PHAROS-AFE-AIMI: Multi-source & Fair Disease Diagnosis

Dimitrios Kollias
Center for Multimodal AI
Digital Environment Research Institute
Queen Mary University of London
d.kollias@qmul.ac.uk

Anastasios Arsenos
National Kapodistrian University Athens
Psachna, Evia, Greece
anarsenos@dind.uoa.gr

Stefanos Kollias
National Technical University Athens
Zografou, Athens, Greece
stefanos@cs.ntua.gr

Abstract

The PHAROS-AFE-AIMI Workshop focuses on the application of trustworthy AI to medical imaging. It advances the AI-MIA Workshop series composed of 4 Workshops in 2021-2024, linking it to the Pharos AI Factory and specifically to its Healthcare vertical. Specific technologies are developed which target explainability, fairness, regularization, continual learning and domain shift analysis. Various medical imaging problems are tackled, including lung disease, cancer diagnosis, MRI, CT scan semantic segmentation and classification. Moreover, a competition was organized, comprising two tracks: (i) Multi-Source COVID-19 Detection Challenge, in which optimal systems were targeted that generalize across acquisition/site variations, (ii) Fair Disease Diagnosis Challenge, targeting CT scan classification to Healthy, Adenocarcinoma, Squamous Cell Carcinoma, or COVID-19, both in male and female categories. A baseline system has been developed employing a unified 3D convolutional encoder with sequence (RNN) aggregation for volumetric context, trained with standardized preprocessing and augmentation.

1. Introduction

Medical image analysis underpins modern diagnostic, prognostic, and therapeutic decision making across a broad spectrum of diseases. High–resolution three–dimensional (3D) chest computed tomography (CT) is central for characterizing pulmonary infections (e.g., COVID-19) and thoracic malignancies (e.g., adenocarcinoma, squamous cell carcinoma), yet routine clinical assessment remains labor–intensive, subject to inter–/intra–observer variability, and challenged by heterogeneous acquisition protocols across institutions. Deep learning (DL) approaches have demonstrated substantial gains in sensitivity, specificity, and multi–class discrimination by leveraging large–scale annotated datasets and representation learning. However, translating these advances into trustworthy, equitable, and generalizable clinical systems requires addressing persistent gaps in (1) robustness to domain shift (scanner models, reconstruction kernels, demographic and epidemiological variations), (2) fairness across patient subgroups, (3) explainability and human–centered transparency, and (4) scalable continual adaptation as data distributions evolve.

Recent regulatory and infrastructure developments, including the emerging European Health Data Space and the EU AI Act, emphasize requirements for transparency, bias monitoring, traceability, and post–deployment governance of high–impact AI systems. The PHAROS AI Factory[1], one of the first seven AI Factories funded by the European Union, embeds these priorities, targeting an applied ecosystem in which research outputs boost innovation, whilst ensuring reproducibility, accountability, and citizen trust. Within this context, the PHAROS Adaptation, Fairness, Explainability in AI Medical Imaging (PHAROS-AFE-AIMI) Workshop advances the AI-MIA series (ICCV 2021 [8], ECCV 2022 [11], ICASSP 2023 [10] and CVPR 2024 [13]), opening pathways for the transparent integration of Generative AI and multi-modal Large Language models (M-LLMs, foundation models) into clinical imaging workflows.

In addition, we have recently developed a combined segmentation and classification approach for 3-D chest CT scans, particularly focusing on Covid-19 detection [14]. Our approach includes vision-language models that seg-

[1] https://grnet.gr/en/business-directory/pharos-ai-factory-en/

EuroHPC Joint Undertaking
Co-funded by the European Union
PHAROS GREECE
HELLENIC REPUBLIC Ministry of Digital Governance

