



D4.6

Final platform for AI dataset benchmarking

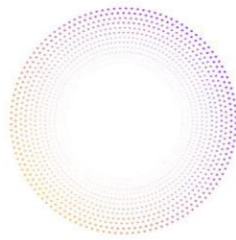
Project Title	AI4Media - A European Excellence Centre for Media, Society and Democracy
Contract No.	951911
Instrument	Research and Innovation Action
Thematic Priority	H2020-EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT) / ICT-48-2020 - Towards a vibrant European network of AI excellence centres
Start of Project	1 September 2020
Duration	48 months



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu

www.ai4media.eu



Deliverable title	Final platform for AI dataset benchmarking
Deliverable number	D4.6
Deliverable version	1.0
Previous version(s)	-
Contractual date of delivery	December 31, 2023
Actual date of delivery	January 16, 2024
Deliverable filename	AI4Media_D4.6.pdf
Nature of deliverable	Demonstrator
Dissemination level	Public
Number of pages	47
Work Package	WP4
Task(s)	T4.6 (Benchmarking of AI Systems)
Partner responsible	UPB
Editor	Mihai Gabriel Constantin
Officer	Evangelia Markidou

Abstract	This document presents the final outcomes of the AI4Media research on benchmarking of AI systems (Task 4.6) and reports on i) the development of the final version of the AI4Media benchmarking platform (i.e. AI4MediaBench), representing the work done between M19 and M40, and ii) the creation and results of several benchmarking competitions, from the beginning of the project up to M40. For the final version of the platform, we present the platform itself, the high-level user and organizer functionalities of the platform, as well as a real-world use case where the platform is used represented by the ImageCLEF2024 competition. For the benchmarking competitions, we present the concepts the different competitions explored, the data they propose, the evolution of the benchmarking tasks along different editions, as well as a set of observations, conclusions, and general trends.
Keywords	Artificial Intelligence (AI), Media, AI benchmarking, benchmarking platform, Evaluation as a Service, benchmarking task, benchmarking initiative, benchmarking competition

Copyright

© Copyright 2024 AI4Media Consortium

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the AI4Media Consortium. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.

All rights reserved.





Authors & Contributors

Name	Organization
Liviu-Daniel Ștefan	UPB
Mihai Gabriel Constantin	UPB

Peer Reviews

Name	Organization
Ioannis Patras	QMUL
Anisa Halimi	IBM

Revision History

Version	Date	Reviewer	Modifications
0.1	07.12.2023	Mihai Gabriel Constantin	First draft with contributions from all partners
0.2	14.12.2023	Mihai Gabriel Constantin	Draft sent to internal reviewers
0.2	12.01.2024	Mihai Gabriel Constantin	Updated version based on internal reviews
1.0	16.01.2024	Mihai Gabriel Constantin	Final version

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.





Table of Abbreviations and Acronyms

Abbreviation	Meaning
AI	Artificial Intelligence
API	Application Programming Interface
CLEF	Conference and Labs of the Evaluation Forum
EaaS	Evaluation-as-a-Service
EEG	ElectroEncephaloGram
EUA	End User Agreement
GAN	Generative Adversarial Network
Grad-CAM	Gradient-weighted Class Activation Mapping
GUI	Graphical User Interface
HTML	Hypertext Markup Language
HTTP	HyperText Transfer Protocol
HTTPS	HyperText Transfer Protocol Secure
mAP	mean average precision
NPM	Node Package Manager
ROC	Receiver Operator Characteristic
SSL	Secure Sockets Layer
URL	Uniform Resource Locator
YAML	YAML Ain't Markup Language





Contents

1	Executive summary	9
2	Introduction	10
3	The AI4Media benchmarking platform	12
3.1	Background	12
3.2	Final user- and organizer-level functions	13
3.2.1	Competition Management	14
3.2.2	User Management	14
3.2.3	Data Management	14
3.2.4	Submission Handling	15
3.2.5	Scoring and Evaluation	15
3.2.6	Interaction Tools	16
3.2.7	Benchmark stages	16
3.2.8	Leaderboard Management	16
3.2.9	Computational efficiency measured via complexity metrics	17
3.2.10	Auditing tools for platform hosts and maintainers	17
3.3	Platform presentation	18
3.3.1	Platform architecture	18
3.3.2	Competition properties	20
3.3.3	User registration	24
3.3.4	Profile settings	25
3.3.5	Running a Competition	25
3.3.6	Participating in a Competition	31
3.4	Use case: ImageCLEF2024	32
3.5	Relevant software, datasets and other resources	36
3.6	Potential for the media industry and beyond	36
4	Benchmarking initiatives and datasets	37
4.1	The ImageCLEF initiative	37
4.2	Predicting Media Interestingness	37
4.2.1	Dataset	37
4.2.2	Benchmarking task	38
4.2.3	Discussion and analysis	38
4.2.4	Relevant publications	38
4.2.5	Relevant software, datasets and other resources	38
4.2.6	Relevance to AI4Media use cases and media industry applications	39
4.3	Predicting Media Memorability	39
4.3.1	Dataset	39
4.3.2	Benchmarking task	40
4.3.3	Discussion and analysis	40
4.3.4	Relevant publications	40
4.3.5	Relevant software, datasets and other resources	41
4.3.6	Relevance to AI4Media use cases and media industry applications	41
4.4	Ensemble learning	41
4.4.1	Dataset	41
4.4.2	Benchmarking task	42





4.4.3	Discussion and analysis	42
4.4.4	Relevant publications	42
4.4.5	Relevant software, datasets and other resources	43
4.4.6	Relevance to AI4Media use cases and media industry applications	43
4.5	GANs usage limitations	43
4.5.1	Dataset	43
4.5.2	Benchmarking task	43
4.5.3	Discussion and analysis	44
4.5.4	Relevant publications	44
4.5.5	Relevant software, datasets and other resources	44
4.5.6	Relevance to AI4Media use cases and media industry applications	44
5	Summary and Conclusions	45





List of Tables

- 1 Development status comparison between the prototype version presented in D4.2 and the current version of the AI4MediaBench platform. 13





List of Figures

1	AI4MediaBench architecture.	19
2	AI4MediaBench registration form.	25
3	AI4MediaBench user profile settings.	26
4	AI4MediaBench competition creation menu.	26
5	AI4MediaBench competition bundle upload.	27
6	AI4MediaBench GUI competition creation form.	28
7	AI4MediaBench benchmark management page.	29
8	Competition page hosted on the AI4MediaBench platform.	29
9	AI4MediaBench competition participants management page.	30
10	AI4MediaBench competition submissions management page.	31
11	AI4MediaBench competition download page.	31
12	AI4MediaBench registering for a competition menu.	32
13	AI4MediaBench uploading a submission for a competition.	33
14	ImageCLEF 2024 Registration Workflow.	36





1 Executive summary

This deliverable summarizes the final results of Task 4.6 “Benchmarking of AI Systems”, part of Work Package 4 “Explainability, Robustness and Privacy in AI”. The deliverable is split into two main parts, presenting the progress made for the development of the final version of the AI4Media benchmarking platform, called AI4MediaBench, and analyzing the results of several multimedia benchmarking competitions, targeting a diverse set of data and annotations.

Section 3 presents the final version of the AI4Media benchmarking platform, developed by UPB. First, we analyze the main high-level user and organizer functions and functionalities, presenting the functionalities and the API functions that are implemented as part of each high-level functionality, as well as their development status and history. Then, we discuss the implementation of these functions in the platform itself, looking at the general software architecture of the platform, open-source software and containers used in the implementation stage, a presentation of how competitions can be created, managed, and ran by competition organizers, and how participants can register and submit their data for a competition. The AI4Media benchmarking platform, called AI4MediaBench, is available at <https://ai4media-bench.aimultimedialab.ro/>. Finally, we analyze the use-case for the AI4MediaBench platform represented by the currently running ImageCLEF2024 benchmarking initiative. This section reflects the work carried out in Task 4.6 during M18–M40 for developing the platform.

Section 4 presents the benchmarking competitions that AI4Media ran and supported as part of Task 4.6, namely the overall ImageCLEF initiative, MediaEval Predicting Media Interestingness, MediaEval Predicting Video Memorability, ImageCLEFfusion, and ImageCLEFmed GANs. For each of these benchmarking competitions, we analyze the concepts targeted by the competition, we describe the dataset, the editions of the competitions, the results of the competition participants, as well as an analysis of the main trends, observations and open questions that the competitions generated throughout their editions.





2 Introduction

Work Package 4 focuses on the research of robustness, explainability, privacy and fairness, as well as the legal and ethical frameworks necessary for trustworthy AI. In this context, Task 4.6 objectives are declared as:

To establish a framework of benchmarking and validating datasets for AI systems with the required legal and ethical constraints in order to ensure the protection of privacy, fairness, and robustness.

As we will show throughout this deliverable, the **AI4Media benchmarking platform** covers these important points by helping benchmarking competition organizers with creating and managing their competitions and data. These aspects represented guiding principles in the research and development stages of the AI4Media benchmarking platform, with the implementation of high-level functions being driven by the desire to ease the workload of competition organizers.

Data privacy is an important component of organizing open-data competitions, with a significant focus on ensuring privacy and anonymity for people that created and annotated the data, as well as improve data access for people that are interested in the hosted tasks. Organizers can ask participants to sign and acknowledge usage agreements, that clearly state their rights and obligations with regards to data access.

Another important component of benchmarking competition organization is ensuring **fairness** towards all the participants and in computing the results. Fairness is ensured by providing tools that help organizers in sharing their data with participants and in providing common evaluation principles that include a common definition, data split, metrics, hidden test set ground truth data, and a common leaderboard composed of participants that adhered to these principles. Furthermore, a reproducibility component can be added by configuring benchmarking tasks so that participants submit docker containers with their methods and these containers are automatically ran over the testing set, ensuring that the proposed systems actually run on the testing data.

Finally, in cases where benchmarking competitions are held for many editions, over many years, the number of proposed systems usually grows significantly, and usually the results get better too as teams improve their methods each year, thus contributing to the **robustness** of the AI models that attempt to solve the task. Furthermore, the high-level observations that organizers can infer with regards to classes of methods or approaches that tend to do better gain more significance as the number of systems gets larger and as results are enforced throughout several editions.

According to the Description of Work, as stated in the Grant Agreement, task T4.6 has the following goals:

- (i) provide annotated data to support the training of the proposed AI systems, (ii) build a community around benchmarking activities to stimulate the innovation and share of resources for better AI, (iii) encourage the development of computationally efficient and effective systems to reduce the power footprint via introducing dedicated metrics for complexity, (iv) foster reproducible systems via re-running the submitted systems in the evaluation phase, (v) building a common repository for sharing the data and to develop approaches for distributed benchmarking with container submission on possibly confidential data (sometimes called Evaluation-as-a-Service – EaaS).

The following two sections cover the work done in T4.6 in order to achieve these goals. Section 3 covers the research and development of the AI4MediaBench platform. We present the main functions of the AI4MediaBench platform, its implementation and the API methods exposed by the platform, covering goal (v), as well as the implementation of computational efficiency metrics





and reproducibility via Docker implementations and in-cloud deployment, covering goals (iii) and (iv). Next, Section 4 shows examples of the benchmarking tasks created as part of T4.6 and the associated datasets, as well as the interest these attract from the scientific community as a measure of participation in tasks, covering goals (i) and (ii).





3 The AI4Media benchmarking platform

This section presents the final version of the AI4Media benchmarking platform. The platform, called AI4MediaBench, was developed by UPB in the context of T4.6 during the period M1-M40. In the following subsections, we briefly remind the reader about the development and main functionalities of the initial version of the platform presented in D4.2 (3.1), offer an overview on the development of the final version of the platform (3.2), present the architecture and final functionalities of the platform (3.3), and, finally, discuss how the platform is used for running the ImageCLEF2024 benchmarking initiative (3.4).

3.1 Background

The previous, prototype version of the AI4Media benchmarking platform was presented in Deliverable D4.2 “Prototype platform for AI dataset benchmarking”. At that point, we defined the main high-level user and organizer functionalities and planned and started the development process. In order to define these functions, we performed an in-depth study of the state-of-the-art on benchmarking platforms, analysing their strengths and weaknesses and identifying a set of requirements for our own platform.

