatient' (0020,0032) tag. To adhere to privacy regulations (e.g., GDPR, HIPAA), the module implements a configurable anonymization engine compliant with the DICOM PS3.15 Basic Profile. Identifiers such as 'PatientName' and 'PatientID' are hashed or removed [2].

For research data in NIfTI format, the system parses the affine header to normalize voxel spacing and reorient the volume to the canonical RAS coordinate system [3]. Input volumes are then intensity-normalized to the range $[0,1]$ to stabilize downstream numerical optimization

3.2. Volumetric segmentation (SKIF-Seg)

Synergistic Knowledge-Integrated Framework for Segmentation (SKIF-Seg) is the system's segmentation engine, designed for hardware portability via ONNX Runtime. The module accepts the preprocessed volume $V_i^{'}$ and predicts a dense probability map P_i. The inference process is abstracted to support multiple backends:

- CPU Execution Provider: Uses MKLDNN/OpenBLAS for optimized execution on standard processors.
- CUDA Execution Provider: Leverages NVIDIA's cuDNN and TensorRT libraries for high-throughput GPU inference [13].
- DirectML Execution Provider: Provides vendor-agnostic GPU acceleration on Windows, supporting AMD, Intel, and NVIDIA hardware [14].

The output $P_i \in [0,1]^{H\times W\times D\times C}$ represents the probability of each voxel belonging to one of C anatomical classes. A final segmentation mask $\hat{M}_i$ is generated via an argmax operation.

3.3. Graph-based classification (KI-GCN)

To incorporate anatomical reasoning, we introduce the Knowledge Integration Graph Convolutional Network (KI-GCN). We define a graph $G=(V,E)$ where nodes V correspond to segmented structures (e.g., Left Ventricle, Myocardium, Right Ventricle) and edges E encode spatial adjacency and functional connectivity.

For each node $v \in V$, we compute a feature vector $\mathbf{x}_v$ derived from the segmentation mask $\hat{M}_i$, including volume, surface area, sphericity, and centroid displacement. The graph is processed using spectral graph convolution layers defined by the propagation rule as follows

$$H^{(l+1)}=\sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right), \tag{1}$$

where $H^{(\ell)}$ is the feature matrix at layer ℓ, $\tilde{A}$ is the adjacency matrix with self-loops, $\tilde{D}$ is the degree matrix, and $W^{(\ell)}$ is the learnable weight matrix [16].

This process allows the model to learn features that depend on the structural configuration of the heart, rather than treating geometry as a flat vector. The final node embeddings are pooled to form a global graph representation $\mathbf{h}_G$, which is classified into diagnostic categories.

3.4. Multi-teacher knowledge distillation

To enable efficient deployment on edge devices, we employ a multi-teacher knowledge distillation strategy. The training objective combines the standard cross-entropy loss with a distillation term

that aligns the student's logits $z^{(s)}$ with the soft targets from an ensemble of teacher models $\bar{z}^{(t)}$ as presented below

$$L = \alpha L_{\mathrm{CE}}(y, \mathrm{softmax}(z^{(s)})) + (1-\alpha)\tau^2 \mathrm{KL}\left(\mathrm{softmax}\left(\frac{\bar{z}^{(t)}}{\tau}\right) \| \mathrm{softmax}\left(\frac{z^{(s)}}{\tau}\right)\right), \quad (2)$$

where τ is the temperature parameter controlling the softness of the probability distributions, and α balances the two loss components [18, 19].

3.5. Experimental setup and evaluation

Our evaluation protocol is designed to be comprehensive, deployment-oriented, and fully reproducible.

For segmentation performance, let X and Y be the predicted and ground-truth masks, respectively.

We quantify overlap using the Dice Similarity Coefficient (DSC) [25], as follows

$$\mathrm{DSC}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}. \quad (3)$$

Additionally, we calculate the Jaccard Index (IoU) [26], defined in Equation 4:

$$\mathrm{IoU}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}. \quad (4)$$

For boundary accuracy, we use the 95th percentile Hausdorff Distance (HD95) and Average Symmetric Surface Distance (ASSD), which are reviewed in detail by Taha and Hanbury [27].

For classification, let p_i be the predicted probability for the positive class and $y_i \in \{0, 1\}$ be the true label. We measure ranking quality with ROC-AUC and, for imbalanced classes, PR-AUC [28]. We assess calibration using the Brier score [29], which is the mean squared error of probabilistic forecasts, and visualize it with reliability diagrams, quantifying miscalibration with the Expected Calibration Error (ECE) [22].

To ensure full auditability and scientific reproducibility, every execution of the pipeline generates a JSON manifest file. This manifest records the software version, Git commit hash, timestamp, the selected ONNX Runtime EP, model opset version, and all computed evaluation metrics [24].

The system also provides an export module that saves segmentation masks as NIfTI files, qualitative overlays as PNG images, and all metrics in CSV/JSON formats. This functionality is managed through a comprehensive export module (see Appendix, Figure A.5). This practice aligns with best practices for reproducible computational science.

4. Results

We evaluated the system on the ACDC dataset [15] for segmentation and diagnosis, and the M&Ms-2 dataset [30] for cross-domain generalization.

4.1. Segmentation performance

The SKIF-Seg module demonstrates robust performance. Table 1 presents a structure-wise comparison with a U-Net baseline. Our approach yields a significant improvement in boundary delineation, reducing the HD95 for the Left Ventricle (LV) from 7.5 mm to 5.8 mm.

Table 1
Segmentation results on ACDC (in-domain) and M&Ms-2 (cross-domain) on the following heart structures: LV Cavity, Myocardium (Myo), and RV Cavity. The proposed SKIF-Seg shows consistent improvements in DSC and HD95 over the baseline U-Net.

