ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240569

A Comprehensive Sustainable Framework for Machine Learning and Artificial Intelligence

Roberto Pagliari[a], Peter Hill[a], Po-Yu Chen [a,b,*], Maciej Dabrowny[a], Tingsheng Tan[a] and Francois Buet-Golfouse[c,d]

[a]J.P. Morgan and Chase
[b]Imperial College London
[c]University College London
[d]Barclays

Abstract. In many applications, regulations or best practices often lead to specific requirements in machine learning relating to four key pillars: fairness, privacy, interpretability and greenhouse gas emissions. These all sit in the broader context of sustainability in AI, an emerging practical AI topic. However, although these pillars have been individually addressed by past literature, none of these works have considered all the pillars. There are inherent trade-offs between each of the pillars (for example, utility vs fairness or utility vs privacy), making it even more important to consider them together. This paper outlines a new framework for Sustainable Machine Learning. It proposes FPIG, a general AI pipeline that allows for simultaneous consideration and a better understanding of the tradeoffs between the pillars. Based on the FPIG framework, we propose a meta-learning algorithm to estimate the four key pillars given a dataset summary, model architecture, and hyperparameters before model training. This algorithm allows users to select the optimal model architecture for a given dataset and a set of user requirements on the pillars. We illustrate the trade-offs under the FPIG model on three classical datasets and demonstrate the meta-learning approach with an example of real-world datasets and models with different interpretability, showcasing how it can aid model selection.

1 Introduction

Artificial Intelligence has become an emerging tool essential for all financial sectors [37, 61, 53, 57]. However, the characterisation of AI extends beyond the realm of technology and permeates into the precincts of infrastructure [26] and ideology [44], leading to an opacity around the concept of AI [40]. This nebulous nature of AI magnifies the challenges of effectively understanding and governing it while underscoring the need for malleability and interdisciplinary dialogue in AI ethics and governance. Consequently, this discourse does not gravitate towards a rigid definition of AI; rather, it embraces its polysemous essence and explores AI as a complex system [16].

The current landscape of AI ethics frameworks [32, 39] is peppered with a proliferation of proposed principles and a conspicuous absence of uniformity across these frameworks. The United Nations Climate Change Conferences drove the initial environmental rights and climate justice movements, and Sustainable Development Goals (SDGs) [6], along with the environmental, social and corporate governance (ESG) frameworks [4]. Unfortunately, while the environmental implications of AI are gradually entering the discourse [56], the broader concept of sustainability in AI appears to be largely overlooked [34]. Recent literature only fostered a narrow vision of sustainable AI [58], neglecting the interconnected nature of various AI governance challenges. A holistic view of sustainable AI should amalgamate three intertwined pillars: social, governance and ecological, necessitating a complex systems approach [27].

Financial institutions have particular duties related to AI that must be paid close attention to. The Information Commissioner's Office (ICO) has strict guidance on AI regarding interpretability, data protection and privacy [8]. Additionally, several recent developments from significant organisations relating to AI regulations have bolstered the importance of sustainable AI's key features. For example, the European Union has proposed the AI Act [7], a European law on AI. The Bank of England has also recently updated their model risk management framework [1], outlining the expectations of the Prudential Regulation Authority (PRA) for banks' management of model risk. They indicate the need for a robust approach to model risk and discuss five key principles within their framework, from governance to validation. The particular focus on model risk mitigants indicates the importance for banks to consider factors such as the interpretability and fairness of their models as part of the model selection stage. In addition, the Financial Conduct Authority (FCA) has also recently updated its consumer duty expectations [3], raising the standards required by financial institutions from previous expectations. With AI's growing role, there is an increased expectation from the FCA for consumers to be at the forefront of model design. The EU AI act [5] has also been proposed recently, focusing on the risk of AI applications, categorising AI use into four risk levels, and imposing increased documentation and validation of models.

In this paper, we first advocate for the adoption of the principles of sustainability science to AI, analysing AI through the lens of an unsustainable system. We then propose a Fair, Private, Interpretable and Green (FPIG) framework to address the above-mentioned pillars. These four features (as illustrated in Fig 1) are tightly associated with the concept of sustainability but, to the best of our knowledge, have yet to be tackled together under a single framework.