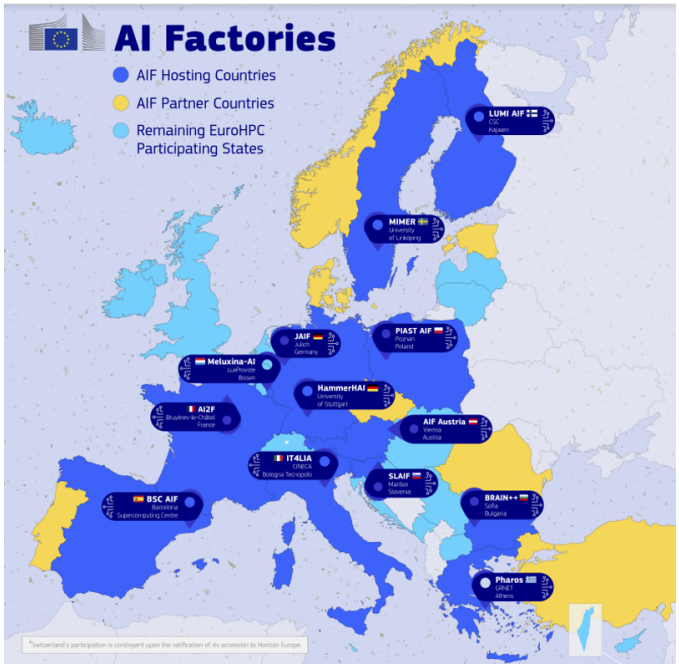

Figure 1. Data samples in Multi-Source COVID-19 Detection Challenge: 3-D scan length histogram

ment the CT scans, followed by a deep neural architecture, named RACNet [9], performing the classification task. This architecture can be used in various applications related to concept detection and relevant classification of images and videos [18], [1]. Segment Anything Model (SAM) and Contrastive Language-Image Pre-Training (CLIP) are two exemplary Vision Foundation Models (VFMs) that have showcased exceptional capabilities in segmentation and zero-shot recognition, respectively. SAM, a prompt-driven segmentation model, excels across diverse domains. SAM has been trained on an extensive dataset of over one billion masks, making it highly adaptable to a wide range of downstream tasks through interactive prompts. It can operate in two distinct modes: segment everything mode and promptable segmentation mode. In our approach, we employ both modes to achieve optimal segmentation results. SAM has shown impressive results in a broad range of tasks for natural images, but its performance has not been state-of-the-art when being directly applied to medical imaging. Conversely, CLIP's training with millions of text-image pairs has endowed it with an unprecedented ability in zero-shot visual recognition.

The developed SAM2CLIP2SAM approach provides CT scan segmentation, leveraging the strengths of both models. At first, SAM produces multiple part-based segmentation masks for each slice in the CT scan; then CLIP selects only the masks that are associated with the regions of interest (ROIs), i.e., the right and left lungs; finally SAM is given these ROIs as prompts and generates the final segmentation mask for the lungs. The method accurately segments the right and left lungs in CT scans, subsequently feeding these segmented outputs into RACNet for classification of COVID-19 and non-COVID-19 cases.

In the following we present the PHAROS AI Factory framework (Section 2), the PHAROS-AFE-AIMI Competition framework (Section 3), the related databases (Section 4), the baseline configuration (Section 5), the experimental results (Section 6) and the future work plan (Section 7).

2. The Pharos AI Factory

The European Union has recently announced the 'Europe: AI Continent Plan' including the following:

- Building a large-scale AI computing infrastructure (13 AI Factories, 5 AI Gigafactories, 200b AI Apply Facility, cloud & AI development act)
- Increasing access to high-quality data (Data Union Strategy, Data Labs across AI Factories)
- Promoting AI in strategic sectors (Apply AI Strategy, Eu-

ropean Digital Innovation Hubs)
- Strengthening AI skills and talents (Educate, Train, Attract, Retain)
- Simplifying the implementation of the AI act.

AI Factories are dynamic ecosystems that build around AI-optimized supercomputers, offering computing resources and support services to the European industry, as well as to the European scientific users for the development of large AI models. The 13 currently approved EU AI Factories are shown in Figure 1.

The Pharos AI Factory capitalizes on existing investments in the area of HPC, AI and data centres to empower on start-ups and SMEs. It prioritizes three verticals, i.e., application fields: Health & Life Sciences, Culture & Language and Sustainability (Energy - Environment - Climate). The services, as well as supporting actions of PHAROS AI Factory to users & developers, are shown in Figure 2.

The Health & Life Sciences vertical targets creating AI models for the analysis of multi-modal biomedical data, domain-specific AI models to support targeted healthcare solutions, AI predictive models for disease progression, including segmentation, classification, generalization, fair and explainable decision making. The PHAROS-AFE-AIMI Workshop and Competition enrich the perfromed research in these directions.