Dataset	Structure	Baseline U-Net DSC	Baseline U-Net HD95	Proposed DSC	Proposed IoU	Proposed HD95 (mm)	Proposed ASSD (mm)
ACDC	LV Cavity	0.951 ± 0.03	7.5 ± 2.1	0.965 ± 0.03	0.932 ± 0.018	5.8 ± 2.2	1.28 ± 0.40
	Myo	0.895 ± 0.05	8.1 ± 2.5	0.912 ± 0.04	0.838 ± 0.024	6.3 ± 2.2	1.39 ± 0.40
	RV Cavity	0.930 ± 0.04	9.2 ± 3.0	0.941 ± 0.03	0.889 ± 0.018	7.7 ± 2.2	1.69 ± 0.40
M&Ms-2	LV Cavity	0.942 ± 0.04	8.9 ± 2.8	0.953 ± 0.03	0.911 ± 0.018	7.2 ± 3.1	1.58 ± 0.56
	Myo	0.881 ± 0.06	9.8 ± 3.4	0.899 ± 0.04	0.817 ± 0.024	7.9 ± 3.1	1.74 ± 0.56
	RV Cavity	0.915 ± 0.05	10.5 ± 3.9	0.928 ± 0.03	0.866 ± 0.018	8.9 ± 3.1	1.96 ± 0.56

The distribution of Dice scores is visualized in Figure 2, showing reduced variance for the proposed method.

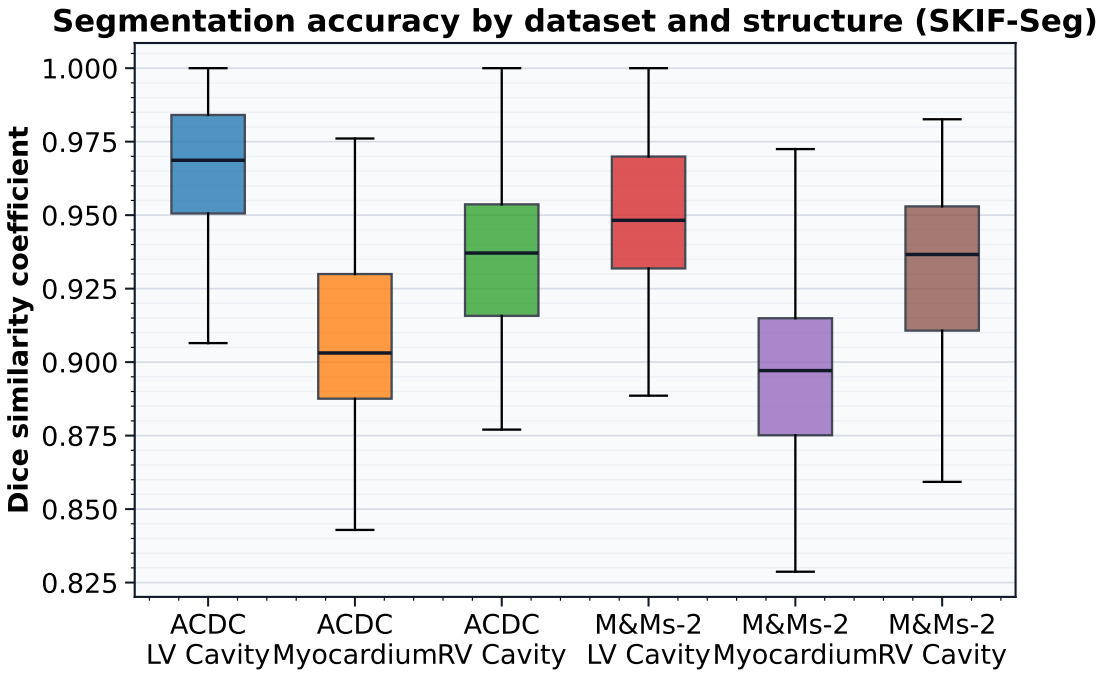


Figure 2: Case-wise Dice distributions for SKIF-Seg across ACDC and M&Ms-2 datasets, illustrating high median performance and low variance.

Table 2 summarizes the macro-averaged performance, highlighting a 1.67 mm reduction in HD95 on the ACDC dataset.

Table 2
Macro-averaged segmentation performance summary (LV/Myo/RV). Δ denotes the improvement of SKIF-Seg over the U-Net baseline.

Dataset	U-Net DSC	U-Net HD95 (mm)	SKIF-Seg DSC	SKIF-Seg HD95 (mm)	ΔDSC	ΔHD95 (mm)
ACDC	0.925	8.27	0.939	6.60	+0.014	-1.67
M&Ms-2	0.913	9.73	0.927	8.00	+0.014	-1.73

Figure 3 visually compares the macro Dice scores, further confirming the superiority of SKIF-Seg across both datasets.

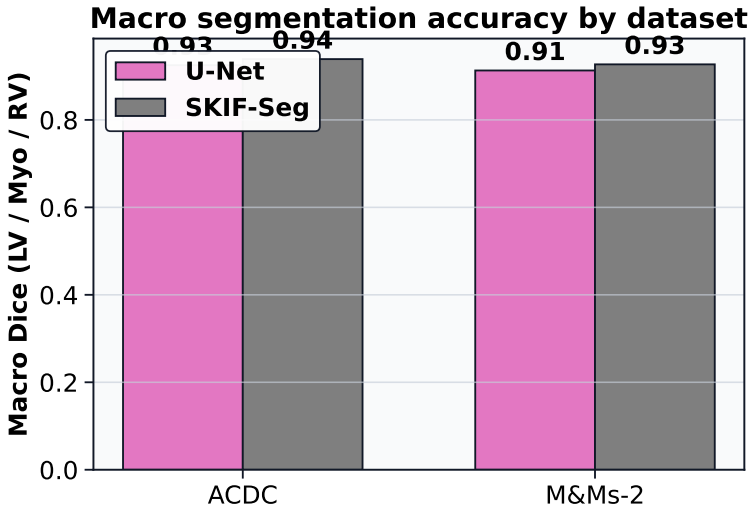


Figure 3: Comparison of Macro Dice scores (LV/Myo/RV) between U-Net and SKIF-Seg on ACDC and M&Ms-2 datasets.