In the FPIG framework, we first propose to integrate fairness [35] into our ML objective function to reduce the loss disparity across

[*] Corresponding Author. Email: po-yu.chen11@imperial.ac.uk.

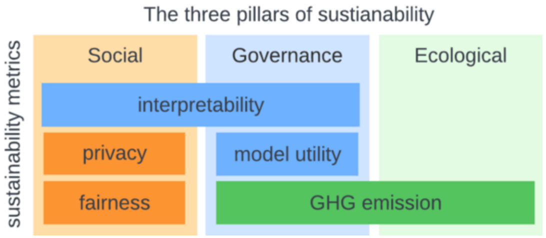


Figure 1: Privacy, fairness, utility, interpretability and GHG emission are the practical sustainability metrics that drive the three pillars of sustainability.

Model	Explainability	Tunable Hyperparameters
Linear Regression	1	Regularisation strength
Tree	1	Max depth
Random Forest	2	Number of estimators Max depth Max rows to subsample
XGBoost	2	Number of estimators Max depth Learning rate Fraction of columns to subsample Max rows to subsample L1 regularisation L2 regularisation Minimum loss reduction for partition Balance between positive and negative samples
Neural Network	3	Number of layers Layer size

Table 1: Models and associated tunable hyperparameters used in our benchmarking study. Note that we consider a subset of the hyperparameters that could be tuned; thus, this is not an exhaustive list. A self-defined measure of explainability from 1 to 3 is included, with 1 being the most and 3 least explainable.

groups. This approach allows us to change the level of fairness required in our training, varying from a standard optimisation (with no additional fairness constraints) to a multi-objective scenario, where trade-offs across different metrics form the Pareto front. Secondly, we integrate the concept of differential privacy during the model training process. Adding noise during training ensures that good models are obtained for all other dimensions across varying degrees of differential privacy.

The carbon dioxide (CO2) emission during model training and inference pipeline is tracked and monitored by an independent software package, CodeCarbon [2]. Ultimately, we propose a new meta-learning algorithm that helps users to find better AI models and hyperparameters (e.g., number of neural network layers) given a dataset and the three sustainability goals plus model interpretability.

To demonstrate the effectiveness of the FPIG framework, we evaluate it with five independent datasets and four distinctive types of machine learning (ML) models, varying in interpretability. We compute the results across the different pillars. Our evaluation reveals common trade-offs from training models using our f rameworks, such as trade-offs between utility, fairness and privacy. We also indicate the significant features when considering a meta-learning approach, allowing us to estimate the impact on utility, fairness, privacy and carbon emissions of training different models without needing to undertake the cost of actually training them. We demonstrate the usefulness of this approach on a particular example.

The rest of this paper is organised as the following. Section 2 gives a brief overview of the related work. We then present the FPIG framework in Section 3. The experiment results with five distinctive datasets are presented in Section 4, and the paper is ultimately concluded in Section 5.

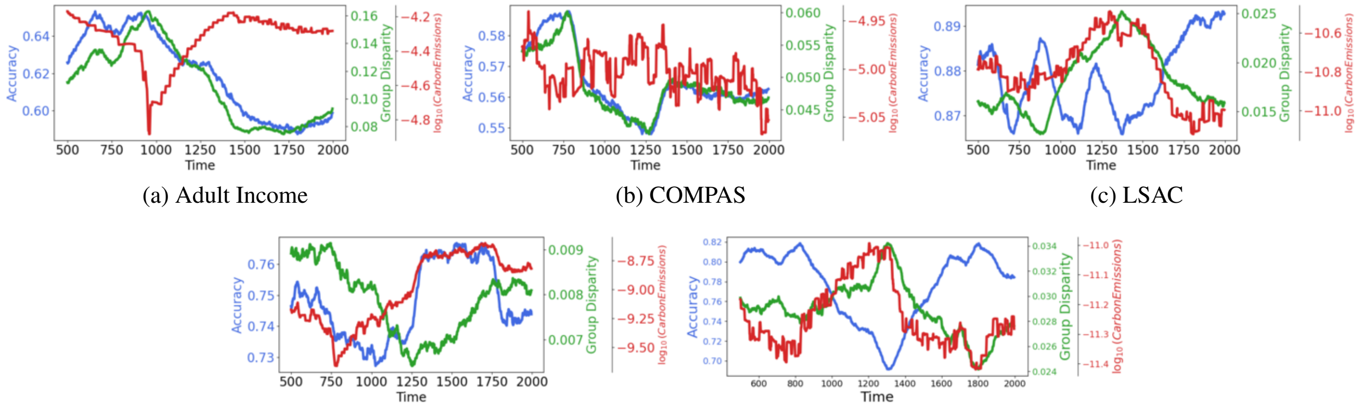
2 Sustainability in Artificial Intelligence and Machine Learning

