SustainML

Application Aware, Life-Cycle Oriented
Model-Hardware Co-Design Framework
for Sustainable, Energy Efficient ML Systems

Working prototype of a ML-exploration tool and evaluation (Initial)

Deliverable D4.4

WP4 - Interaction and User Studies

This project has received funding from the European Union's Horizon Europe research and innovation programme (HORIZON-CL4-2021-HUMAN-01) under grant agreement No 101070408

eProsima The Middleware Experts

Project

Title:	SustainML: Application Aware, Life-Cycle Oriented Model-Hardware Co-Design Framework for Sustainable, Energy Efficient ML Systems
Acronym:	SustainML
Coordinator:	eProsima
Grant agreement ID:	101070408
Call:	HORIZON-CL4-2021-HUMAN-01
Program:	Horizon Europe
Start:	01 October 2022
Duration:	36 months
Website:	https://sustainml.eu
E-mail:	sustainml@eprosima.com
Consortium:	**eProsima (EPROS)**, Spain **DFKI**, Germany **Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau (RPTU)**, Germany **University of Copenhagen (KU)**, Denmark **National Institute for Research in Digital Science and Technology (INRIA)**, France **IBM Research GmbH**, Switzerland **UPMEM**, France

Deliverable

Number:	**D4.4**
Title:	**Working prototype of a ML-exploration tool and evaluation (Initial)**
Month:	24
Work Package:	WP4 - Interaction and User Studies
Work Package leader:	Inria
Deliverable leader:	Inria
Deliverable type:	DEM
Dissemination level:	Confidential (CO)
Date of submission:	2025-01-31
Version:	v4.0
Status:	Finished

Version history

Version	Date	Responsible	Author/Reviewer	Comments
v1.0	01-09-2024	Inria	Inria	
v2.0	15-11-2024	Inria	Inria	
v3.0	10-01-2025	Inria	Inria	
v4.0	31-01-2025	Inria	Inria	

Executive summary

SustainML project aims to develop a design framework and an associated toolkit, so-called SustainML, that will foster energy efficiency throughout the whole life-cycle of Machine Learning (ML) applications: from the design and exploration phase that includes exploratory iterations of training, testing and optimizing different system versions through the final training of the production systems (which often involves huge amounts of data, computation and epochs) and (where appropriate) continuous online re-training during deployment for the inference process. The framework will optimize the ML solutions based on the application tasks, across levels from hardware to model architecture. It will also collect both previously scattered efficiency-oriented research, as well as novel Green-AI methods. Artificial Intelligence (AI) developers from all experience levels can make use of the framework through its emphasis on human-centric interactive transparent design and functional knowledge cores, instead of the common black box and fully automated optimization approaches.

To promote energy efficiency and avoid AI waste throughout the lifecycle of ML applications, developers must be able to make environmentally conscious decisions from the start. Our examination of current model selection practices and sustainability awareness demonstrates how HCI techniques can be used to rethink the life-cycle of machine learning models by promoting sustainable practices through human-computer cooperation. This deliverable offers information about the background for evaluating such tools and the role of model selection in machine learning practice, which will be covered in subsequent deliverables.

Contents

Acronyms

AI	Artificial Intelligence.
ML	Machine Learning.

1 Introduction

Rapid advances in machine learning (ML) have enabled the development of numerous powerful interactive applications. Large language models such as ChatGPT have attracted significant attention doubling the number of weekly users to over 200 million in less than a year [1]. Public release of models such as LLama3 and API access to ChatGPT has led to widespread adoption of these tools for both research and everyday applications.

The field of Human-Computer Interaction (HCI) has adopted many of these models, creating a surge of advanced interactive tools, with a corresponding rise in research on human-artificial intelligence (AI) collaboration. However, these developments come at an increasing environmental cost [2]. The HCI and ML communities have been studying sustainability for over two decades [3, 4], but today's increased complexity and commercial use of ML models has increased awareness of this pressing issue.

However, optimizing individual algorithms is clearly inadequate in the face of the widespread adoption of ML in everyday applications such as chatbots, wearables, and personal health apps. Research shows that the inference costs of large ML models, i.e. the everyday use of LLM's or Diffusion models, quickly overgrow their training costs, thus shifting the sustainability costs further down the machine-learning lifecycle [5]. This results in an expected doubling of the energy consumption of data centers, AI, and cryptocurrency by 2026, which would result in an energy consumption equivalent to the energy consumed by all of Japan today [6]. We argue that a transition towards a more holistic view of the use of ML and AI is necessary. This includes a deeper reflection on the ML models' requirements with regard to the application domain, user needs, and the impact on the entire ML life-cycle and sustainability in general.

One of the main objectives of the SustainML project is to help developers of ML applications make more informed decisions when selecting an ML model, both in terms of suitability and sustainability.

The **initial problem definition and model selection** in the ML life-cycle plays a critical role here. During this early phase, ML developers investigate the characteristics of the problem, consider several model alternatives, and establish prioritizing criteria like interpretability and processing speed. The relevance of sustainability in this model trade-off evaluation, as well as how to enhance the reflection on alternatives in terms of sustainability and long-term environmental effects for ML model application, are still unclear.