Given our analysis of the state-of-the-art we identified several interesting ideas with regards to what our platform can offer compared with the 20 other Evaluation-as-a-Service platforms we studied. We identified the following differentiating ideas and functions for the AI4MediaBench platform:

- **D1** To the best of our knowledge, no platform offers a method of integrating computational complexity metrics in the analysis of participant system. Our platform proposes an easy to use, integrate, and deploy method of doing this, providing an execution-based method of computing complexity, as well as allowing competition organizers the option to easily integrate their own complexity metrics via an API, as presented in Section 3.2.9.
- **D2**: While some EaaS platforms may provide some options for reproducibility, such as API integration or the use of containers for submitting methods, few platforms offer both options for integrating participant runs or systems. In this regards, we chose to implement both methods, creating an architecture that uses containers in order to create a plug-and-play environment that is easily adaptable to the requirements the task organizers may have, as presented in Section 3.3.1.
- **D3**: Finally, we emphasised the importance of having an EU-based AI benchmarking platform, that would allow AI4Media and interested parties to concentrate and develop features considered as top-priority at a European level by lawmakers, media agencies, and industrial partners.

In order to monitor and understand the current development status and future expectation for the AI4MediaBench platform and its main functions and functionalities, we will define the following six development statuses:

- **(1) Development not started** Development for the proposed functionality not started yet, or in its planning phase;
- **(2) Development started** Development started, but no testable version achieved yet;
- **(3) Prototype finished** A prototype is finished, containing a first version of the modules composing the functionality, and an initial testing phase is finished for this prototype, resulting in a list of bugs and feature requests;
- **(4) Development finished** Final development and major and critical debugging finished, testing still ongoing;





High-level function	Previous status	Current status
Competition Management	3	5*
User Management	3	5*
Data Management	2	5*
Submission Handling	3	5
Scoring and Evaluation	3	5*
Interaction Tools	2	5
Benchmark stages	3	5
Leaderboard Management	3	5
Computational efficiency measured via complexity metrics	1	5*
Auditing tools for platform hosts and maintainers	2	5*

Table 1. Development status comparison between the prototype version presented in D4.2 and the current version of the AI4MediaBench platform.

- **(5) Final version finished** Final development, debugging and thorough testing finished, no further updates expected;
- **(5*) Final version finished, in continuous maintenance** Final development, debugging and thorough testing finished, with likely updates planned for open-source packet updates, security updates, API and cloud implementation updates, as well as implementation of new functions for our ImageCLEF use case, as well as future use cases.

Given these statuses, Table 1 compares the current version of the platform with the previous prototype version presented in Deliverable D4.2. The main high-level functions are either in status 5 or 5*, representing the final version of the platform. During this time, we concentrated our efforts on finishing the development of the platform, thoroughly testing its features and functions, and debugging. While some functions will likely require continuous maintenance, the platform is currently deployed and can be accessed at the following link: <https://ai4media-bench.aimultimedialab.ro/>. Furthermore, the AI4MediaBench platform is open source, and its source code is available at: <https://github.com/AIMultimediaLab/AI4Media-Bench>.

3.2 Final user- and organizer-level functions

AI4MediaBench offer a comprehensive set of APIs (Application Programming Interfaces) that allow users to interact programmatically with various aspects of the benchmarking platform. These APIs provide a standardized way to manage competitions, datasets, submissions, and other functionalities. Users can create, read, update, and delete competitions, datasets, and submissions using the corresponding API endpoints. Additionally, there are APIs for managing user profiles, organizations, participant information, and more. The methods offered by these APIs include GET (retrieve information), POST (create new entries), PUT (update existing entries), PATCH (partially update entries), and DELETE (remove entries). These APIs empower users to automate tasks, integrate external tools, and efficiently engage with the AI4MediaBench platform, fostering flexibility and ease of use for participants, organizers, and developers alike.

Next, we outline all the attributes and user-centric functionalities intended for the platform's final version and provide an update on the current development status, using the categories described in Section 3.1. Below, we introduce the attributes and functions. In the following, we present all the modules of the platform alongside their corresponding functions.





3.2.1 Competition Management

Description: Organizers can create competitions, define the problem statement, objectives, and evaluation criteria. They can customize the competition settings to suit the specific requirements of their research or challenge. Competitions can be created through a GUI wizard or through a bundle (a self-contained unit that encapsulates the essential components of a competition). Organizers can adjust competition rules and settings dynamically, allowing for real-time adaptations based on the evolving needs of the competition or unforeseen circumstances. Organizers can also provide comprehensive documentation and resources for participants. This includes guidelines, starting kits, and any additional information necessary for understanding the competition task and requirements.

Relevant APIs: The following APIs interact with this module:

- List competitions: GET /competitions/
- Create a competition: POST /competitions/
- Read competition details: GET /competitions/id/
- Update competition details: PUT /competitions/id/
- Partially update competition details: PATCH /competitions/id/
- Delete a competition: DELETE /competitions/id/
- Toggle publish status: POST /competitions/id/toggle_publish/
- View competition front page: GET /competitions/front_page/
- View public competitions: GET /competitions/public/
- Get competition results: GET /competitions/id/results/
- Register for a competition: POST /competitions/id/register/

Current status: Final version finished, in continuous maintenance.

3.2.2 User Management

Description: The platform allows organizers to manage participants, including the registration process, team formation, and communication with participants.

Relevant APIs: The following APIs interact with this module:

- List participants: GET /participants/
- Create a participant: POST /participants/
- Read participant details: GET /participants/id/
- Update participant details: PUT /participants/id/
- Partially update participant details: PATCH /participants/id/
- Delete a participant: DELETE /participants/id/
- Create an organization: POST /organizations/
- Validate organization invite: POST /organizations/validate_invite/
- Update organization details: PUT /organizations/id/
- Partially update organization details: PATCH /organizations/id/
- Delete organization member: DELETE /organizations/id/delete_member/
- Delete an organization: DELETE /organizations/id/delete_organization/
- Invite users to an organization: POST /organizations/id/invite_users/
- Update organization member group: POST /organizations/id/update_member_group/

Current status: Final version finished, in continuous maintenance.

3.2.3 Data Management

Description: Organizers can manage and distribute datasets associated with the competition.

Relevant APIs: The following APIs interact with this module:



- List datasets: GET /datasets/
- Create a dataset: POST /datasets/
- Read dataset details: GET /datasets/id/
- Update dataset details: PUT /datasets/id/
- Partially update dataset details: PATCH /datasets/id/
- Delete a dataset: DELETE /datasets/id/

Current status: Final version finished, in continuous maintenance.

3.2.4 Submission Handling

Description: Participants can submit their solutions to the competition, and the platform provides an isolated environment for executing and evaluating them, addressing the dependencies as per user requirements and ensuring the reproducibility of participants' submissions. The platform supports various file formats and allows organizers to define specific submission requirements. Results and competition data can be exported for further analysis.

Relevant APIs: The following APIs interact with this module:

- List submissions: GET /submissions/
- Create a submission: POST /submissions/
- Read submission details: GET /submissions/id/
- Update submission details: PUT /submissions/id/
- Partially update submission details: PATCH /submissions/id/
- Delete a submission: DELETE /submissions/id/
- Cancel a submission: GET /submissions/id/cancel_submission/
- Get detailed result for a submission: GET /submissions/id/get_detail_result/
- Get submission details: GET /submissions/id/get_details/
- Re-run a submission: POST /submissions/id/re_run_submission/
- Create submission leaderboard connection: POST /submissions/id/submission_leaderboard_connection/
- Delete submission leaderboard connection: DELETE /submissions/id/submission_leaderboard_connection/
- Toggle submission visibility: GET /submissions/id/toggle_public/
- Update submission fact sheet: PATCH /submissions/id/update_fact_sheet/

Current status: Final version finished.

3.2.5 Scoring and Evaluation

Description: The platform offers a transparent and standardized framework for scoring and evaluating submissions. Organizers can define evaluation metrics and use them to objectively assess the performance of participants' solutions. Organizers can also define multiple evaluation criteria for submissions, allowing for a comprehensive assessment of different aspects of participants' solutions. This granularity provides a more nuanced understanding of performance.

Relevant APIs: The following APIs interact with this module:

- List tasks: GET /tasks/
- Create a task: POST /tasks/
- Read task details: GET /tasks/id/
- Update task details: PUT /tasks/id/
- Partially update task details: PATCH /tasks/id/
- Delete a task: DELETE /tasks/id/
- List submission scores: GET /submission_scores/



- Create submission scores: POST /submission_scores/
- Read submission scores details: GET /submission_scores/id/
- Update submission scores details: PUT /submission_scores/id/
- Partially update submission scores details: PATCH /submission_scores/id/
- Upload submission scores: POST /upload_submission_scores/submission_pk/
- Delete submission scores: DELETE /submission_scores/id/

Current status: Final version finished, in continuous maintenance.

3.2.6 Interaction Tools

Description: The platform includes features to facilitate communication and collaboration among participants, organizers and platform developers. This includes discussion forums attached to each competition which the organizers can moderate, and messaging systems that allow the organizers to compose and send emails to various groups of participants directly from the platform. Furthermore, we have integrated osTicket¹, an open source support ticket system to allow interaction between the platform users and the platform administrators. The AI4MediaBench customer support platform² allows users to submit tickets related to the platform usage, feature requests, or a support request related to the competition hosted on AI4MediaBench.

Relevant APIs: The following APIs interact with this module:

- Send email to a participant: POST /participants/id/send_email/
- Send email to all the participants POST /competitions/id/email_all_participants/

Current status: Final version finished.

3.2.7 Benchmark stages

Description: The benchmark stages represent different tasks or challenges within the overall competition. For example, a competition might have a data exploration phase, a model development phase, and a final evaluation phase. Organizers may release specific datasets or challenges at the beginning of each phase, guiding participants through a step-by-step process.

Relevant APIs: The following APIs interact with this module:

- List phases: GET /phases/
- Create a phase: POST /phases/
- Read phase details: GET /phases/id/
- Update phase details: PUT /phases/id/
- Partially update phase details: PATCH /phases/id/
- Delete a phase: DELETE /phases/id/
- Get phase leaderboard: GET /phases/id/get_leaderboard/
- Manually migrate a phase: POST /phases/id/manually_migrate/
- Rerun submissions for a phase: GET /phases/id/rerun_submissions/

Current status: Final version finished.

3.2.8 Leaderboard Management

Description: The platform automatically generates and updates leaderboards based on the competition's evaluation metrics. A leaderboard is a dynamic and public ranking of participants based on their performance in the competition. It serves as a central hub where participants can view how well their models or solutions are performing relative to other competitors. This real-time

¹<https://github.com/osTicket/osTicket>

²<https://support.aimultimedialab.ro/open/>





feedback provides participants with insights into their standing and motivates ongoing engagement. Organizers have the option to disable the leaderboards until a specific date, having the freedom of releasing the official results according to a specific schedule.

Relevant APIs: The following APIs interact with this module:

- List leaderboards: GET /leaderboards/
- Create a leaderboard: POST /leaderboards/
- Read leaderboard details: GET /leaderboards/id/
- Update leaderboard details: PUT /leaderboards/id/
- Partially update leaderboard details: PATCH /leaderboards/id/
- Delete a leaderboard: DELETE /leaderboards/id/

Current status: Final version finished.

3.2.9 Computational efficiency measured via complexity metrics

Description: AI4MediaBench empowers users to define complexity measures aligned with the competition’s objectives. Participants submit their solutions, subject to evaluation against user-defined metrics using designated evaluation scripts. The absence of universally applicable complexity metrics necessitates a tailored approach, with appropriateness contingent on the specific goals, tasks, and nature of each competition, which may vary widely. Therefore, the evaluation of algorithmic complexity is often undertaken based on criteria customized to the competition’s requirements, illustrating the platform’s flexibility.

In this context, the computational resources, encompassing the processing power of CPUs and available RAM, significantly influence the speed and effectiveness of task execution. Therefore, we developed a system that allows tasks to be executed in sequential order. This approach ensures fair resource allocation among participants, mitigating potential biases arising from variations in resource availability at different times. This commitment to uniform computational conditions is particularly vital in competitions prioritizing fairness and reproducibility as essential principles.

To establish a foundational framework for organizers, the platform provides competition examples that measure time complexity of the submitted runs.

Relevant APIs: The following APIs interact with this module:

- List queues: GET /queues/
- Create a queue: POST /queues/
- Read queue details: GET /queues/id/
- Update queue details: PUT /queues/id/
- Partially update queue details: PATCH /queues/id/
- Delete a queue: DELETE /queues/id/

Current status: Final version finished, in continuous maintenance.

3.2.10 Auditing tools for platform hosts and maintainers

Description: A complete set of auditing tools is available for platform hosts and maintainers, that could provide important data for helping task organizers getting their tasks online.

Relevant APIs: The following APIs interact with this module:

- List data groups: GET /data_groups/
- Create a data group: POST /data_groups/
- Read data group details: GET /data_groups/id/
- Update data group details: PUT /data_groups/id/
- Partially update data group details: PATCH /data_groups/id/
- Delete a data group: DELETE /data_groups/id/



- View analytics data: GET /analytics/
- List user quota cleanup: GET /user_quota_cleanup/

Current status: Final version finished, in continuous maintenance.