The notion of "sustainable AI" has been proposed by a variety of researchers and practitioners [25, 54] intending to emphasise the interconnection between AI and sustainability [58]. Nevertheless, "sustainable" is frequently interpreted as synonymous with "environmentally friendly." For instance, the "Sustainable AI" manifesto issued by Facebook AI [63] is solely focused on diminishing carbon emissions from AI systems whilst vowing to "advance the field of AI in an environmentally responsible manner". This exemplifies the challenge of encouraging stakeholders to embrace a multifaceted perspective on sustainability rather than confining it merely to environmental aspects.

Fairness is one of the most crucial sustainability metrics according to SDGs. Fairness in ML models refers to the absence of bias or discrimination in the predictions and decisions made by the models [50, 59, 15]. Technically speaking, fairness involves identifying and addressing biases in the data used to train the models and the algorithms themselves. Techniques such as pre-processing the data to remove biased patterns, using specialised algorithms that explicitly consider fairness constraints during training, and employing fairness metrics to evaluate model performance can help achieve fairness in ML models [17, 18]. Recent research has defined different fairness metrics for AI [50, 59, 15]. Among these definitions, group fairness metrics such as demographic parity [19], equalised odds [35], and social fairness ensure that different groups based on a protected attribute are treated equally. The impossibility theorems on fairness [41, 23] show that these definitions cannot all be satisfied at once. There is also individual fairness and counterfactual fairness [42], which are proposed to ensure fairness at individual levels. Fairness can be considered in both supervised and unsupervised learning settings. Recent research also investigated in incorporating multiple fairness objectives into ML models given a desired level of fairness, using group functionals [17, 18].

Privacy in ML models pertains to preserving the confidentiality and security of sensitive data used for training and inference [55]. Privacy protection involves implementing mechanisms that prevent unauthorised access, use, or disclosure of personal information. Techniques like data anonymisation, encryption, and secure multi-party computation can be employed to protect privacy in ML models. Differential privacy (DP) has also been regarded as the gold standard in academia as it provides a well-defined theoretical guarantee. It can be applied to ensure that individual data points are not distinguishable, thereby safeguarding privacy while maintaining the utility of the models [9]. Adding noise during training is one way to incorporate DP into ML models. For example, Differentially private stochastic gradient descent (DP-SGD) [30] makes deep learning models differentially private by modifying the mini-batch stochastic optimisation process during gradient descent [21, 20, 38]. Other approaches incorporate DP to data synthesis [45, 60] and than train the model on the DP synthetic data, making these models more robust against DP attacks with exponentially many queries.

Interpretability in ML models refers to understanding and explaining the reasoning behind the model's predictions or decisions [51]. Interpretability techniques involve feature importance analysis, rule-based approaches, and model-agnostic techniques [28]. These techniques provide insights into the factors influencing the model's



(a) Adult Income

(b) COMPAS

(c) LSAC

(d) Loan Default

(e) Support 2

Figure 2: The moving average (we consider a rolling 500 trial period.) of Accuracy, Group Disparity, and Emissions metrics over time (number of trials), for (a) Adult income, (b) COMPAS, (c) LSAC, (d) Loan Default, (e) Support2 datasets.

output and enable humans to comprehend and validate the decision-making process. Also, layer-wise relevance propagation (LRP) or attention mechanisms can help identify relevant features or parts of the input contributing to the model's predictions. A few works investigated the trade-offs between computational efficiency and model explainability [33, 46].

Greenhouse Gas (GHG) emissions associated with ML models refer to the carbon footprint generated during the model's lifecycle, including data processing, training, inference, and deployment [36]. At a technical level, reducing GHG emissions involves optimising the computational resources used for training by employing energy-efficient hardware and algorithms. Techniques like model compression, which reduces the model's size or complexity, can also contribute to lower energy consumption during inference [22]. Additionally, adopting hardware acceleration techniques, using distributed computing, and leveraging renewable energy sources can help minimise the environmental impact of ML models [43]. [62] consider the ecological impact of AI's growth across the whole pipeline, from data to system hardware.