In previous deliverables, we reported our insights from interviews with ML professionals and their considerations when searching and selecting an ML model *(Deliverable 4.1)*. In the following work, we highlighted our first interface design based on these insights and how it is placed within the SustainML pipeline *(Deliverable 4.2)*. In this deliverable, we extend this work by first presenting selected literature related to model selection and reflection in ML and HCI literature (see Section 2). We continue with the **prototype for the initial problem definition** (see Section 4) and the study we are planning to conduct (see Section 4.2). We then **introduce the rational for our model selection prototype** (see Section 5) and further outline the study we are planning to conduct (see Section 5.2).

2 Related Work

2.1 Understanding ML Model selection practice

Model selection is the first phase of the ML life-cycle in which developers translate user and model goals into ML model requirements. It is critical for the ML application as a well-chosen model to ensure that the ML application meets user expectations effectively and efficiently, setting a foundation for the overall success of the project. A sub-optimal model choice can result in unnecessary training, noisy findings, misleading conclusions, or poor performance [7].

In fact, the selection process profoundly affects the features available in the system [8], as well as its maintenance and re-usability [7].

The decision of what constitutes a suitable model is often made by the developer and relies on a variety of factors. Some developers prioritize models that have demonstrated good performance and strong outputs [9] using metrics such as accuracy and loss [10] to measure their effectiveness, depending on the specific task [11]. For others a good model is interpretable and explainable [12], one that uses less energy [13] or better fits the available data. Ultimately, the majority believe that the model's fit for the task at hand is crucial but the final decision always depends on the developer's needs, knowledge, and preferences [14]. Some researchers argue that ML model selection is ultimately exploratory and experimental rather than confirmatory [7], which means without experimental validation, developers cannot be sure a ML model is the most suitable for their goal. It often involves trial and error with training several models using different sets of hyper-parameter values on different data samples to determine the best one [15] rather than confirming the decisions with full confidence from the beginning.

One critical aspect is the adequacy of data relative to the chosen model. The lack of high-quality features can significantly reduce a model's effectiveness, resulting in poor performance regardless of the model's capabilities [8]. To address this, many researchers strongly emphasize the importance of evaluating models based on robust performance metrics, selecting those that demonstrate optimal accuracy, precision, and generalization to unseen data [11] often resulting in more complex models than necessary. Beyond performance metrics, the computational cost associated with running and training models has become a significant criterion more recently, particularly in the context of energy efficiency. This has led some practitioners to prefer models that, while potentially less complex, achieve a balance between performance and computational needs [13]. The increase of Human-AI interaction introduced consideration that ML models should not just be efficient and high-performing, but also interpretable and explainable [12]. Explainability is increasingly recognized as essential, especially in high-risk domains where understanding the decision-making process of a model and its results is crucial [16].

Researchers have explored a variety of strategies to support model selection ranging from traditional statistical methods [17] to more advanced techniques like automated machine learning (AutoML) [18]. AutoML, a system that automates the selection and optimization of ML algorithms, has transformed model selection and hyper-parameter tuning, allowing these tasks to be automated without direct human oversight [19]. While AutoML offers efficiency and can simplify the model selection process, it has sparked debate among researchers as it frequently ignores the understanding and intuition that humans offer to the process often resulting in rather standard and too complex model suggestions. They suggest a ”human-in-the-loop” approach where interactive, iterative model selection benefits from human prioritizing and context knowledge about the application domain and end users could improve model results [20].

There is a significant gap in understanding how ML experts resolve trade-offs including model accuracy against interpretability, balancing computational efficiency and resource consumption, and how well a model fits with the application, requirements, and limits. Understanding the role of human knowledge about context, needs, and preferences in the model selection procedure can, however, help to develop better model selection approaches that take advantage of this expertise.

2.2 The role of ML models selection in sustainable ML

While ML researchers have discussed related sustainability issues since a few decades [21, 22], the increase of complex ML systems used daily has created awareness and concern within and outside the community [4].

Recent work defines sustainable ML as “a movement to foster change in the entire life-cycle of AI products (i.e. idea generation, training, re-tuning, implementation, governance) towards greater ecological integrity and social justice” [23] It represents a combination of its environmental footprint combined with ethical concerns [24], model reliability, and considerations of longevity [25]. This includes reducing model sizes [26], optimized use of hardware, as well as improvements in data protection and accessibility to maintain privacy and security while decreasing energy consumption [27].

Most work in the sustainable ML community though focuses on the optimization of resource-intensive activities [25, 22] and the development of new models that consume less energy during training while achieving equivalent results [28]. It includes addressing trade-offs between accuracy and energy consumption in ML proposing novel sustainability indicator that allows assessing the feasibility of ML models early on [29], other showing how known approaches like transfer learning can be used to reduce overall energy consumption with comparable results [30].

Research shows that the inference costs of large ML models, e.g. ChatGPT and Dall•E, can quickly overgrow their training costs shifting the sustainability costs further down the ML life cycle. A recent 2024 study that evaluated the impact of ML model type on energy consumption concluded that "using multi-purpose models for discriminative tasks is more energy-intensive compared to task-specific models for these same tasks" [5]. The authors highlight how this contradicts the current trend of using complex, readily available multi-purpose models such as ChatGPT as a one-for-all model and warn about the potential consequences of such trends. At the same time, a recent comparative analysis started quantifying the impact of LLMs on sustainability compared to human consumption. The authors highlight that the environmental and economic impacts of using LLMs to create content of a set text length could outperform human labor in terms of energy consumption, water usage, carbon emissions, and economic costs [31]. However, these estimates do not consider the cost of prompt engineering and contextualization, nor the need, relevance, or usefulness of a multi-purpose model for the results. This trend of using complex models leads to worrying forecasts of energy consumption of data centers, AI, and cryptocurrencies by the International Energy Agency, which implies that within three years consumption might exceed energy demands similar to Japan today [6].