3. Competition Overview

A variety of technologies have been developed for early diagnosis of COVID-19 based on medical image analysis, with special emphasis on 3D chest CT scans. Recent methods often employ combined segmentation and classification strategies, targeting abnormalities such as consolidation, ground-glass opacities, and interlobular septal thickening primarily under pleura [20].

The 2025 PHAROS-AFE-AIMI Workshop continues the tradition established by prior competitions, such as the COV19D Competitions organized within the formerly organized workshops. To advance the community's development along the crucial dimensions of robustness and fairness, this edition introduces two new competition tracks, each employing standardized datasets and evaluation protocols.

Multi-Source COVID-19 Detection Challenge. This track provides annotated 3D chest CT volumes aggregated from four distinct hospitals or medical centers, each identified by source labels (0–3). Each CT scan has been meticulously manually annotated as either COVID-19 positive or non-COVID-19. The aggregated dataset is partitioned into training, validation, and test sets. Participants receive access to the annotated training and validation datasets to develop AI/ML/DL models capable of robust and accurate COVID-19 prediction. The primary challenge and goal of this competition track is achieving robust generalisation across different acquisition conditions, scanner models, imaging protocols, and institutional contexts inherent to the provided datasets. Consequently, participants are encouraged to develop models that generalise well, delivering uniformly strong performance across diverse sites rather than overfitting to particular datasets. Performance evaluation is rigorously conducted using a per-source macro F1 score, which is computed independently for each hospital/medical center source and subsequently averaged. This evaluation protocol explicitly reveals each model's cross-institutional robustness, highlighting strengths and weaknesses across distinct institutional conditions.

Fair Disease Diagnosis Challenge. This multi-class classification challenge encompasses four diagnostic classes: Healthy, Adenocarcinoma, Squamous Cell Carcinoma, and COVID-19. Each CT scan is accompanied by metadata specifying patient sex (male/female). The dataset provided is partitioned into clearly defined training, validation, and test sets. Participants have access to the annotated training and validation sets to develop robust AI/ML/DL classification models. The central goal of this challenge is to ensure fairness across gender groups in diagnostic outcomes. Fairness is operationalised by calculating the macro F1 score separately for each sex subgroup over the four diagnostic categories and then averaging these gender-specific scores. This evaluation framework aims explicitly to prevent models from optimizing performance disproportionately for majority or easier subgroups, highlighting and mitigating any gender-related performance discrepancies. The resulting fairness-oriented metric serves as a transparent, scalar indicator of balanced and equitable clinical performance, promoting fairness and reducing demographic biases in diagnostic accuracy.

4. Database Description

4.1. Multi-Source COVID-19 Detection Database

The database for the Multi-Source COVID-19 Detection Challenge is based on the COV19-CT-DB [9] and contains 3-D chest CT scans, aggregated from four distinct hospitals/medical centers, identified by source labels (0–3). The database consists of 3,020 CT scans, of which 1,035 are COVID-19 cases and 1,985 are non-COVID-19 cases. All scans are anonymized and manually annotated to ensure high-quality labels.

Table 1 summarises the distribution of scans across training, validation, and test partitions.

Each 3-D scan consists of a sequence of 2-D CT images, ranging typically between 50 and 700 [2] slices per scan, maintaining a consistent image resolution of 512×512 pix-