3.3 Platform presentation

AI4MediaBench³ takes center stage in fostering innovation through collaborative data science competitions. These competitions provide a standardized environment where researchers and data scientists contribute to the same computational narrative. Organizers can design and structure competitions, including defining problem formulation and evaluation metrics to guide participants, having all the tools needed to manage participant’s submissions and track the progress of the competitions. Participants can submit code or results, which are executed or evaluated in a controlled setting, ensuring fairness and promoting transparency. Moreover, the platform serves as a dynamic ground for collaboration. Both organizers and participants can share insights, methodologies, and code that allow them to build upon each other’s ideas. AI4MediaBench is based on the open-source Codalab⁴ instance which utilizes the Apache License 2.0⁵, permitting the use, modification, distribution, and sublicensing of the software (more details on the selection of Codalab as the baseline can be found in D4.2).

3.3.1 Platform architecture

AI4MediaBench consists of an orchestrated network of Docker⁶ containers connected and managed through docker-compose. The diagram of the platform is depicted in Figure 1. Next, we provide a detailed breakdown of each container and its role within the platform:

1. Django Container: This central container hosts the Django⁷ project, a high-level Python web framework. It serves as the core for various utility functions, including administrative tasks, backups, and manual alterations through the Python Django shell. The Gunicorn web server internally serves content on port 8000.
2. Caddy Container: Acting as the HTTP/HTTPS web server, Caddy⁸ functions as a reverse proxy for the Django container. It manages SSL/HTTPS functionality and other web server configuration options, serving content on the standard HTTP port 80.
3. Postgres: The default database container contains the Postgres⁹ database used by AI4MediaBench. The specifics, such as database name, user, and password, are determined by the environment file.
4. Compute Worker Container: This container executes submissions for the AI4MediaBench instance. It operates on associated queues, with default workers linked to the default queue.
5. Site Worker Container: Responsible for various tasks related to the Django container, such as unpacking and processing competition bundles.

³<https://ai4media-bench.aimultimedialab.ro/>

⁴<https://github.com/codalab>

⁵<https://www.apache.org/licenses/LICENSE-2.0.html>

⁶<https://www.docker.com/>

⁷<https://www.djangoproject.com/>

⁸<https://caddyserver.com/>

⁹<https://www.postgresql.org/>



6. Minio Container: The storage solution container runs a Minio instance. Minio¹⁰ is an object storage server compatible with the S3 API, providing scalable and distributed storage for AI4MediaBench. The port it serves on is defined by settings in the environment file.
7. Create Buckets Container: This helper container for Minio is designed to create the specified buckets defined in the environment file. Once the buckets are created, this container typically exits, having fulfilled its purpose.
8. Builder Container: This container is tasked with building RiotJS¹¹ tags into a single, unified tag that can be mounted. It utilizes NPM (Node Package Manager) to manage and execute the build process.
9. Rabbit Container: Serving as the task and message management container, Rabbit¹² organizes queues for Celery Tasks and compute workers. It provides the infrastructure for handling asynchronous tasks and coordination between different components.
10. Flower Container: Flower¹³ serves as an administrative utility container specifically for monitoring Celery¹⁴ tasks and queues. It provides a web-based interface for observing and managing Celery processes.

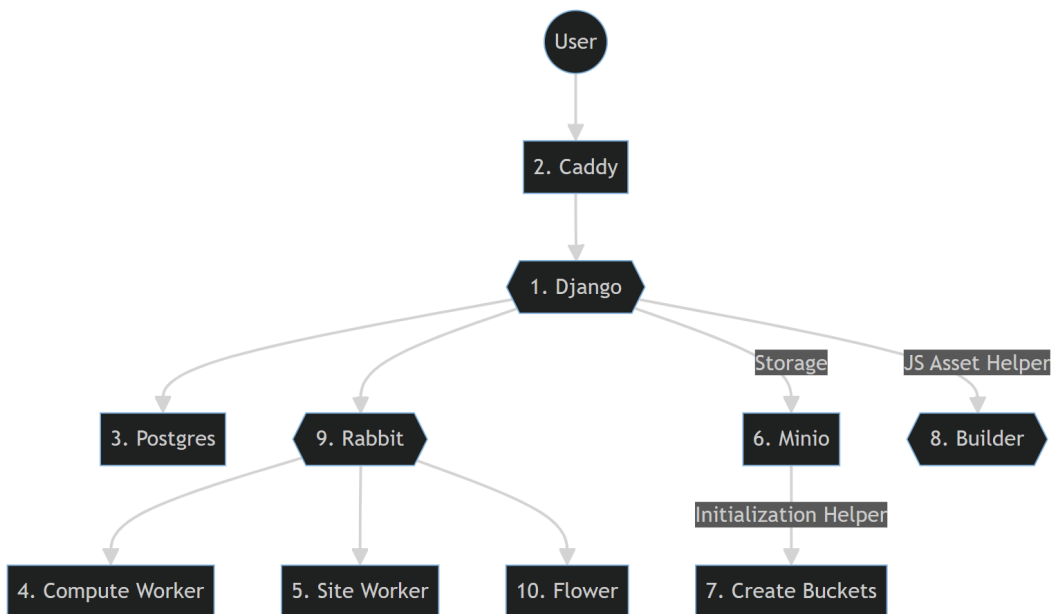


Figure 1. AI4MediaBench architecture.

¹⁰<https://min.io/>

¹¹<https://riot.js.org/>

¹²<https://www.rabbitmq.com/>

¹³<https://flower.readthedocs.io/en/latest/>

¹⁴<https://docs.celeryq.dev/en/stable/>





3.3.2 Competition properties

This section provides a detailed description of attributes within the AI4MediaBench competition definition language, which uses YAML¹⁵. This language is utilized for creating configuration files for competitions in AI4MediaBench. The following are all the parameters that can be set in a competition:

Details

The "Details" section contains fundamental information about the competition, including the title, logo, description, environment, the utilized queue, or the contact person.

Required

- **title**: String literal indicating the competition title.
- **image**: String literal indicating the path to the competition logo file.
- **terms**: String literal indicating the path to the Markdown or HTML page containing the End User Agreement.

Optional:

- **description**: String literal describing the competition.
- **registration_auto_approve**: Boolean indicating whether participation requests are automatically approved (**True**) or require manual approval (**False**).
- **docker_image**: URL link to the docker image for the competition environment.
- **make_programs_available**: Boolean indicating the sharing preferences for ingestion and scoring programs. When set to **True**, it signifies a choice to share the programs with the participants; when set to **False**, it indicates a preference for keeping these program confidential and not accessible to participants.
- **make_input_data_available**: Boolean indicating the sharing preferences for input data. When set to **True**, it signifies a choice to share the input data with participants; when set to **False**, it indicates a preference for keeping the input data confidential and not accessible to participants.
- **queue**: URL for the destination of the queue submissions; specify the compute workers for the competition.
- **enable_detailed_results**: Boolean indicating whether to watch and store detailed results. When set to **True**, it signals an intention to track the outcomes; when set to **False**, it denotes a decision not to track or store detailed results.
- **contact_email**: String literal containing the contact email for organizers.
- **reward**: String literal indicating the reward of the competition.

¹⁵<https://yaml.org/>





Pages

In the "Pages" section, the user can include extra content to share with competition participants. This content will be presented as pages accessible through tabs on the competition detail page.

Required:

- **title:** String literal indicating the title of the page.
- **file:** String literal indicating the file path to a markdown or HTML page.

Phases

A "phase" in a competition signifies a specific stage or period wherein participants engage in submitting their work and competing for performance in a designated task or set of tasks. Competitions are typically structured into multiple phases, providing a timeline and objectives for participants. In a standard machine learning or data science competition, participants progress through three primary stages: registration, development, and testing. During the registration phase, participants sign up on the competition platform, acquainting themselves with the competition's rules and objectives. The development stage takes center stage, where participants actively train their algorithms and enhance their models or solutions using a provided dataset. Frequent submissions are made to assess solution effectiveness. The testing (or evaluation) stage represents the concluding phase, during which participants submit their final entries on a new dataset (unseen by the models during development). These submissions undergo meticulous evaluation, with scores computed based on predefined metrics. Ultimately, the competition winners are determined based on their performance in this conclusive stage.

Required:

- **name:** String literal indicating the name of the phase. If indexes are not provided, the order will be determined by the declaration sequence.
- **start:** Datetime string representing the start of the competition in ISO 8601 format. Phases should be in a sequential, non-overlapping order.
- **end:** Datetime string representing the end of the competition in ISO 8601 format. Phases should be in a sequential, non-overlapping order. Optional for the last phase only. If not specified for the final phase, it remains ongoing indefinitely.
- **tasks:** Array of numerical values indicating the index of defined tasks relative to this phase (refer to the task layout provided below).

Optional:

- **index:** Integer specifying the order of phases.
- **max_submissions:** Positive integer indicating the maximum number of submissions allowed per user per phase. If set to 0, the phase does not permit the uploading of submissions.
- **max_submissions_per_day:** Positive integer indicating the maximum number of submissions allowed per user per day. If set to 0, the phase does not permit the uploading of submissions.





- `auto_migrate_to_this_phase`: Boolean indicating if re-submission of all successful entries from the prior phase to the current phase is triggered at the phase start (True) or not (False). This configuration is not applicable to the first phase of the competition.
- `execution_time_limit`: Numerical data indicating the submission execution time limit, measured in seconds.
- `hide_output`: Boolean indicating whether to conceal output from non-admin users (True) or make it visible (False).
- `starting_kit`: String literal indicating the path to the starting kit provided for participants.
- `public_data`: String literal indicating the path to public data provided for participants.

Tasks

A competition is structured around one or multiple phases. Each phase is associated with one or more tasks. A task represents the problem that submissions aim to solve, essentially making submissions solving a task equivalent to a solution. Each task is characterized by reference data, input data, a scoring program, and an ingestion program.

Required:

- `index`: Positive integer representing the reference id for the task, referenced by solutions (refer to the **Solution** layout) and phases (refer to the **Phase** layout).
- `name`: String literal indicating the name of the task.
- `scoring_program`: String literal indicating the file path for the location of a .zip file or an unzipped directory containing the scoring program.

Optional:

- `description`: String literal indicating the description of the task.
- `input_data`: String literal indicating the path to data provided during the prediction step.
- `reference_data`: String literal indicating the path to data provided to the scoring program.
- `ingestion_program`: String literal indicating the path to ingestion program files.
- `ingestion_only_during_scoring`: Boolean indicating whether the ingestion program should run concurrently with the scoring program (True), allowing communication via a shared directory, or not (False).

Solutions

A solution is a resource linked to a task intended to serve as an illustrative example of a successful submission, be it a code submission or a result submission. The primary objective of a solution is to validate the functionality of a task and provide a baseline for the participants.





Required:

- **index**: Positive integer representing the reference id for the solution.
- **tasks**: Array of internally referenced tasks to which this solution applies.
- **path**: String literal indicating the path to .zip or directory containing the solution data.

Fact Sheet

JSON format metadata contains details related to each submission at the time of its upload. This format enables organizers to define custom fields for participants to complete when submitting a run, such as team name, run description, or other metadata.

Optional:

KEY: Positive integer representing the id for a response field. **QUESTION TYPE**:

- **"checkbox"**: Gives the user a checkbox to select a value from:
 - **Required SELECTION**: [true, false]
- **"text"**: Gives the user a text field to insert a string:
 - **Required SELECTION**: " "
 - **"is_required"**: Boolean indicating the choice of the users to not submit a response (False), or requiring the users to insert a response (True).
- **"select"**: Gives the user a dropdown to select a value from:
 - **SELECTION**: Array of values separated by commas, allowing users to choose from options such as ["Value1", "Value2", "Value3", ..., "ValueN"].
- **is_on_leaderboard**: Boolean indicating whether the corresponding response will be visible on the leaderboard alongside the user's submission (true) or hidden (false).

Leaderboards

The "Leaderboards" section configures the competition results table, offering the flexibility to force or allow manual submissions, display the best score per participant or all the submissions scores per participant, enable anonymous leaderboard presentation, implement automatic migration of submissions between phases, switch between public and private leaderboard modes, handle multiple datasets and multiple custom scoring functions, and incorporating detailed results in an HTML file, with the option to sort columns based on a specified selection.

Required:

- **title**: String literal indicating the title of the leaderboard.
- **key**: String literal indicating the leaderboard identifier.
- **columns**: Array of columns (refer to the column layout provided below).