To quantify this trend and help to increase awareness, recent tools such as CodeCarbon [32], a lightweight software package, calculate the carbon dioxide (CO_2) emissions caused by the cloud or personal computing resources needed to run code. Also, Tracarbon[33], a Python library, and Green Algorithms [34] measure device energy use and calculate carbon emissions to enable users to estimate and report the carbon footprint.

In similar efforts, the Green Web Foundation [35], not only measures emissions and provides resources to identify greener hosting providers that rely on renewable energy, but also provides frameworks and data to articulate the importance of sustainable internet practices to movement leaders and policymakers. Their initiatives contribute to shaping the environmental impact of digital services in a way that emphasizes the need for systemic changes, such as encouraging data centers to adopt renewable energy sources and promoting digital equity.

Despite continuous progress in the sustainable ML community through optimization techniques and algorithmic improvements, there is a significant gap in attention regarding how ML developers consider sustainability impacts when selecting a model, as well as how to support these considerations projected across the entire ML life-cycle. Our interviews in the initial Deliverable 4.1 shed light on this process and we presented our suggested design rationales for supporting both, the problem definition as well as the final model selection, in Deliverable 4.2. In this deliverable, we outline how this impacted our prototypes further and how we intend to study our insights.

2.3 Sustainable HCI perspectives on ML model selection

Sustainable HCI started over two decades ago [3] and aims to make the interaction process between humans and machines less taxing and more efficient [36] and emphasizes the sustainability of the computational process itself when real-time feedback is involved [37]. It focused on the multidimensional impact of HCI and its environmental, economic, and societal implications [38].

Various approaches have been proposed by the community to address sustainability, beginning with working on individual behavior [39], applying persuasive technologies [40], to a more general 'Green Policy informatics' aiming to leverage a more traditional HCI skillset in addressing sustainability issues [41]. Recent work also aimed to increase awareness and provide guidance towards more sustainable behavior in a variety of domains. This includes establishing eco-design standards that, for example, highlight the use of renewable and biodegradable materials [42], as well as reflections on how the field of HCI can support sustainable innovations [43]. Others have addressed how incorporating collaborative modeling into the design process can help system designers to reduce rebound effects and promote more sustainable HCI practices overall [44].

Especially early work on individual behavior and persuasive systems for sustainability considered users as rational consumers [41] whom technology can (and should) support to make more sustainable decisions. Methodologies commonly used in HCI, such as rapid prototyping, user involvement, and iterative design, can rather allow for early and increasing feedback when developing interactive systems and increase awareness regarding sustainability impacts [45]. In a recent review of existing work on sustainable HCI and its critiques, the authors concluded that the HCI community "should challenge the narrative that we can rely on technology to save us, just as we challenged the narrative that climate change is an individual behavior change problem" [41].

In the context of ML, the impact of such work is reflected in the INTRPRT guidelines, which recommend formative user research and human-centered design principles to understand user demands and domain requirements to improve transparency in ML models [46].

It is also seen in the increased discussion and calls within the HCI community on more responsible, transparent, accessible, and safe ML models [47] which is echoed within the ML community and resulted in a requirement for high-risk applications in the recent European AI Act [16]. Efforts in HCI include making explainable AI more accessible using tangible and embodied interactions for improving the understanding of ML models [48] or a new technique that explains the predictions of classifier in an interpretable and more faithful manner [49].

While there is an increasing awareness of the need for explainability, interpretability, and safety of ML models for interactive systems, there is unfortunately little work on supporting more critical reflection on the necessity and scale of technology embedded in them, as well as how to support more sustainable decision making for developers of and with ML models.

3 SustainML platform from a User's Perspective

In order to contextualize the integration of the interfaces, we show an overview of the pipeline in Fig. 1 again. We distinguish between two phases of user interaction: 1) Supporting the definition of ML model parameters, which are then send to the pipeline (see Section 4), and 2) Supporting the selection of ML models, which are retrieved from the pipeline (see Section'5). While the ultimate goal is to combine both interfaces, we design and evaluate them independently. As it is planned for this initial deliverable (*Deliverable D4.4 - Working prototype of a ML-exploration tool and evaluation (Initial)*), we will in the following sections first introduce the design of the prototypes, by describing a user walk-through and then introduce the study protocol that we aim to conduct in the following months. The results will be discussed in the following deliverable.

