Guide for Applicants

dAIEDGE 2nd Open Call

Collaborative Projects

Submission starts on **10th January 2025 at 9:00** (Brussels Time),

with deadline on **13th March 2025 at 15:00** (Brussels Time).

Funded by
the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

dAIEDGE has received funding from the European Union's Horizon Europe research and innovation programme, under Grant Agreement No. 101120726.

Acronyms and definitions

Acronym	Meaning
AI	Artificial Intelligence
EC	European Commission
EU	European Union
FAQ	Frequently Asked Questions
GfA	Guide for Applicants
IPR	Intellectual Property Rights
NoE	Network of Excellence
OC	Open Call

History of changes

No	Date	Changes
1	03.03.2025	Section 3.1. Footnote #3 - Specify Switzerland is not eligible in the footnote on eligible countries.
2	03.03.2025	Section 4.3. - Change from “each application will be evaluated by 3 external independent experts” to “each application will be evaluated by 2 external independent experts”
3	03.03.2025	Section 4.4 - Change from 3 experts joining the Consensus Meeting to 2 experts joining the Consensus Meeting.
4	03.03.2025	Section 4.2 - Change in the Scope of in/out of scope screening to include: Alignment with challenge, European dimension, Project focus and TRL level.

CONTENTS

1. Basic Info about dAIEDGE 2
2. What do we offer? 4
3. Eligibility Criteria 5
    3.1. Who are we looking for? 5
    3.2. What types of activities can be funded? 6
    3.3. Ground rules 7
4. How will we evaluate your proposal? 10
    4.1. First Check 10
    4.2. In/Out Scope screening 11
    4.3. Independent Individual Evaluation 11
    4.4. Consensus Meeting 13
    4.5. Formal Check, Sub-Grant Agreement Preparation and Signature 14
5. The Collaborative Projects Support Programme and Payment Arrangements 15
6. Contact us 17
7. Final provisions 18
8. Extra hints before you submit your proposal 18

ANNEX 1. Challenges

ANNEX 2. VLab

1. Basic Info about dAIEDGE

dAIEDGE is a Network of Excellence (NoE) for distributed, trustworthy, efficient and scalable AI at the Edge.

dAIEDGE seeks to strengthen and support the development of a dynamic European cutting-edge Artificial Intelligence (AI) ecosystem under the umbrella of the European Lighthouse for AI, and to sustain the development of advanced AI. dAIEDGE fosters the exchange of ideas, concepts, and trends on cutting-edge next generation AI, creating links between ecosystem actors to help both the European Commission (EC) and the European Union (EU) identify strategies for future developments in Europe.

dAIEDGE main objective is to advance Europe's innovation and technology base by developing a comprehensive policy and governance approach to AI.

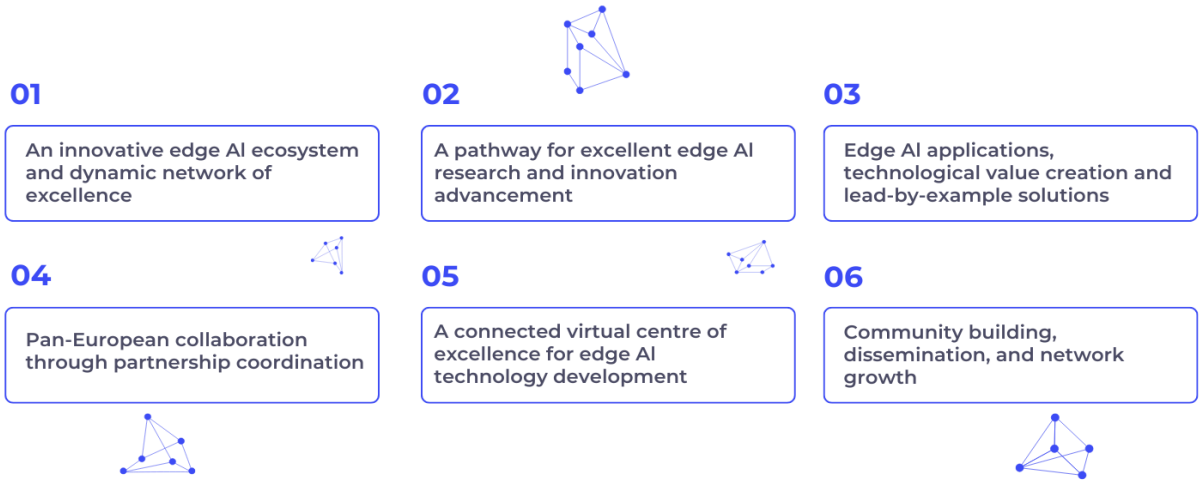


Figure 1 – dAIEDGE objectives

dAIEDGE project partners are a multidisciplinary mix of European-level experts with complementary expertise aligned with the project objectives. Find on dAIEDGE website and at Frequently Asked Questions document more information about the 35 partners from 15 European Countries coordinated by Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI).

This document summarises the main points of the **dAIEDGE 2nd Open Call for Collaborative Projects**, which will be open **from the 10th of January 2025 at 9:00 (Brussels Time) to the 13th of March 2025 at 15:00 (Brussels Time).**

You can submit your application to this open call and find more information at:
https://daiedge-2oc.fundingbox.com/

If you have any technical problems or doubts, tell us at: dAIEDGE.help@fundingbox.com

2. What do we offer?

The 2nd dAIEDGE Collaborative Projects Open Call aims to distribute up to 600.000 € among up to 10 Proposals to solve the Industrial Challenges defined by the dAIEDGE Consortium, with the possibility of using resources from the dAIEDGE Virtual Lab or adding new resources to it or to re-use developments of the dAIEDGE project or dAIEDGE use-cases (see detailed definition of the Challenges in the Annex 1 to this GfA and dAIEDGE VLab details in the Annex 2).

Participants can submit a proposal for research and development activities to solve or push forward the research in the identified Challenges.

Each third-party project selected under dAIEDGE Collaborative Projects Open Call can receive a **maximum amount of up to €60 000.** The Support Programme will last up to **7 months** and include the following stages:

- Stage 1: **Plan Phase**, during which the beneficiary should deliver an Individual Mentoring Plan (including planned activities, KPIs, milestones and budget). The duration of this stage is **up to 1 month** and, after completing it, the beneficiary will receive a **maximum amount of up to €9 000**.

- Stage 2: **Implementation Phase**, which should lead to a Proof of Demonstration. The duration of this stage is **up to 6 months** and, after completing it, the beneficiary will receive a **maximum amount of up to €51 000.**

During the two stages of the dAIEDGE Collaborative Projects Support Programme the beneficiaries will receive **technical mentoring from dAIEDGE technical experts**. They will be supported with the project's development and demonstration of the added value of their solutions to address the dAIEDGE Industrial Challenges.

3. Eligibility Criteria

To participate in the dAIEDGE Support Programme, the applicant has to meet all the criteria described in Section 3 of this Guide, positively pass the evaluation process and finally sign the Sub-Grant Agreement with the dAIEDGE Consortium.

The proposals that do not comply with the criteria described in this section will be excluded. The eligibility criteria will be checked during the whole evaluation process.

3.1. Who are we looking for?

dAIEDGE Collaborative Projects Open Call looks for **legal entities**, applying **individually** or as a **consortium of up to 2** entities.

The eligible entities are:

i. Research and Technology Organisations (RTO),

ii. Academia or

iii. SMEs[1], including Startups, legally registered as a company at the moment of the application submission to this Open Call,

registered in one of the following eligible countries:

- EU Member States[2] and its Overseas Countries and Territories (OCT), or
- Horizon Europe Associated Countries[3].

Applicants subject to EU restrictive measures under Article 29 of the Treaty on the European Union (TEU) and Article 215 of the Treaty on the Functioning of the EU (TFEU)[4] are not eligible to participate in this Open Call.

[1] An **SME** will be considered as such if it complies with the European Commission's Recommendation 2003/361/EC. As a summary, the criteria defining an SME are:
• Headcount in Annual Work Unit (AWU) less than 250;
• Annual turnover less or equal to €50 million OR annual balance sheet total less or equal to €43 million.
Note that the figures of partners and linked enterprises should also be considered as stated in the SME user guide. For detailed information check EU recommendation:
https://ec.europa.eu/growth/smes/business-friendly-environment/sme-definition_en

[2] Following the Council Implementing Decision (EU) 2022/2506, as of 16th December 2022, no legal commitments can be signed with Hungarian public interest trusts established under Hungarian Act IX of 2021 or any entity they maintain. Affected entities may continue to apply to calls for proposals. However, in case the Council measures are not lifted, such entities are not eligible to participate in the dAIEDGE Open Call. In case of consortium, co-applicants will be invited to remove or replace that entity. Tasks and budget may be redistributed accordingly

[3] AC as of 12.12.2024: Albania, Armenia, Bosnia and Herzegovina, Canada, Faroe Islands, Georgia, Iceland, Israel, Kosovo, Moldova, Montenegro, North Macedonia, Norway, Serbia, Türkiye, Tunisia, United Kingdom, Ukraine, for the most up-to-date list please first part of this document. For the avoidance of doubt, New Zealand and Switzerland are not eligible in this open call.

[4] Please note that the EU Official Journal contains the official list and, in case of conflict, its content prevails over that of the EU Sanctions Map

dAIEDGE Consortium partners, their affiliated entities, board members, employees or permanent collaborators CANNOT be beneficiaries of dAIEDGE Open Call.

3.2. What types of activities can be funded?

The activities that qualify for financial support under dAIEDGE Collaborative Projects Open Call are the **research and development activities** to solve or push forward the **research in the Industrial Challenges defined by the dAIEDGE Consortium**, with the possibility of using resources from dAIEDGE Virtual Lab or adding new resources to it, or to re-use developments of the dAIEDGE project or dAIEDGE use-cases.

Each single applicant or consortium must select and reply in its proposal to **one** of the following Industrial Challenges (details in Annex 1):

1. Mobile concealed weapon detection
2. Continual Learning for edge-based Robotics: 3D Environment Exploration and Reconstruction using gradient-free Variational Techniques.
3. Smart and Adaptive AI Agents at the Edge: real-time Active Inference for IoT and Robotics
4. Integration of an AI-powered Edge Device to the Vlab for remote benchmarking
5. Cloud-to-Edge AI-Powered Solution for Plant Disease and Pest Detection with BioClip Integration
6. Energy-Efficient Deployment of AI Models on Edge Devices Using FPGAs
7. SAFE AI - Enhancing AI Model Predictions Through Synthetic Data and Domain Knowledge Integration
8. Very low consumption of speech enhancement AI algorithm at the edge
9. Off-Chip Weights for Streaming Architectures
10. 2D Vehicle Detection Network on PYNQ-Z1
11. Real-Time Vision Transformer (ViT) on FPGA SoC for Image Classification
12. Spiking Neural Networks (SNNs) on the edge
13. Energy efficient Text Embeddings inference at the edge
14. Automated Edge deployment, tuning and Performance Evaluation of Binary Neural Networks targeting FPGA enabled platform.
15. Deploying LLMs on Low-Memory Edge Devices
16. Developing an Early Warning System for Harmful Algal Blooms Using Earth Observation Data
17. Edge AI framework for web browsers empowering Federated Machine Learning applications addressing Classification problems over multiple format data (text and images)
18. Machine Learning Benchmarks for On-Board Processing in Space Applications
19. Incentivisation framework on blockchain for optimizing AI development
20. Planning and Scheduling for Space Observations at the Extreme Edge

The activities within the dAIEDGE Project should start at a **Technological Readiness Level** (TRL) of **2-3 and to achieve a TRL 4-5**. You can also check the FAQ document for a detailed explanation about TRL.

Your project should have a clear **European Dimension**, meaning fostering projects that generate a substantial positive impact on the European society and economy.

3.3. Ground rules

When applying to the dAIEDGE Open Call, please also note that:

- **Be on time and use our system:** only proposals submitted through the online form before the deadline of 13th March 2025 at 15:00 (Brussels Time) will be evaluated and considered for funding. If you submit the form correctly, the system will send you a confirmation of your submission. Get in touch with us if it is not the case.

- **English Language:** proposals must be written in English in all mandatory parts to be eligible. Only parts written in English will be evaluated. If the mandatory parts of the proposal are in any other language, the entire proposal will be rejected. If only non-mandatory parts of a proposal are submitted in a language different from English, those parts will not be evaluated but the proposal is still eligible.

- **Every question deserves your attention:** all mandatory sections - generally marked with an asterisk - of the proposal must be completed. The data provided should be actual, true, and complete and should allow assessment of the proposal. Additional material, not specifically requested in the online application form, will not be considered for the evaluation.

- **Be exhaustive:** applicants have to verify the completeness of the form, as it won't be possible to add any further information after the deadline. After the proposal is submitted, the form can be modified until the deadline.

- **Conflicts of interest:** the existence of a potential conflict of interest among applicants and one or more Consortium partners will be taken into consideration. Consortium partners, their affiliated entities, employees, or persons treated as personnel (ex. working under B2B contract) cannot take part in the dAIEDGE programme. All cases of potential conflict of interest will be assessed case by case. See EC definition of Conflict of Interest[5] and check our FAQ for more info.

[5] EC definition of Conflict of Interest: https://commission.europa.eu/strategy-and-policy/eu-budget/protection-eu-budget/conflict-interest_en

- **Healthy finances and a clean sheet are mandatory:** we don't accept entities that are under liquidation or are an enterprise under difficulty[6] according to the Commission Regulation No 651/2014, art. 2.18, or that are excluded from the possibility of obtaining EU funding under the provisions of both national and EU law, or by a decision of both national or EU authority. We also don't accept entities that are meeting national regulations regarding bankruptcy.

- **It is your proposal:** Proposals should be based on original work or the right to use the IPR must be clear. Any work related to the implementation of the project described in the application may not violate the IPR of third parties, and the IPR of the application project may not be the subject of a dispute or proceedings for infringement of third-party IPR.
  dAIEDGE Consortium encourages the use of Open Source Software and Hardware Solutions.
  IPR and confidentiality issues will be assessed case by case and provisions included in each Sub-Grant Agreement if necessary.

- Applicants must ensure that their proposals have an exclusive focus on **civil applications**. Military use is not allowed and such projects will not be funded by the dAIEDGE.

- **Each single entity or consortium can submit up to 3 applications (maximum one per Industrial Challenge):** If more than one proposal per Challenge is identified, only the last proposal which has been submitted in order of time will be evaluated.
  BUT **neither team members nor any legal entities can be funded twice** by dAIEDGE. If an applicant submits applications for different Challenges and more than one proposal (with any similar team members or from the same organisation) will be among the selected projects, **only the ONE with a higher position on the ranking list created per each of the Industrial Challenges can be selected for funding.**

---

[6] An enterprise will be considered an undertaking in difficulty if more than half of the capital has disappeared. This refers to the loss of "subscribed share capital". If profit and loss reserves deficit more than 50% of share capital, there is a potential problem with the company. (Article 2, item 18 point a) and b)

(a) In the case of a limited liability company (other than an SME that has been in existence for less than three years […]), where more than half of its subscribed share capital has disappeared as a result of accumulated losses. This is the case when deduction of accumulated losses from reserves (and all other elements generally considered as part of the own funds of the company) leads to a negative cumulative amount that exceeds half of the subscribed share capital. For the purposes of this provision, 'limited liability company' refers in particular to the types of company mentioned in Annex I of Directive 2013/34/EU (1) and 'share capital' includes, where relevant, any share premium
.
(b) In the case of a company where at least some members have unlimited liability for the debt of the company (other than an SME that has been in existence for less than three years […]), where more than half of its capital as shown in the company accounts has disappeared as a result of accumulated losses. For the purposes of this provision, 'a company where at least some members have unlimited liability for the debt of the company' refers in particular to the types of company mentioned in Annex II of Directive 2013/34/EU.