Optional:

- **submission_rule**: String literal indicating the behavior of the leaderboard regarding new submissions. One of:
 - **Add**: Allows adding a single submission to the leaderboard. It will replace the existing submission (if one exists) from the leaderboard. This option allow the user to select which submission to send to the leaderboard.
 - **Add_And_Delete**: Similar to the **Add** option allowing also the deletion of the submission from the leaderboard. This option allow the user to replace an existing submission from the leaderboard.
 - **Add_And_Delete_Multiple**: Allows adding multiple submissions to the leaderboard and remove those submission from the leaderboard.
 - **Force_Last**: Allows only adding the last submission to the leaderboar. The previous submission added to the leaderboard will be replaced with the current one.
 - **Force_Latest_Multiple**: Force adding all the submission to the leaderboard (multiple entries).
 - **Force_Best**: Adds only the best submission to the leaderboard.
- **hidden**: Boolean indicating whether to hide the leaderboard from non-admin users (True) or display it (False).

Column Details (Required):

- **title**: String literal indicating the title of the column.
- **key**: String literal indicating the identifier for the scoring program.
- **index**: Positive integer indicating the order of the column on the leaderboard.

Optional:

- **sorting**: String literal indicating the sorting order for the column. One of:
 - **desc**.
 - **asc**.
- **computation**: String literal indicating the operation to perform. One of:
 - **sum**
 - **avg**
 - **min**
 - **max**
- **computation_indexes**: An array of indexes representing the columns to which the computation should be applied.
- **precision**: Positive integer specifying the number of digits to which the score should be rounded.

3.3.3 User registration

Participating in a benchmark initiative on the AI4MediaBench platform requires a user account, which can be created with the sign-up page as depicted in Figure 2. During the registration process, essential personal information, namely the participant username and a valid email address, must be provided, while acknowledging and accepting the provided Terms and Conditions.



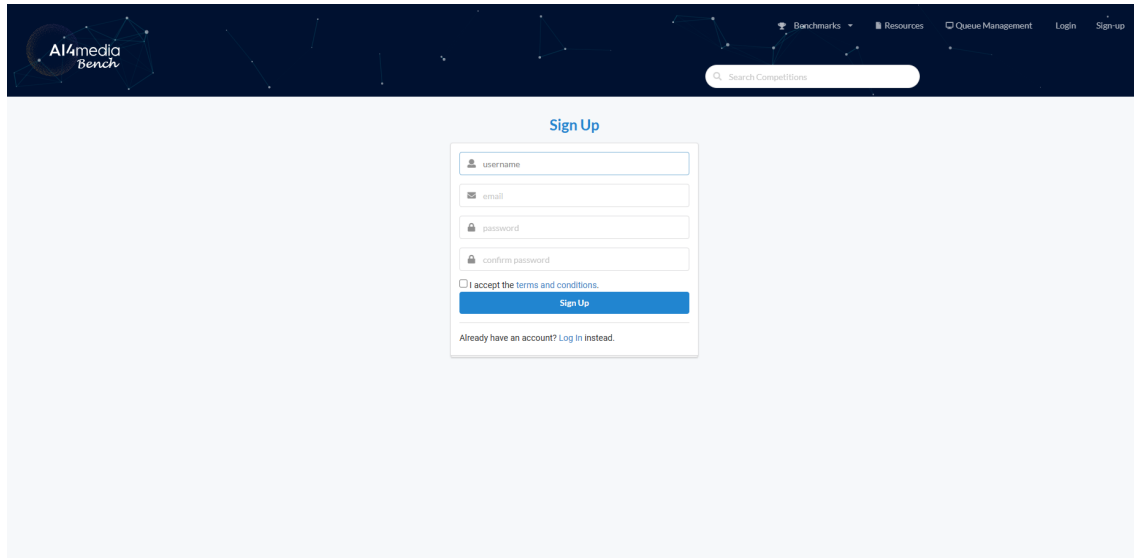



Figure 2. AI4MediaBench registration form.

3.3.4 Profile settings

The platform allows the user to modify his password by clicking on his username located in the upper right-hand corner of the interface. Additionally, within the same menu, the user can also adjust his account settings (update biographical information, manage affiliations with organizations on AI4MediaBench, upload or change his profile picture, change his account password, set personal information such as full name, location, and social media profiles, or modify his email address associated with the AI4MediaBench account), refine notification preferences, and establish organizations—each of which comes with corresponding administrative settings, as shown in Figure 3.

3.3.5 Running a Competition

A competition is a structured event or challenge where individuals or groups participate to demonstrate their skills, knowledge, or abilities in a specific domain. A competition can be created in two ways: using the competition creation form or uploading a competition bundle. Figure 4 illustrates the competition creation menu.

Competition Creation Form

The Competition Creation Form allows users to define and set up a new competition in the AI4MediaBench platform. It serves as a user interface through which competition organizers can provide detailed information and configuration settings for the competition they are creating. Section 3.3.2 describes all the settings that can be set for a competition hosted on AI4MediaBench platform. Figure 6 depicts the GUI competition creation form.

Bundle upload

A competition bundle is essentially a compressed file (.zip file) that consolidates all the components of a competition. It includes:



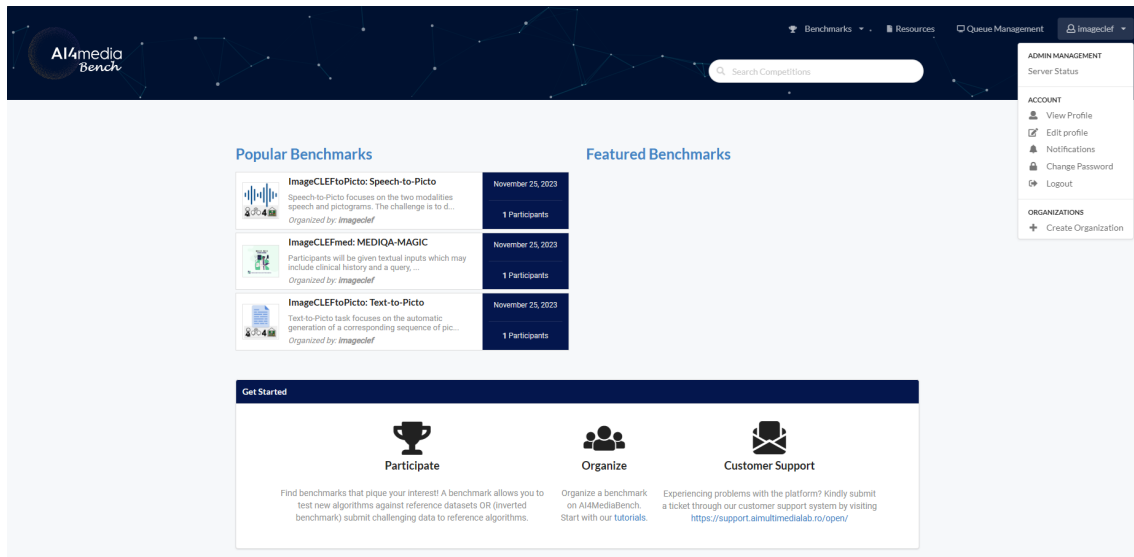


Figure 3. AI4MediaBench user profile settings.

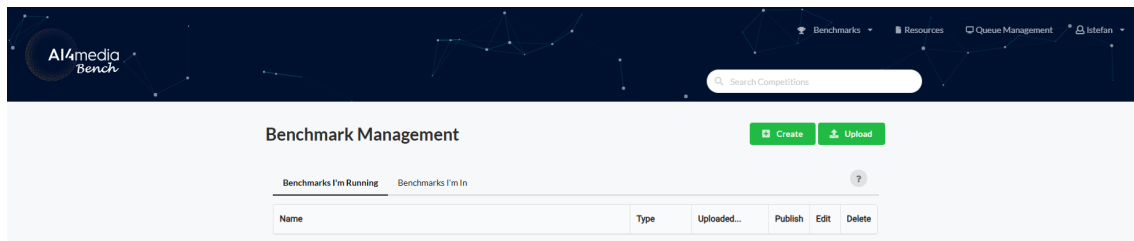


Figure 4. AI4MediaBench competition creation menu.

1. The competition configuration file: The competition configuration file, designated as `competition.yaml` serves as a configuration hub, defining all aspects of the competition. It acts as a central link to various resources needed for competition organization, including HTML files, data sets, and programs. The options are described in Section 3.3.2.
2. HTML Pages: These pages contain descriptive text and participant instructions, providing essential information about the competition in a user-friendly format.
3. Program Files: These files are the building blocks of the competition and include the ingestion program, scoring program, and starting kit.
4. Data Files: These files house training data and reference data, offering participants the necessary datasets for their involvement in the competition.

Competitions necessitate, at a minimum, a scoring program to assess submissions by comparing their results against ground truth. Additionally, an ingestion program is required to execute code submissions in a controlled manner, adhering to an organizer-defined API. Figure 5 illustrates the competition bundle upload page.

Steps to:



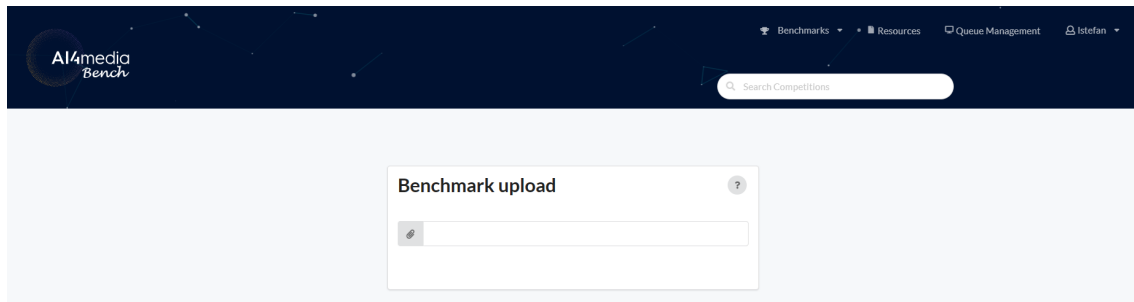


Figure 5. AI4MediaBench competition bundle upload.

Create a competition on AI4MediaBench

1. Navigate to the "Benchmarks" section and select "Management" at the top of the page.
2. Choose one of the following:
 - Click the "Create" button (to create the competition using the GUI Form), or
 - Click the "Upload" button. Use the Open dialog to select the competition bundle (.zip) and click "Open".
3. Return to the dashboard to check the competition. From the dashboard, one can edit, publish, and delete the competition, as well as manage participants and submissions.

Edit a Competition

After creating a competition, the user can later edit the settings. To further update the settings:

1. Navigate to the "Benchmarks" section and select "Management" at the top of the page.
2. Access the "Competitions I'm Running" tab.
3. Click the "Edit" button for the competition to be modified.
4. Make the desired changes, then amend the changes at the bottom of the page by clicking "Submit".

Note: To modify a dataset or program, it must be uploaded in the "Datasets and programs" page under the "Resources" Tab.

Publish a Competition

Publishing a competition makes it visible to the public. Before publishing, competitions are solely visible to the competition organizer:

1. Navigate to the "Benchmarks" section and select "Management" at the top of the page.
2. Access the "Competitions I'm Running" tab.
3. Click the "Publish" button for the desired competition.