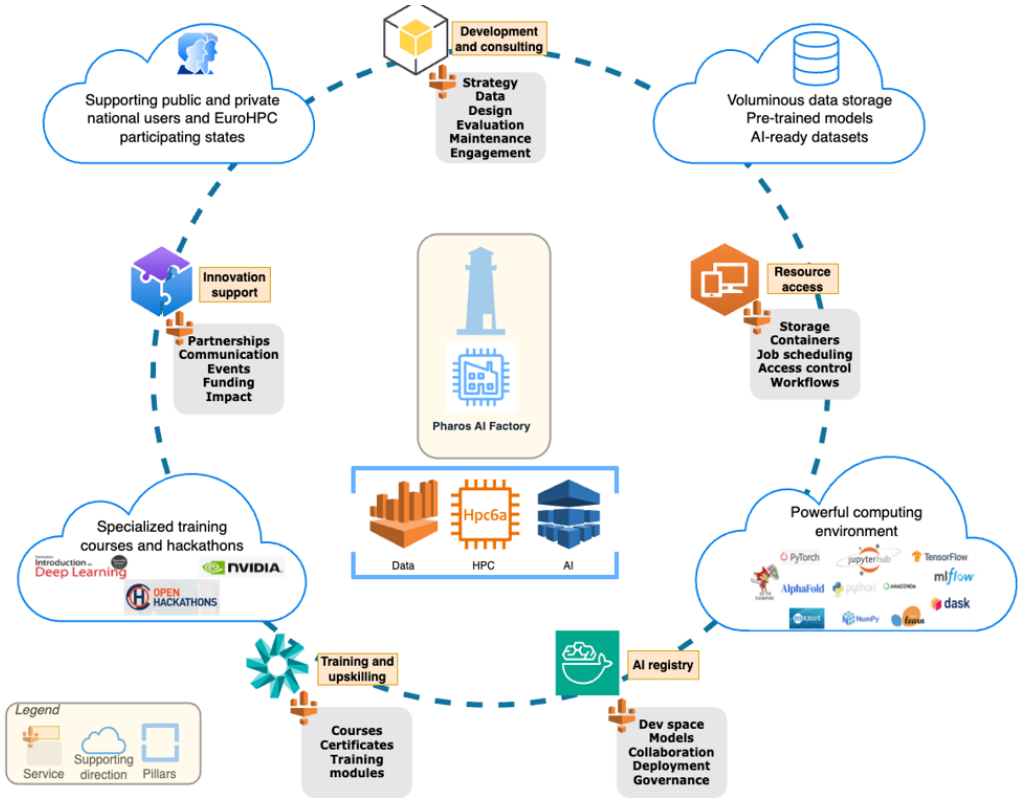

Figure 2. Data samples in Multi-Source COVID-19 Detection Challenge: 3-D scan length histogram

els. Table 2 presents a summary of the main elements of train and validation sets in Multi-Source COVID-19 Detection Challenge.

Table 1. Data samples in each Set in Multi-Source COVID-19 Detection Challenge

Set	COVID-19	Non-COVID-19	Total
Train	564	660	1,224
Validation	128	180	308
Total	**692**	**840**	**1532**

Table 2. Data samples in Multi-Source COVID-19 Detection Challenge : main elements

Elements	Values
number of 3-D CT scans	1,035 COVID 1,985 non-COVID
number of 2-D images	424,273 COVID 1,098,238 non-COVID
number of images in scan series	50 - 700
size of images	512×512

4.2. Fair Disease Diagnosis Database

The Fair Disease Diagnosis Challenge dataset comprises 3-D chest CT scans annotated into four diagnostic categories: Healthy (normal), Adenocarcinoma (A), Squamous Cell Carcinoma (G), and COVID-19. Each scan includes metadata specifying patient sex (male/female). The dataset consists of 1,254 CT scans, divided into clearly defined training, validation, and test partitions.

Table 3 summarizes the detailed categorical distribution across train and validation partitions.

Similarly, each 3-D scan in this dataset includes between 50 and 700 slices per scan, with an image resolution of 512×512 pixels. The provided demographic metadata enables comprehensive fairness assessments across gender groups.

Figure 3 analyzes the length of the CT scan series, presenting their histogram. This shows the differences regarding the length of 3-D CT scans in Data samples in Multi-Source COVID-19 Detection Challenge; these are caused by various reasons, including the requested resolution analysis, or the specific features of the used equipment.

Figure 4 shows a series of slices from a COVID-19 case, Figure 5 shows a series of slices from a normal case, Figure 6 shows a series of slices from an adenocarcinoma case

Table 3. Data samples in each Set in Fair Disease Diagnosis Challenge

Set	Sex	Healthy	Adenocarcinoma	Squamous Cell	COVID-19
Train	Female	100	125	5	100
	Male	100	125	79	100
Validation	Female	20	25	13	20
	Male	20	25	12	20
Total	Female	**120**	**150**	**18**	**120**
	Male	**120**	**150**	**91**	**120**

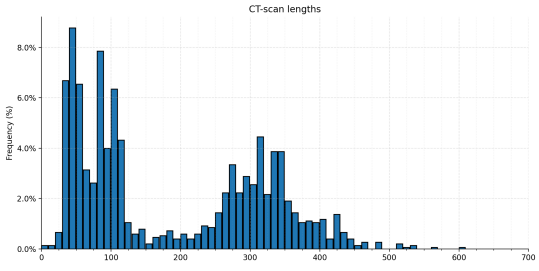

Figure 3. Data samples in Multi-Source COVID-19 Detection Challenge: 3-D scan length histogram

and Figure 7 shows a series of slices from a squamous cell carcinoma case.