While extensive literature is working on improving model efficiency, leading to lower GHG emissions, explainable AI (xAI) has become an emerging area that attracts research [28, 51]. A few works investigated the trade-offs between computational efficiency and model explainability [33, 46]. Nevertheless, to our knowledge, this paper is the first to introduce a framework encompassing all four sustainability goals, including fairness, privacy, model interpretability and low GHG emissions.

3 The FPIG Framework

We propose a framework incorporating the four sustainability features into the model training pipeline. Specifically, we are optimizing along different dimensions (e.g., model performance, explainability, carbon emission and fairness with some level of privacy).

3.1 Single-Objective Optimization

Traditionally, the hyper-parameters of a machine learning model are tuned to maximize one metric of interest, for example, the area under the curve. Once the metric of interest is defined, the objective is to minimize the quantity in Eq. (1):

$$\begin{aligned} &\text{minimize} && f(\mathbf{x}) \\ &\text{subject to} && \mathbf{x} \in \mathbf{X} \end{aligned} \tag{1}$$

where $\mathbf{x} \in \mathbf{X} \subseteq \mathbb{R}^d$ is the set of d hyper-parameters, $\mathbf{X} \subseteq \mathbb{R}^d$ is the search space and $f(\cdot)$ is the objective function, for example, a loss function to be minimized.

In the case of a single-objective scenario, such as the one shown in Eq. (1), the Tree-structured Parzen Estimator (TPE), initially proposed for neural networks [14], is widely used for optimizing a wide number of machine learning models. The key idea is to separate the likelihood function $p(\mathbf{x}|y)$ in two components to identify which region the best hyper-parameters are likely to be in:

$$p(\mathbf{x}|y) = \begin{cases} l(\mathbf{x}) & \text{if } y < y^* \\ g(\mathbf{x}) & \text{if } y \geq y^* \end{cases} \tag{2}$$

where y^* is usually a quantile of the observed values y (e.g, 80%), and $l(\cdot)$ and $g(\cdot)$ are the probability density functions formed using the observations $\{\mathbf{x}^{(i)}\}$ below and above y^*, respectively. This methodology begins with several random observations $\{\mathbf{x}\}^{(i)}$ and proceeds iteratively by adding one observation at a time such that the expected improvement is maximized. As shown in [14], the expected improvement is proportional to

$$\mathrm{EI}_{y^*}(\mathbf{x}) \propto \left[y^* + \frac{g(\mathbf{x})}{l(\mathbf{x})}(1 - y^*) \right]^{-1}.$$

In other words, the aim is to sample with higher probability under $l(\mathbf{x})$ (i.e., the portion of the density with the most promising hyper-parameters) and lower probability under $g(\mathbf{x})$.

3.2 Multi-Objective Optimization

In a multi-objective scenario, we are interested in the minimization (or maximization) of many objectives that usually conflict. The optimization problem is defined as follows:

$$\begin{aligned} &\text{minimize} && f(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})) \\ &\text{subject to} && \mathbf{x} \in \mathbf{X} \end{aligned} \tag{3}$$

where, as in Eq. (1), $\mathbf{x} \in \mathbf{X} \subseteq \mathbb{R}^d$ is the set of d hyper-parameters, $\mathbf{X} \subseteq \mathbb{R}^d$ is the search space. The difference is that this time, we define a vector of cutoffs $\mathbf{Y}^* = (y_1^*, \ldots, y_n^*)$, that is, a cutoff for every objective, and the TPE estimator is generalized a follows:

$$p(\mathbf{x}|\mathbf{y}) = \begin{cases} l(\mathbf{x}) & \text{if } \mathbf{y} \succ \mathbf{Y}^* \cup \mathbf{y} \parallel \mathbf{Y}^* \\ g(\mathbf{x}) & \text{if } \mathbf{Y}^* \succeq \mathbf{y} \end{cases} \tag{4}$$

where the $\succ$, $\succeq$ and $\parallel$ operators denote dominant, weakly dominant and non-comparable relationships, respectively, as per [52]. This time, as shown in Eq. (4), the most "promising" solutions are those that dominate the cutoff $\mathbf{Y}^*$ or those that are not comparable to $\mathbf{Y}^*$. The less promising models are those that are weakly dominated by $\mathbf{Y}^*$ and, thus, contribute to forming the density function of the least "promising" hyper-parameters $g(\mathbf{x})$. Splitting the data into two sets is, in this instance, achieved via the Hype method [12]; however, the methodology is, in principle, the multi-objective equivalent of Eq. (1).