Please note that, if your SME exist from less than three years, you won`t be considered as undertaking any difficulties.

Beneficiaries from the dAIEDGE 1st Open Call can not be beneficiaries of the 2nd OC (neither team members nor any legal entities can be funded twice by dAIEDGE).

- **Gender Equality Plan:** public bodies, higher education institutions, and research organisations from EU countries and associated countries must have a Gender Equality Plan (GEP)[7].

- **Acceptance of the Open Call rules:** to apply for this Open Call, applicants have to accept its rules and regulations detailed in this Guide for Applicants.

dAIEDGE is planning some dissemination/information activities about this Open Call. They will be announced at the dAIEDGE OC website and dAIEDGE Social Media channels.

If extra hints on how to prepare the application are needed, applicants can check out the section: extra Hints to submit your proposal.

[7] For more details please check https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/democracy-and-rights/gender-equality-research-and-innovation_en

4. How will we evaluate your proposal?

dAIEDGE evaluation process is transparent, fair and equal to all participants, with a clearly defined complaint procedure (see section Complaints). dAIEDGE will evaluate the proposals in 4 phases, as shown in the following graphic:

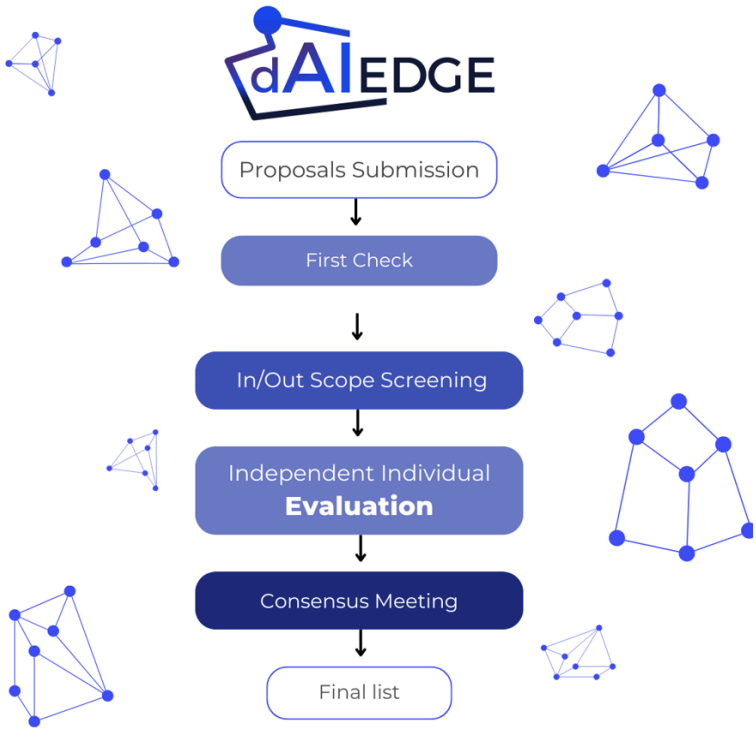


Figure 2 – dAIEDGE evaluation process

For this call, dAIEDGE is looking for the best fit for the dAIEDGE project, and a high volume of applications is anticipated. Since the primary concern is quality over quantity, applicants are encouraged to put effort into presenting their applications in the best possible way, offering as much detail as possible. This will assist us in evaluating the applications and identifying how the proposals align with the overall scope of dAIEDGE.

4.1. First Check

After the closure of the Open Call, the proposals will be reviewed to ensure they meet the conditions outlined in Section 3. This assessment will be based on the statements provided in the proposal.

At this stage, the eligibility criteria are checked against the Declaration of Honour or self-declarations included in the application form, and they will be continuously verified throughout the evaluation process, including the final formal check.

Projects that do not comply with the above-mentioned criteria will be rejected.

4.2. In/Out Scope screening

In case of a high number of applications or special needs of the project, the Selection Committee may decide to implement an additional evaluation step - In/Out Scope Screening.

To maximise the impact within the framework of dAIEDGE, proposals must be aligned with the scope of activities outlined in Section 3.2 and provide a clear demonstration of this alignment.

For this reason, one partner from the Selection Committee will review the following aspects of the proposals:

- Scope. The objectives of the proposal must fit within the scope of the project, in particular, described in Section 3.2 of this GfA:
  - ➢ Alignment with the challenge selected. Projects should solve or push forward the research in the Industrial Challenges defined by the dAIEDGE Consortium.
  - ➢ The TRL declared. Projects should start at a Technological Readiness Level (TRL) of 2-3 and aim to achieve a TRL 4-5.
  - ➢ Focus on research and innovation. Projects should have a practical application. Research that is purely theoretical, without a concrete technological application or development, is out-of-scope of this program.
  - ➢ European Dimension. The project should have a European dimension, as described in Section 3.2 of this GfA.

The partners from the Selection Committee will assess if the proposals match the criteria mentioned above following a "Yes/No" approach. Each application will be assessed by one member of the Selection Committee. The partners will justify their assessment in case of rejection of the proposal. The Selection Committee will review the partner's assessment and verify the validity of each rejection, generating an 'In Scope List'.

Please note that proposals that do not comply with the criteria described above will be rejected. Only those meeting all the criteria will proceed to the Independent Individual evaluation phase.

Applicants will be informed about the results of the first check and the in/ out scope screening.

4.3. Independent Individual Evaluation

In this phase, each application will be evaluated by 2 external independent experts.

The applications will be evaluated within the following awarding criteria:

(1). EXCELLENCE will evaluate:

Ambition: Applicants have to demonstrate how their proposal contributes to the project scope, have a European dimension and is beyond the State of the Art. They need to describe the innovative approach behind the project (like ground-breaking objectives, novel concepts and approaches, new products, services or business).

Innovation: Applicants should provide information about the level of innovation and about the degree of differentiation that their project will bring.

Soundness of the approach and credibility of the proposed methodology.

Gender dimension: Applicants have to demonstrate to what extent the gender dimension has been integrated into the research and innovation content.

(2). IMPACT will analyse:

Market opportunity: Applicants have to demonstrate a clear idea of what they want to do and whether the new/improved product has market potential.

Competition: Applicants have to provide information about the degree of competition of their product/service and why their idea is disruptive and breaks the market. They should explain why their products/services can be differentiated from the competition.

Commercial Strategy and Scalability: Applicants must demonstrate the level of scalability of the new/improved product, meaning by not addressing a specific problem but being able to solve a structural problem in a specific sector/process/etc.

Environmental and social impact: Applicants should demonstrate their project contribution towards environmental, social and economic impacts to contribute to sustainable development, Green Deal and other European policies.

(3). IMPLEMENTATION will consider:

Team: Applicants have to demonstrate the team's management and leadership qualities, their ability to take a concept from ideas to market, their capacity to carry through their ideas and understand the dynamics of the market they are trying to tap into. The team should be a cross-functional team, with a strong background and skills base and considering its gender balance.

Resources. Applicants should demonstrate the quality and effectiveness of the resources assigned to get the objectives proposed.

The evaluators will score each criterion on a scale from 0 to 5:

0 - **The proposal fails to address the criterion** or it cannot be assessed due to missing or incomplete information.

1 - **Poor** – The criterion is inadequately addressed or there are serious inherent weaknesses.

2 - **Fair** – The proposal broadly addresses the criterion, but there are significant weaknesses.

3 - **Good** – The proposal addresses the criterion well but a number of shortcomings are present

4 - **Very good** – proposal addresses the criterion very well but a small number of shortcomings are present.

5 - **Excellent** – The proposal successfully addresses all relevant aspects of the criterion. Any shortcomings are minor.

Each evaluator will produce an **Individual Evaluation Report**. Once the Individual Evaluation Reports are submitted, **the final score for each individual criterion** will be calculated as the average of the scores provided by each evaluator. The **final score per each application** will be calculated as the sum of the scores for each individual criterion.

For each section, **the threshold for individual criteria is 3 out of 5** points. The total maximum score is 15 points, with a **minimum total threshold of 10 points**.

Once the initial ranking is obtained, **ties** (if any) will be solved using the following criteria in order of priority:

- The highest score in the Excellence Section.
- Gender balance (prioritizing teams with more female team members) among the personnel responsible for carrying out the activities.
- The highest score in the Impact Section.

As a result of the Independent Individual Evaluation, a **'Ranking List per Challenge'** and a **"General Ranking List"** will be produced. All proposals that reach the threshold or are scored above the threshold, will pass to the next phase.

Please note that we need time to process all the proposals in this phase, so you probably won't hear back for a while.

4.4. Consensus Meeting

Following the General Ranking List and the Ranking List per Challenge, the Selection Committee, with the support of 2 Experts who participate in the Independent Individual Evaluation, will decide, at this stage, the Provisional List of beneficiaries and Reserve List. The experts will take part in an

advisory capacity (i.e. with a voice but without a vote). The decision will be based on the 'Ranking List per Challenge' and "General Ranking List" obtained as a result of the previous step.

Whilst normally the highest ranked proposals will be selected for funding, the Selection Committee might have fair reasons for objecting to a specific proposal, like the alignment with dAIEDGE goals and scope, the ability to achieve the highest impact possible, as well as the existence of significant ethical concerns or a potential conflict of interest. In this case, the choice may pass to the next-ranked proposal.

dAIEDGE aims to select 10 Proposals for 10 different Industrial Challenges, but the exact number of proposals approved will be decided based on the overall **quality** of the proposals. All decisions are taken by consensus or a minimum of 2/3 majority votes

After the Consensus Meeting, the results will be communicated to the applicants.

4.5. Formal Check, Sub-Grant Agreement Preparation and Signature

Before the selected applicants get started with the dAIEDGE Collaborative Projects Programme, they need to sign the Sub-Grant Agreement with the dAIEDGE Consortium.

Before signing the Agreement, applicants should provide documents regarding their formal status. The dAIEDGE Consortium will verify them to prove the eligibility.

Applicants should be extremely vigilant to:

1. **The nature of the documents requested.** If the documents provided do not prove the eligibility, applicant participation will end there.

2. **The deadlines for submission of these documents.** If an applicant does not deliver the requested documents on time, without a clear and reasonable justification, we will have to exclude that proposal from further formal assessment. Another applicant from the Reserve list will then take the place.

5. The Collaborative Projects Support Programme and Payment Arrangements

Once the applicant's eligibility has been confirmed following the formal check and the Sub-Grant Agreement signed, the applicant will become an official beneficiary of the dAIEDGE Programme.

Beneficiaries will enrol in an up to 7 months Technical Support Program, receiving technical mentoring from dAIEDGE experts for the development of their projects. Each third-party project can receive a maximum lump sum of up to €60 000.

The lump sum is a simplified method of settling expenses in projects financed from Horizon Europe Programme funds. It means that the grantee is not required to present strictly defined accounting documents to prove the cost incurred (e.g. invoices) but is obliged to demonstrate that the implementation of the project is in line with the milestones set for it.

Simply speaking, it means that the Selection Committee will carefully assess beneficiaries' progress and the quality of their work during Interim Reviews after each stage of the Programme. The milestones (deliverables, KPIs, and ethical recommendations, as well as the budget) will be fixed in the **'Individual Mentoring Plan'** elaborated at the beginning of the programme.

The lump sum method does not exempt beneficiaries from collecting documentation to confirm the costs under fiscal regulations.

The grant will be paid after the Milestone's Review of each stage, following this scheme:

Stage	Name	Duration	Deliverable	Funding
Stage 1	Plan Phase	1 month	Individual Mentoring Plan	15% of the grant, up to €9 000
Stage 2	Implementation Phase	6 months	Proof of Demonstration	85% of the grant, up to €51 000
			TOTAL	up to €60 000

A delayed payment mechanism could be applied to the payments if decided by the dAIEDGE Consortium. This mechanism implies that up to 15% of each tranche will be paid to the beneficiaries who completed the given stage once the whole dAIEDGE Project is completed. This should happen approximately 9 months after the end of the Project. The expected end of the dAIEDGE Projects is September 2026. Relevant provisions will be included in the Sub-grant Agreements. Please bear in mind that dAIEDGE project might be extended.

Milestones Review

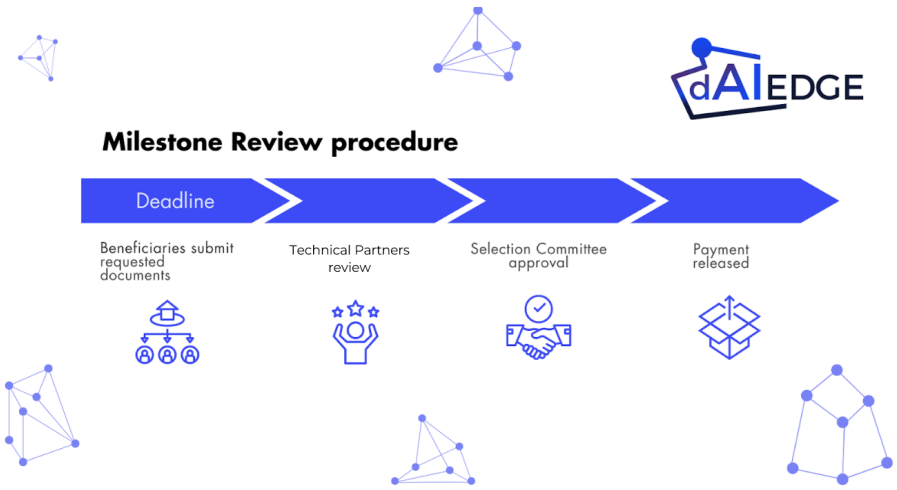


Figure 3 – dAIEDGE milestone review and payment process

Beneficiaries performance during the Exchange Programme will be reviewed at the Milestone Review (established every time a payment is due), according to the following criteria:

- Deliverable quality (30%).
- Technical performance indicators (60%).
- Deadline Compliance (10%).

Each criterion will be scored from 0 to 10 and, based on the weight of each criterion, the final score will be calculated.

According to this final score, beneficiaries over the threshold (7 points) will successfully receive the corresponding part of the grant and continue the programme. The beneficiaries who have not reached the threshold (7 points) will be invited to leave the programme without receiving the corresponding payments.

The Selection Committee will review and validate the evaluations, paying special attention to the 'under threshold' cases, if any, by taking into consideration all possible objective reasons for underperformance (i.e. external factors which might have influenced the beneficiaries' performance). The dAIEDGE Selection Committee will make the final decision and approve the payments or invite beneficiary projects which have not reached the threshold to leave the programme.

6. Contact us

How can we help you?

If you have any questions regarding our application process, feel free to email us at dAIEDGE.help@fundingbox.com

Responses to any questions are provided on an individual basis, do not constitute a change to this Guide for Applicants, and are provided for informational purposes.

In case of any technical issues or problems, please include the following information in your message:

- your username, phone number and email address;
- details of the specific problem (e.g. error messages you encountered, bug description, i.e. if a dropdown list isn't working, etc.); and
- screenshots of the problem.

Complaints

If you believe that a mistake has been made after receiving the results of one of the evaluation phases (when foreseen), you may submit a complaint. To do so please email us your complaint in English at dAIEDGE.help@fundingbox.com and include the following information:

- your contact details (including email address);
- the subject of the complaint;
- information and evidence regarding the alleged breach.

You have 3 calendar days to submit your complaint, starting from the day after the communication was sent. We will review your complaint within seven calendar days from its reception. If we need more time to assess your complaint, we will inform you about the extension by email. We will not review anonymous complaints as well as complaints with incomplete information.

Please take into account that the evaluation is run by experts in the relevant field, and we do not interfere with their assessment. Therefore, we will not evaluate complaints related to the results of the evaluation other than those related to procedural or technical mistakes.