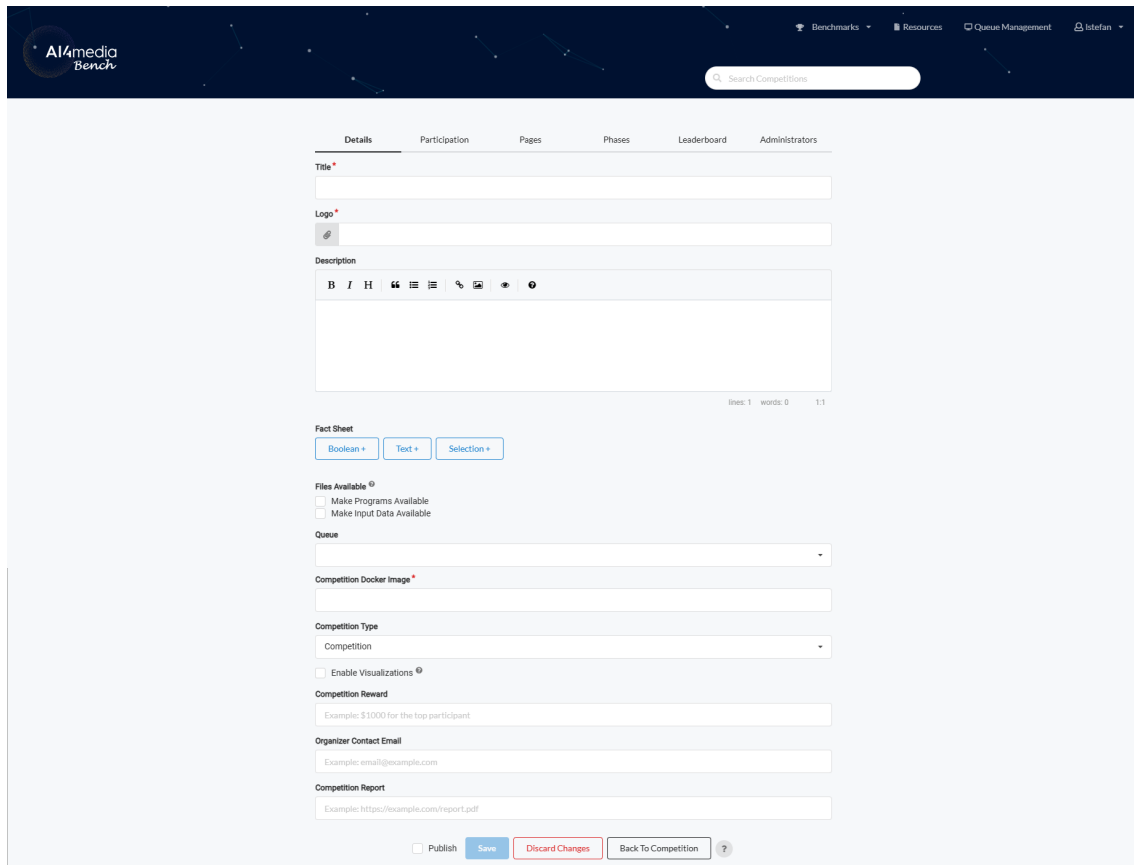



Figure 6. AI4MediaBench GUI competition creation form.

Un-Publish a Competition

Un-publishing a competition removes its public visibility, making it viewable only by the competition organizer:

1. Navigate to the "Benchmarks" section and select "Management" at the top of the page.
2. Access the "Competitions I'm Running" tab.
3. Select a competition and uncheck the "Publish" checkbox.

Delete a Competition

To delete a competition:

1. Navigate to the "Benchmarks" section and select "Management" at the top of the page.
2. Access the "Competitions I'm Running" tab.
3. Click the "Delete" button for the competition to be delete. Confirm the deletion. **Note:** If the competition was previously published, it must be un-published first (See "Un-Publishing a Competition").



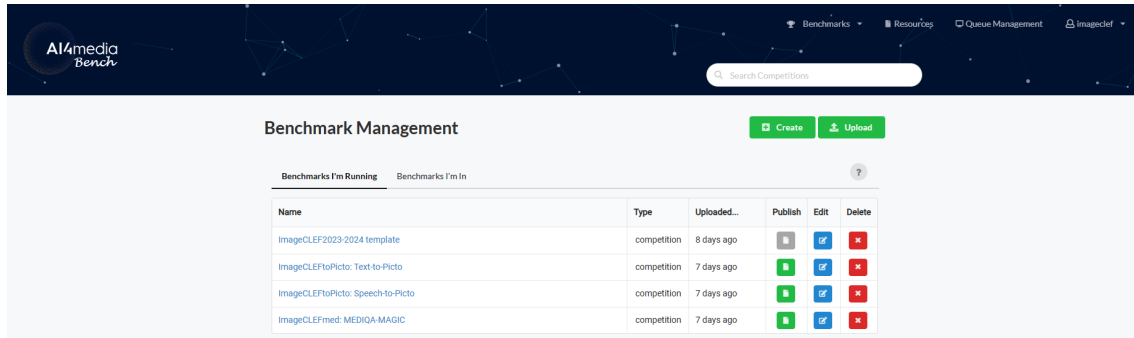


Figure 7. AI4MediaBench benchmark management page.

Figure 7 depicts the benchmark management page, illustrating the processes of creating, editing, publishing, unpublishing, and deleting a competition within the AI4MediaBench framework.

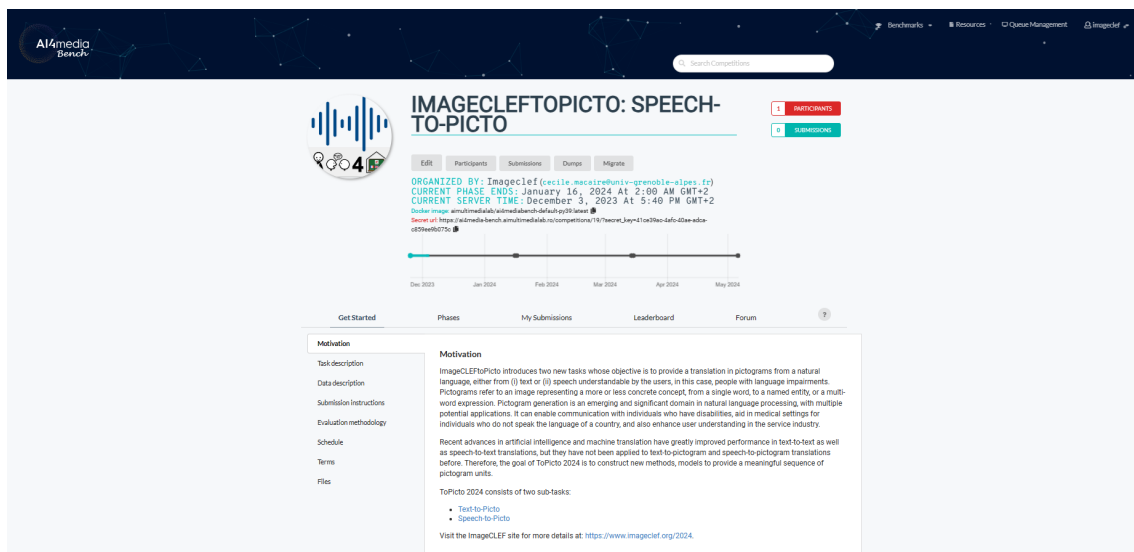


Figure 8. Competition page hosted on the AI4MediaBench platform.

View Participants

Upon participant registration, they are added to a participant list. Follow these steps to view participants for a competition.

1. Navigate to the "Benchmarks" section and select "Management" at the top of the page.
2. Access the "Competitions I'm Running" tab.
3. Select a competition, and click the "Participants" button to view the participant list.

Figure 8 illustrates the settings for managing a competition on AI4MediaBench.



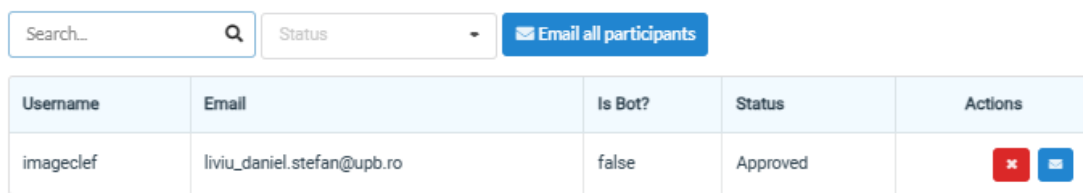


Approve/Deny Participants

The competition organizer must approve or deny each participant. To manage participant approvals or denials:

1. Follow the instructions to view participants for a competition. Pending participants are listed at the top.
2. For each participant, select a status of "Approve" or "Deny" from the drop-down menu. Optionally, provide a short message to the participant in the "Reason" field.
3. Click the "Process" button to complete the approval process.

After initial acceptance, participant permissions can be revoked/denied from the list of participants. Figure 9 displays the page for managing competition participants in AI4MediaBench.





Username	Email	Is Bot?	Status	Actions
imageclef	liviu_daniel.stefan@upb.ro	false	Approved	 

Figure 9. AI4MediaBench competition participants management page.

View Submissions

The competition organizer can review all submissions, displayed in a table with the following details:

- Participant ID
- Filename (click to download the competition bundle for that submission). This option opens another set of options:
 - Download the submission
 - View standard output and error logs
 - Download evaluation output from the prediction and scoring steps
- Username of the owner
- The phase identifier
- Time and date of submission
- The submission status. One of: cancelled, failed, finished, preparing running, scoring, submitted, submitting)
- Action menu:
 - re-run submission
 - delete submission
 - publish or unpublish submission to or from the leaderboard
 - make the submission public

Figure 10 illustrates the page for managing competition submissions in AI4MediaBench.

Download a Competition

To download a competition:





<input type="checkbox"/> All	ID #	File name	Owner	Phase	Date	Status	Score	Actions
<input type="checkbox"/>	19	sample_result_submission.zip	imageclef	Registration	2023-12-03 15:42	Finished	0.33	<input type="button" value="Refresh"/> <input type="button" value="Delete"/> <input type="button" value="Download"/> <input type="button" value="Share"/>

Figure 10. AI4MediaBench competition submissions management page.

1. Navigate to the "Benchmarks" section and select "Management" at the top of the page.
2. Access the "Competitions I'm Running" tab.
3. Click the "Dumps" button for the competition to be downloaded.
4. Click the "Create Dump" button.
5. After clicking "Create competition dump," refresh the page until it is ready, and you can download the competition bundle.

Dump with keys
Dump with files
<input type="button" value="Download"/> Bundle: imageclef - competition_bundle
<input type="button" value="Download"/> Dump: ImageCLEF2023-2024 template Dump #1 Created 2023-11-25 20:12:03

Figure 11. AI4MediaBench competition download page.

Figure 11 illustrates the page for downloading the competition settings.

3.3.6 Participating in a Competition

To contribute to a benchmark challenge, one will be prompted to acknowledge and comply with the benchmark rules, as well as complete the registration for the specific challenge. Upon sending the registration request, the benchmark organizers will be notified, and they will assess and approve the registration request. Certain competitions offer the option to collaborate as an organization and will display the organization's designation on the leaderboard when submitting entries. In specific instances, participation as an organization may necessitate a formal organization registration process. Figure 12 illustrates the competition registration page.

A competition is displayed using a tab navigation system allowing to transit between distinct areas of interest and information within the competition framework. It contains the following tabs:

- **Get Started:** Contains supplementary content for the competition participants such as dedicated pages and files. These pages are accessible through tabs.



- Phases: Contains a diagram list with details on each phase in the order in which they're active.
- Participation: Prompts to accept the rules and register to that competition.
- My Submissions: This view presents a comprehensive table that compiles all of one's submitted entries while also providing with the capability to upload new submissions. The benchmark allows for two discrete submission categories: code or results. Code submissions encompass a metadata file specifying the execution command, whereas result submissions provide the solution to the problem, without executing any code on the platform.
- Results: This section displays the leaderboard, offering a comprehensive view of benchmark standings and outcomes. It should be noted that certain benchmarks may exclusively unveil their results upon the conclusion of the benchmarking campaign.
- Forum: This view contains the official forum of the competition.



Figure 12. AI4MediaBench registering for a competition menu.

Steps to upload a solution for a competition

1. Sign in to AI4MediaBench.
2. Select a benchmark from the available benchmarks list.
3. Navigate to "Participation" and register to the selected benchmark.
4. Navigate to the "My Submissions" tab.
5. Click on the paper clip logo, then select the solution bundle for submission.

The "My submissions" tab allows making new submissions and inspect prior submissions corresponding to each stage of the competition. Figure 13 illustrates the menu for uploading a submission for a competition.

3.4 Use case: ImageCLEF2024

In this section, we present a use case for our platform, represented by the ImageCLEF 2024 benchmarking campaigns¹⁶. With this use case, we show a set of fundamental features of the platform:

¹⁶<https://www.imageclef.org/2024>



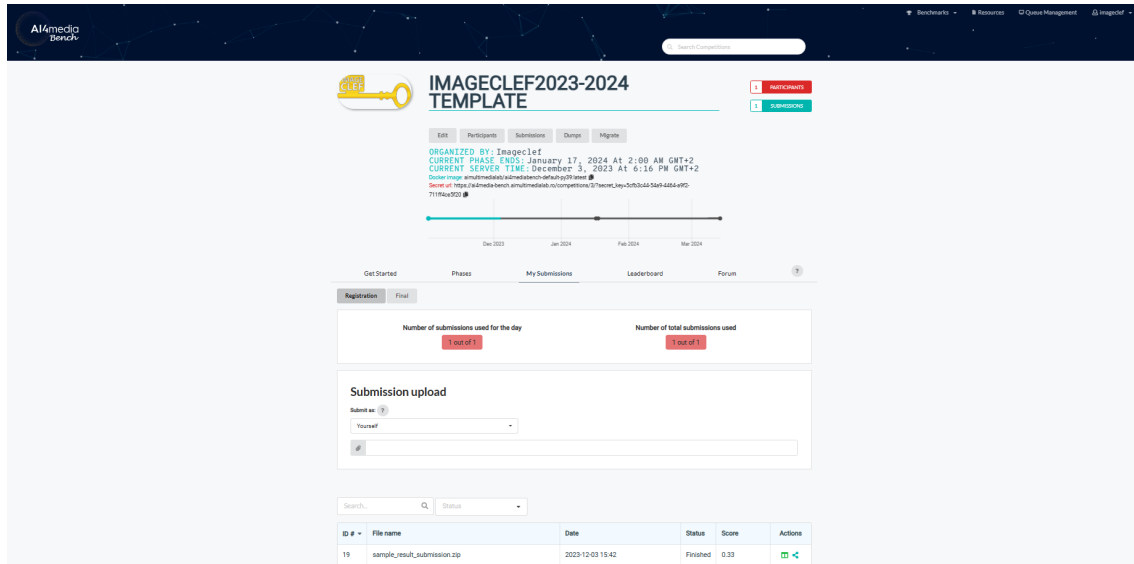


Figure 13. AI4MediaBench uploading a submission for a competition.