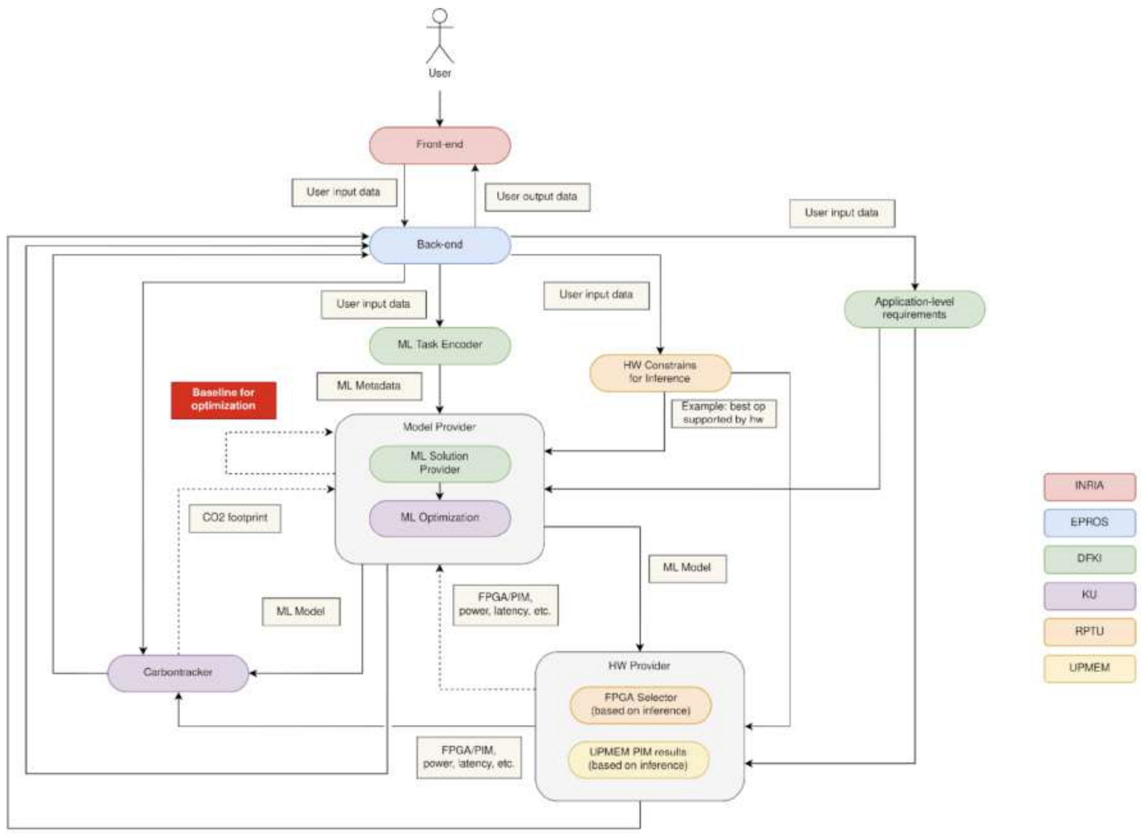

Figure 1: SustainML platform interconnections

4 Supporting the definition of ML model

The initial interface supports users to define their initial model requirements. The user interface needs to input information to the Model Provider to evaluate the current task, the application-level requirements to find the best hardware, and finally to the carbon tracker to identify the source of energy used by the model training and application, providing insights beyond energy consumption.

The user interface will adapt to the information needs of the ML task encoder using the knowledge graph provided by DFKI. To be more specific, depending on the project goal, such as classification or prediction, different information is required to make an appropriate recommendation. The interface, which includes suggestions, alternatives, and explanations, will be tailored to these requirements.

4.1 Application Walk-through

Bob is a developer for a finance company and has to identify customer groups based on their recent money transaction behavior. As the company aims to reduce its overall CO_2 footprint throughout their processes. He enters the SustainML interface to get recommendations for suitable models and their CO_2 footprint. He sees an overview of the information required by the system to suggest suitable models. He starts with the data that he has received from the operations department. He uploads a subset of the available data into the system by dragging the CSV file into the interface. He sees that the table next to it starts filling with the information that he received. As he is not yet familiar with the data, he first selects the columns he thinks are relevant for the classification by ticking the checkboxes in the head column. The system highlights that some values are missing and some outliers were detected. He clicks on the more option setting suggested to him to automatically identify the missing value rows and remove them from the dataset. He also looks at the detected outlier but decides that these might be relevant for the later classification. He returns to the main interface and checks that all extracted metadata are correct. He corrected the number of data points as he only uploaded a subset of the available data.

After completing the data part, he sees that parts of the goal and domain sections are already pre-filled and highlighted. He specifies that the problem definition is the classification and checks that the problem and data type correspond with his selected information. He also sees the option to specify the target domain, and the help box attached explains that this could help to suggest models commonly used in the target domain and could help to follow state-of-the-art and domain-specific regulations. He specifies the domain – finance – and moves on to the final part regarding the environment. He specifies the country he is connected from as well, and due to company requirements, the final model has to run locally.

After Bob fills out this information, the interface estimates the CO_2 footprint of each of the three decision domains. He explores alternatives that he identifies as reasonable and observes their impact on the CO_2 numbers. When he finalized the selection he clicked in the central button to receive model suggestions.

The interface updates and three models are suggested, including parameter and system environment details along with the predicted carbon emission. Bob chooses the model with the least emissions and directly downloads it from Huggingface by clicking on it.