7. Final provisions

Any matters not covered by this Guide will be governed by Polish law and rules related to the Horizon Europe Programme and general rules of EU grants.

Please take into account that we make our best effort to keep all provided data confidential; however, for the avoidance of doubt, you are solely responsible for indicating your confidential or sensitive information as such.

Your Intellectual Property Rights (IPR) will remain your property.

For the selected grantees, the Sub-grant agreement will include a set of obligations towards the European Commission (for example: promoting the project and giving visibility to the EU funding, maintaining confidentiality, understanding potential controls by the EC/ECA, EPPO and OLAF).

The dAIEDGE Consortium might cancel the call at any time, change its provisions or extend it. In such a case we will inform all applicants about such change. The signature of the Sub-grant agreement is an initial condition to establish any obligations among applicants and any Consortium partners (with respect to the obligation of confidentiality of the application).

8. Extra hints before you submit your proposal

A proposal takes time and effort and we know it. Here are a few crucial points you should read before submitting your proposal.

- Is your proposal in line with what dAIEDGE is looking for? You are not sure? You can consult Section 3.1 and Section 3.2.
- Did you present your project in a way that will convince evaluators? Not sure if you did? Go back to Section 4.
- Is your project fulfilling all eligibility requirements described in the Guide? Check again Section 3.
- Are you sure you can cope with our process of the Sub-grant agreement signature and payment arrangements for selected proposals? You may want to go over Section 4.5.
- Did you check our Sub-grant agreement template? You didn't? Check it here.
- Do you need extra help? Contact us.

Good luck!

ANNEX 1:

dAIEDGE Challenges

ANNEX 2:

dAIEDGE VLab

daiedge.eu

dAIEDGE 2nd Open Call. COLLABORATIVE PROJECTS.

Challenges

1. Mobile concealed weapon detection
2. Continual Learning for edge-based Robotics: 3D Environment Exploration and Reconstruction using gradient-free Variational Techniques.
3. Smart and Adaptive AI Agents at the Edge: real-time Active Inference for IoT and Robotics
4. Integration of an AI-powered Edge Device to the Vlab for remote benchmarking
5. Cloud-to-Edge AI-Powered Solution for Plant Disease and Pest Detection with BioClip Integration
6. Energy-Efficient Deployment of AI Models on Edge Devices Using FPGAs
7. SAFE AI - Enhancing AI Model Predictions Through Synthetic Data and Domain Knowledge Integration
8. Very low consumption of speech enhancement AI algorithm at the edge
9. Off-Chip Weights for Streaming Architectures
10. 2D Vehicle Detection Network on PYNQ-Z1
11. Real-Time Vision Transformer (ViT) on FPGA SoC for Image Classification
12. Spiking Neural Networks (SNNs) on the edge
13. Energy efficient Text Embeddings inference at the edge
14. Automated Edge deployment, tuning and Performance Evaluation of Binary Neural Networks targeting FPGA enabled platform.
15. Deploying LLMs on Low-Memory Edge Devices
16. Developing an Early Warning System for Harmful Algal Blooms Using Earth Observation Data
17. Edge AI framework for web browsers empowering Federated Machine Learning applications addressing Classification problems over multiple format data (text and images)
18. Machine Learning Benchmarks for On-Board Processing in Space Applications
19. Incentivisation framework on blockchain for optimizing AI development
20. Planning and Scheduling for Space Observations at the Extreme Edge

CHALLENGE 1. Mobile concealed weapon detection

SHORT DESCRIPTION: Advances in computer vision and thermal imaging technology have opened the possibility of creating a system that detects concealed handguns from thermal images. Handguns attenuate the body heat signal in a way that could be automatically detected with advanced computer vision techniques. This challenge aims at proving that such application is feasible on a mobile device, like in the form of bodyworn cameras carried by police officers. For this, a smartphone app (Android) can be used connected to a miniature thermal camera, although other form factors and devices can be explored. Privacy and GDPR issues should be considered.

EXPECTED DELIVERABLE: The expected deliverables are: 1) Report, 2) Software and 3) Dataset.

REFERENCE INFO. There are two reference public thermal image datasets for concealed weapon detection:

1. Ozan Veranyurt and C Okan Sakar. Concealed pistol detection from thermal images with deep neural networks. Multimedia Tools and Applications, 82(28):44259–44275, November 2023.
2. Muñoz, Juan Daniel; Ruiz-Santaquiteria, Jesus; Deniz, Oscar (2024), "Thermal Image Dataset for Concealed Handgun Detection", Mendeley Data, V1, doi: 10.17632/b6rpgr6nrh.1

Larger datasets with more variability can be collected as thermal cameras have become small and affordable, like the TOPDON TC001.

dAIEDGE VLab infrastructure can be used.

DIAGRAM & TECHNICAL SPEC: The idea is to capture and process, on a mobile device, a thermal image and detect any concealed handguns:

REQUISITES: Mandatory: Knowledge in Computer vision and deep learning.

EXPECTED RESULT: Above 85% accuracy in detection with at least 10 frames per second on a mobile device such as a smartphone. The resolution must be at least 256x192 (TOPDON's TC001 camera resolution). (Note that higher working resolutions are preferred, as in general the detection range will be larger with higher resolutions). The report must include performance metrics in Qualcomm processors.

CHALLENGE 2: Continual Learning for edge-based Robotics: 3D Environment Exploration and Reconstruction using gradient-free Variational Techniques.

SHORT DESCRIPTION: This project aims to advance robotic capabilities in 3D environment exploration and navigation by addressing the challenge of continual learning in dynamic settings. Current state-of-the-art solutions often suffer from catastrophic forgetting when processing continuous data streams. In contrast, gradient-free variational learning techniques on edge devices, such as Variational Bayes Gaussian Splatting (VBGS), enable efficient model updates from partial, sequential observations without relying on replay buffers.

The project shall implement and evaluate various gradient-free Bayesian methods for VBGS, including approaches like the No-U-Turn Sampler (NUTS), Black-Box Variational Inference (BBVI), and Coordinate Ascent Variational Inference (CAVI) using models based on Gaussian Mixture Models (GMMs). Targeted for edge platforms like Nvidia Jetson, this work aims to provide a comparative analysis of probabilistic machine learning techniques optimized for low-latency and resource-constrained robotic applications.

EXPECTED DELIVERABLE:
Implementation:

a. Implementation of the various approaches (NUTS, BBVI, CAVI) for VBGS while tailoring them for resource-constrained devices by incorporating optimizations for memory usage and computational efficiency.
b. Implementation of VBGS in a continual learning setting.
c. Methods preferably implemented in JAX.

Evaluation: Comprehensive evaluation of various benchmark datasets, comparing the models in terms of predictive accuracy, computational efficiency, and uncertainty quantification.
Edge Device Deployment: Successful deployment of the models on Nvidia Jetson devices, demonstrating its feasibility and performance in real-world edge computing scenarios.
Documentation: Report with clear and concise documentation of the implementation, evaluation, and deployment processes.

REFERENCE INFO
Research Papers

1. Van de Maele, Toon, et al. "Variational Bayes Gaussian Splatting." arXiv preprint arXiv:2410.03592 (2024).
2. Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. J. Mach. Learn. Res., 15(1):1593–1623, 2014.
3. Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In Artificial intelligence and statistics, pages 814–822. PMLR, 2014.
4. Heins, C., Wu, H., Markovic, D., Tschantz, A., Beck, J., & Buckley, C. (2024). Gradient-free variational learning with conditional mixture networks. arXiv preprint arXiv:2408.16429.
5. David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. Journal of the American Statistical Association, 112(518):859–877, 2017. doi: 10.1080/01621459. 2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

Sample Implementations

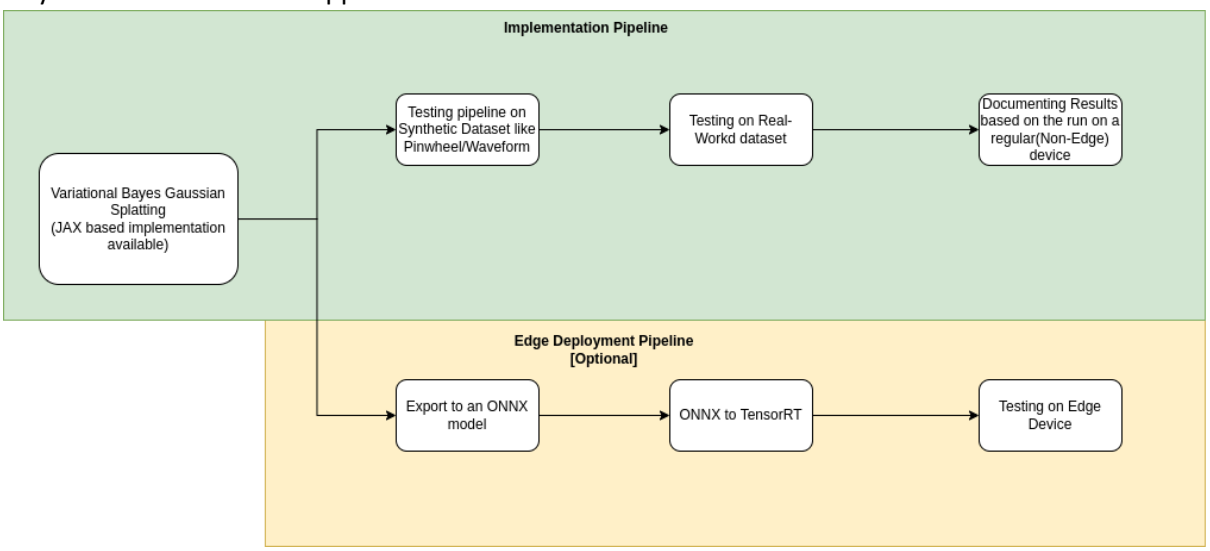
1. VBGS: https://github.com/VersesTech/vbgs
2. NUTS: https://github.com/kasparmartens/NUTS
3. BBVI: https://github.com/jamesvuc/BBVI
4. CAVI-CMN: https://github.com/VersesTech/cavi-cmn

Datasets:

- 3D reconstruction and exploration
  - Tiny ImageNet: https://zenodo.org/doi/10.5281/zenodo.10720916
  - Blender 3D models: https://arxiv.org/abs/2003.08934
  - Habitat scenes: https://arxiv.org/abs/1904.01201
- Variational Inference modelling
  - Synthetic
    a. Pinwheel Dataset: https://www.researchgate.net/publication/311668483_Robust_Local_Scaling_using_Conditional_Quantiles_of_Graph_Similarities https://github.com/VersesTech/cavi-cmn?tab=readme-ov-file#pinwheel-dataset
    b. WaveformDomains: https://archive.ics.uci.edu/dataset/107/waveform+database+generator+version+1
  - Real-World:
    c. UCI: https://archive.ics.uci.edu/datasets

DIAGRAM & TECHNICAL SPEC

Proposed system for the Robotics application:



Input spec

- RGB-D image + Camera Poses (position & orientation)

Output spec

- A Gaussian Splat: i.e. the parameters of a Gaussian Mixture model on color and space (space means, space covariances, color means, color covariance matrices, and mixture weights). This could be the same format as VBGS or the original 3DGS repository.

REQUISITES:

- Mandatory
  - Open-source implementation
  - Deployment on an edge device
- Preferential
  - Methods preferably implemented in JAX
  - Portability to ONNX

EXPECTED RESULT:

a. Implementation of the methods on an edge device
b. Comparative analysis report
c. Optimize and quantify the scalability in feature dimension size, e.g., employ a high-dimensional feature space instead of only RGB color)
d. Optimize and quantify performance scaling w.r.t. the number of updated components, for instance, by selecting a subset of the components to update.

CHALLENGE 3. Smart and Adaptive AI Agents at the Edge: real-time Active Inference for IoT and Robotics

SHORT DESCRIPTION: Edge computing is essential for real-time processing in industrial applications and autonomous systems, where fast, local decision-making is critical. However, traditional AI systems often struggle in dynamic environments, especially when the data encountered differs from the distributions learned during training. This challenge stresses the need for intelligent agents to operate autonomously at the edge, continuously adapting the underlying system to changing conditions. Active Inference models, which allow agents to sense, reason, and act based on real-time environmental feedback, offer a promising First Principles approach to overcoming these limitations.

This project aims to optimize intelligent, active inference agents for robotics, IoT, and Industry 4.0 applications on resource-constrained embedded devices. The project shall explore deployment techniques tailored for such platforms, focusing on low latency, energy efficiency, scalability, and adaptability to dynamic environments.

EXPECTED DELIVERABLE:

Implementation:

a. Evaluation and selection of suitable models for robotics and industrial IoT use cases based on the latest SOTA (1-month)
b. Implementation, fine-tuning, and deployment of the models in pymdp - a python package for simulating Active Inference agents - incorporating hardware-in-the-loop optimizations for memory usage and computational efficiency.
c. Demo showcasing the deployment in a real-world scenario.

Evaluation:

Comprehensive evaluation of different optimization methods in terms of predictive accuracy, computational efficiency, and memory usage on at least one dataset and two embedded devices.

Edge Device Deployment:

Successful deployment of the methods on two embedded devices:

1. Mid/high-end embedded device, e.g., Raspberry Pi or Nvidia Jetson
2. Micro-controller Unit, e.g. STM32 H7

Documentation:

Report with clear and concise documentation of the implementation, evaluation, and deployment processes.

REFERENCE INFO

Research Papers:

1. Sedlak, Boris, et al. "Active inference on the edge: A design study." 2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops). IEEE, 2024.
2. Hamburg, Sarah, et al. "Active Inference for Learning and Development in Embodied Neuromorphic Agents." Entropy 26.7 (2024): 582.
3. Heins, Conor, et al. "pymdp: A Python library for active inference in discrete state spaces." arXiv preprint arXiv:2201.03904 (2022).
4. Pezzato, Corrado, et al. "Active inference and behavior trees for reactive action planning and execution in robotics." IEEE Transactions on Robotics 39.2 (2023): 1050-1069.

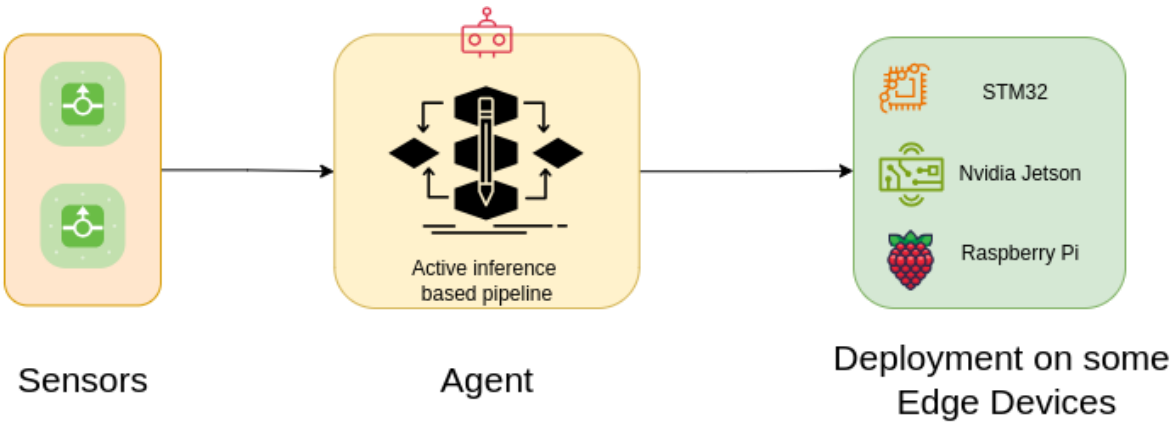
5. de Prado, Miguel, et al. "Automated design space exploration for optimized deployment of dnn on arm cortex-a cpus." IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 40.11 (2020): 2293-2305.