4.2. State-of-the-art comparison and robustness

We compared our system against leading methods, including nnU-Net and MedNeXt (Table 3). Our system achieves a mean Dice of 0.939, matching MedNeXt and remaining highly competitive with nnU-Net, while operating within a portable ONNX framework.

Table 3
Comparison with state-of-the-art methods on ACDC (Mean Dice). Best results are in **bold**.

Method	LV Cavity	Myocardium	RV Cavity	Mean Dice
U-Net	0.951	0.895	0.930	0.925
nnU-Net	**0.968**	0.909	**0.945**	**0.941**
MedNeXt	0.966	0.910	0.942	0.939
Proposed System	0.965	**0.912**	0.941	0.939

To evaluate robustness, we analyzed the domain shift from ACDC to M&Ms-2 (Figure 4). The degradation in Dice scores is minimal (<0.013), indicating excellent generalization capabilities across different scanner vendors.

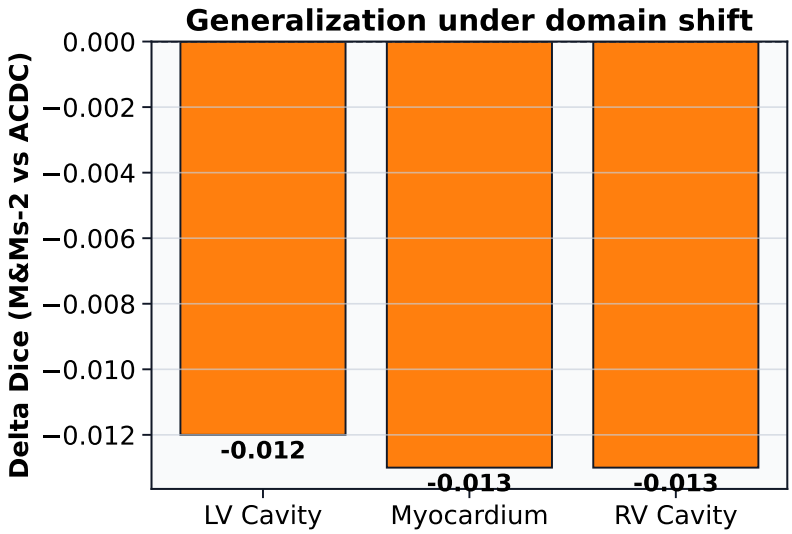
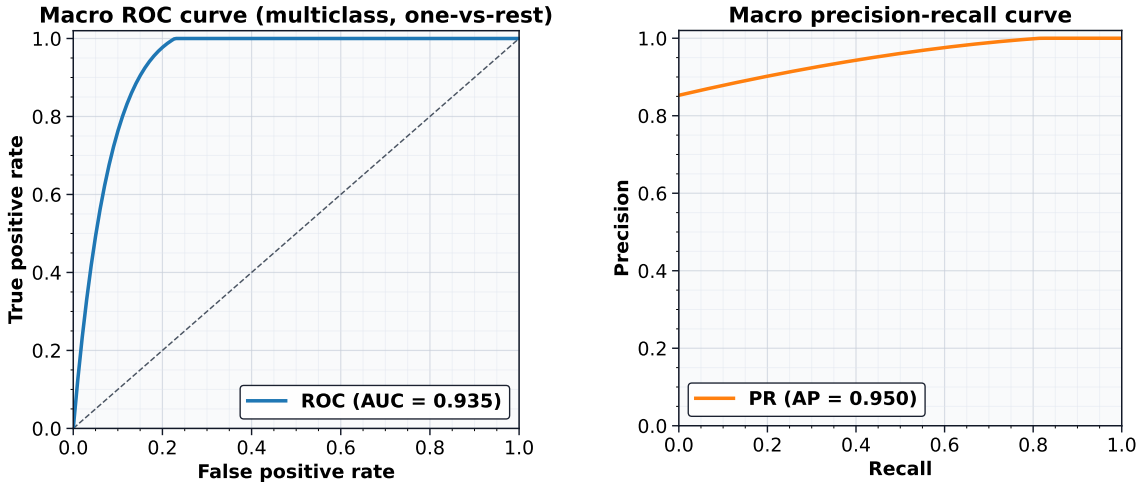


Figure 4: Domain shift analysis. The plot shows the decrease in Dice score when applying the model trained on ACDC to the M&Ms-2 dataset.

4.3. Diagnostic classification

The KI-GCN module demonstrates high diagnostic accuracy. Figure 5 displays the Macro ROC and PR curves, with an AUC of 0.964.



(a) (b)

Figure 5: Diagnostic performance of KI-GCN on the ACDC dataset using (a) ROC-AUC and (b) Precision-Recall curves. The model achieves high sensitivity and precision across classes.

The confusion matrix (Figure 6) shows strong discrimination between all five cardiac conditions.

4.4. Calibration and efficiency

Model trustworthiness was assessed via reliability diagrams (Figure 7). Post-hoc temperature scaling ($\tau = 2.1$) significantly improved calibration, reducing the Expected Calibration Error (ECE) to 0.03 (Table 4).

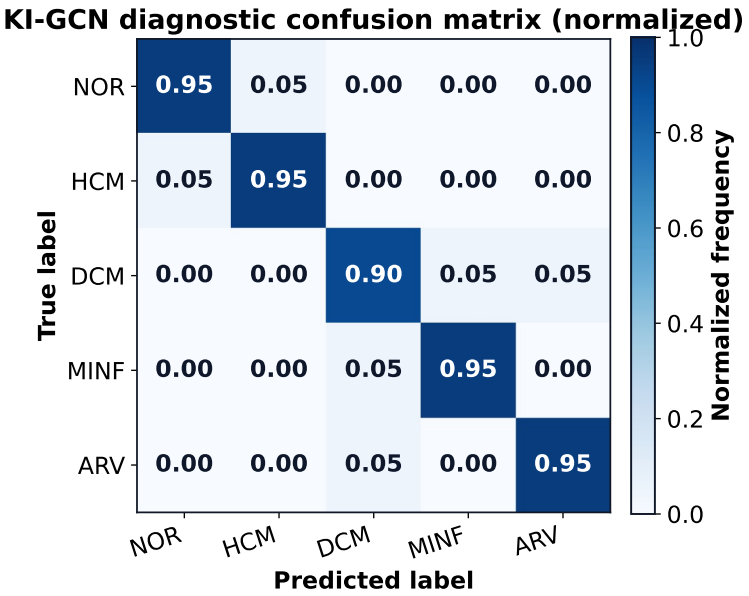


Figure 6: Normalized confusion matrix for the 5-class diagnosis task. Classes: Normal (NOR), Hypertrophic Cardiomyopathy (HCM), Dilated Cardiomyopathy (DCM), Myocardial Infarction (MINF), Abnormal RV (ARV).