(1) custom API for private registration system, (2) the results evaluation mode, (3), transparency guaranteed by code submission, (4) reproducibility guaranteed by docker, (5) flexibility of benchmark bundles, and (6) customized computational resources.

Background

ImageCLEF, part of the Conference and Labs of the Evaluation Forum (CLEF)¹⁷, represents a continuous evaluation campaign that started in 2003, designed to stimulate the evaluation of cutting-edge technologies specifically focused on annotating, indexing, and retrieving multimodal data. ImageCLEF’s core mission is to stimulate advancements that enhance information access within large and diverse data collections, catering to a spectrum of usage scenarios and domains through systematic evaluation of information access systems, primarily through experimentation on shared tasks. The 2024 edition of the ImageCLEF organizes the following three main tasks, each one consisting of one or more sub-tasks:

- **ImageCLEFmedical:** The task targets the generation of knowledge for medical images. In this context, it hosts four main challenges:
 - *Caption*¹⁸: Seeks solutions for automatically identifying individual components from which captions are composed in Radiology Objects in COnText images. The task comprises two sub-tasks: (i) Concept Detection Task—involving the identification and localization of relevant concepts within a large collection of medical images. These concepts serve as fundamental elements for constructing captions, representing individual components contributing to the overall scene understanding, and (ii) Caption Prediction Task—involving the generation of coherent captions for entire images. Building on the concept vocabulary identified in the Concept Detection task and leveraging visual information about how these concepts interact within the image, participating systems aim to compose meaningful and contextually relevant captions. This task emphasizes

¹⁷<https://clef2024.imag.fr/>

¹⁸<https://www.imageclef.org/2024/medical/caption>





understanding the interplay and relationships among visible elements in the image. The success of this task is determined by the system's ability to generate captions that capture the holistic interpretation of the visual content.

- *ImageCLEFmed VQA*¹⁹: Aims to harness artificial intelligence for generating medical images based on textual input, utilizing optimal prompts for off-the-shelf diffusion models. Building on the dataset from the first edition of MEDVQA-GI²⁰, the task's ultimate goal is to enhance the diagnosis and classification of real medical images through AI-generated imagery. The task is divided into two sub-tasks: (i) Image Synthesis—involving the exploration of text-to-image diffusion models to create a diverse dataset of medical images derived from textual prompts. Examples include generating images of different pathologies based on text descriptions, like creating an image for "An early-stage colorectal polyp" from the corresponding textual description. (ii) Optimal Prompt Generation—focuses on creating optimal textual prompts guiding off-the-shelf diffusion models in producing realistic medical images. These images span various modalities, from MRIs and CT scans to endoscopic imagery depicting different medical conditions. For instance, participants, given a medical condition like "late-stage stomach ulcer" and an off-the-shelf diffusion model, must generate an optimal textual prompt guiding the model to produce an accurate and realistic image of the condition.
- *GANs*²¹: Focuses on investigating the hypothesis that GANs generate medical images containing "fingerprints" from the real images used during generative network training. If confirmed, artificial biomedical images may face the same sharing and usage restrictions as real sensitive medical data. Conversely, if the hypothesis is disproven, various generative networks could potentially be employed to create extensive datasets of biomedical images without ethical and privacy concerns. Participants will evaluate the hypothesis on two levels: identifying the source dataset used for training and exploring the problem of detecting, and potentially isolating, image regions in generated images that inherit patterns from the original ones.
- *MEDVQA-MAGIC*²²: The task focuses on the problem of Multimodal And Generative TelemedICine (MAGIC) in the area of dermatology. Inputs include text which give clinical context and queries, as well as one or more images. The challenge consists of generating textual response to queries.
- **toPicto**²³: The toPicto task focuses on providing a translation in pictograms from a natural language, either from text or speech understandable by the users, in this case, people with language impairments. ToPicto consists of two sub-tasks: (i) Text-to-Picto²⁴ – focusing on the automatic generation of a corresponding sequence of pictogram terms from a French text. This challenge can be seen as a translation problem, where the source language is French, and the target language is French pictogram terms, (ii) Speech-to-Picto²⁵ – focusing on the two modalities, speech and pictograms. The challenge is to directly translate speech to pictogram terms without going through the transcription dimension, which is the focus of the speech community with current spoken language translation systems.
- **ImageCLEF recommending**²⁶: The task focuses on addressing a critical challenge for researchers and heritage professionals related to Europeana, a digital platform with over 53

¹⁹<https://www.imageclef.org/2024/medical/vqa>

²⁰<https://www.imageclef.org/2023/medical/vqa>

²¹<https://www.imageclef.org/2024/medical/gans>

²²<https://ai4media-bench.aimultimedialab.ro/competitions/20/>

²³<https://www.imageclef.org/2023/topicto>

²⁴<https://ai4media-bench.aimultimedialab.ro/competitions/18/>

²⁵<https://ai4media-bench.aimultimedialab.ro/competitions/19/>

²⁶<https://www.imageclef.org/2024/recommending>





million records. In this context, the task calls for participants to develop recommendation methods and systems. Using a dataset sourced from Europeana, participants are required to implement these methods to provide recommendations for both individual items and editorials.

Implementation

The 2024 edition of the ImageCLEF benchmark campaign involves tasks where participants submit results, each task being associated with unique datasets. AI4MediaBench employs a flexible competition structure, comprising various phases, each tailored to different tasks, and evaluating specific datasets. Within the platform, the benchmark organizers can specify the docker image by indicating its docker hub name and tag, encapsulating all software dependencies into a lightweight virtual image. The provided docker by the benchmark organizer ensures that the scoring program runs within an environment identical to the one with the installed packages. This docker is fetched every time a benchmark's scoring program is executed. Different benchmarks use distinct dockers, either supplied by organizers or built by platform administrators based on organizers' specifications. A default docker is available for more general benchmarks.

In addition to submitting results, certain tasks necessitate participants to provide the source code of their solution along with the result files. Participants using AI4MediaBench have the capability to compress their code along with the result file. Simultaneously, the task's scoring program is set up to extract the required file from the compressed archive, calculate the metrics, and update the results on the platform leaderboard. The platform provides task organizers with the flexibility to download participants' zip files at any point in time.

An additional requirement within the ImageCLEF benchmark campaign is the implementation/usage of a registration system aimed at collecting diverse participant data, distinct from the information configurable by users in AI4MediaBench. In the context of ImageCLEF, a participant is officially registered for an ImageCLEF task upon signing an End User Agreement (EUA) validated by the ImageCLEF task organizers. In the event of necessary registration modifications, the applicant should be allowed to submit an alternate EUA according to the organizers' stipulations. AI4MediaBench facilitates the addition or customization of registration forms within the task's terms and conditions section. A custom API can be set by the platform administrators to allow the organizer retrieve the applicants data via an API key and a form ID. This feature allow organizers to gather all necessary data from users.

To validate the EUAs, we have set up a workflow involving automating actions based on the data submitted through a designated form—a workflow is triggered when a form is submitted. Upon completion of the registration form, the participant will receive an email containing a summary of the provided information, accompanied by an acknowledgment that his registration is currently in a "pending" state. Form validators will be notified via email about the requisite action for an ImageCLEF registration, prompting them to review the participant's submitted information. Throughout this phase, the workflow status defaults to ACTIVE, signifying that a decision regarding acceptance or denial is pending. Two alternative states, namely APPROVED and DENIED, are available for selection by the validators. In the event of a status transition from ACTIVE to DENIED, the participant will receive a follow-up email containing his information and a justification for the denial. The participant will be instructed to resubmit the information based on the guidelines provided in the denial message. Conversely, if the status changes from ACTIVE to APPROVED, the participant will be notified via email that his registration has been accepted, granting him access to the platform for the tasks he has registered for. Figure 14 illustrates the ImageCLEF 2024 workflow. Furthermore, the organizers can utilize the user management function on the platform to either approve or deny applicants' requests based on the status of their End





User Agreement (EUA).

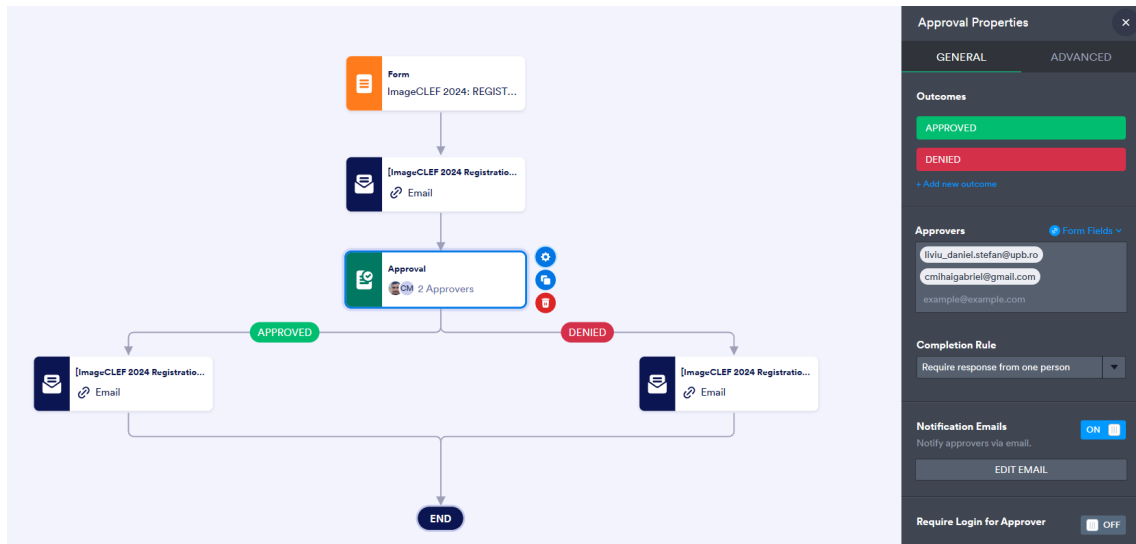


Figure 14. ImageCLEF 2024 Registration Workflow.

Finally, ImageCLEF comprises three core tasks, each encompassing various sub-tasks. Sub-tasks within a main task leverage similar resources, prompting the preference for reusing pre-configured competitions rather than starting anew with configuration steps. The versatile design of the AI4MediaBench bundle allows for the effortless adaptation of an existing competition to align with the specific demands of a new sub-task. All other configuration files remain reusable. This capability enables the straightforward cloning and expansion of the initial benchmark to create similar benchmarks adapted to distinct data types, scoring programs, or leaderboard structures.

3.5 Relevant software, datasets and other resources

- Online and functioning version of the platform, currently running ImageCLEF 2024: <https://ai4media-bench.aimultimedialab.ro/>
- Source code for the platform: <https://github.com/AIMultimediaLab/AI4Media-Bench>

3.6 Potential for the media industry and beyond

The platform is currently deployed and used in the ImageCLEF 2024 benchmarking initiative, hosting four main tasks, each one comprised of several smaller subtasks. Furthermore, the platform can be used to host various media or non-media benchmarking tasks, either for research purposes, when creating an open benchmarking competition, or for industrial purposes, when creating an environment for internal testing.





4 Benchmarking initiatives and datasets

Throughout the duration of the project, several benchmarking tasks and initiatives have been organized, sponsored or supported by AI4Media as part of Task 4.6. In general, these benchmarking tasks provide a common evaluation platform for AI models and approaches, where participants have access to the same definitions and interpretations of the studied concepts, data and data splits, pre-computed features, annotations and ground truth data, and use case scenarios. These tasks therefore represent an important venue for bringing attention to a certain topic in media data processing, thus creating and maintaining a community around the respective topic. This section presents the benchmarking tasks developed throughout the project's duration as part of Task 4.6.