3.3 Incorporating Fairness

Several metrics exist for quantifying machine learning models' fairness and algorithmic bias. Amongst them, the most popular and often considered are equalized odds, equal opportunity, and demographic parity [49]. Without loss of generality, in our framework, we optimized for demographic parity, which is satisfied when the condition below holds true:

$$P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1).$$

In other words, the protected attribute A (e.g., sex or age) does not influence the model's outcome. In reality, demographic parity can never be exactly zero because the protected attribute A usually correlates with other features the model uses. Hence the objective is minimizing group disparity $f(\hat{Y}, A)$ as

$$f(\hat{Y}, A) = \left| P(\hat{Y}|A = 0) - P(\hat{Y}|A = 1) \right|. \tag{5}$$

3.4 Incorporating Privacy

To include privacy in our framework, we consider the concept of differential privacy [29, 30], which we use to include privacy guarantees in the model.

Definition 1. *A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies (ϵ, δ)-differential privacy if for any two adjacent inputs $d, d' \in \mathcal{D}$, and any $S \subset \mathcal{R}$ fulfil the inequality below:*

$$\mathbb{P}(\mathcal{M}(d) \in S) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(d') \in S) + \delta. \tag{6}$$

To incorporate Differential Privacy into diverse model architectures, we didn't take the common route where privacy is applied via training, such as the popular Differentially-Private Stochastic Gradient Descent (DP-SGD) [9]. Instead, we exploit the idea that converts data into differentially private synthetic data, which can be exploited by different model architectures universally. Various DP data synthesis approaches were proposed for different modalities. For example, [60] and [47] can generate DP synthetic data for tabular and images, respectively.

In this work, we adopted DPView [45], a state-of-the-art DP-aware high-dimensional data synthesis for tabular data. For a given privacy requirement ϵ, it utilises the domain size of attributes and the correlation among attributes to analytically optimise both privacy budget allocation and consistency in producing synthetic data points. Their evaluation demonstrated that the approach is versatile (when applied to tabular data from vest applications) and can effectively preserve model utilities. Compared to traditional gradient-based approaches (e.g., DP-SGD [9]) where privacy can still be breached by querying the model multiple times and the total privacy budgets are required to split across users, DPView is more robust as it does not suffer from the same issues since noises are directly applied to data instead of the model during training.

3.5 Incorporating Interpretability

In this framework, we introduce interpretability into the objective function by assigning different model types to an ordinal value to illustrate their interpretability. The lower the value, the more interpretable the model is considered to be. The objective is to minimise this value during the training process.

3.6 Evaluating GHG Emissions

We included a GHG emission tracking with CodeCarbon [48] throughout our training and inference pipeline to monitor the carbon emissions from each model training. It helps to track the overall carbon emission by accumulating the power consumption of individual hardware components and converting it into GHG emission based on the energy mixture of local power grids. At the end of the training, the overall GHG emission amount is output along with the model parameters.

4 Evaluation

This section presents an experimental analysis of an AI pipeline implementing the proposed FPIG framework. All the code is implemented in Python 3.9, and the experiments are performed on an **16 CPU** instance consisting of **64GB** memory. The evaluation is divided into two parts. Firstly, we run 2000 trials for each dataset using different models and hyperparameters using the FPIG framework. We then identify the relationships between the key metrics (accuracy, fairness, emissions) using results across all trials. Secondly, we use the results of the trials to develop a *sustainable meta learning* algorithm, aiming to learn the relationship between the meta-features of the model and dataset used, and the outputted accuracy, fairness and emissions.