Sample Implementations:

1. Pymdp: https://github.com/infer-actively/pymdp
   https://pymdp-rtd.readthedocs.io/en/latest/index.html

DIAGRAM & TECHNICAL SPEC

Proposed system:



Input spec:

- Input sensor data

Agent:

- Sense, reason

Output spec:

- Action on the environment

REQUISITES:

- Mandatory
  - Open-source implementation
  - Methods implemented using pymdp
  - Deployment on two edge devices
- Preferential
  - Methods preferably implemented in JAX
  - Portability to ONNX and other inference engines, e.g., tensorRT, arm compute library, STM32Cube.AI

EXPECTED RESULT:

1. Implementation of the methods on two edge devices.

2. Comparative analysis report with a quantitative reduction in latency, memory, and energy consumption while maintaining accuracy. Target KPIs will be defined during 1st month of the project (model evaluation and selection stage).

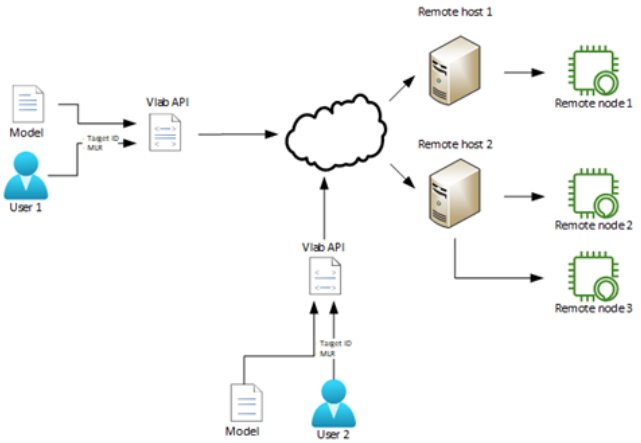
CHALLENGE 4. Integration of an AI-powered Edge Device to the Vlab for remote benchmarking

SHORT DESCRIPTION: Current implementation of dAIEdge VLab assists users who may not have expertise in embedded programming or access to the target embedded board by enabling real-time model benchmarking on a remote embedded board. It also allows owners of AI-powered edge devices to seamlessly integrate their hardware into the VLab infrastructure so that external users can run AI experiments remotely on the supported edge devices. This challenge proposes to the owners of AI-powered edge devices to integrate them to the VLab infrastructure. RISC-V based AI-powered devices and embedded boards integrating neural network accelerators are welcome.

EXPECTED DELIVERABLE: 1) Report; 2) Software; 3) Registration of the AI-powered Edge Device (Remote Node) to the VLab

REFERENCE INFO: AI-powered edge devices' owners interested in adding their boards to the VLab can find the guidelines under [dAIEdge Documentation (he-arc.ch)]

DIAGRAM & TECHNICAL SPEC: The figure below illustrates the overall architecture of the VLab. The objective pursued by this challenge is the integration of a new Remote Node to the current VLab infrastructure.



REQUISITES: Knowledge on Edge@AI and embedded programming

EXPECTED RESULT: New AI-powered Edge Device registered to the VLab and capable of running model benchmarking triggered by remote users.

CHALLENGE 5. Cloud-to-Edge AI-Powered Solution for Plant Disease and Pest Detection with BioClip Integration

SHORT DESCRIPTION: Agriculture faces mounting challenges from pests and plant diseases due to changing climates and evolving ecosystems. Traditional monitoring methods are labor-intensive and error-prone, requiring expert knowledge. This challenge calls for participants to develop an AI-powered solution that integrates BioClip, a vision foundation model for biological data, to detect plant diseases and insect pests using images acquired via mobile devices. The solution will leverage a cloud-to-edge distributed AI framework to ensure real-time detection and monitoring directly on smartphones while enabling more intensive processing and learning on the cloud. This challenge emphasizes sustainability and precision agriculture, aiming to help farmers make timely, data-driven decisions for managing pests and diseases while minimizing pesticide and fungicide usage.

The objectives of this challenge are:

- BioClip Integration for Detection of Plant Diseases and Pests. Develop AI models that utilize BioClip, a vision foundation model trained on diverse biological datasets, to accurately detect and identify both plant diseases and insect pests. The model should handle a variety of plant species and pests in real-world agricultural environments, addressing challenges like lighting variation, occlusion, and the presence of mixed pest and disease signals.
- Cloud-to-Edge Distributed AI for Real-Time Monitoring. The solution must be designed as a cloud-to-edge distributed application, where lightweight models run on edge devices (e.g., smartphones) for real-time inference, and more computationally intensive tasks are handled in the cloud. The edge device should provide instant feedback to farmers, while the cloud processes larger datasets for continuous model refinement and learning.
- Integration of Environmental and Contextual Data. Incorporate additional contextual data such as local temperature, humidity, and soil moisture to refine pest and disease prediction. The solution should offer actionable insights, enabling farmers to make informed decisions about interventions like pesticide application and irrigation adjustments.
- Scalability and Sustainability. Design the system to be scalable, supporting both small-scale farmers with basic mobile devices and larger farming operations with advanced IoT and cloud infrastructure. The solution should promote sustainability by enabling precise interventions that reduce unnecessary chemical treatments and improve crop health. The impact of climate change on viticulture has increased the unpredictability of pest dynamics, making pest management critical to sustainable wine production. Sticky traps are widely used to monitor pests, but current methods are labor-intensive and reliant on expert knowledge for species identification.

As pests and plant diseases continue to threaten crop yields worldwide, there is an urgent need for automated, reliable monitoring tools. BioClip, a vision foundation model built for biological data, provides a significant opportunity to improve the accuracy of AI models for detecting both pests and plant diseases by leveraging its deep hierarchical understanding of the tree of life. However, deploying such models in real-world agricultural settings requires efficient edge-based processing to provide farmers with real-time feedback, as well as cloud-based infrastructure for continuous learning and refinement.

This challenge builds on the concept of **edge AI for agriculture** by integrating BioClip and a cloud-to-edge architecture. The goal is to provide **real-time, mobile-based disease and pest detection** for crops, while ensuring the system can scale to handle large datasets in the cloud for future improvements.

This challenge also emphasizes sustainability by reducing the need for broad-spectrum pesticides through more precise pest control recommendations based on the collected data.

The challenge will proceed along the following steps:

- Dataset Preparation and Preprocessing. Participants will use or generate datasets containing images of crops showing symptoms of diseases or signs of insect infestations. These datasets should reflect real-world conditions such as variable lighting, occlusions from leaves or fruits, and a mixture of plant diseases and pests. Datasets can include well-known diseases (e.g., powdery mildew, leaf blight) and European pests.
- Model Development Using BioClip. Develop AI models that leverage BioClip to detect plant diseases and insect pests. Participants must fine-tune the BioClip model to improve accuracy on their specific dataset. The solution should generalize well to unseen species and diseases using BioClip's hierarchical taxonomy representations.
- Cloud-to-Edge Deployment. Design the solution as a cloud-to-edge distributed AI system. On edge devices, lightweight models should provide fast, on-device inference using smartphone cameras, while the cloud component should manage more resource-intensive tasks, such as training and model updates. The system must enable synchronization between edge and cloud for improved accuracy over time.
- Real-Time Monitoring and Decision Support. Integrate environmental data from sensors or external sources (e.g., IoT devices) to refine the prediction results. Based on detected pests or diseases and current environmental conditions, the system should offer real-time suggestions for interventions, helping farmers optimize their use of pesticides and water resources.

EXPECTED DELIVERABLE:

- Mobile AI Model: A fully-deployable AI model for plant disease / pest detection and counting, optimized for real-time inference on smartphones.
- Report: A detailed evaluation report comparing the performance of different models in terms of accuracy, speed, and power efficiency, with specific focus on edge deployment.

REFERENCE INFO:

Participants can use existing agricultural datasets or generate new ones. Datasets should include images of crop diseases and pests under various conditions (e.g., lighting, occlusion).

The solution will be evaluated on its accuracy (precision, recall, F1 score), latency of inference on edge devices, energy efficiency, and the quality of recommendations for pest and disease management.

DIAGRAM & TECHNICAL SPEC

Input Specification: Mobile-acquired images of plants showing symptoms of disease or pest presence, supplemented with environmental data such as temperature or humidity from sensors or APIs.

Output Specification:Real-time detection of plant diseases or pests with confidence scores, alongside treatment recommendations based on environmental conditions.

REQUISITES:

Mandatory: Experience with edge AI, cloud computing, and real-time agricultural applications.

EXPECTED RESULT:

- High Detection Accuracy: Achieve at least 85% accuracy in identifying common plant diseases and pests across multiple crops.
- Efficient Edge Processing: Ensure that the model can run in real-time on smartphones, processing an image within 5 seconds with minimal battery drain.
- Improved Sustainability: Demonstrate a reduction in unnecessary pesticide and fungicide use through precise, data-driven interventions.

CHALLENGE 6. Energy-Efficient Deployment of AI Models on Edge Devices Using FPGAs

SHORT DESCRIPTION: The increasing demand for edge AI in fields like agriculture, solar panel inspection, and drone-based infrastructure monitoring requires innovative approaches to deploy machine learning models that are both highly efficient and effective. This challenge aims to harness the potential of Network Architecture Search (NAS) in conjunction with hardware acceleration via Field Programmable Gate Arrays (FPGAs) to optimize deep learning models for such edge AI applications.

Participants will be tasked with developing an automated, hardware-aware NAS framework that designs energy-efficient AI models tailored for FPGA platforms, with a strong focus on reducing power consumption while maintaining performance in real-time applications with minimal engineering effort.

The objectives of this challenge are:

1. Hardware-Aware NAS Framework: Develop an automated NAS framework capable of optimizing neural network architectures specifically for FPGA-based platforms, with a primary focus on significantly reducing energy consumption.
2. Optimization of Performance Metrics: Fine-tune models to improve on-device inference performance, targeting key performance indicators such as latency, accuracy, and power efficiency across different edge AI tasks.
3. Automated Workflow with Minimal Engineering Effort: Streamline the development pipeline to minimize manual intervention, making it easier to adapt AI models for FPGAs without significant engineering overhead.
4. Real-World Testing with Industry-Related Datasets: Validate the proposed solutions using datasets from agriculture, solar panel inspection, and underwater structure monitoring, demonstrating real-world applicability.

Recent advancements in hardware-aware NAS have shown promise in generating deep learning models that balance accuracy with hardware-specific constraints such as power and latency. FPGAs, compared to traditional GPUs, offer advantages such as lower energy consumption and better timing predictability, which are crucial for edge applications requiring real-time processing. By combining NAS techniques with FPGA characteristics, this challenge explores new frontiers in efficient AI deployment at the edge.

The challenge will proceed along the following steps:

1. Dataset selection
2. Implement a two-stage NAS process that first identifies promising architecture candidates and then refines them based on FPGA-specific constraints
3. Deploy the optimized architectures on FPGAs (e.g. Xilinx SoC FPGA) to evaluate their performance in terms of throughput and energy efficiency and compare them with embedded GPUs (e.g. Nvidia Jetson) fabricated with the same technology
4. Measure accuracy, inference latency, and power consumption across different architectures and datasets

EXPECTED DELIVERABLE:

- Software: A robust framework for hardware-aware NAS tailored for FPGA platforms.
- Hardware: Optimized neural network architectures that demonstrate superior performance in edge AI applications especially in energy efficiency.
- Report: Comprehensive evaluations that provide insights into the effectiveness of FPGA acceleration compared to traditional methods.

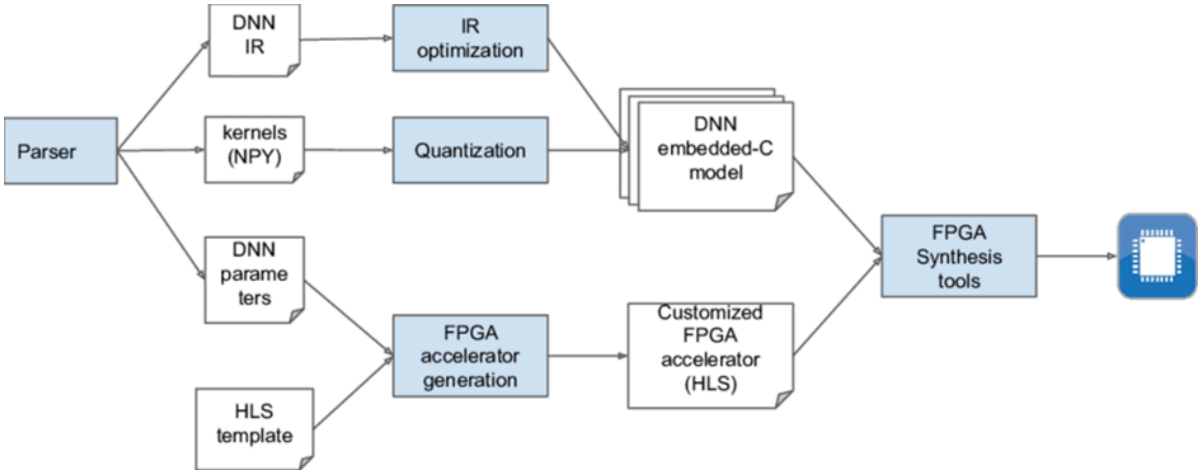
REFERENCE INFO :

The proposal should define industry relevant datasets for common image-based tasks of:

- Image classification
- Segmentation
- Object detection

Benchmarks should consist of accuracy drop/loss versus baseline, MAC compression ratio, memory compression ratio, latency, and energy consumption.

DIAGRAM & TECHNICAL SPEC:



REQUISITES:

<u>Mandatory:</u> The NAS workflow must work in an automated manner and not require significant engineering effort to obtain a satisfactory outcome for the FPGA.

EXPECTED RESULT:

- <u>Maintain Accuracy:</u> Ensure model accuracy remains on par with or better than embedded GPU implementations.
- <u>2x Improvement in Energy Efficiency:</u> Achieve a 2x reduction in energy consumption per inference compared to embedded GPUs (e.g., Nvidia Jetson).
- <u>10x Faster FPGA Development Time:</u> Deliver a 10x acceleration in hardware development time compared to traditional manual RTL design methods.

CHALLENGE 7. SAFE AI - Enhancing AI Model Predictions Through Synthetic Data and Domain Knowledge Integration

SHORT DESCRIPTION: Models often struggle to integrate domain-specific knowledge, leading to challenges in making reliable predictions in high-stakes environments, such as safety-critical tasks. This challenge aims to create efficient embedding structures that can be foundational for specific applications, especially in safety-critical environments. By leveraging domain-specific knowledge and synthetic data, the objective is to increase model robustness to noise and attacks, accelerate learning, and improve prediction stability and generalization. The resulting embeddings should facilitate the development of neural networks capable of tackling complex tasks with reduced retraining.

The objectives of this challenge are:

- Efficient Embedding Structures. Develop embeddings that encapsulate domain-specific knowledge and allow them to be reused across applications. These structures should support efficient learning and integration into neural networks for complex tasks.
- Leveraging Synthetic Data for Domain Knowledge. Use synthetic datasets to infuse domain-specific metadata into embeddings, enhancing the ability of the model to generalize from controlled scenarios to real-world applications.
- Resilience and Robustness. Focus on building models that are robust to noise and adversarial attacks, ensuring reliable performance in safety-critical tasks.
- Improved Generalization and Explainability. Ensure that the embeddings developed contribute to models that generalise well across domains and can provide improved explainability for their predictions.

Embedding structures are essential for efficiently capturing domain-specific knowledge in AI models. By using synthetic data with precise metadata, these embeddings can be optimized for specific applications, reducing the need for extensive retraining. This approach not only enhances model robustness, but can also improves explainability, making it suitable for high-stakes environments where reliability is paramount.