Other database that we have considered was the Lung-PET-CT-Dx [16] database, comprising CT and PET-CT DICOM images of lung cancer subjects, including XML annotations indicating tumor locations with bounding boxes. The dataset includes patients retrospectively acquired who underwent standard-of-care lung biopsy and PET/CT scans due to suspected lung cancer. Patients are categorized based on tissue histopathological diagnosis into Adenocarcinoma (names/IDs containing 'A'), Small Cell Carcinoma ('B'), Large Cell Carcinoma ('E'), and Squamous Cell Carcinoma ('G'). Images are provided in mediastinum (window width, 350 HU; level, 40 HU) and lung (window width, 1,400 HU; level, –700 HU) settings. Reconstructions include 2mm-slice thickness with CT slice intervals varying from 0.625 mm to 5 mm, encompassing plain, contrast-enhanced, and 3D reconstruction scanning modes.

5. The baseline configurations

5.1. Multi-Source COVID-19 Detection & Fair Disease Diagnosis baselines

The baseline architecture adopted for both the Multi-Source COVID-19 Detection Challenge and the Fair Disease Diagnosis Challenge is a CNN-RNN architecture [3, 7, 9, 12].

Each 3-D CT scan has been padded to achieve a uniform length t, ensuring all scans consistently contain exactly t slices. The entire unsegmented sequence [19] of 2-D slices from each CT scan is first processed through the SAM2CLIP2SAM method for segmentation purposes. It is then fed to the CNN component of the CNN-RNN classifier. This CNN component performs localized feature extraction on a slice-by-slice basis, emphasizing relevant features primarily from lung regions. This localized analysis mirrors expert medical annotation methods, capturing diagnostic features throughout the complete 3-D scan series.

Following this, the RNN component sequentially analyzes the CNN-generated features from slice 0 to slice $t-1$, effectively capturing the volumetric context across the scan. The outputs from the RNN are then passed to a Fully Connected (FC) layer, and finally to an output layer employing a softmax activation function to produce the classification results. To mitigate overfitting and enhance generalization, a Dropout layer is integrated immediately before the FC layer.

For the Multi-Source COVID-19 Detection Challenge, the output layer consists of two units representing the binary COVID-19 classification (COVID vs. Non-COVID). This binary classification design directly corresponds to clinical practice, distinguishing positive COVID-19 cases from other pulmonary conditions and healthy states. We trained the baseline with a source-wise Loss Averaging - a Domain-Balanced Loss. Instead of computing the loss over all samples jointly (which can bias learning toward sources with more data) the method computes the loss separately for each data source and then averages these losses. This ensures that each source contributes equally to the gradient updates, promoting balanced performance across domains. The CNN-RNN baseline emphasizes robust generalization across diverse medical centers and scanner configurations by effectively integrating localized and volumetric features from heterogeneous sources.

In the Fair Disease Diagnosis Challenge, the output layer comprises four units, each corresponding to one of the diagnostic categories: Healthy (normal), Adenocarcinoma, Squamous Cell Carcinoma, and COVID-19. This multi-class setup addresses the clinical requirement of discriminating between multiple prevalent lung diseases. Similarly as in the previous Challenge, we trained the baseline with a source-wise Loss Averaging - a Domain-Balanced Loss. Instead of computing the loss over all samples jointly (which can bias learning toward the gender with more data) the method computes the loss separately for each gender