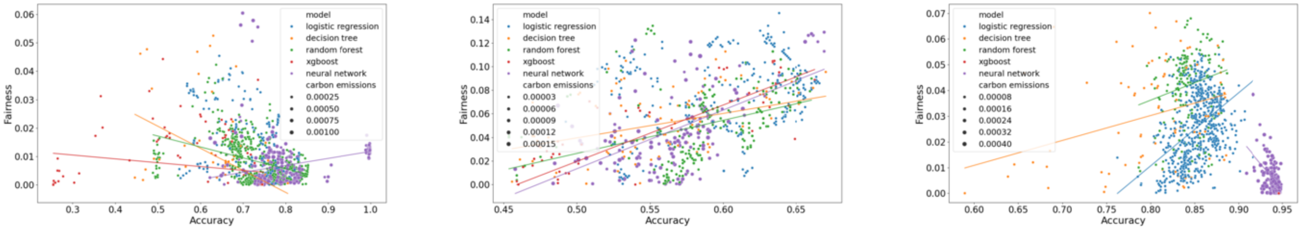
4.1 Dataset, Differential Privacy and Protected Attributes

We included the five public datasets in our evaluation:

- **Adult Income** [13] is a public multivariate *social* dataset for annual income classification (i.e., if annual income is above 50K). It comprises 48,842 records with 14 attributes such as education, occupation, and work class.
- **COMPAS Recidivism Racial Bias** [11] is a popular commercial algorithm judges and parole officers use to score criminal defendants' recidivism likelihood. This dataset compares the algorithm outputs and the ground truths, which shows that the algorithm is biased in favour of white defendants and against black inmates, based on a 2-year follow-up study.
- **LSAC** [24] is a public dataset originally collected for a 'LSAC National Longitudinal Bar Passage Study' study. It includes background information and if (and how) candidates passed the bar exam to become lawyers in the United States.

Dataset	Best Model w.r.t.	Model Architecture	Accuracy	Group Disparity	Differential Privacy	Explainability	Carbon Emissions
Adult income [13]	Accuracy	xgboost	**0.861**	0.167	10.5	2.0	3.92×10^{-6}
	Fairness	xgboost	0.767	**0.000**	10.0	2.0	3.18×10^{-6}
	Carbon Emissions	decision tree	0.726	0.446	10.5	1.0	$\mathbf{2.26 \times 10^{-6}}$
	Equal Importance	xgboost	0.767	0.000	10.0	2.0	3.05×10^{-6}
COMPAS [11]	Accuracy	decision tree	**0.671**	0.095	10.5	1.0	1.65×10^{-7}
	Fairness	xgboost	0.460	**0.000**	0.5	2.0	6.59×10^{-7}
	Carbon Emissions	decision tree	0.630	0.079	1.5	1.0	$\mathbf{7.00 \times 10^{-8}}$
	Equal Importance	logistic regression	0.614	0.007	0.5	1.0	5.02×10^{-7}
LSAC [24]	Accuracy	neural network	**0.949**	0.003	10.5	3.0	9.71×10^{-5}
	Fairness	xgboost	0.946	**0.000**	2.0	2.0	7.58×10^{-7}
	Carbon Emissions	xgboost	0.946	0.000	2.0	2.0	$\mathbf{7.58 \times 10^{-7}}$
	Equal Importance	xgboost	0.946	0.000	2.0	2.0	7.58×10^{-7}
Loan Default [65]	Accuracy	neural network	**0.999**	0.013	10.5	3.0	2.56×10^{-4}
	Fairness	decision tree	0.745	**0.000**	10.5	1.0	3.59×10^{-6}
	Carbon Emissions	decision tree	0.745	0.000	10.5	1.0	$\mathbf{3.59 \times 10^{-6}}$
	Equal Importance	random forest	0.848	0.0003	10.5	2.0	1.44×10^{-5}
Support2 [65]	Accuracy	xgboost	**0.977**	0.0259	10.5	2.0	3.01×10^{-6}
	Fairness	xgboost	0.255	**0.000**	2.5	2.0	1.64×10^{-6}
	Carbon Emissions	decision tree	0.913	0.021	10.5	1.0	$\mathbf{6.16 \times 10^{-7}}$
	Equal Importance	decision tree	0.956	0.005	10.5	1.0	7.02×10^{-7}

Table 2: The best models across different datasets, concerning different sustainability metrics. We consider accuracy, fairness, carbon emissions and an equal importance approach, as defined in Section 4.3. The results shown are on an out-of-sample test set.