The challenge will proceed along the following steps:

1. Synthetic Dataset Generation and Selection. Create or select synthetic datasets enriched with domain-specific metadata to train models on diverse scenarios and enhance embedding structures.
2. Embedding Development. Implement adequate techniques based on the state-of-the-art to develop efficient embeddings that capture essential application domain characteristics.
3. Robustness Enhancement. Develop strategies to ensure the resilience of models to noise and adversarial attacks, exploiting the robustness of the embedding structures.
4. Explainability and Generalization Verification. Evaluate the proposed models against state-of-the-art techniques, focusing on their ability to generalize across domains and provide clear explanations for their predictions.

EXPECTED DELIVERABLE:

- Efficient Embedding Structures: Creation of reusable embedding structures and domain-specific embeddings, facilitating rapid adaptation of neural networks to complex tasks.
- **Performance Evaluation Report:** Evaluation of model performance against current state-of-the-art methods, emphasising robustness, generalization, and explainability.

REFERENCE INFO:

Tasks: Real-time and with application in safety critical applications.

Benchmarks: Comparison against established techniques using metrics such as accuracy, resilience, explainability, and computational efficiency.

REQUISITES:

Mandatory:

- Experience in the analysis and optimized implementation of neural networks on a wide range of reconfigurable parallel accelerators (e,g. FPGAs), from the very small suitable for power-constrained embedded systems to high performance computing.
- Experience in knowledge distillation aimed at extracting the essential features to minimize size and improve generalization without sacrificing accuracy.

Preferential:

- Experience in generating synthetic data for neural network training.
- Experience with efficient embedding structures and embedding manipulation.

EXPECTED RESULT:

- Higher Model Accuracy: Achieve significant improvements in prediction accuracy compared to models trained without domain-specific embeddings.
- Enhanced Robustness: Demonstrate superior resilience to noise and adversarial attacks compared to existing techniques.
- **I**mproved Explainability and Generalization: Ensure that model predictions are more interpretable and generalize well across different domains.

CHALLENGE 8. Very low consumption of speech enhancement AI algorithm at the edge

SHORT DESCRIPTION: To promote multimodal data exploitation in dAIEDGE, we propose to enrich the algorithm library developed in the project (Whereas the algorithms of this library would be mainly focused on images, we intend here to contribute to it also with algorithms on audios which could complement also fit overall purpose (also some of the Use Cases envisioned) with other modalities, in particular here with audio data).
This is highly relevant to release full potentiality of multimodality in a variety of Use Cases targeting AI at the Edge such as the Smart Warehouse monitoring use case depicting hereafter where multimodality could be used to significantly enrich intrusion detection capabilities based on multimodal data including audio and image data.

For this open call, we are looking for:

1) A HW platform on top of which a speech AI enhanced algorithm could be embedded and run at very low consumption (< 50 mW).
2) The development of one speech AI enhanced algorithm. LSTMs, Transformers or Mamba could be good candidates. The list is not exhaustive.

The whole system should be able to get on-boarded on a drone.

EXPECTED DELIVERABLE: On the software level, one must be able to input an ONNX speech enhancement model and a few audio files to the virtual lab and get metrics desired on the outputs of the algorithm: audio metrics (i.e. PESQ, SNR) and hardware metrics (i.e. inference time, consumption).

REFERENCE INFO

State of the art.
The current state of the art, in the context of implementation in very low consumption constraints, is the use of small LSTMs, which has its limitations. Meantime more recent results on Transformers and Mamba-type architectures have proved their worth “off-edge” and appear ready to move to the edge.

Datasets
Public datasets could be used for this open call: Librispeech, Voxceleb for example. Moreover, for some datasets, improvements have been done with the inclusion of noise to test the robustness of the algorithms.

Benchmarks
Experiments on some of these datasets are also available with the testing of denoising algorithms such as WaveUnet, DCCRN and TinyDenoiser). Results could be used as baselines for the benchmarks of the new algorithms and hardware implementation that should be done in this open call.

dAIEDGE VLab infrastructure can be used.

REQUISITES:

- Speech enhancement algorithms that will improve baselines defined in section 4: Mandatory
- Very low consumption speech enhancement pipeline: Mandatory
- Integration of the pipeline with ONNX: Preferential
- Integration in the dAIEdge-Vlab: Preferential
- Benchmark process automation

EXPECTED RESULT:

- Total time computation with specific constraints (to be defined)
- Total energy consumed with specific constraints (to be defined)

CHALLENGE 9. Off-Chip Weights for Streaming Architectures

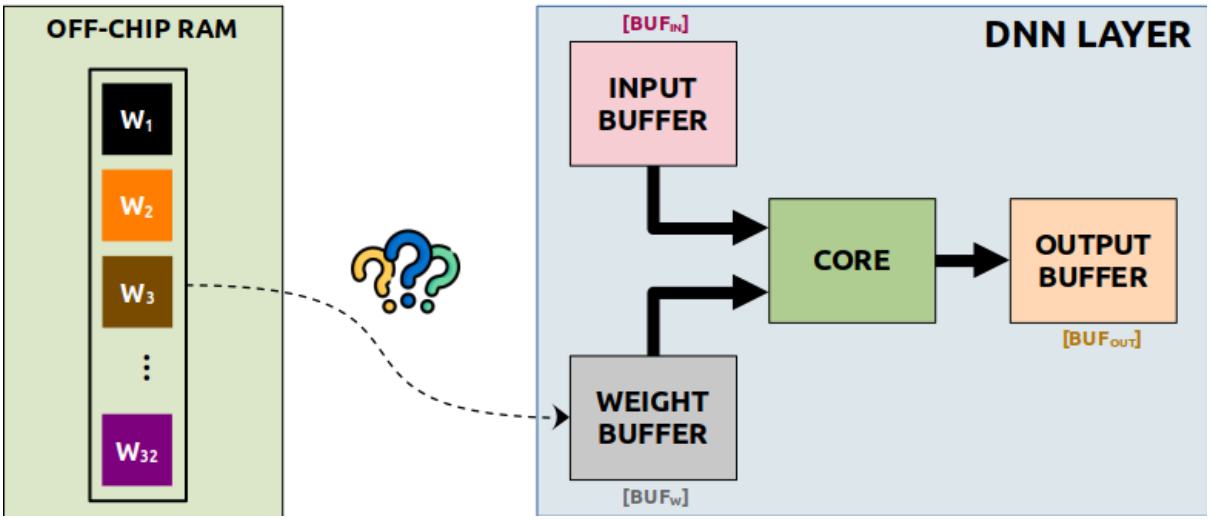
SHORT DESCRIPTION:

It is often challenging to fit large models directly on-chip due to their substantial memory requirements. System-on-Chip (SoC) architectures provide a flexible solution, enabling customized data handling and processing to accommodate such large networks. Streaming architectures, a popular approach in SoC designs, help manage data flow efficiently, allowing for continuous data processing rather than batch loading, which can minimize latency and optimize bandwidth usage. When the entire neural network cannot be stored on-chip, off-chip memory becomes essential for storing large sets of model weights. By strategically utilizing off-chip memory and transferring only the required weights to the on-chip processors when needed, streaming architectures ensure that data movement is minimized and managed efficiently. This approach maximizes processing throughput while keeping power consumption low, addressing the constraints of both hardware resources and real-time processing demands.

The **objective** of this challenge is to determine the optimum **size** and **order** of parameter chunks that should be transferred from the off-chip memory onto the on-chip fabric so that the latency is minimal, and the energy inefficiency is avoided.

EXPECTED DELIVERABLE:

- An algorithm that manages off-chip weight storage and dynamically loads only required weights onto on-chip memory, minimizing latency and optimizing memory usage.
- The algorithm should be benchmarked on common deep neural networks such as MobileNet, ResNet and AlexNet.
- Comprehensive documentation outlining the algorithm design, memory access strategy, setup instructions, and potential areas for further optimization.



CHALLENGE 10. 2D Vehicle Detection Network on PYNQ-Z1

SHORT DESCRIPTION:

This challenge aims to implement a real-time vehicle detection network on the PYNQ-Z1 board using a compact convolutional neural network (CNN) optimized for the resource constraints of the FPGA-based platform. Leveraging the flexibility of the PYNQ-Z1's System-on-Chip (SoC) architecture, the network will be partitioned to run both on the ARM Cortex-A9 processor and on the programmable logic (FPGA) for acceleration. The network could use off-chip memory efficiently to store model weights, ensuring seamless data transfer through DMA channels and optimizing detection speed for real-time applications.

EXPECTED DELIVERABLE:

- A quantized (and pruned) vehicle detection CNN model (e.g., YOLO or MobileNet SSD), tailored for deployment on the PYNQ-Z1.
- A quantized hardware-accelerated CNN implemented as custom IP cores or using Xilinx's Vitis AI library deployed on the PYNQ-Z1.
- A complete software stack on PYNQ, including model partitioning, DMA for A functional real-time detection demo showcasing real-time vehicle detection on live video input or sample images, running on the PYNQ-Z1 board.
- Comprehensive documentation covering the setup, model optimization, hardware-software integration, and usage instructions for deploying the vehicle detection network on PYNQ-Z1.

Datasets

COCO, KITTI Vehicle Detection, or PASCAL VOC.

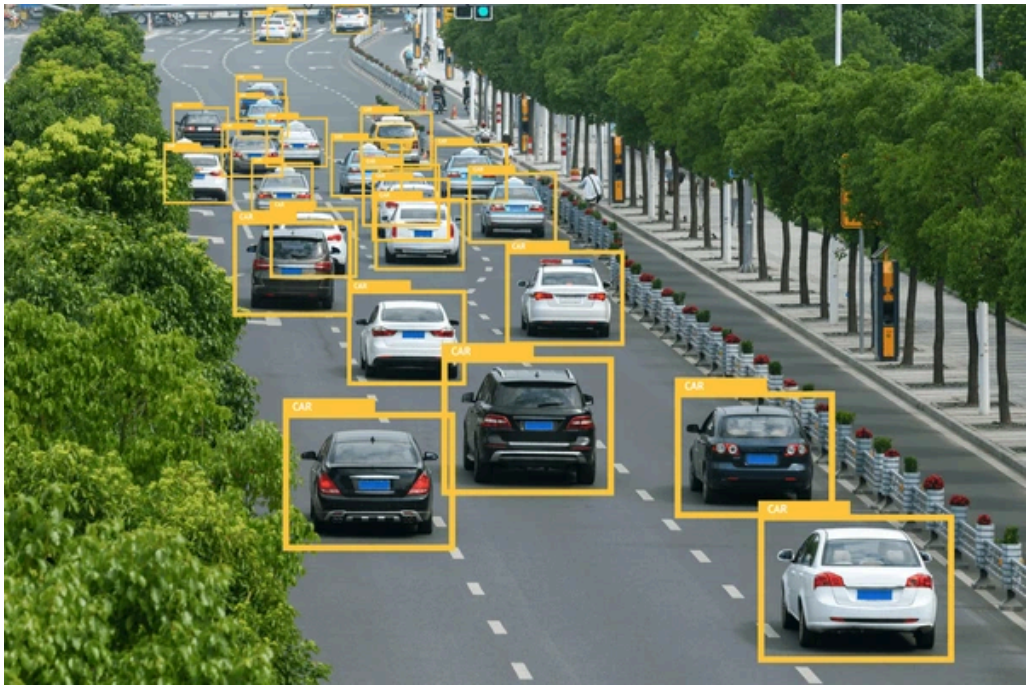


Image source: https://www.mouser.mx/blog/mystery-of-vehicle-detection

CHALLENGE 11. Real-Time Vision Transformer (ViT) on FPGA SoC for Image Classification

SHORT DESCRIPTION:

This challenge focuses on deploying a Vision Transformer (ViT) model on an FPGA SoC to perform image classification tasks. Vision Transformers leverage attention mechanisms instead of traditional convolutional layers, making them ideal for handling visual tasks that require global context. Implementing ViT on an FPGA will involve optimizing the model's matrix operations, attention layers, and patch embeddings for efficient FPGA execution. This approach is designed to accelerate processing while maintaining model accuracy, making it suitable for real-time image classification on edge devices.

EXPECTED DELIVERABLE:

- Optimized Transformer Model: A streamlined, quantized version of the Transformer model, with attention layers and feedforward networks tailored to the FPGA SoC architecture.
- Custom Hardware Modules: FPGA-accelerated modules for matrix multiplication, attention mechanisms, and softmax, designed as reusable IP cores.
- Efficient Memory Management: Implementation of an optimized memory transfer system (e.g., using DMA) to handle large token embeddings and model weights stored off-chip, with data streaming to the FPGA.
- Demo Application: A working demo showcasing the Transformer model performing tasks like text classification or summarization on a sample input dataset.
- Documentation: Detailed documentation on model optimization, hardware design, memory management strategy, and usage instructions for deploying the Transformer on FPGA SoC.

Datasets

CIFAR-10, ImageNet Subset, or Tiny ImageNet.

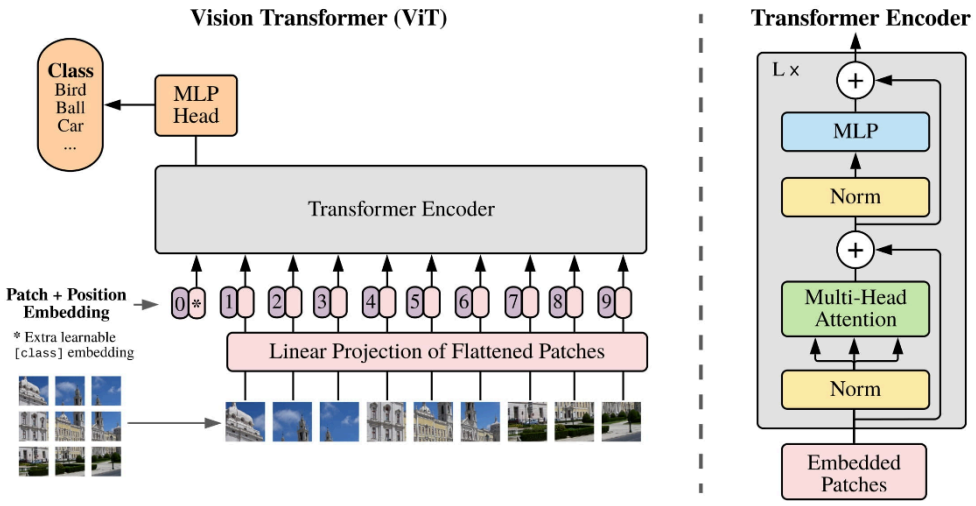


Image source: https://arxiv.org/pdf/2010.11929

CHALLENGE 12. Spiking Neural Networks (SNNs) on the edge

SHORT DESCRIPTION:

Test the capacity of SNNs on the edge, possibly (but not restricted) on speech enhancement tasks.
SNN are particularly interesting for dAIEDGE because they use energy only when “spikes” are emitted. It means that, compared to their neural networks equivalent, the computation graph is much more sparse whereas standard neural networks compute the interactions of all the weights with all the activations of the previous layer. As activations are often the “ReLU”, one can expect a 50% sparsity from standard neural networks whereas SNN can achieve very much higher sparsity (>95% or more).

EXPECTED DELIVERABLE:

The SNN-compatible hardware is expected to be accessible by the dAIEdge-VLab and, on the software level, one must be able to input SNN model (through ONNX or PyTorch) and a few audio files to the virtual lab and get metrics on the outputs of the algorithm : audio metrics (PESQ, SNR) and hardware metrics (inference time, consumption).

REFERENCE INFORMATION:

We are interested in different modalities (audio, image, etc.) so this challenge can be applied to one or more modalities. One needs a comparison between an equivalent non-SNN baseline (a standard neural network) and your solution.

REQUISITES:

<u>Mandatory:</u>

- An implementation running on the dAIEdge-VLab of your solution (hardware + software).
- A report detailing your solution and results in terms of the KPI (described below).