Figure 4. Slices from a COVID-19 case Fair Disease Diagnosis Challenge

Figure 5. Slices from a Normal case Fair Disease Diagnosis Challenge

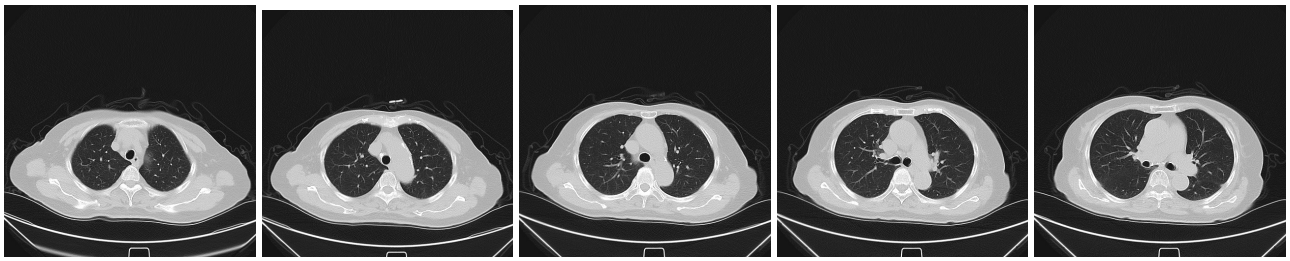

Figure 6. Slices from a Adenocarcinoma case Fair Disease Diagnosis Challenge

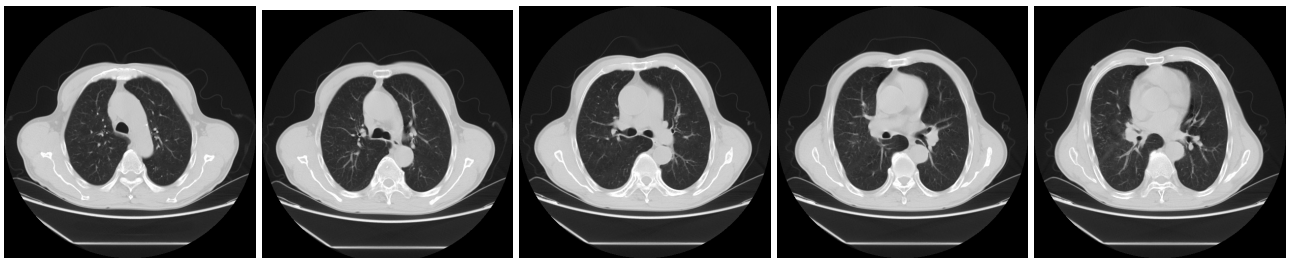

Figure 7. Slices from a Squamous Cell Carcinoma case Fair Disease Diagnosis Challenge

and then averages these losses. Additionally, demographic metadata (patient sex) is explicitly considered during the evaluation stage. Specifically, the baseline model's performance is assessed separately across gender subgroups, enabling detailed analysis of performance disparities. This evaluation strategy aims to uncover potential biases and ensures equitable diagnostic accuracy, aligning model predictions closely with fair and balanced clinical outcomes.

5.2. Pre-Processing & Implementation Details

In the pre-processing stage, all 2-D CT slices were extracted from respective DICOM images. Next, voxel intensity values were computed using a lung window with a width of 350 Hounsfield units (HU) and a level of -1150 HU, and subsequently normalized to the range $[0, 1]$. Data augmentation techniques were applied to enhance the dataset's variability and robustness, including random rotation within the range $[-10°, 10°]$ and horizontal flipping [6, 22]. These augmentations were specifically designed to

Table 4. Multi-Source COVID-19 Detection Challenge: Performance in terms of Macro F1 Score (in %);

Teams	Total	Source 0	Source 1	Source 2	Source 3
ACVLAB[15]	77.6	94.2	74.2	51.2	90.9
FDVTS [21]	77.6	96.2	72.9	49.0	92.1
baseline	70.1	85.8	65.3	44.2	85.1

Table 5. Fair Disease Diagnosis Challenge: Performance in terms of Macro F1 Score (in %);

Teams	Total	Female	Male
FDVTS [17]	70.4	78.3	62.5
baseline	62.3	68.7	55.9

emphasize region-of-interest extraction, primarily focusing on lung areas in the 2-D images.

Regarding the baseline model's implementation, we utilized the ResNet50 architecture as the CNN component. Following this CNN backbone, we incorporated a global average pooling layer, a batch normalization layer, and a dropout layer with a keep probability of 0.8. For sequential modeling, a single unidirectional GRU RNN layer with 128 neurons was employed. Each input CT scan was resized from the original dimensions of $512 \times 512 \times 3$ to $224 \times 224 \times 3$.

Training was conducted with a batch size of 5, meaning each iteration processed five CT scans simultaneously, and the sequence length 't' was set to 700, corresponding to the maximum number of slices across all scans. We adopted the softmax cross-entropy loss function for both challenges. The Adam optimizer was chosen with a learning rate of 10^{-4}. All training processes were executed on a Tesla V100 GPU with 32GB memory, ensuring efficient model convergence and performance.