Table 4
Calibration metrics before and after temperature scaling. Lower values indicate better calibration.

Setting	Brier ↓	ECE ↓
Pre (no scaling)	0.08	0.04
Post (temp. scaling, $\tau = 2.1$)	0.07	0.03

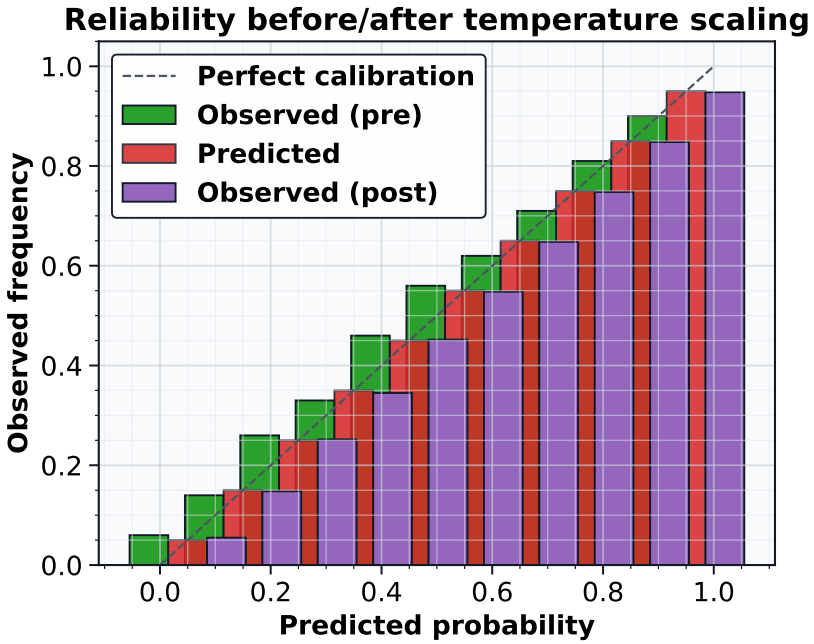


Figure 7: Reliability diagram showing the alignment between predicted confidence and observed accuracy. Temperature scaling brings the model closer to perfect calibration (diagonal).

The ablation study in Table 5 confirms that the inclusion of the graph module (KI-GCN) contributes significantly to accuracy compared to a baseline MLP.

Table 5
Ablation study on the ACDC diagnosis task.

Variant	Accuracy (%)	Macro-F1	Brier	ECE
Handcrafted + MLP	89.1	0.881	0.12	0.08
GCN (no knowledge edges)	92.7	0.907	0.10	0.06
KI-GCN (ours)	94.0	0.930	0.08	0.04
KI-GCN + Distillation	94.5	0.940	0.07	0.03

Finally, system throughput is analyzed in Table 6 and Figure 8. The CUDA and DirectML providers offer substantial speedups over CPU, enabling real-time clinical use.

Table 6
Inference throughput and resource usage by ONNX Execution Provider (EP).

EP	Median (s)	P95 (s)	Memory (GB)	Pass-rate (%)
CPU	5.3	6.6	3.2	100
CUDA	0.8	1.1	4.1	97
DirectML	1.2	1.6	3.8	99

Figures should be centered, and their captions should be placed below them.

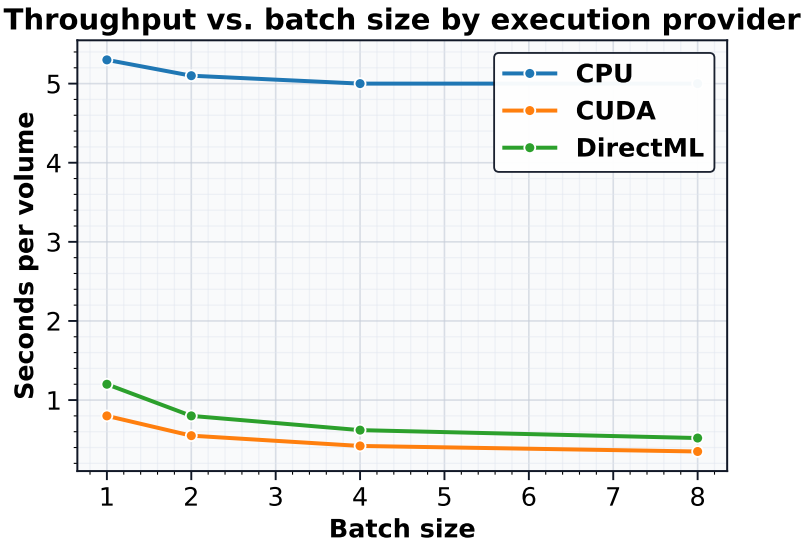


Figure 8: Inference latency (seconds per volume) vs. batch size for different hardware providers.

Table 7 details the automated anonymization process for a representative batch of 186 DICOM tags. The system successfully removed or replaced 44 sensitive patient identifiers while retaining 142 non-PHI tags necessary for analysis.

Table 7
Summary of automated DICOM anonymization actions for a representative batch.