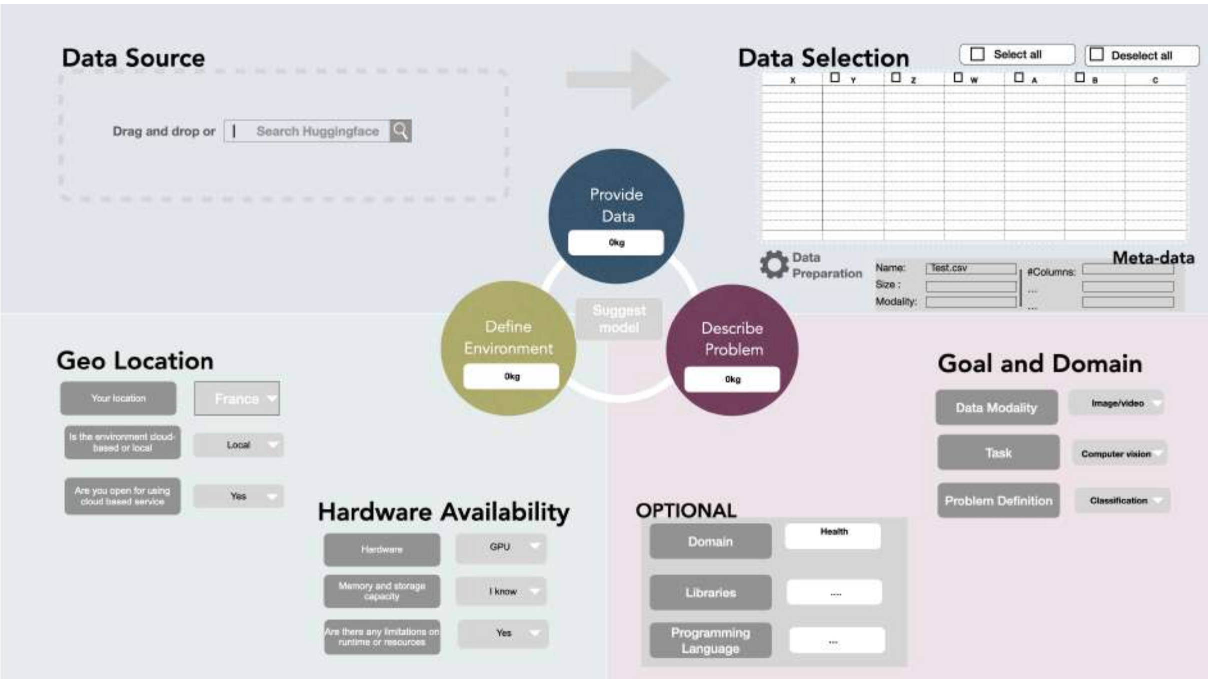

Figure 2: Overview of Design of the interface

4.2 Evaluation study

We plan to conduct an experiment, using a within-participant design, to investigate the potential of our initial guided interface regarding it's ability to: a) support ML developers make more sustainable decisions; b) enable ML developers to better understand the aspects that impact the CO_2 emission; and c) support model selection efficiency.

We will compare three conditions, in which the participants receive the task to identify the most suitable model with the least CO_2 impact.

These three conditions compare the following tool:

1. BASELINE Huggingface website
2. LLM LLM-based interface developed by DFKI
3. GUIDED Guided interface presented in this deliverable

4.3 Procedure

The experiment will last approximately one hour *with at least 12 participants who are experienced ML developers.* Participants will be asked to follow a talk aloud protocol, where they describe their intentions and actions as they explore the potential models. The session is organized into three main tasks, each representing a different interaction approach. At the end of each task, participants answer questions and fill out a short Likert-type questionnaire about their experiences with each task. We will further collect additional information such as time to complete the task, used energy during the exploration and average final model consumption. For the LLM condition we will also record the rate of hallucinations.

Tasks: Participants begin with an introduction session where they sign the consent form and are informed about the goal of the study. They fill out a pre-questionnaire to identify their level of knowledge regarding ML model selection and sustainability. They then perform Task 1, which acts as the baseline in which participants aim find suitable model for the task using the Huggingface website. Tasks 2 and 3 are counterbalanced for order across participants. Task 2 presents an LLM interface developed by DFKI within the SustainML project, that allows users to explain in text their model needs, whereas task 3 provides our guided interface with a comparable task.

Conditions: We provide comparable tasks for each condition. Participants in the *baseline* condition use the official Huggingface website. As both the LLM and GUIDED-conditions are trained on the Huggingface database, it will alow us to compare them more realistically. Participants in the *LLM* condition use the LLM interface developed by DFKI. The LLM then queries a graph-database based on Huggingface to identify a suitable model and outputs the most suitable model based on this evaluation. In the *guided* condition, participants receive a comparable task and are encouraged to define and explore the presented interface as described in the walk-through above. The provided information are then also queried to the Huggingface database allowing us to provide comparable models for our final evaluation.

4.3.1 Data collection

We will collect both quantitative and qualitative data.

Quantitative data:
We will record log data from the application during each condition, including the time to complete the task, used energy during the exploration and average final model consumption. This will allow us to compare the effect of the different interfaces on the sustainability and efficiency of the decision making process.

Qualitative data:
We will audio record participants' descriptions of their thoughts and actions during the tasks. We will also collect their responses to the questionnaire and open-ended interview questions about their experiences with each condition, their level of control and their satisfaction with their results.

5 Supporting the selection of ML model

As the goal of the SustainML platform is to avoid AI-waste overall, we intend to provide as much information and decision support before a model is trained and eventually even data are collected. For this reason the user interface will operate based on estimated information we receive from the connected modules, which will be updated over time allowing to provide pre-training information without the need for running full model simulations.