4.1 The ImageCLEF initiative

Contributing partners: [UPB](#), [HES-SO](#)

ImageCLEF²⁷ is a long running conference and benchmarking initiative for evaluation tasks, dealing with diverse topics on cross-language annotation and retrieval of media items, with its first edition being held in 2003. AI4Media has been a supporter and sponsor of ImageCLEF since 2022²⁸ and continues to support the initiative in its current 2024 edition²⁹. During this time, ImageCLEF hosted AI benchmarking tasks in various domains, including medical data processing, generative networks, recommendation systems, and ensemble learning. Furthermore, in its current 2024 edition, ImageCLEF provides the use case for the AI4Media benchmarking platform, as presented in the previous section of this deliverable.

4.2 Predicting Media Interestingness

Contributing partners: [UPB](#), [IDF](#)

The Interestingness10k dataset [1] is the final result of the two editions of the MediaEval Predicting Media Interestingness task held in 2016³⁰ and 2017³¹. While the benchmarking tasks ran outside AI4Media, the analysis of the participating teams, their results, prediction methods and overall observations have been part of the AI4Media project. The task targets the prediction of image and video interestingness, defined according to a use case scenario deployed at Technicolor France³², where an automated system should be able to select the most interesting image or video sequence for an underlying movie [2].

4.2.1 Dataset

The final version of the dataset, as published in [1] is divided into two parts, namely image interestingness prediction, where key-frames are extracted from segments of videos, and video interestingness prediction, where video segments are extracted from longer movies. While the initial version of the dataset uses 5,054 image and video samples as the training set and 2,342 samples as the testing set, the final version is extended, with 7,396 images and videos in the training set, and 2,435 samples in the testing set.

Annotations were performed manually by 270 trusted assessors, i.e., master and doctoral students and faculty staff with a good understanding of the given task and its definition. Annotations

²⁷<https://www.imageclef.org/>

²⁸<https://www.imageclef.org/2022>

²⁹<https://www.imageclef.org/2024>

³⁰<https://www.multimediaeval.org/mediaeval2016/mediainterestingness/>

³¹<https://www.multimediaeval.org/mediaeval2017/mediainterestingness/>

³²https://www.interdigital.com/data_sets/interestingness-dataset





were carried out with the help of a custom online platform, that uses a pairwise comparison approach, i.e., assessors are shown pairs of images or videos, and are asked to identify the sample in the given pair that is more likely to make them watch the entire source movie. Two metrics are used for measuring the performance of automatic interestingness prediction systems, namely mean average precision (mAP) and mean average precision over the top 10 items (mAP@10).

In order to aid researchers that may not be from the computer vision domain, as well as provide help for junior researchers, a set of features are computed and distributed for this dataset: Dense SIFT [3], HoG [4], LBP [5], GIST [6], Color Histogram, layers extracted from the AlexNet [7] and C3D [8] deep neural networks.

4.2.2 Benchmarking task

As previously mentioned, the benchmarking task that validated this dataset ran for two years, in 2016 and 2017, as part of the MediaEval Benchmarking Initiative ³³. While the 2016 edition gathered a number of 27 runs for both image and video prediction, the 2017 edition attracted more interest, with 33 systems being submitted for image prediction, and 42 for video interestingness prediction. Final results show a top performance of mAP = 0.3125 for image prediction, and a mAP value of 0.2228 for video prediction, with a promising increase in top performance between the two years of the competition, namely 25.75% and 22.75% for image and video prediction respectively.

4.2.3 Discussion and analysis

Given the high number of systems submitted during the two editions of the benchmarking tasks, some interesting trends are detected, that can be summarized as follows:

- Interestingness entails a high degree of annotator subjectivity;
- What is interesting in an image? Analysis of annotator data reveals some specific patterns such as colored and aesthetic frames, and presence of people;
- System performance for prediction is much lower than for more objective tasks, such as object detection or scene classification. Even humans, while significantly surpassing machine performance, do not achieve perfect prediction;
- Current state-of-the-art deep neural networks, while achieving good performance, they are not the top prediction performers;
- What do deep neural networks learn? Grad-CAM analysis shows an explicit focus on the main subject, but also on the area around. The presence of people triggers activation also around the faces;
- Late fusion and ensemble systems represent a good option with implicit higher performance than single systems of any type.

4.2.4 Relevant publications

- Constantin, M. G., Ștefan, L. D., Ionescu, B., Duong, N. Q., Demarty, C. H., Sjöberg, M. (2021). Visual interestingness prediction: a benchmark framework and literature review. *International Journal of Computer Vision*, 129, 1526-1550.

4.2.5 Relevant software, datasets and other resources

- Dataset: https://www.interdigital.com/data_sets/interestingness-dataset

³³<https://multimediaeval.github.io/>





4.2.6 Relevance to AI4Media use cases and media industry applications

Media interestingness is one of the most important concepts related to the subjective perception of data in general, and of media data in particular. Thus, this dataset, as well as the repository of knowledge created thanks to the associated benchmarking tasks represent an important resource for developing AI models and approaches that predict media interestingness. This research can represent the dataset and knowledge base for an interestingness prediction system that would best be integrated into Task 6.6 - “Measuring and Predicting User Perception of Social Media”, and best be applied to UC3 “AI for Vision”, feature 3C2-13 - “Modality-dependent sentiment analysis”.

4.3 Predicting Media Memorability

Contributing partners: UPB

The Predicting Video Memorability task is a long running benchmarking task that is hosted at the MediaEval Benchmarking initiative, running from 2018³⁴ to 2022³⁵, and being continued with the newest edition in 2023³⁶. This task is supported by AI4Media since its 2020 edition [9]. The task targets the prediction of short- and long-term video memorability, using several different datasets and modalities, being focused on short-form social media shared videos.

4.3.1 Dataset

During its six editions, three different memorability datasets were used, namely the Memento10k [10], the VideoMem [11], a portion of the TRECVID 2019 Video-to-Text annotated for memorability [12], as well as the EEGMem dataset using physiological data based on EEG recordings [13].

The Memento10k dataset is comprised of 10,000 3-second long videos depicting in-the-wild scene, annotated for short-term memorability, while also containing a set of user generated descriptions for each video. The dataset is split into 7,000 videos for the training set, 1,500 for the validation set, and 1,500 for the testing set. The VideoMem dataset also has 10,000 longer soundless videos (7 seconds on average), with short- and long-term memorability annotations associated, and a set of user generated descriptions, split into 7,000 videos for training, 1,000 for validation, and 2,000 for the testing set. The TRECVID memorability dataset uses a subset of the popular TRECVID dataset, featuring 6,000 Twitter Vine videos, with 4,384 videos belonging to the training set, 1,116 to the validation set, and 500 for the testing set, annotated for short- and long-term memorability. Finally, the EEGMem dataset, contains EEG samples and features extracted during the memorability annotation phase. Two main metrics were used throughout the task, namely Spearman’s Rank Correlation for the prediction and generalization tasks, and Area under the ROC Curve for the EEG task.

All datasets are annotated using the same protocol. A sequence of videos are shown to human assessors, using an online annotation tool. For accurately measuring the memorability of the videos on the short-term, videos are repeated after a few minutes from the retention point, and viewers are supposed to press the space bar whenever they recognize a video. On the other hand, long-term memorability entails another viewing session that is programmed after 24 to 72 hours have passed.

Several features are extracted from the videos in all datasets and provided to participants as a starting baseline, namely: (i) image-level features: features extracted from the AlexNet [7], VGG [14], DenseNet121 [15], ResNet50 [16], and EfficientNetB3 [17] deep neural networks, HOG [4], HSV and RGB histograms, LBP [5]; and (ii) video-level features extracted from the C3D [8] network.

³⁴<https://www.multimediaeval.org/mediaeval2018/memorability/>

³⁵<https://multimediaeval.github.io/editions/2022/tasks/memorability/>

³⁶<https://multimediaeval.github.io/editions/2023/tasks/memorability/>





4.3.2 Benchmarking task

The six editions of the task featured different setups with regards to the tasks and datasets deployed. Three main directions were proposed during the editions of the memorability task, namely: (i) memorability prediction, where participants are asked to use the data extracted from the same dataset for all stages of training, validation, and testing; (ii) generalization, where participants are asked to use different datasets for training and testing the systems, with the role of checking whether systems are overfitting on the training dataset or actually learning memorability-defining characteristics of the targeted concepts; (iii) EEG-based prediction, where participants are asked to use EEG data to infer memorability. Overall, short-term memorability prediction ran for five years during 2018 and 2022, long-term memorability prediction ran for four years between 2018 and 2021, the generalization task for three years between 2021 and 2023, and the EEG task also for three years between 2021 and 2023.

Given its long history, an impressive number of runs have been submitted to this task. In total, 358 runs have been submitted, including those for the 2022 edition of the task. From these, 207 deal with short-term memorability prediction, 122 with long-term memorability prediction, 24 with generalized prediction, and 5 runs belong to the EEG task. Results vary greatly, depending on the task and dataset, however maximum results seem to plateau around a Spearman's value of 0.7–0.75 for short-term memorability prediction, and 0.25–0.3 for long-term prediction, while maximum values for generalization and EEG data are still an open research question, with tasks associated with them still running this year.

4.3.3 Discussion and analysis

Unlike the previous section dealing with media interestingness, no paper that analyzes the competitions throughout their six editions has been written yet, although one such paper is in our plans. However, some conclusions we can draw at this point would be:

- Short-term memorability scores are generally easier to predict by automated systems compared with long-term scores;
- While memorability itself is not as subjective as other concepts like interestingness, there still is a significant degree of subjectivity;
- Ensemble and early or late fusion systems seem to perform better on average.

4.3.4 Relevant publications

- Garcia Seco De Herrera, A., Savran Kiziltepe, R., Chamberlain, J., Constantin, M. G., Claire-Hélène, D., Doctor, F., ..., Smeaton, A. F. (2020). Overview of MediaEval 2020 Predicting Media Memorability task: What does it Make a Video Memorable?. In Working Notes Proceedings of the MediaEval 2020 Workshop (Vol. 2882). CEUR Workshop Proceedings.
- Savran Kiziltepe, R., Constantin, M. G., Demarty, C. H., Healy, G., Fosco, C., Garcia Seco De Herrera, A., ..., Sweeney, L. (2021, January). Overview of The MediaEval 2021 Predicting Media Memorability Task. In CEUR Workshop Proceedings (Vol. 3181).
- Sweeney, L., Constantin, M. G., Demarty, C. H., Fosco, C., de Herrera, A. G. S., Halder, S., ..., Sultana, M. (2022). Overview of the MediaEval 2022 predicting video memorability task. arXiv preprint arXiv:2212.06516.
- Kiziltepe, R. S., Sweeney, L., Constantin, M. G., Doctor, F., de Herrera, A. G. S., Demarty, C. H., ..., Smeaton, A. F. (2021). An annotated video dataset for computing video memorability. Data in Brief, 39, 107671.
- de Herrera, A. G. D., Constantin, M. G., Demarty, C. H., Fosco, C., Halder, S., Healy, G., ..., Sweeney, L. (2022). Experiences from the MediaEval Predicting Media Memorability Task.



arXiv preprint arXiv:2212.03955.

4.3.5 Relevant software, datasets and other resources

- Latest version of the benchmarking competition and dataset: <https://multimediaeval.github.io/editions/2023/tasks/memorability/>

4.3.6 Relevance to AI4Media use cases and media industry applications

While not as subjective as other concepts related to media data and their effect on human viewers, memorability is one of the defining concepts when dealing with the creation and dissemination of information and media. This dataset, as well as the results of the benchmarking tasks associated with it greatly contributed to the popularization of this concept throughout the computer vision domain, garnering considerable attention and a significant number of contributing papers. This research can represent the dataset and knowledge base for a memorability prediction system that would best be integrated into Task 6.6 - “Measuring and Predicting User Perception of Social Media”, and best be applied to UC3 “AI for Vision”, feature 3C2-13 - “Modality-dependent sentiment analysis”.

4.4 Ensemble learning

Contributing partners: UPB

The ImageCLEFfusion benchmarking competition³⁷ tasks participants with creating ensemble learning methods and schemes that would allow improving the overall performance when compared with single-system approaches. This task was hosted at ImageCLEF and ran for two editions, in 2022 [18] and 2023 [19]. During the two editions we tasked participants with using data from diverse domains, including data for image interestingness prediction [1], diverse social image retrieval [20], and medical image captioning [21].

4.4.1 Dataset

Given the nature of this task, we want to give participants equal opportunities when it comes to assessing the results of their ensemble methods. Therefore, we do not give access to the original images and videos that make up the datasets that represent the three tasks. Instead, we choose to give the outputs and predictions of systems that attempt to solve the three tasks. These sets of predictions (also called inducers) can then be used directly by participants as inputs for their ensembling methods. Furthermore, participants are not allowed to create additional inducers – they must only use the ones we provide.