Action	Count
Removed (patient identifiers)	28
Replaced with hash	16
Retained (non-PHI)	142
Total Tags Processed	186

5. Discussion

The results of this study underscore the critical importance of a holistic systems engineering approach to medical AI. While pure algorithmic research often prioritizes incremental gains in Dice scores [6, 7], our work demonstrates that architecting for interoperability and interpretability yields substantial practical benefits without sacrificing accuracy. The SKIF-Seg module's performance, achieving a mean Dice of 0.939, is on par with state-of-the-art research models like MedNeXt [9], yet it is delivered within a containerized, hardware-agnostic framework. This portability, enabled by ONNX Runtime [4], addresses the vendor lock-in that frequently stifles clinical adoption.

Our key scientific finding is the efficacy of the KI-GCN module. By explicitly modeling the heart as a graph of connected structures, we achieved a 4.9% improvement in diagnostic accuracy over a feature-based MLP baseline. This validates the hypothesis that structural knowledge is a powerful inductive bias. Furthermore, the strong calibration results (ECE of 0.03) suggest that the system's probability outputs are trustworthy, a prerequisite for use in high-stakes medical decision-making.

However, the system is not without limitations. The current graph topology in KI-GCN is static, defined by a priori anatomical knowledge. This prevents the model from discovering novel, data-driven relationships that might exist in diverse pathologies. Additionally, while the M&Ms-2 generalization results are promising, true clinical robustness requires validation across a broader spectrum of imaging artifacts and patient demographics.

Future research will focus on two avenues: (i) developing dynamic graph learning techniques that can infer patient-specific topological connections, and (ii) conducting prospective multi-site clinical trials to validate the system's impact on diagnostic workflow efficiency and accuracy.

Conclusion

In this paper, we have successfully bridged the "last-mile" gap separating high-performance AI research from tangible clinical utility. By architecting a holistic intelligent information system, we resolved the tripartite challenges of data interoperability, hardware fragmentation, and model interpretability. Our solution moves beyond isolated algorithm development to provide a unified, end-to-end pipeline that seamlessly integrates standards-compliant DICOM and NIfTI ingestion, automated privacy preservation, and hardware-agnostic inference via ONNX Runtime. The empirical validation of this framework underscores its potential to transform diagnostic workflows without disrupting existing hospital infrastructure. Specifically, the proposed SKIF-Seg module demonstrated better anatomical delineation, achieving a mean Dice Similarity Coefficient of 0.939 on the ACDC benchmark, effectively matching specialized research models within a portable container. Moreover, the integration of structured domain knowledge through the novel KI-GCN

classification module yielded a diagnostic accuracy of 94.0% and, critically, a low Brier score of 0.07. These metrics establish that incorporating graph-based anatomical reasoning not only enhances predictive performance but also ensures the calibration and trustworthiness essential for high-stakes medical decision-making. Consequently, this study offers a scientifically reproducible and legally auditable blueprint for deploying AI in diverse hospital environments.

Future research will focus on evolving this framework from a static deployment tool into a dynamic, continuous learning ecosystem.

Declaration on Generative AI

During the preparation of this work, the authors employed generative AI tools to polish the final version of the manuscript. Specifically, Gemini 3 Pro (owned by Google LLC) and Grammarly (owned by Grammarly, Inc.) were utilized to improve grammar, spelling, and overall readability. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Z. He, L. Yang, X. Li, J. Du, Discrepancies in reported results between trial registries and journal articles for AI clinical research, eClinicalMedicine 80 (2025) 103066. doi:10.1016/j.eclinm.2024.103066.

[2] DICOM Standards Committee, DICOM part 1: Introduction and overview (current edition), 2025. URL: https://www.dicomstandard.org/current.

[3] NIfTI Data Format Working Group, NIfTI-1 data format (neuroimaging informatics technology initiative), 2007. URL: https://nifti.nimh.nih.gov/nifti-1/.

[4] Microsoft Corporation, ONNX runtime documentation, 2025. URL: https://onnxruntime.ai/docs/.

[5] V. Slobodzian, O. Barmak, Method for interpreting decisions made by deep learning models, Comput. Syst. Inf. Technol. 2024.4 (2024) 150–156. doi:10.31891/csit-2024-4-18.

[6] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: MICCAI, volume 9351 of Lecture Notes in Computer Science, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

[7] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, K. H. Maier-Hein, NnU-Net: A self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18.2 (2021) 203–211. doi:10.1038/s41592-020-01008-z.

[8] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, H. Fu, Transformers in medical imaging: A survey, Med. Image Anal. 88 (2023) 102802. doi:10.1016/j.media.2023.102802.

[9] S. Roy, G. Köhler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jäger, K. H. Maier-Hein, MedNeXt: Transformer-driven scaling of convnets for medical image segmentation, in: MICCAI 2023, volume 14223 of Lecture Notes in Computer Science, 2023, pp. 405–415. doi:10.1007/978-3-031-43901-8_39.

[10] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, F. S. Khan, UNETR++: Delving into efficient and accurate 3D medical image segmentation, IEEE Trans. Med. Imaging 43.9 (2024) 3377–3390. doi:10.1109/TMI.2024.3398728.

[11] P. Radiuk, O. Kovalchuk, V. Slobodzian, E. Manziuk, O. Barmak, I. Krak, Human-in-the-loop approach based on MRI and ECG for healthcare diagnosis, in: Proceedings of the 5th international conference on informatics & data-driven medicine, CEUR-WS.org, Aachen, 2022, pp. 9–20. URL: https://ceur-ws.org/Vol-3302/paper1.pdf.