Once the user has provided a thorough project description, the platform will forward the user data to the relevant stakeholders. Their findings will provide a range of recommendations derived from realistic simulations, which can then be presented to the developer for a better-informed final decision.

5.1 Application Walk-through

Bob is a developer for a financial organization. He is responsible for identifying consumer groups based on how they have behaved in their previous money transactions. He uses the new SustainML platform as the corporation aims to decrease its total CO_2 emissions across all of its operations.

As a result, he inputs the necessary information into the SustainML platform and receives suggestions for an appropriate model and its CO_2 footprint. When he goes to the comparison page, he sees his anticipated model preferences in the bottom left corner. He raises the bar for accuracy while reducing training time. He further specifies that the model training must only take place during his working hours. Above the priority section, there is a summary of all the possible models that match the description he provided. The models are shown in several dimensions, such as ecological effects or requests per day. He chooses the more explainable and interpretable models in the chart since his colleagues should utilize them to support restructuring procedures, by selecting the upper area of the chart. The model description section is updated, and Bob chooses the model that has the smallest ecological impact, which is shown by a number in a green box, which is highlighted in red. He then uses the link to Huggingface to download the model.

5.2 Evaluation of ML model selection

We plan to conduct an experiment, using a within-participant design, to investigate the potential of our adaptive selection interface regarding it's ability to: a) support ML developers make more sustainable decisions; b) enable ML developers to better understand the aspects that impact the CO_2 emission; and c) support sustainable model selection in the long term.

We will compare two conditions, in which the participants receive the task to identify the most suitable model with the least CO_2 impact.

These two conditions compare the following tool:

1. BASELINE Huggingface website
2. ADAPTIVE interface presented in this deliverable

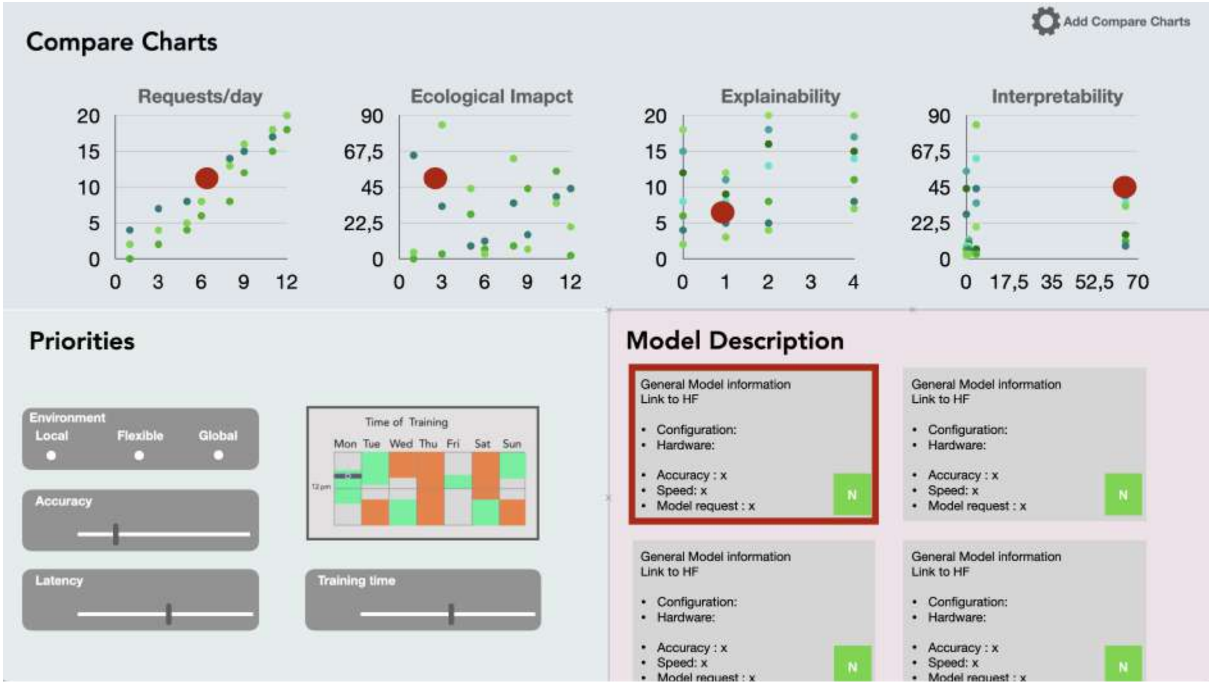

Figure 3: Overview of Design of the Compare interface

5.3 Procedure

The initial experiment will last approximately one hour *with at least 12 participants who are experienced ML developers*. Participants will be asked to follow a talk aloud protocol, where they describe their intentions and actions as they explore the potential models. The session is organized into two main tasks, each representing a different interaction approach. At the end of each task, participants answer questions and fill out a short Likert-type questionnaire about their experiences with each task. We will further collect additional information such as time to complete the task, and average final model consumption. We will conduct a follow-up interview and questionnaire with participants two weeks later to find out how our prototype affected their work processes and how they reflected on sustainable model choices.

Tasks: Participants begin with an introduction session where they sign the consent form and are informed about the goal of the study. They fill out a pre-questionnaire to identify their level of knowledge regarding ML model selection and sustainability. They then perform task 1, which acts as the baseline in which participants aim to find suitable model for the task using the Huggingface website. Task 2 presents our adaptive interface with a comparable task.