EXPECTED RESULTS:

For audio speech enhancement: improvement in the audio quality (PESQ of processed files must be better than PESQ of noisy files). For image or other modalities, we’re open to various metrics as long as they are representative to the overall quality performance of your solution. About energy efficiency, one expects a very high FLOPs/s over watts efficiency (coherent to SNN promises), one expects one or more orders of magnitude efficiency compared to standard neural networks.

CHALLENGE 13. Energy efficient Text Embeddings inference at the edge.

SHORT DESCRIPTION:

Text embeddings inference is the process of transforming text data into numerical vectors that capture semantic meaning, enabling machines to understand and analyse the text effectively.

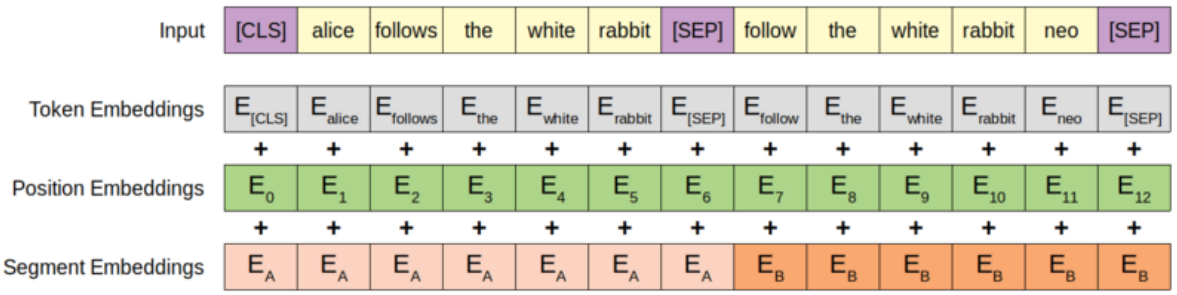


Image Source: https://commons.m.wikimedia.org/wiki/File:BERT_input_embeddings.png

This technique is crucial for various natural language processing tasks such as sentiment analysis, information retrieval, and machine translation and it is often the core of Retrieval Augmented Generation systems. Embeddings computation can heavily decrease the transferred data between the edge and the cloud; thus, it is pivotal to compute text embeddings efficiently at the edge.

EXPECTED DELIVERABLE:

- Report on the average energy consumption of the inference model for Retrieval task in English language
- Detailed explanation of the approach adopted to benchmark at least 3 hardware accelerators on 8 different sentence transformer models
- (Preferable) Open-source code used for the report data generation for wide reproducibility

REFERENCE INFORMATION:

Benchmark:

- Massive Text Embedding Benchmark should be used to assess the results. Target application should be the Retrieval one, but more can be selected and used. GitHub - embeddings-benchmark/mteb: MTEB: Massive Text Embedding Benchmark
- Models should be picked between the non-Proprietary models available on the mteb leaderboard with a memory usage < 2GB for English language and for the Sentence Transformers task: https://huggingface.co/spaces/mteb/leaderboard
- Hardware platform should be selected between the available systems in the Virtual Lab

REQUISITES:

- Mandatory:
  - A report on the average power consumption on different hardware of the sentence transformer task. The analysis should focus on at least 8 models on at least 3 hardware platforms
- Preferential:
  - Open-source repository with the code used in the evaluation phase to guarantee reproducibility of the results.
  - Mode models can be included (up to 20) and more hardware models available in the virtual lab
  - Overall guide on model selection for edge execution of sentence transformer tasks.

CHALLENGE 14. Automated Edge deployment, tuning and Performance Evaluation of Binary Neural Networks targeting FPGA enabled platform

SHORT DESCRIPTION:

Binarization is essentially a form of extreme (1-bit) quantization of neural networks values of weights, activations or both. This compression technique allows heavy floating point math operations to be replaced by lightweight bitwise XNOR and Bitcount operations. As a result, binary neural networks offer several hardware-friendly advantages that fits well with the constraints of embedded and edge platforms: reduced memory consumption, improved power efficiency, and significant speed enhancements.

The pioneering research results on BNN and XNOR-Net networks have demonstrated the effectiveness of binarization, achieving up to 32× memory savings and 58× speed improvements on CPUs for 1-bit convolution layers. This has sparked extensive research in computer vision and machine learning, with applications in tasks like image classification and detection. Additionally, binarization allows for easy validation of a layer's importance by switching between full-precision and 1-bit representations.

FPGA based hardware is natively suited for BNN implementation from resources perspective (i.e. configurable logic operators) however the design, implementation, validation and integration process is relatively lengthy and tedious.

CETIC is developing and edge middleware (DMWay) aiming at enhancing the productivity of edge AI computation developers by providing means to quickly configure and interface these computations, the associated In/out data streams and their set-up and execution parameters. It also enables a better automation of bechmarking and performance evaluation scenarios.

The objective of this challenge is to develop, prototype and evaluate a dedicated approach for the automation of BNN deployment, tuning and performance evaluation on Zynq FPGA platform, leveraging the DMWay middleware capabilities.

The challenge will proceed along the following steps:

4) Selection of a suitable target set of BNN models (and data sets) for the challenge based on the SotA literature on the subject, the availability of exploitable baseline or reference implementations for Zynq FPGA and the level of “tunability” of these models.
4) Define relevant, generic and reusable scenarios for the automation of performance evaluation and parameters exploration of these models execution on the FPGA targets, leveraging the Pynq environment (Python for Zynq) for Zynq FPGA
4) Set-up, develop and showcase the automation process implementation on different BNNs, different hardware targets, leveraging DMWay middleware capabilities. Direct and continuous support from the middleware team is guaranteed.
4) Based on the previous steps, define a reusable set of BNN related primitives and operations that could be automated, monitored and controlled by the middleware

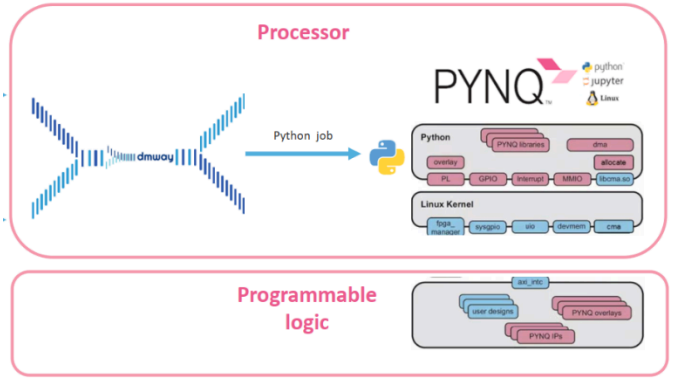
EXPECTED DELIVERABLE:

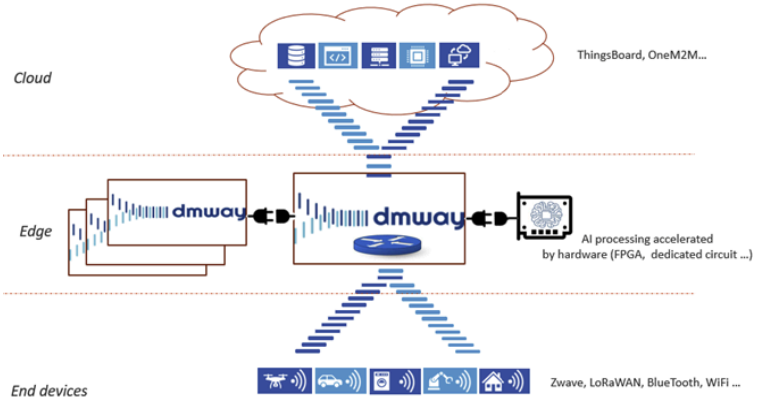
- Working hardware implementations of selected BNNs on Zynq FPGA
- Definition of automation scenarios for deployment and evaluation of BNNs on Zynq FPGA leveraging DMWay middleware
- Demonstrator showcasing the forementioned scenarios and the automation capabilities developed

REFERENCE INFO:

- BNNs : Survey (2020) Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, Nicu Sebe, Binary neural networks: A survey, Pattern Recognition,Volume 105,2020, 107281, ISSN 0031-3203,https://doi.org/10.1016/j.patcog.2020.107281.
  https://www.researchgate.net/publication/339411073_Binary_Neural_Networks_A_Survey
- Pynq framework for Zynq FPGA : https://www.pynq.io/
- DMWay middleware:
  - L. Deru, A. Achour and L. Guedria, "Multi-protocols and Data Manager for IoT Gateways: A smart-building use-case demo," *2021 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, Bordeaux, France, 2021, pp. 722-723.
  - https://asset.cetic.be/en/dmway/

DIAGRAM & TECHNICAL SPEC:





REQUISITES:

- Python programming
- Decent practical FPGA/embedded computing experience
- AI & neural networks background

EXPECTED RESULT:

- Higher level of automation of BNNs implementation and qualification on FPGA Zynq platforms
- 3+ implemented and showcased tuning scenarios of BNNs on FPGA using DMWay middleware

CHALLENGE 15. Deploying LLMs on Low-Memory Edge Devices

SHORT DESCRIPTION:

LLM inference on resource-constrained edge devices becomes impossible due to the limited memory capacity to load LLM weights before inference. Additionally, specialised hardware accelerators are used alongside the host processor to accelerate the compute-intensive operations within LLMs, but due to the nature of LLMs, the performance gain from these accelerators is capped by the overall bandwidth between main memory and accelerator.

Our challenge is to solve this problem by enabling dynamic loading of LLM weights into main memory without significantly hampering performance.
Our goal for this challenge is to develop an architectural solution that incorporates software and hardware innovations to improve the LLM inference performance.
The solution should focus on loading weight data from storage to the main memory and eventually to the accelerator memory efficiently.

EXPECTED DELIVERABLE:

- Technical report of the proposed architectural solution and evaluation.
- The architectural solution should incorporate FPGA-based accelerators to improve generation performance of LLM inference.
- Demonstration of LLMs inference of models larger (10-30%) than memory capacity of the target device (Preferably PYNQ Z1 or Kria KV260).
- Open-source code and documented design process ideally utilising the SECDA methodology.

REFERENCE INFO (datasets, AI/SW/HW models, benchmarks):

- The SECDA methodology: "SECDA: Efficient Hardware/Software Co-Design of FPGA-based DNN Accelerators for Edge Inference." SBAC-PAD 2021.
- The SECDA-TFLite toolkit: "SECDA-TFLite: A toolkit for efficient development of FPGA-based DNN accelerators for edge inference". JPDC 2023. https://github.com/gicLAB/SECDA-TFLite
- SECDA-LLM toolkit: "Designing Efficient LLM Accelerators for Edge Devices", ARC-LG @ ISCA 2024.

REQUISITES:

- Mandatory:
  - The technical report should include an initial performance/roofline model to estimate the expected performance for the target device before and after the proposed solution.
  - Evaluation of performance in terms of latency, throughput and power efficiency across at least three different LLMs.

- Preferential:
  - SystemC model of the architectural solution.
  - Interactive demo on real hardware which enables simple Q&A chatbot with LLM running locally on target device.

EXPECTED RESULT:

We should be able to run and execute LLM models larger than (10-30%) the main memory capacity of target devices.
We should achieve reasonable performance in terms of latency (10 token/s).

CHALLENGE 16. Developing an Early Warning System for Harmful Algal Blooms Using Earth Observation Data

SHORT DESCRIPTION:

Harmful Algal Blooms (HABs) are rapid overgrowths of toxic algae in marine, freshwater, and brackish ecosystems. These blooms disrupt local ecosystems, kill aquatic life, and pose health risks to humans, resulting in widespread economic consequences. In the United States alone, HABs cause annual economic losses of between $10 million and $100 million, with some major events resulting in tens of millions of dollars in damages. Internationally, the impact has been similarly severe: in 2019, a HAB event in northern Norway killed eight million salmon, leading to economic losses exceeding 850 million NOK (Davidson et al., 2020). In Chile, a HAB event in 2016 caused the mortality of 39 million salmon, with an economic toll of approximately USD $800 million (Anderson and Rensel, 2016).

Traditional in situ testing for HABs relies on physical sampling and laboratory analysis, which are labor-intensive, costly, and geographically limited. In contrast, Earth Observation (EO) data offers a scalable, faster solution that can enhance HAB detection through satellite-based remote sensing, potentially even providing early warnings. Real-time satellite insights processed onboard can make HAB detection more efficient, achieving greater geographical coverage globally. The primary challenge is to develop a dataset of EO data featuring HAB events to train a detection model that is compatible with CogniSAT6 (Rijlaarsdam et al., 2024), an advanced satellite platform with frequency bands relevant to HAB detection.

An initial training dataset will be sourced from the Tick Tick Bloom competition, focusing on inland water bodies due to the scarcity of high-quality open water HAB datasets. To enhance the dataset further, an innovative approach will use the model's probability scores for HAB events, cross-referencing these with known HAB event data and verification from human checks, to iteratively build a robust open-water HAB dataset compatible with the specifications of CogniSAT6. This capability will enable precise, large-scale HAB monitoring, allowing targeted in situ measurements and advancing global HAB detection efforts.

EXPECTED DELIVERABLE:

- <u>HAB Dataset and CNN Model for Inland Waters:</u> Development of a convolutional neural network (CNN) model trained on a dataset specifically designed for detecting harmful algal blooms (HABs) in inland water bodies.

- Expansion to Open Waters: Augmentation of the model and dataset to include detection capabilities for HABs in open water environments, by using the "inland" model predictions on open water data and human validation.
- CogniSAT6 Compatible Model: Creation of a model that is compatible with the frequency bands and specifications of CogniSAT6 for effective HAB monitoring and early warning systems.

REFERENCE INFORMATION:

https://www.drivendata.org/competitions/143/tick-tick-bloom/page/649/

David Rijlaarsdam, Tom Hendrix, Pablo T Toledano González, et al. The Next Era for Earth Observation Spacecraft: An Overview of CogniSAT-6. *TechRxiv.* February 22, 2024.

Davidson, K., Jardine, S., Martino, S., Myre, G., Peck, L., Raymond, R., et al. (2020). "The Economic Impacts of Harmful Algal Blooms on Salmon Cage Aquaculture," in *GlobalHAB. Evaluating, Reducing and Mitigating the Cost of Harmful Algal Blooms: A Compendium of Case Studies*, ed. V. L. Trainer 84–94.

Anderson, D., and Rensel, J. (2016). *Harmful Algal Blooms Assessing Chile's Historic HAB Events of 2016 A Report Prepared for the. Glob. Aquac. Alliance 19.* Available Online at: https://www.aquaculturealliance.org/wp-content/ (accessed March 24, 2021).

REQUISITES:

- Mandatory: Dataset of open water raw EO data with HAB and non-HAB events labelled.

- Preferential: A trained CNN model compatible with the CS6 data

CHALLENGE 17. Edge AI framework for web browsers empowering Federated Machine Learning applications addressing Classification problems over multiple format data (text and images)

SHORT DESCRIPTION:

Federated learning (FL) is a decentralized approach to training machine learning models that gives advantages of privacy protection, data security, and access to heterogeneous data over the usual centralized machine learning approaches. We can obtain more accurate and generalizable models through FL without having the data leave the client devices. The three main strategies to perform FL are Centralized FL, Decentralized FL, and Heterogeneous FL with popular FL algorithms such as FedSGD, FedAvg, and FedDyn.

Centralized federated learning requires a central server. It coordinates the selection of client devices in the beginning and gathers the model updates during training. The communication happens only between the central server and individual edge devices. While this approach looks straightforward and generates accurate models, the central server poses a bottleneck problem—network failures can halt the complete process. Decentralized federated learning does not require a central server to coordinate the learning. Instead, the model updates are shared only among the interconnected edge devices. The final model is obtained on an edge device by aggregating the local updates of the connected edge devices. Heterogeneous federated learning involves having heterogeneous clients such as mobile phones, computers, or IoT (Internet of Things) devices. These devices may differ in terms of hardware, software, computation capabilities, and data types.