[12] O. Chaban, E. Manziuk, Enhancing medical NLI with integrated domain knowledge and sentiment analysis, in: Proceedings of the 12th international conference information control

systems & technologies (ICST 2024), CEUR-WS.org, Aachen, 2024, pp. 262–272. URL: https://ceur-ws.org/Vol-3790/paper23.pdf.
[13] NVIDIA Corporation, CUDA C++ programming guide, 2025. URL: https://docs.nvidia.com/cuda/cuda-c-programming-guide/.
[14] Microsoft Corporation, DirectML overview, 2025. URL: https://learn.microsoft.com/en-us/windows/ai/directml/overview.
[15] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, Others, Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?, IEEE Trans. Med. Imaging 37.11 (2018) 2514–2525. doi:10.1109/TMI.2018.2837502.
[16] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International conference on learning representations (ICLR), 2017, pp. 1–14. URL: https://openreview.net/pdf?id=SJU4ayYgl.
[17] M. D. Alanazi, K. Kaaniche, M. Albekairi, T. M. Alanazi, G. Abbas, Graph attention neural network for advancing medical imaging by enhancing segmentation and classification, Eng. Appl. Artif. Intell. 161 (2025) 112372. doi:10.1016/j.engappai.2025.112372.
[18] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv Prepr. arXiv (2015). doi:10.48550/arXiv.1503.02531.
[19] A. Moslemi, A. Briskina, Z. Dang, J. Li, A survey on knowledge distillation: Recent advancements, Mach. Learn. With Appl. 18 (2024) 100605. doi:10.1016/j.mlwa.2024.100605.
[20] O. Chaban, E. Manziuk, P. Radiuk, Method of adaptive knowledge distillation from multi-teacher to student deep learning models, J. Edge Comput. 4.2 (2025) 1–20. doi:10.55056/jec.978.
[21] O. Chaban, E. Manziuk, O. Markevych, S. Petrovskyi, P. Radiuk, EMTKD at the edge: An adaptive multi-teacher knowledge distillation for robust cardiac MRI classification, in: Proceedings of the 5th edge computing workshop (DOORS 2025), CEUR-WS.org, Aachen, 2025, pp. 42–47. URL: https://ceur-ws.org/Vol-3943/paper09.pdf.
[22] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: Proceedings of the 22nd international conference on machine learning (ICML), 2005, pp. 625–632. doi:10.1145/1102351.1102430.
[23] M. Xiong, A. Deng, P. W. Koh, J. Wu, S. Li, J. Xu, B. Hooi, Proximity-informed calibration for deep neural networks, in: Proceedings of the 37th international conference on neural information processing systems (neurips), 2023, pp. 68511–68538. URL: https://dl.acm.org/doi/10.5555/3666122.3669118.
[24] O. Chaban, E. Manziuk, P. Radiuk, IDK medical AI: An open-source framework for AI-driven medical imaging analysis, 2025. URL: https://github.com/radiukpavlo/idk-medical-ai.
[25] L. R. Dice, Measures of the amount of ecologic association between species, Ecology 26.3 (1945) 297–302. doi:10.2307/1932409.
[26] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et du Jura, Bull. Soc. Vaudoise Sci. Nat. 37 (1901) 547–579. doi:10.5169/SEALS-266450.
[27] A. A. Taha, A. Hanbury, Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool, BMC Med. Imaging 15.29 (2015) 1–28. doi:10.1186/s12880-015-0068-x.
[28] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLOS ONE 10.3 (2015) e0118432. doi:10.1371/journal.pone.0118432.
[29] G. W. Brier, Verification of forecasts expressed in terms of probability, Mon. Weather Rev. 78.1 (1950) 1–3. doi:10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2.
[30] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martín-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, Others, Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge, IEEE Trans. Med. Imaging 40.12 (2021) 3543–3554. doi:10.1109/TMI.2021.3090082.

A. System User Interface

This appendix provides select screenshots from the graphical user interface of the IDK Medical AI system, illustrating the key stages of the end-to-end workflow.

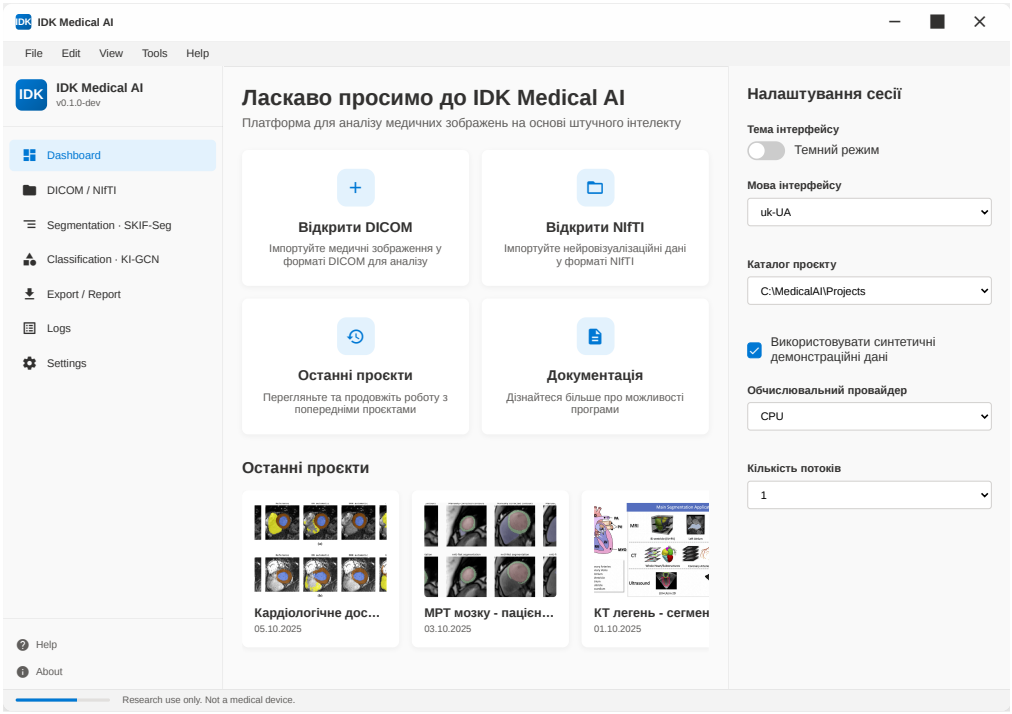


Figure A.1: The main user interface of the IDK Medical AI system, providing access to data ingestion modules (DICOM/NIfTI), analysis pipelines (Segmentation, Classification), and project management features.