We provide 29 inducers for the interestingness task, 56 for the diverse retrieval task, and 84 inducers for the medical captioning task. It is important to note that these three tasks represent different types of machine learning approaches. The most simple is the interestingness data, which represents a simple one-class regression approach, where each image in the dataset is annotated or must be predicted by using a simple $[0, 1]$ value that represents image interestingness for the given sample. The diversity task represents an information retrieval task, where the two main measures of prediction success are represented by the relevance and the diversity of the provided list of outputs. Finally, the medical caption task represents a multi-class labeling approach, where each image in the dataset can be annotated or predicted for one or more medical labels.

³⁷<https://www.imageclef.org/2023/fusion>





For the interestingness task, we used a 1,877 image prediction outputs in the training set and 558 in the testing set. The diversity task used query responses as data samples, and we provided the outputs of 60 queries in the training data and 63 in the testing data. Finally, for the medical caption task we provided label predictions for 6,101 medical images in the training set and 1,500 in the testing set. We used the same metrics for these three tasks as the ones used in their original form, namely: mean average precision at 10 (mAP@10) for interestingness, F1 at 20 (F1@20) and Cluster Recall at 20 (CR@20) for the diversity task, and F1 for medical captioning.

4.4.2 Benchmarking task

The first edition of ImageCLEF fusion has featured only the interestingness and the diversity tasks, while in the second edition we added the medical captioning task. This diversity in data and studied concepts was needed as it ensures a high diversity with regards to the machine learning tasks that are being targeted (single-class regression, information retrieval, and multi-class labeling). In total, 62 runs were submitted by participants, with 27 belonging to the interestingness task and 35 to the diversity task, while unfortunately no competitors were attracted to the medical caption task.

While the metrics for individual task performances are presented in the previous section, we believe the most important metric of success can be computed by measuring the differences between the results of the ensemble methods submitted by our participants and the performance of the top inducers in the ensemble. In this case, we noticed the following top performances: for interestingness prediction, the top participant recorded an impressive 131.7% increase over the baseline inducer performance, while for the diversity task the increase was 18.7%.

4.4.3 Discussion and analysis

So far, during the two editions of the task, we can derive a set of interesting conclusions, observations, and general trends:

- Ensemble systems help in increasing the performance of single-system approaches, by exploiting the knowledge from each individual inducer they use as input;
- So far it would seem that the best approach is represented by using deep neural networks as the main ensemble engine;
- We are pleased to see that a diverse set of approaches have been submitted by participants, ranging from statistical approaches to machine and deep learning approaches, even using in some cases ensembles of ensembles;
- The optimization of these types of systems is still an open question – while deep learning ensemble engines seem to have the best performance so far, they must use the entire set of inducers (or a very large portion of that set) in order to reach those performances. Can an optimization scheme be applied to the input space in order to lower the number of inducers used by the ensemble methods and therefore reduce the high processing and energy requirements for each samples?

4.4.4 Relevant publications

- Ştefan, L. D., Constantin, M. G., Dogariu, M., Ionescu, B. (2022). Overview of imageclef-fusion 2022 task-ensembling methods for media interestingness prediction and result diversification. In CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Bologna, Italy.
- Ştefan, L. D., Constantin, M. G., Dogariu, M., Ionescu, B. (2023, September). Overview of imagecleffusion 2023 task-testing ensembling methods in diverse scenarios. In Experimental



IR Meets Multilinguality, Multimodality, and Interaction. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (pp. 18-21).

4.4.5 Relevant software, datasets and other resources

- Latest version of the benchmarking competition and dataset: <https://www.imageclef.org/2023/fusion>

4.4.6 Relevance to AI4Media use cases and media industry applications

Ensemble systems represent a trade-off between hardware requirements and overall performance. While using more than one system in processing multimedia data can sometimes significantly increase the final accuracy of the predictors, it comes at the expense of having to run each inducer at training and inference time. However, these types of approaches have shown their worth in numerous setups, including but not limited to: (i) multimodal setups, where each inducer system must analyze one modality; (ii) complex tasks which are particularly hard to predict accurately, therefore needing more than one processing branch in order to increase accuracy; (iii) critical systems, where every increase in performance is more important than an increase in processing needs. Therefore, this benchmarking task and the methods and observations it generated would be useful in a diverse range of industrial applications.

4.5 GANs usage limitations

Contributing partners: UPB, HES-SO

The ImageCLEFmed GANs competition³⁸ tasks participants with creating automated systems that test the hypothesis that generative networks produce artificial images that still contains the “fingerprints” of the real set of images used in the training phase, and would therefore be limited in their use by privacy and ethical regulations. This task ran for one edition in 2023 [22], and will be continued at the 2024 edition of ImageCLEF, and this hypothesis is tested on medical images, given the particular emphasis on data privacy that this type of data imposes.

4.5.1 Dataset

The medical data used for this task consisted of real and artificially generated CT scans of lung tuberculosis patients. The training set is composed of 500 GAN-generated images and 160 real-world images, with 80 of the 160 images being used in the creation of the GAN-generated samples. For the testing set, 10,000 artificial images were used, along with 200 real-world images, without disclosing the proportion of non-used and used real images at GAN training time. The main official metric for this dataset is the F1 metric, with Precision, Recall and Accuracy also used as additional metrics.

4.5.2 Benchmarking task

Given that the task has only ran for one edition so far, an impressive number of systems were submitted by the participating teams. In total, 40 runs were submitted by participants, with a top performance of $F1 = 0.802$.

³⁸<https://www.imageclef.org/2023/medical/gans>





4.5.3 Discussion and analysis

While we are still early in this task’s history, several important observations can be drawn at this point:

- Given the high results achieved by participants, we can infer that the hypothesis, in this particular medical setup, is confirmed, meaning that artificially generated images present a set of “fingerprints” of the original real-world data that they were trained on;
- A high diversity of methods was proposed by participants to this task, with many types of approaches having a good performance on the proposed data.

4.5.4 Relevant publications

- Andrei, A., Radzhabov, A., Coman, I., Kovalev, V., Ionescu, B., Müller, H. (2023, September). Overview of ImageCLEFmedical GANs 2023 task-identifying training data “Fingerprints” in synthetic biomedical images generated by GANs for medical image security. In CLEF2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS. org, Thessaloniki, Greece (pp. 18-21).

4.5.5 Relevant software, datasets and other resources

- Latest version of the benchmarking competition and dataset: <https://www.imageclef.org/2023/medical/gans>.

4.5.6 Relevance to AI4Media use cases and media industry applications

While this task has been geared towards medical data, we consider the work and ideas explored during this benchmarking task are of high interest to the multimedia community. On one hand, it represents an interesting exploration of the capabilities of GANs in general, with possible applications in content generation use cases like 3C2-9 (Synthetic Video Generation from Single Semantic Label Map), with ties to tasks like T5.2 (Media content production), while on the other hand these approaches can represent important work for privacy and ethical regulations in applying GANs on multimedia data in general.





5 Summary and Conclusions

This deliverable provides an overview of the work done for Task 4.6 “Benchmarking of AI Systems”, presenting an update on the final version of the AI4Media benchmarking platform, as well as an overview of the benchmarking initiatives that were supported by AI4Media in Task 4.6 along their various editions.

We show the progress achieved on the AI4Media benchmarking platform, analyzing the main high-level functionalities, from the perspectives of organizers and of participants, show their API-level implementations in the final version of the platform, as well as analyze their current development status. We show the implementation of these functionalities in the platform, and analyze the main use case of this platform, namely the ImageCLEF2024 benchmarking initiative.

We also analyze the benchmarking tasks that were supported by AI4Media through Task 4.6, dealing with various multimedia-centric subjects like media interestingness, video memorability, ensemble learning, and medical GANs. We summarize information regarding the data these tasks propose, the various editions and incarnations of the benchmarking tasks, as well as look at some high-level observations and conclusions.

As presented in this Deliverable, the AI4MediaBench Evaluation-as-a-Service platform covers a wide variety of requirements for hosting benchmarking tasks, and aiding competition organizers in deploying their tasks and sharing them with the participants. We described three main advantages that this platform brings to the current multimedia benchmarking landscape, as follows: (i) providing a comprehensive and easy-to-implement API collection that can aid competition organizers in deploying computational complexity-related metrics, while also providing an implemented time complexity metric that organizers can either use as-is or use as an implementation example for their own metrics; (ii) using both API integration and containerization-based integration for submitting participant methods, thus offering several options for competition organizers to check and implement reproducibility for the proposed AI models; (iii) offering an important EU-based benchmarking platform, that can be focused towards common AI goals for the European Community. Furthermore, the collaboration with ImageCLEF helps us understand novel requirements from competition organizers and lets us know which of the functionalities of the platform will need constant updating, covering special particularities that are task or conference-related. This, while the platform is in its stable and final version, updates and maintenance work on the platform will continue, as new and interesting features requests from competition organizers appear, open source packages are updated by their creators or maintainers, or new security measures must be taken.

We wish to continue our collaboration with ImageCLEF in the following years, thus ensuring a constant presence in the multimedia environment and a great method of exposure to the public for our platform. Furthermore, we plan to present this platform for next year’s MediaEval Benchmarking Initiative task organizers. This may create additional exposure for our platform, targeting its implementation in another important multimedia benchmarking initiative.





References

- [1] M. G. Constantin, L.-D. Ștefan, B. Ionescu, N. Q. Duong, C.-H. Demarty, and M. Sjöberg, “Visual interestingness prediction: A benchmark framework and literature review,” *International Journal of Computer Vision*, vol. 129, pp. 1526–1550, 2021.
- [2] C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, M. Gygli, and N. Q. Duong, “Mediaeval 2017 predicting media interestingness task,” in *MediaEval workshop*, 2017.
- [3] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, pp. 91–110, 2004.
- [4] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, Ieee, vol. 1, 2005, pp. 886–893.
- [5] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
- [6] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International journal of computer vision*, vol. 42, pp. 145–175, 2001.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012.
- [8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.
- [9] A. Garcia Seco De Herrera, R. Savran Kiziltepe, J. Chamberlain, M. G. Constantin, D. Claire-Hélène, F. Doctor, B. Ionescu, and A. F. Smeaton, “Overview of mediaeval 2020 predicting media memorability task: What does it make a video memorable?” In *Working Notes Proceedings of the MediaEval 2020 Workshop*, CEUR Workshop Proceedings, vol. 2882, 2020.
- [10] R. Cohendet, C.-H. Demarty, N. Q. Duong, and M. Engilberge, “Videomem: Constructing, analyzing, predicting short-term and long-term video memorability,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2531–2540.
- [11] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, and A. Oliva, “Multimodal memorability: Modeling effects of semantics and decay on video memorability,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, Springer, 2020, pp. 223–240.
- [12] G. Awad, A. A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, *et al.*, “Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval,” *arXiv preprint arXiv:2009.09984*, 2020.
- [13] L. Sweeney, A. Matran-Fernandez, S. Halder, A. G. S. de Herrera, A. Smeaton, and G. Healy, “Overview of the eeg pilot subtask at mediaeval 2021: Predicting media memorability,” *arXiv preprint arXiv:2201.00620*, 2021.
- [14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.



- [15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [17] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.
- [18] L.-D. Ștefan, M. G. Constantin, M. Dogariu, and B. Ionescu, “Overview of imagecleffusion 2022 task-ensembling methods for media interestingness prediction and result diversification,” 2022.
- [19] L.-D. Ștefan, M. G. Constantin, M. Dogariu, and B. Ionescu, “Overview of imagecleffusion 2023 task-testing ensembling methods in diverse scenarios,” 2023.
- [20] B. Ionescu, M. Rohm, B. Boteanu, A. L. Gînscă, M. Lupu, and H. Müller, “Benchmarking image retrieval diversification techniques for social media,” *IEEE Transactions on Multimedia*, vol. 23, pp. 677–691, 2020.
- [21] J. Rückert, A. Ben Abacha, A. Garcia Seco De Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, and C. M. Friedrich, “Overview of imageclefmedical 2022–caption prediction and concept detection,” in *CEUR Workshop Proceedings*, CEUR Workshop Proceedings, vol. 3180, 2022, pp. 1294–1307.
- [22] A. Andrei, A. Radzhabov, I. Coman, V. Kovalev, B. Ionescu, and H. Müller, “Overview of imageclefmedical gans 2023 task-identifying training data “fingerprints” in synthetic biomedical images generated by gans for medical image security,” in *CLEF2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS. org, Thessaloniki, Greece*, 2023, pp. 18–21.