Federated ML Framework - The participants should develop a pioneering framework leveraging federated learning (FL), WebAssembly (Wasm), and WebGPU to facilitate training and deploying deep neural networks (DNNs) directly in web browsers. By combining FL's privacy-preserving distributed learning approach with Wasm and WebGPU's performance capabilities, we aim to create a powerful, accessible, and secure environment for real-time machine learning on client-side devices. This framework will enable decentralized model training without centralizing sensitive data, making it particularly suited for applications where data privacy and security are paramount.

Use case development - This framework can be valuable for building applications in areas such as: 1) *Healthcare Diagnostics:* Real-time image and signal analysis for remote health diagnostics without sharing patient data externally. 2) *Personalized E-Learning:* Adaptive learning systems that adjust in-browser educational content in real-time based on user interaction. 3) *Smart Cities:* Decentralized processing for edge devices, enhancing urban management services, such as traffic monitoring, without centralized data storage. 4) *IoT Device & Sensing Optimization*: Localized model training on IoT-enabled / networked sensor devices, ensuring data remains on-device while enhancing capabilities such as predictive maintenance, event forecasting and identification. 5) *Augmented Reality and Virtual Assistance:* Real-time personalization and contextual awareness in AR/VR environments based on user interaction patterns, all executed within the browser. One such test application will be developed (an existing use case can be adapted appropriately) that will leverage federated learning based on local user or other agent type interactions over the browser. Decentralized approach ensures that users retain full control of their data while benefiting from a truly personalized and contextual user experience that dynamically adapts to foster engagement and improve specific outcomes for each use case. FL framework built for computer vision applications will be utilised, combined with AI/NLP resources where needed.

Despite the ongoing research on scaling FL systems, certain limitations still need to be addressed. It's necessary to improve communication efficiency, protect data privacy, and incorporate the heterogeneity present at systems and statistical levels.

EXPECTED DELIVERABLE:

a. Report on the FL framework architecture
b. Software implementation of the enhanced browsers
c. Model Trained using the Federated ML framework
d. Development of an application use case utilizing the Federated ML framework

REFERENCE INFO :

[1] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106775, Mar. 2021, doi: 10.1016/j.knosys.2021.106775.

[2] S. Kakati and M. Brorsson, "WebAssembly Beyond the Web: A Review for the Edge-Cloud Continuum", 2023 3rd International Conference on Intelligent Technologies (CONIT), Karnataka, India. June 23-25, 2023 https://orbilu.uni.lu/bitstream/10993/57810/1/WebAssembly_Beyond_the_Web_A_Review_for_the_Edge-Cloud_Continuum.pdf

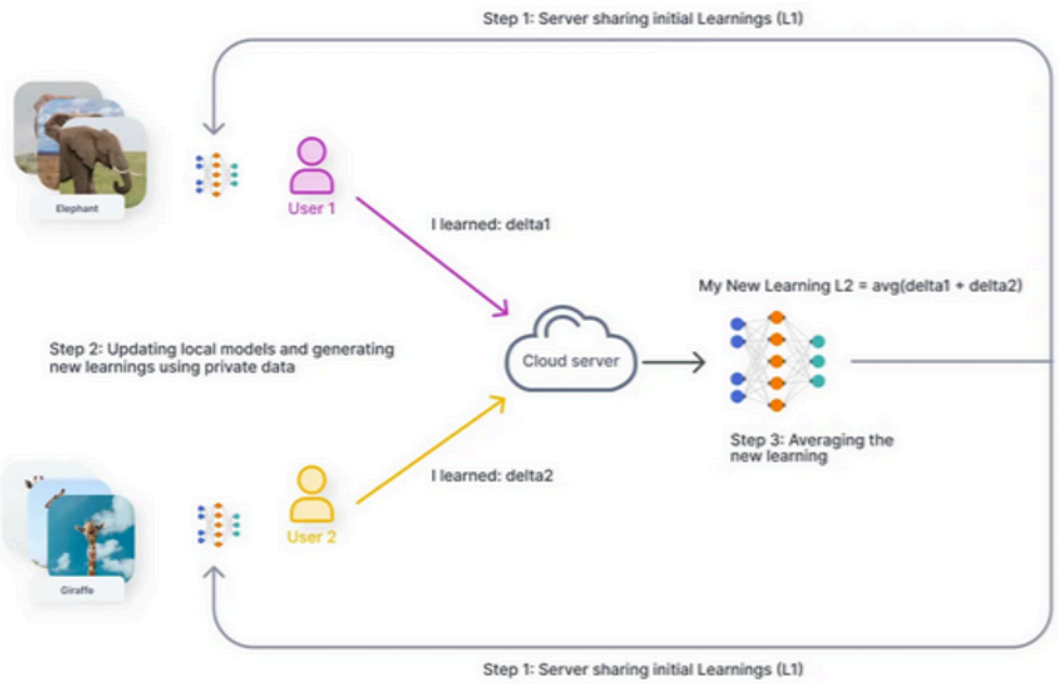
[3] ONNX Runtime Web (https://onnxruntime.ai/)

[4] Apache TVM for WebGPU (**https://tvm.apache.org/**): Apache TVM, a deep learning compiler, now supports WebGPU and WebAssembly, making it possible to achieve near-native GPU performance within browsers.

Datasets: https://universaldependencies.org/

DIAGRAM & TECHNICAL SPEC:

A generic baseline model is stored at the central server. The copies of this model are shared with the client devices, which then train the models based on the local data they generate. Over time, the models on individual devices become personalized and provide a better user experience.



In the next stage, the updates (model parameters) from the locally trained models are shared with the main model located at the central server using secure aggregation techniques. This model combines and averages different inputs to generate new learnings. Since the data is collected from diverse sources, there is greater scope for the model to become generalizable.

Once the central model has been re-trained on new parameters, it's shared with the client devices again for the next iteration. With every cycle, the models gather a varied amount of information and improve further without creating privacy breaches.

REQUISITES:

- Machine Learning
- NLP

EXPECTED RESULTS:

Benchmarking of the federated ML framework, in terms of:

- Flexibility in terms of use case development
- Performance
- Scalability
- Accuracy

CHALLENGE 18. Machine Learning Benchmarks for On-Board Processing in Space Applications

SHORT DESCRIPTION:

As space missions become increasingly reliant on on-board processing (OBP) systems for critical tasks like image processing, radar processing, and RF signal compression, the need for standardized performance benchmarks has grown. Existing commercial benchmarks often fail to reflect the unique demands of space applications or account for highly parallel processors such as GPUs, FPGAs, and AI accelerators, which are critical for space-based machine learning tasks.

This challenge invites participants to leverage and extend the OBPMarkML suite, a set of open-source computational performance benchmarks, to design and benchmark AI and machine learning models for space applications. The models should target space-relevant use cases such as cloud screening, object detection, 6DoF, and event classification, with a focus on ensuring efficiency in highly constrained space environments. The benchmarking solution should be optimized for hardware diversity, including FPGAs, GPUs, and AI accelerators, while delivering high-performance machine learning on limited power and computational resources.

The objectives of this challenge are:

- Develop and Optimize AI Models for On-Board Processing (OBP) Machine Learning Benchmarks. Build and optimize AI and machine learning models tailored to space applications, using benchmarks from the OBPMark-ML suite. The solution should focus on space-specific tasks such as cloud screening, ship detection, and solar event classification. Each model should be efficient, capable of performing real-time processing, and deployable on constrained hardware platforms like FPGAs and AI accelerators ranging from Int8 to Float32.

- Benchmark and Compare Computational Performance. Utilize the OBPMarkML suite to benchmark the computational performance of the developed models across different hardware platforms. The solution should include comprehensive comparisons of processing speed, power consumption, and inference accuracy, ensuring that the solution meets space-specific requirements for on-board processing.

Space missions rely on sophisticated on-board processing (OBP) systems to handle a variety of tasks, such as real-time data reduction, signal processing, and decision-making. These systems must be optimized for performance, power efficiency, and reliability due to the constraints of space environments. Current benchmarking methods are limited in scope, often focusing on specific devices or proprietary applications, which restricts reproducibility and comparability across different hardware platforms.

To address this, OBPMark was developed as an open-source suite of computational benchmarks, covering a wide range of space-relevant applications including image processing, radar, encryption, and machine learning. This challenge builds on the OBPMark-ML subset, focusing on benchmarks for machine learning tasks in space, such as cloud screening and object detection.

The challenge will proceed along the following steps:

1. Benchmark Selection and Dataset Preparation. Select relevant benchmarks from the OBPMark-ML suite, such as cloud screening, ship detection, or solar event classification. Prepare the necessary datasets for benchmarking, utilizing publicly available space datasets or generating synthetic data as required.
2. Model Development and Optimization. Develop and optimize machine learning models using standard neural network architectures, such as U-Net for cloud screening and YOLO for object detection. These models

should be optimized for low-power, high-performance processing on space-grade hardware such as FPGAs and AI accelerators.

3. Benchmarking Across Multiple Hardware Platforms. Benchmark the developed models on different hardware platforms, including FPGAs, GPUs, and AI accelerators. Use the OBPMarkML suite to measure key performance metrics such as processing speed, power consumption, and model accuracy. Provide detailed performance comparisons across these platforms.

EXPECTED DELIVERABLE:

- Benchmarking Reports. Detailed reports comparing the computational performance (e.g., FLOPS, energy efficiency, latency) of the developed models across different devices (e.g., FPGA, AI accelerators), based on the OBPMarkML benchmarks.

- Open-Source Benchmark Source Code Implementation. An open-source repository containing the implemented benchmarking scripts and results. The repository should follow OBPMark's open-source philosophy, enabling future extensions and optimizations by the community.

REFERENCE INFO:

- Datasets
  Participants can use space-relevant datasets, such as Earth observation images for cloud screening or satellite images for ship detection. Public datasets from past ESA missions, as well as simulated or synthetic data, are also acceptable.
- Benchmarks
  The solution will be evaluated using the OBPMark-ML benchmarks for machine learning tasks in space applications. Metrics will include processing speed, energy efficiency, model accuracy, and scalability.

DIAGRAM & TECHNICAL SPEC:

- Input Specification: Image data from space missions (e.g., Earth observation, satellite data) along with reference labels for training and testing. Participants should also consider additional input like environmental data or real-time sensor outputs.
- Output Specification: Real-time detection of clouds, objects, or solar events, with confidence scores and alerts for event-based processing. The solution should also include benchmarking metrics for computational performance across different hardware.

REQUISITES:

- Mandatory: Solutions must leverage OBPMark-ML benchmarks and be optimized for space-relevant hardware platforms (FPGAs, AI accelerators).
- Preferential: Experience with space-based machine learning and FPGA programming.

EXPECTED RESULT:

- High Processing Efficiency: The solution should demonstrate significant computational efficiency (processing speed, energy consumption) on space hardware platforms.
- Real-Time Performance: Models should achieve real-time or near-real-time performance, suitable for on-board processing during space missions.
- Broad Scalability: The solution should be adaptable to different missions and reusable across a variety of space applications.

CHALLENGE 19. Incentivisation framework on blockchain for optimizing AI development

SHORT DESCRIPTION: Training and running AI models is compute-intensive and costly – and does not compensate AI talent creators fairly for their work. Can decentralized technology provide a solution? For instance, LLMs consume large amounts of compute power and data, which are currently not correctly incentivized. The question is how this resource allocation will be decided. Blockchain's immutability constructs a fruitful environment for creating high quality, permanent and growing datasets for deep learning, and thus the fusion of blockchain technology with AI can be explored in order to create a properly incentivized environment for the evolution of LLMs or other AI applications. The objective is to develop the right incentives and payments for data, models and compute resources, as blockchain can facilitate low-fee micropayments for use of a generative AI model using a stable coin. A smart contract could allow revenue sharing across multiple co-owners of a model in a decentralized fashion.

EXPECTED DELIVERABLE: 1) Report, 2) Software and 3) Demo.

REFERENCE INFO

There are a number of papers on the topic of incentives for AI in distributed systems:

- Artificial Intelligence Implementations on the Blockchain. Use Cases and Future Applications by Konstantinos Sgantzos and Ian Grigg. https://doi.org/10.3390/fi11080170
- Incentive techniques for the Internet of Things: A survey, July 2022 Journal of Network and Computer Applications.
- ViSDM: A Liquid Democracy based Visual Data Marketplace for Sovereign Crowdsourcing Data Collection, 2023, Venkata Satya Sai Ajay Daliparthi Nurul Momen Kurt Tutschku.

DIAGRAM & TECHNICAL SPEC:

The idea is to develop an incentive mechanism for a part or all of the AI development process, and demonstrate how it can be deployed on a main public blockchain.

REQUISITES:

Mandatory: Knowledge in machine learning and blockchain.

EXPECTED RESULT:

Demo for a blockchain based framework where one part or more of the AI development process (data, training, compute) is optimised using a replicable incentivization framework. Has to demonstrate gains from legacy approaches.

CHALLENGE 20: Planning and Scheduling for Space Observations at the Extreme Edge

SHORT DESCRIPTION:

Increasing developments consider Space Situation Awareness, Earth Monitoring and Surveillance using spaceborne observations. Development of new spacecraft constellation has become affordable for private companies and nations, and where Edge AI can be applied. However, those systems require resources and performance optimisation, but also resilient and reactive behaviours while limiting vulnerabilities.
We propose a challenge to study distributed modelling and solving approach for allocating on-line space observations under temporal and resource constraints. Observations will consider earth pole monitoring.
The approach shall take advantage of the state of the art in distributed algorithms for the Edge based on (Distributed Constraint Optimisation Problems, Distributed Planning and Scheduling, Argumentation methods, Distributed Agreements) as well as various distributed problem-solving paradigms. Contribution consists in a distributed solving method with adequate performance on edge processor and without single point of failure.

EXPECTED DELIVERABLE:

- Report on the distributed solving technique
- Software
- Demonstration over Benchmarks

REFERENCE INFO:

[1] Daniel Cellucci, Nick B Cramer, and Jeremy D Frank. 2020. Distributed spacecraft autonomy. In ASCEND 2020. 4232.
[2] Nicholas Cramer, Daniel Cellucci, Caleb Adams, Adam Sweet, Mohammad Hejase, Jeremy Frank, Richard Levinson, Sergei Gridnev, and Lara Brown. 2021. Design and testing of autonomous distributed space systems. In 35th Annual Small Satellite Conference.
[3] Michel Lemaître, Gérard Verfaillie, Frank Jouhaud, Jean-Michel Lachiver, and Nicolas Bataille. 2002. Selecting and scheduling observations of agile satellites. Aerospace Science and Technology 6, 5 (2002), 367–381.
[4] Shreya Parjan and Steve A Chien. 2023. Decentralized Observation Allocation for a Large-Scale Constellation. Journal of Aerospace Information Systems 20, 8 (2023), 447–461.
[5] Gauthier Picard. 2022. Auction-based and distributed optimization approaches for scheduling observations in satellite constellations with exclusive orbit portions. arXiv preprint arXiv:2106.03548 (2022).
[6] Itai Zilberstein, Ananya Rao, Matthew Salis, and Steve Chien. 2024. Decentralized, Decomposition-Based Observation Scheduling for a Large-Scale Satellite Constellation. In 34th International Conference on Automated Planning and Scheduling
[7] F. Hochard, C. Guettier, J. Turi, W. Ajour 2024. A Distributed Modelling Approach for Solving Constellation Observations Problems at the Edge. 1st DASS Workshop - IEEE SMC-IT_SCC 2024

DIAGRAM & TECHNICAL SPEC:

The observation that the satellite must perform are defined as a set of observation requests defined by a minimal utility to achieve, and a deadline. The goal of the scheduler is to find an allocation that achieve the best observation schedule in an arbitrary receding horizon.
In addition to that, the proposed solution must optimize the observation schedule by optimizing one or several of the following (but not limited to) criterion, at the constellation level: makespan, robustness, overall utility, resilience. The scheduler must deal with additional constraints on the capability of each spacecraft to perform the observation requests.