For the Fair Disease Diagnosis Challenge, a detailed fairness assessment was performed by evaluating predictions separately for male and female subgroups, identifying and addressing any potential performance imbalances.

6. Experimental Results

This section describes the results of the two Challenges, reporting the performance of the winning teams versus the baseline configurations, taking into account that there exists only a single label for the whole CT scan and no labels for each CT scan slice [9].

6.1. Evaluation of the developed Models

In the Multi-Source COVID-19 Detection Challenge, the performance measure (P) is the average macro F1 score achieved across all four sources:

$$P = \frac{1}{4}\sum_{i=0}^{3}\left(\frac{F1^{i}_{\text{covid}} + F1^{i}_{\text{noncovid}}}{2}\right)$$

In the Fair Disease Diagnosis Challenge, the performance measure (P') is the average of per-gender macro F1-scores. To calculate this, one needs to first split the set by gender (Subset A: all male samples; Subset B: all female samples) and then compute the macro F1 on each. Therefore, the performance measure is:

$$P' = \frac{1}{2}\left(\text{F1}^{\text{macro}}_{\text{male}} + \text{F1}^{\text{macro}}_{\text{female}}\right)$$

The macro F1 score is defined as the unweighted average of the class-wise/label-wise F1-scores.

6.2. Competition Results

Table 4 shows the performance of the best submission of the two winning methods in the Multi-Source COVID-19 Detection Challenge, versus each other and versus the baseline approach.

It can be seen that the two methods that achieved the highest performance score achieved more than 10 % improved performance over the baseline approach. However, their performance varied between 50 % and 96 % across different data sources. This illustrates the need for continual learning of the developed systems. Continual learning is an issue which is further examined in papers of the PHAROS-AFE-AIMI Workshop.

Table 5 shows the performance of the best submission of the winning method in the Fair Disease Diagnosis Challenge versus the baseline one.

It can be seen that the method that achieved the highest performance score achieved more than 12 % improved performance over the baseline approach. However, its performance differed by about 20 % between female and male patients. This illustrates the need for fairness of the developed systems. Fairness is also an issue which is further examined in papers of the PHAROS-AFE-AIMI Workshop.

7. Conclusions and Future Work

In this paper we presented the goals and results of the PHAROS-AFE-AIMI Workshop and of the two Challenges

that it contained: the first on multi-source COVID-19 detection and the second on fair disease ndiagnosis. We provided a short description of the PHAROS AI Factory and its health & life sciences application domain, which constitutes the framework in which the results of the PHAROS-AFE-AIMI Workshop will be further examined and fertilized. We presented the databases, extracts from which were used in the two Workshop Challenges. We also presented the developed baseline approaches and their performance in the Challenges.

We also provided a comparison of the performance of the winning methods with the respective baselines, showing that they outperformed the baselines in both Challenges by more than 10-12 %.

These results illustrate the ability of the deep learning and AI enabled methods to appropriately handle segmentation and classification problems in medical imaging. Moreover, they illustrate that Domain Adaptation and Generalization [4] can be a valuable approach for tackling the diversity of datasets obtained across different hospitals and medical centers; this should take place in parallel with explainability and fairness of the developed approaches. Deployment of a recently developed solution has been developed for user-friendly medical usage [5].

8. Acknowledgment

Pharos AI Factory, a project coordinated by GRNET S.A. – National Infrastructures for Research and Technology, operating under the auspices of the Greek Ministry of Digital Governance, has received funding from the Horizon Europe programme: 50% from the EuroHPC Joint Undertaking, Grant Agreement Number 101234269, and 50% from National Resources (Ministry of Digital Governance). The project duration is 36 months: 1/04/2025 – 31/03/2028.