Conditions: We provide comparable tasks for each condition. Participants in the *baseline* condition use the official Huggingface website with preset parameters to find a suitable and sustainable model for the presented task. As the adaptive interface is also based on the Huggingface database, it will allow us to compare them more realistically. Participants in the *adaptive* condition use the new interface and are encouraged to define and explore the presented interface as described in the walk-through above.

5.3.1 Data collection

We will collect both quantitative and qualitative data.

Quantitative data:

We will record log data from the application during each condition, including the time to complete the task, the average final model consumption and the models explored. This will allow us to compare the effect of the different interfaces on the sustainability and efficiency of the decision making process.

Qualitative data:
We will audio record participants' descriptions of their thoughts and actions during the tasks. We will also collect their responses to the questionnaire and open-ended interview questions about their experiences with each condition, their level of control and their satisfaction with their results. The follow-up interview and questionnaire will further allow us to compare these evaluations on the longer term.

6 Summary and next steps

In this deliverable, we provide an update on our work in establishing the key interactive systems that will help enable more sustainable decisions during the machine learning development process. We further provided additional details on our approach to studying these interfaces.

We intend to assess these interfaces next month using a two-stage procedure. Following the findings of the initial study, we will enhance the proposed interface design, which will serve as the foundation of the SustainML pipeline for end-users. Ultimately, we will integrate both interfaces to enhance the interactivity of the SustainML system.

The findings of both research will be detailed in a forthcoming deliverable and are expected to result in an academic article.

References

[1] Juby Babu. *OpenAI says Chatgpt's weekly users have grown to 200 million.* Ed. by ArunEditor Koyyur. 2024. URL: https://www.reuters.com/technology/artificial-intelligence/openai-says-chatgpts-weekly-users-have-grown-200-million-2024-08-29/.

[2] Jörg Kienzle et al. "Toward model-driven sustainability evaluation". In: *Communications of the ACM* 63.3 (2020), pp. 80–91.

[3] Lon Åke Erni Johannes Hansson, Teresa Cerratto Pargman, and Daniel Sapiens Pargman. "A Decade of Sustainable HCI: Connecting SHCI to the Sustainable Development Goals". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445069. URL: https://doi.org/10.1145/3411764.3445069.

[4] David Rolnick et al. "Tackling climate change with machine learning". In: *ACM Computing Surveys (CSUR)* 55.2 (2022), pp. 1–96.

[5] Sasha Luccioni, Yacine Jernite, and Emma Strubell. "Power Hungry Processing: Watts Driving the Cost of AI Deployment?" In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.* FAccT '24. Rio de Janeiro, Brazil: Association for Computing Machinery, 2024, pp. 85–99. ISBN: 9798400704505. DOI: 10.1145/3630106.3658542. URL: https://doi.org/10.1145/3630106.3658542.

[6] IEA. *Electricity 2024: Analysis and forecast to 2026.* 2024. URL: https://www.iea.org/reports/electricity-2024.

[7] Jie Ding, Vahid Tarokh, and Yuhong Yang. "Model selection techniques: An overview". In: *IEEE Signal Processing Magazine* 35.6 (2018), pp. 16–34.

[8] Jiayi Wang et al. "Coresets over multiple tables for feature-rich and data-efficient machine learning". In: *Proceedings of the VLDB Endowment* 16.1 (2022), pp. 64–76.

[9] Yair Weiss and William T Freeman. "What makes a good model of natural images?" In: *2007 IEEE conference on computer vision and pattern recognition.* IEEE. 2007, pp. 1–8.

[10] Matthew Parry, A Philip Dawid, and Steffen Lauritzen. "Proper local scoring rules". In: (2012).

[11] Zezhen He and Yaron Shaposhnik. "Visualizing the Implicit Model Selection Tradeoff". In: *Journal of Artificial Intelligence Research* 76 (2023), pp. 829–881.

[12] Tor Sporsem. "Discovering Explainability Requirements in ML-Based Software". In: *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings.* 2024, pp. 170–172.

[13] Rafiullah Omar. "Energy-Efficient Development of ML-Enabled Systems: A Data-Centric Approach". In: *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI.* 2024, pp. 259–263.

[14] Jens de Hoog et al. "Selection of image classifiers for noisy images through metalearning". In: *Proceedings of the 2024 7th International Conference on Machine Vision and Applications.* 2024, pp. 92–99.

[15] Gavin C Cawley and Nicola LC Talbot. "On over-fitting in model selection and subsequent selection bias in performance evaluation". In: *The Journal of Machine Learning Research* 11 (2010), pp. 2079–2107.

[16] Council of European Union. *Council regulation (EU) no 2024/1689 – AI Act.* 2024. URL: http://data.europa.eu/eli/reg/2024/1689/oj.

[17] Maria Maddalena Barbieri and James O Berger. "Optimal predictive model selection". In: (2004).

[18] Rafael Barbudo, Sebastián Ventura, and José Raúl Romero. "Eight years of AutoML: categorisation, review and trends". In: *Knowledge and Information Systems* 65.12 (2023), pp. 5097–5149.