REQUISITES:

Mandatories:

· Performances will be evaluated using a problem generator provided by SAFRAN Electronics and Defense.
· The proposed solution must not exhibit single point of failure
· The proposer must propose an algorithm with an objective function to optimise

EXPECTED RESULT (target KPIs):

· Time-utility performances: the required time utility functions will be provided by SAFRAN Electronics and Defense
· Scalability: the performances will be assessed depending on the number of observation requests and number of spacecraft. A communication model will also be provided by SAFRAN Electronics and Defence

The dAIEdge-VLab Description

The **dAIEdge-VLab** aims to implement a collaborative platform that enables Researchers and Developers to conduct experiments and research across various edge AI domains and edge AI devices. This will be achieved by sharing resources on a distributed virtual lab.

The dAIEdge-VLab is designed to help users who lack expertise in embedded programming or do not have direct access to specific hardware. It will enable them to conduct real-time AI experiments on a remote farm of embedded boards. Currently, it allows the launch of AI models benchmarking using randomly generated data.

In terms of hardware compatibility, the dAIEdge-VLab is built to support a wide range of embedded boards, spanning from high-performance MPUs and GPGPUs to energy-efficient MCUs and specialized NPUs. On the software side, it accommodates both Linux-based and real-time operating systems, as well as bare-metal solutions. Additionally, the platform will be compatible with widely used vendor-agnostic inference runtimes along with proprietary AI engines.

A virtual lab for online benchmarking of edge AI applications typically consists of several architectural components designed to facilitate remote experimentation, testing, and optimization of AI models across different edge hardware platforms. Below is an outline of its overall structure together with the main functionalities:

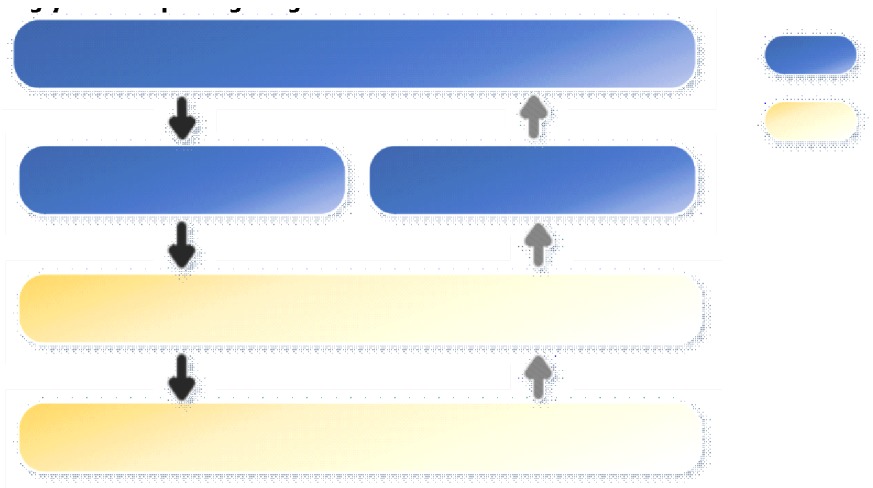


Figure 1. dAIEdge-VLab five-layer model

1. User Interface (UI) Layer

- **Web Interface/Dashboard**: The virtual lab provides a user-friendly web-based dashboard that allows data scientists and developers to interact with the system. This includes uploading AI models, selecting target hardware platforms, configuring benchmarking parameters, and monitoring results.
- **CLI API**: An API-based interface allows a seamless integration with custom AI pipelines.
- **AI Application Management**: Users can upload pre-trained AI models, select from a model zoo, or choose a complete AI application. The system also supports model versioning, enabling easy retrieval and comparison of different versions.

2. Orchestration Layer

- **Resource Management and Scheduling**: This component handles the orchestration of resources across multiple platforms, ensuring efficient allocation of hardware for testing and benchmarking. It manages queues, schedules jobs, and optimizes resource usage based on platform availability.
- **Virtualization and Containerization**: Models and benchmarking tasks are often containerized (e.g., using Docker) to ensure compatibility across diverse hardware. This enables easy deployment across various edge devices, regardless of their underlying operating system or architecture. This layer deals with the management of containers on edge devices running general purpose operating systems.
- **Benchmark Configuration**: Users define test cases, including hardware configurations, datasets, and performance metrics such as latency, power consumption, and accuracy. The ultimate goal is to define the technical requirements to compare benchmarking results among different configurations.

3. Edge Device Layer (Board farm)

- **Multi-Platform Hardware Pool**: The system integrates a variety of edge devices, ranging from low-power IoT devices (like STM32 MCUs, RISC-V based platforms, etc.) to high-performance systems (such as Raspberry Pi or Jetson Orin Nano or AGX). This diversity ensures comprehensive testing across different edge environments.
- **Remote Access & Control**: The lab allows remote access to real, physical hardware. Developers can deploy AI models directly to these devices for testing under real-world conditions, avoiding the limitations of simulated environments.

4. Data and Model Processing Layer

- **Data Preprocessing and Inference**: This layer handles the preprocessing of input data and manages the inference execution on edge devices. It ensures that the models are efficiently adapted to the target hardware's limitations, such as memory and computational power.
- **Performance Metrics Collection**: During benchmarking, this layer monitors critical performance metrics like inference time, throughput, power consumption, and memory usage, which are fed back into the system for analysis.

5. Analytics & Reporting Layer

- **Results Analysis & Visualization**: The benchmarking results are processed and presented through interactive dashboards, allowing users to compare metrics across different hardware platforms. Detailed reports include insights into energy efficiency, processing latency, accuracy, and other relevant performance factors.

The proposed dAIEdge-VLab platform features a modular architecture that allows for easy integration of remote embedded boards.

Figure 2 illustrates the overall architecture of the dAIEdge-VLab. The key components of this architecture are:

- **Remote User**: A user initiating an AI experiment on a remote node submits a request via a web-based user interface.
- **dAIEdge-VLab Server**: The dAIEdge-VLab Server receives the user request and forwards it to the corresponding Remote Host that is connected to the target Remote Node. All available Remote Nodes are registered to the dAIEdge-VLab Server.
- **Remote Host**: Located at the facility of the Remote Node owner, the Remote Host is responsible for executing node-specific scripts. It compiles and deploys the AI application to the Remote Node, retrieves the benchmarking results, and sends them back to the user through the dAIEdge-VLab Server.

- **Remote Node**: This refers to the embedded board where the AI experiments, such as model benchmarking, are executed. A single Remote Host can manage multiple Remote Nodes at the owner's site.

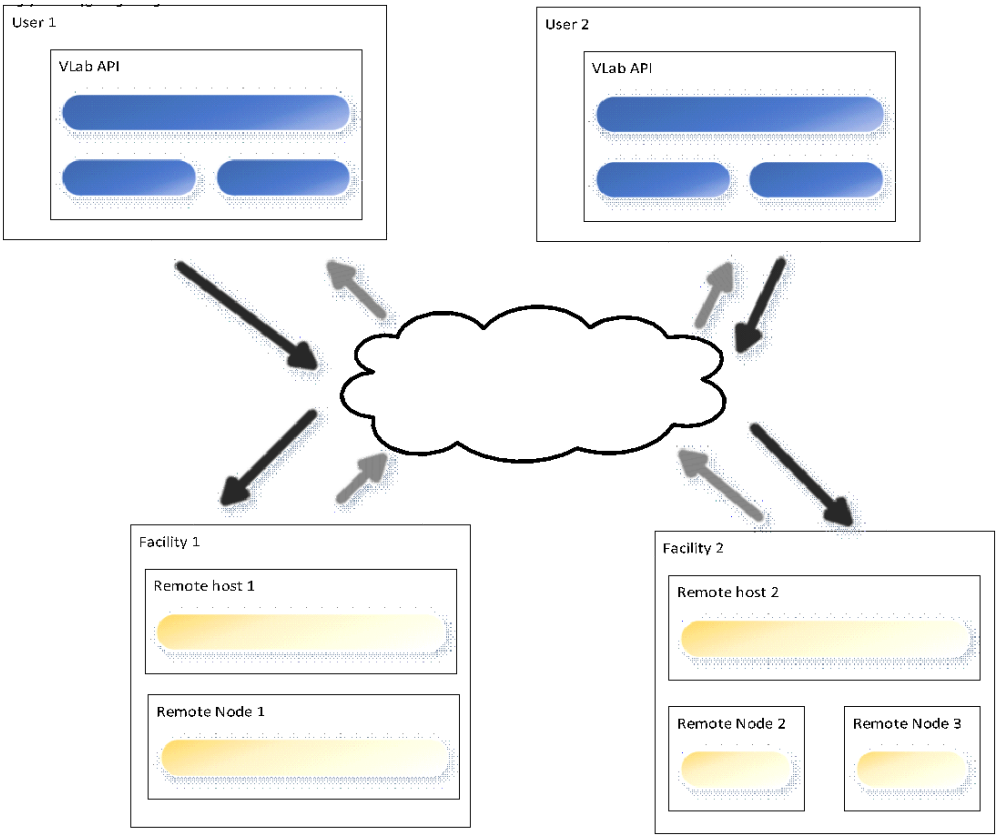


Figure 2. dAIEdge-VLab Architecture

Below are screenshots showcasing the web interface dashboards for both input selection (Figure 3) and visualization of benchmarking results (Figure 5). A history of previously executed experiments (Figure 4) is saved locally in the browser cache, allowing users to review and compare different benchmarking experiments in the future.

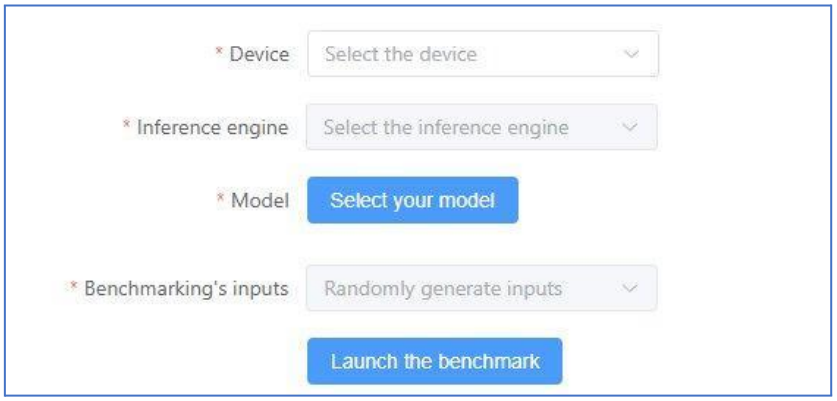


Figure 3. User Input Selection through Web Interface

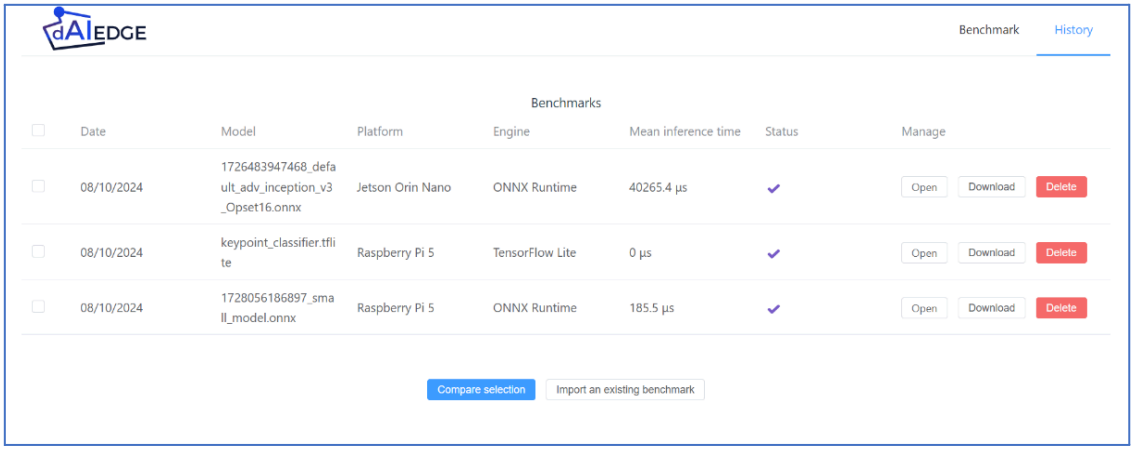


Figure 4. List of Benchmarking Results

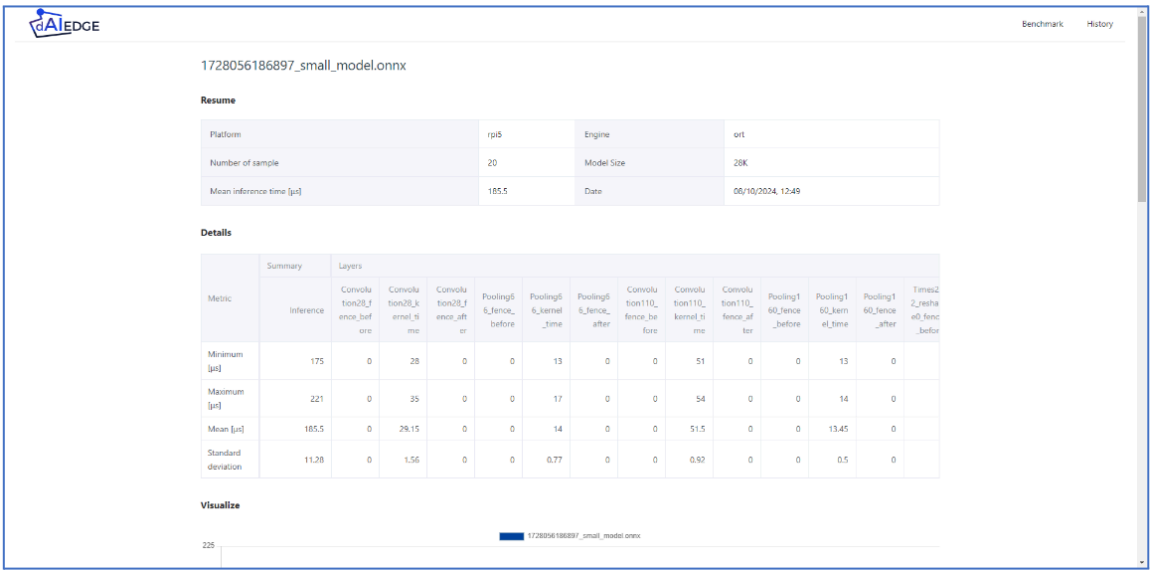


Figure 5. Visualisation of Model Benchmarking Results

The current implementation of **dAIEdge-VLab** supports a wide range of Linux-based MPUs, including Raspberry Pi 4B and Raspberry Pi 5, as well as GPGPUs such as Nvidia Jetson Xavier and Nvidia Jetson Orin Nano. It also accommodates MCU+NPU platforms like the STM32MP257, along with bare-metal MCUs such as STM32L4R9 and NXP LPC55S69. Additionally, it includes support for vendor-agnostic Machine Learning runtimes, such as ONNX Runtime, TensorFlow Lite, and TensorFlow Lite for microcontrollers, as well as vendor-specific Machine Learning runtimes like CUBE-AI and TensorRT.

dAIEdge-VLab currently assists users who may not have expertise in embedded programming or access to the target embedded board by enabling real-time model benchmarking on a remote embedded board (see above the list of supported boards).