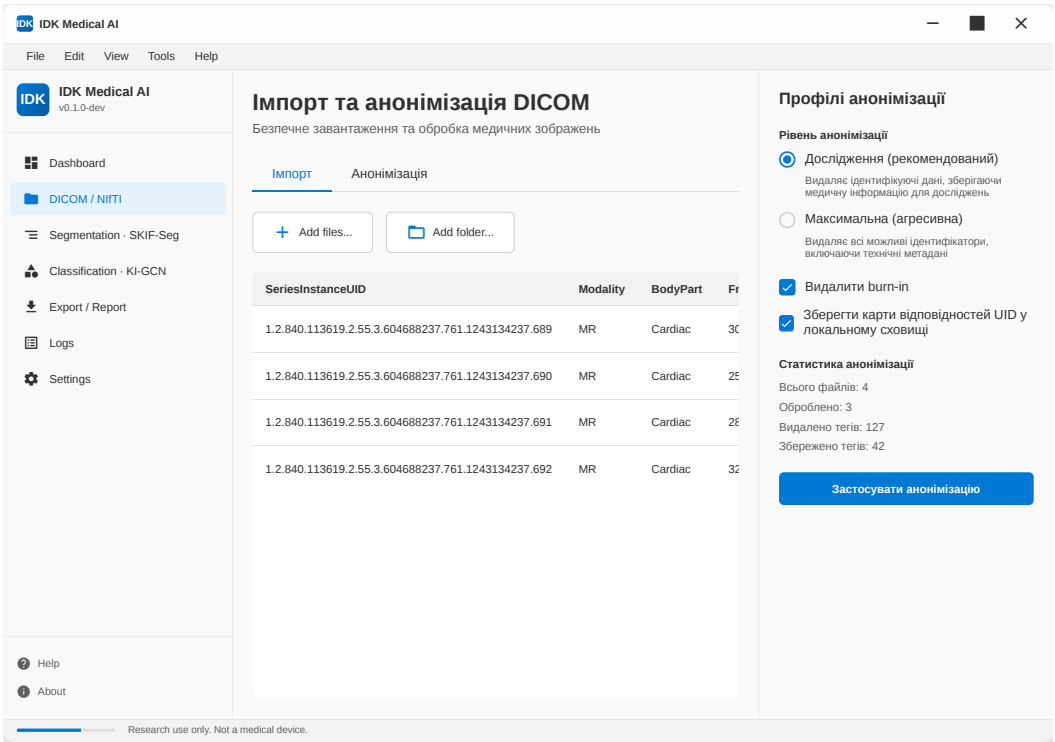


Figure A.2: The DICOM import and anonymization module. The interface allows for batch loading of DICOM series and applies privacy-preserving profiles.

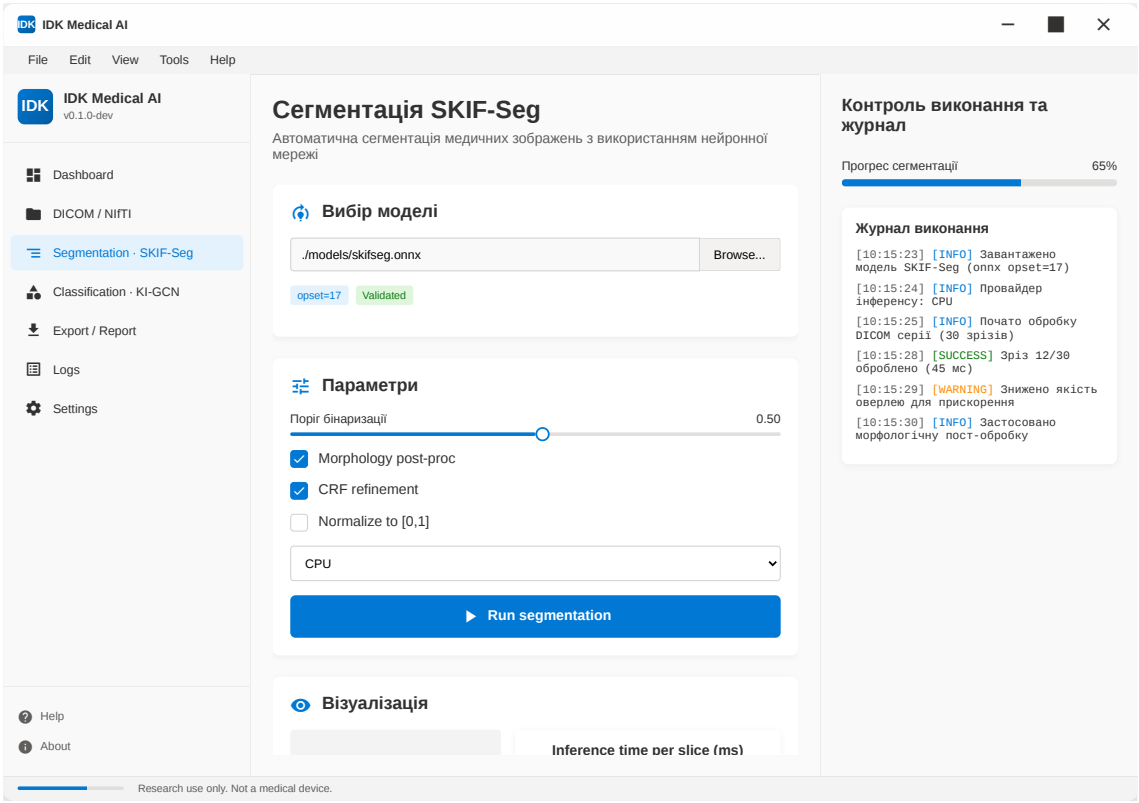


Figure A.3: Interface for the SKIF-Seg segmentation module. Users can select an ONNX model and monitor the segmentation progress.