References

[1] Face detection in color images and video sequences. In *2000 10th Mediterranean Electrotechnical Conference. Information Technology and Electrotechnology for the Mediterranean Countries. Proceedings. MeleCon 2000 (Cat. No. 00CH37099)*, pages 498–502. IEEE, 2000. 2

[2] Anastasios Arsenos, Dimitrios Kollias, and Stefanos Kollias. A large imaging database and novel deep neural architecture for covid-19 diagnosis. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2022. 3

[3] Anastasios Arsenos, Andjoli Davidhi, Dimitrios Kollias, Panos Prassopoulos, and Stefanos Kollias. Data-driven covid-19 detection through medical imaging. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE, 2023. 5

[4] Anastasios Arsenos, Dimitrios Kollias, Evangelos Petrongonas, Christos Skliros, and Stefanos Kollias. Uncertainty-guided contrastive learning for single source domain generalisation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6935–6939. IEEE, 2024. 8

[5] Demetris Gerogiannis, Anastasios Arsenos, Dimitrios Kollias, Dimitris Nikitopoulos, and Stefanos Kollias. Covid-19 computer-aided diagnosis through ai-assisted ct imaging analysis: Deploying a medical ai system. *arXiv preprint arXiv:2403.06242*, 2024. 8

[6] Lu Huang, Rui Han, Tao Ai, Pengxin Yu, Han Kang, Qian Tao, and Liming Xia. Serial quantitative chest ct assessment of covid-19: a deep learning approach. *Radiology: Cardiothoracic Imaging*, 2(2):e200075, 2020. 6

[7] Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos Kollias, and Georgios Tagaris. Deep neural architectures for prediction in healthcare. *Complex & Intelligent Systems*, 4(2):119–131, 2018. 5

[8] Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian, and Stefanos Kollias. Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 537–544, 2021. 1

[9] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. A deep neural architecture for harmonizing 3-d input data analysis and decision making in medical imaging. *Neurocomputing*, 542:126244, 2023. 2, 3, 5, 7

[10] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Ai-enabled analysis of 3-d ct scans for diagnosis of covid-19 & its severity. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE, 2023. 1

[11] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Ai-mia: Covid-19 detection and severity analysis through medical imaging. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 677–690. Springer, 2023. 1

[12] Dimitrios Kollias, Karanjot Vendal, Priyankaben Gadhavi, and Solomon Russom. Btdnet: A multi-modal approach for brain tumor radiogenomic classification. *Applied Sciences*, 13(21):11984, 2023. 5

[13] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Domain adaptation explainability & fairness in ai for medical image analysis: Diagnosis of covid-19 based on 3-d chest ct-scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4907–4914, 2024. 1

[14] Dimitrios Kollias, Anastasios Arsenos, James Wingate, and Stefanos Kollias. Sam2clip2sam: Vision language model for segmentation of 3d ct scans for covid-19 detection. *arXiv preprint arXiv:2407.15728*, 2024. 1

[15] Chia-Ming Lee, Bo-Cheng Qiu, Ting-Yao Chen, Ming-Han Sun, Fang-Ying Lin, Jung-Tse Tsai, I-An Tsai, Yu-Fan Lin, and Chih-Chung Hsu. Multi source covid-19 detection via kernel-density-based slice sampling. *arXiv preprint arXiv:2507.01564*, 2025. 7

[16] Ping Li, Shuo Wang, Tang Li, Jingfeng Lu, Yunxin HuangFu, and Dongxue Wang. A large-scale CT and PET/CT dataset for lung cancer diagnosis, 2020. 5

[17] Qingqiu Li, Runtian Yuan, Junlin Hou, Jilan Xu, Yuejie Zhang, Rui Feng, and Hao Chen. Advancing lung disease diagnosis in 3d ct scans. *arXiv preprint arXiv:2507.00993*, 2025. 7

[18] Phivos Mylonas, Evaggelos Spyrou, Yannis Avrithis, and Stefanos Kollias. Using visual context and region semantics for high-level concept detection. *IEEE Transactions on Multimedia*, 11(2):229–243, 2009. 2

[19] Natalia Salpea, Paraskevi Tzouveli, and Dimitrios Kollias. Medical image segmentation: A review of modern architectures. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 691–708. Springer, 2023. 5

[20] Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng, et al. A deep learning algorithm using ct images to screen for corona virus disease (covid-19). *European radiology*, pages 1–9, 2021. 3

[21] Runtian Yuan, Qingqiu Li, Junlin Hou, Jilan Xu, Yuejie Zhang, Rui Feng, and Hao Chen. Multi-source covid-19 detection via variance risk extrapolation. *arXiv preprint arXiv:2506.23208*, 2025. 7

[22] Chuansheng Zheng, Xianbo Deng, Qing Fu, Qiang Zhou, Jiapei Feng, Hui Ma, Wenyu Liu, and Xinggang Wang. Deep learning-based detection for covid-19 from chest ct using weak label. *MedRxiv*, 2020. 6