[19] Aniththa Umamahesan and Deepak Mukunthu Iyappan Babu. "From zero to AI hero with automated machine learning". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 2020, pp. 3495–3495.

[20] Liangde Li, Supun Nakandala, and Arun Kumar. "Intermittent human-in-the-loop model selection using cerebro: a demonstration". In: *Proceedings of the VLDB Endowment* 14.12 (2021).
[21] Randy H Katz et al. "Sustainable Computing: Informatics and Systems". In: (2010).
[22] Christoph Becker et al. "Sustainability design and software: The karlskrona manifesto". In: *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*. Vol. 2. IEEE. 2015, pp. 467–476.
[23] Aimee Van Wynsberghe. "Sustainable AI: AI for sustainability and the sustainability of AI". In: *AI and Ethics* 1.3 (2021), pp. 213–218.
[24] Carsten Felix Draschner, Hajira Jabeen, and Jens Lehmann. "Ethical and Sustainability Considerations for Knowledge Graph based Machine Learning". In: *2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. 2022, pp. 53–60. DOI: 10.1109/AIKE55402.2022.00015.
[25] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and policy considerations for deep learning in NLP". In: *arXiv preprint arXiv:1906.02243* (2019).
[26] Armin Moin et al. "Supporting AI Engineering on the IoT Edge through Model-Driven TinyML". In: *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*. 2022, pp. 884–893. DOI: 10.1109/COMPSAC54236.2022.00140.
[27] Peter Henderson et al. "Towards the systematic reporting of the energy and carbon footprints of machine learning". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 10039–10081.
[28] Roy Schwartz et al. "Green ai". In: *Communications of the ACM* 63.12 (2020), pp. 54–63.
[29] Vasileios Perifanis et al. "Towards Energy-Aware Federated Traffic Prediction for Cellular Networks". In: *2023 Eighth International Conference on Fog and Mobile Edge Computing (FMEC)*. 2023, pp. 93–100. DOI: 10.1109/FMEC59375.2023.10306017.
[30] Paul Walsh et al. "Sustainable AI in the Cloud: Exploring Machine Learning Energy Use in the Cloud". In: *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*. 2021, pp. 265–266. DOI: 10.1109/ASEW52652.2021.00058.
[31] Shaolei Ren et al. "Reconciling the contrasting narratives on the environmental impact of large language models". In: *Scientific Reports* 14.1 (2024), p. 26310.
[32] Wilson Friedler. *Track and reduce CO2 emissions from your computing*. https://codecarbon.io/. 2021.
[33] Valeye. *Tracarbon — Track your device's carbon footprint*. https://medium.com/@florian.valeye/tracarbon-track-your-devices-carbon-footprint-fb051fcc9009/. 2022.
[34] Loıc Lannelongue, Jason Grealey, and Michael Inouye. "Green algorithms: quantifying the carbon footprint of computation". In: *Advanced science* 8.12 (2021), p. 2100707.
[35] Chris Adams et al. *Green web foundation*. https://www.thegreenwebfoundation.org/about/. 2024.
[36] Todd Kulesza et al. "Why-oriented end-user debugging of naive Bayes text classification". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1.1 (2011), pp. 1–31.
[37] Jianbo Chen et al. "Learning to explain: An information-theoretic perspective on model interpretation". In: *International conference on machine learning*. PMLR. 2018, pp. 883–892.
[38] Paul Dourish. "HCI and environmental sustainability: the politics of design and the design of politics". In: *Proceedings of the 8th ACM conference on designing interactive systems*. 2010, pp. 1–10.
[39] Bran Knowles, Oliver Bates, and Maria Håkansson. "This changes sustainable HCI". In: *Proceedings of the 2018 CHI Conference on human factors in computing systems*. 2018, pp. 1–12.
[40] Carl DiSalvo, Phoebe Sengers, and Hrönn Brynjarsdóttir. "Mapping the landscape of sustainable HCI". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010, pp. 1975–1984.
[41] Christina Bremer, Bran Knowles, and Adrian Friday. "Have We Taken On Too Much?: A Critical Review of the Sustainable HCI Landscape". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–11.

[42] Sofía Guridi, Matteo Iannacchero, and Emmi Pouta. “Towards More Sustainable Interactive Textiles: A Literature Review on The Use of Biomaterials for eTextiles.” In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–19.

[43] Olivia Doggett, Kelly Bronson, and Robert Soden. “HCI research on agriculture: competing sociotechnical imaginaries, definitions, and opportunities”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–24.

[44] Laetitia Bornes, Catherine Letondal, and Rob Vingerhoeds. “Systemic Sustainable HCI: Integrating Collaborative Modeling into a Design Process to Address Rebound Effects”. In: *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 2024, pp. 3046–3062.

[45] Jon Froehlich, Leah Findlater, and James Landay. “The design of eco-feedback technology”. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010, pp. 1999–2008.

[46] Haomin Chen et al. “Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review”. In: *NPJ digital medicine* 5.1 (2022), p. 156.

[47] Albrecht Schmidt et al. “Artificial intelligence for humankind: a panel on how to create truly interactive and Human-Centered AI for the benefit of individuals and Society”. In: *IFIP Conference on Human-Computer Interaction*. Springer. 2021, pp. 335–339.

[48] Maliheh Ghajargar et al. “From” explainable ai” to” graspable ai””. In: *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*. 2021, pp. 1–4.

[49] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “” Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