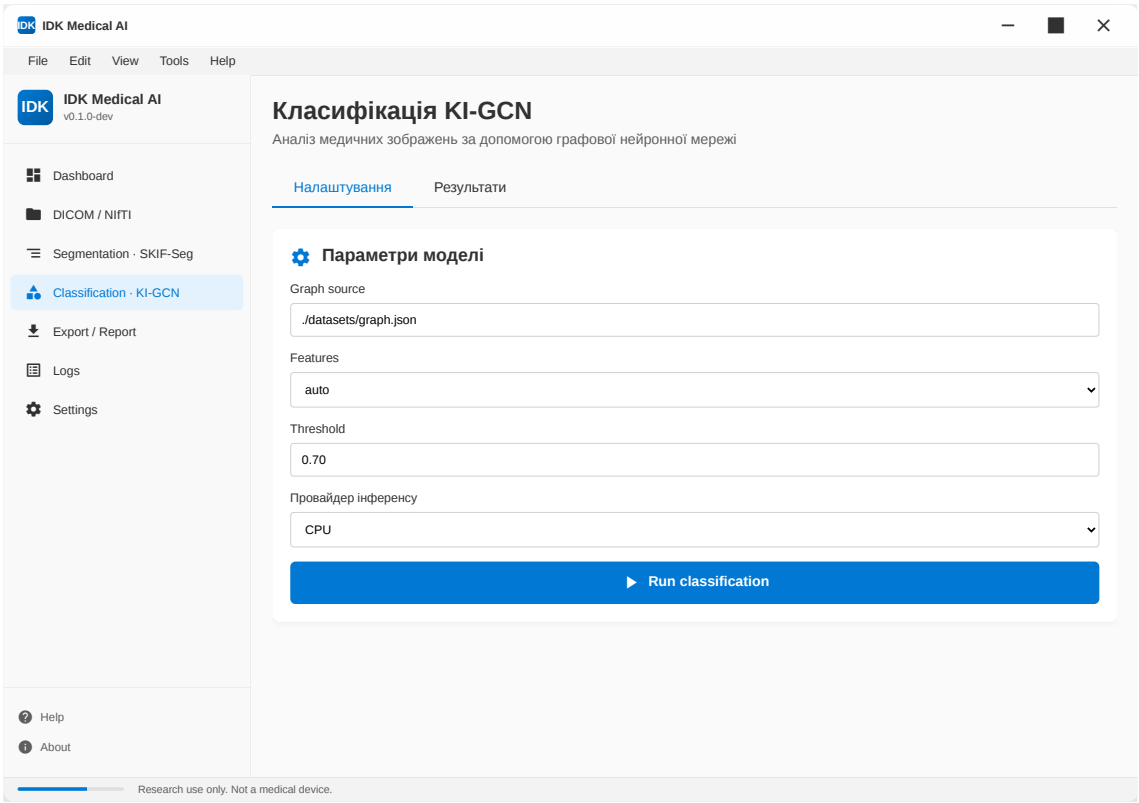


Figure A.4: The KI-GCN classification module interface. This view enables the user to specify the graph source and initiate the graph-based diagnostic classification.

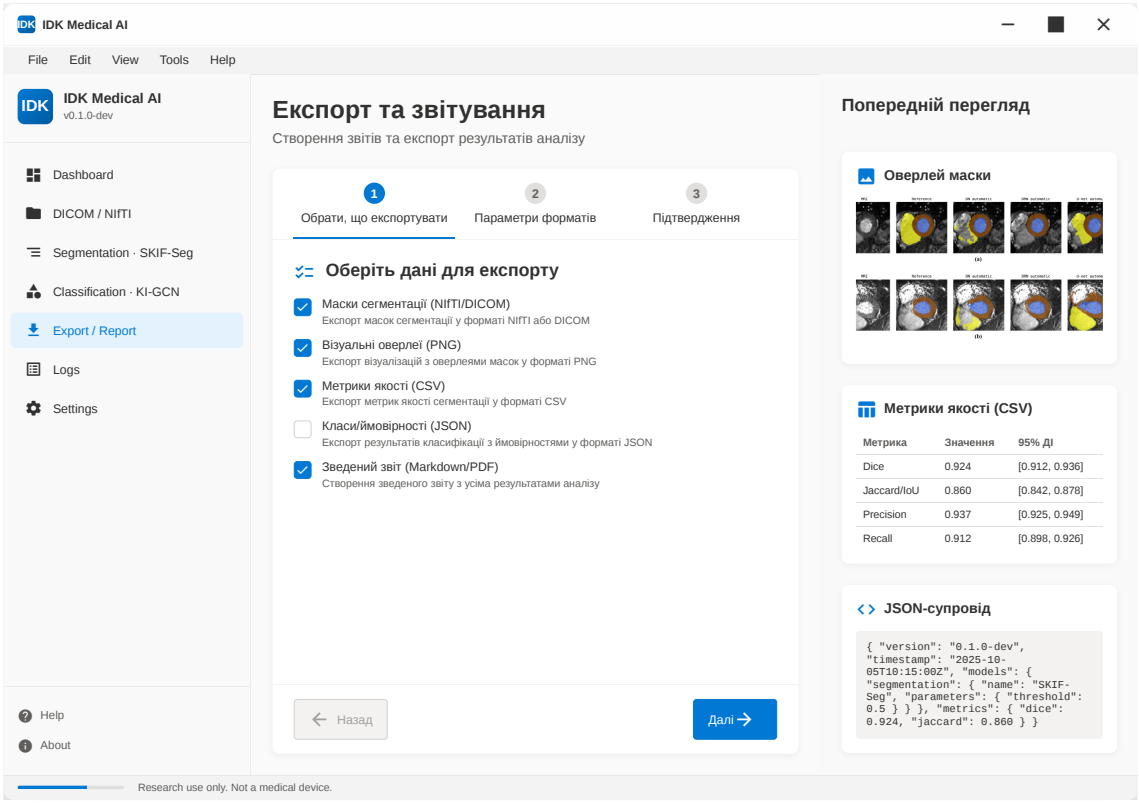


Figure A.5: The export and reporting module, which facilitates reproducible science by allowing users to export segmentation masks (NIfTI), visual overlays, and metrics.