(a) Loan (b) COMPAS (c) LSAC

Figure 3: Scatter plots of accuracy against group disparity for different datasets for all 2000 trials. Different models are shown in different colours, and different levels of carbon emissions are shown using different sizes.

- **Loan Default** [65] is a public multivariate *financial* dataset for loan default classification. It includes 139,202 records with 34 attributes such as income, gender, and loan purpose. Note that we randomly sampled 40,000 records from this dataset in our evaluation.
- **Support2** [31] is a public multivariate *health* dataset for predicting survival over a 180-day period for seriously ill hospitalized adults. It comprises 9,105 records and 42 attributes, such as age, sex, and follow-up days.

Each dataset is split into a training set (70%) and a test set (30%). We then apply DPView [45] to each training set and generate twenty Differentially Private (DP) synthetic data (having the same number of data points as the training sets) with different DP levels $\epsilon = [0.5, 10.0]$, where smaller ϵ indicates higher DP. Also, we only consider one protected attribute, that is, *gender*, for all five datasets. We only consider binary protected attributes (thus generating two groups in each case) and measure fairness via disparity, i.e., the absolute difference in the average loss between both groups. This is for simplicity, but our approach can easily carry over to other metrics and more complex settings.

4.2 Models and their Parameters

We built the FPIG framework using various models with varying degrees of complexity. We consider a range of hyperparameters for each model, as detailed below. Varying the model complexity and parameters will give different performance metrics and, better or worse, fairness and GHG emissions. Complex models like Neural Networks are expected to worsen fairness due to overfitting. The trade-offs between fairness and accuracy and GHG emissions should always be considered when designing new models. Table 1 illustrates our study's models and associated hyperparameters. We also include a self-defined measure of each model's explainability. This ranges from 1 to 3, with 1 being the most explainable model and 3 least explainable.

4.3 Search Space and Pareto-Front Analysis

We exploited Optuna [10], a state-of-the-art hyper-parameter optimization package, as our optimization engine to find the Pareto Fronts of the objectives defined in the FPIG framework. For each dataset, we run 2000 trials. For each trial, we select a value between 0.5 and 10 for differential privacy, where 10 indicates lower differential privacy and vice versa. We then use the respective differentially private dataset based on this value. We also include the option of no differential privacy. The objectives are listed below:

- **Fairness**: we exploited demographic parity [49] as our fairness metrics. The objective is to minimise the group disparity f in Equation (5).

- **Interpretability**: ML models are assigned an ordinal value in $\{0, 1, 2, 3\}$ to illustrate their interpretability as in Table 1, the lowest indicating the simplest models. Models range from decision logistic regression to decision trees, random forests, XGBoost, and neural networks.
- **Carbon emission**: CodeCarbon [48] is used to measure GHG emission in practical settings. The GHG emission is reported in the unit of the kilogram.
- **Accuracy**: given that all the five tasks are classification, we use classification accuracy $[0, 1]$ as the performance metric evaluating model utility.
- **Equal Importance**: To find the best models concerning accuracy, fairness and carbon emissions, we consider a naive approach for trading off each of them with equal importance. To define this, we scale the values of each metric to be between 0 and 1 to obtain $v^D_{i,j}$, the scaled value for each trial $i \in \{1, \cdots, 2000\}$ for each metric $j \in \{\text{Accuracy, Fairness, Carbon Emissions}\}$, for each dataset D. The best trial is defined to be

$$i^D = \text{argmin}_i \Big((1-v^D_{i,\text{Accuracy}})+v^D_{i,\text{Fairness}}+v^D_{i,\text{Carbon Emissions}}\Big), \tag{7}$$

as we wish to maximise Accuracy, whilst minimising Fairness and Carbon Emissions.

4.4 Optimisation under the FPIG Framework

We use Optuna to search through the hyperparameter space and find the Pareto-Frontier. Table 2 summarises the best models and their overall performances concerning the performance metrics - Accuracy, Fairness, Carbon Emissions - and their corresponding model Explinabilities and Differential Privacy levels. We further illustrate (in a rolling average of 500 trials) the trade-offs between the three performance metrics over time in Figure 2. Below, we summarise our observations.

Although the degree may vary, the trade-offs between objectives always exist. The results showed that Accuracy can always trade for Fairness across all five datasets. This observation aligns with the heuristic, where better fairness usually leads to worse model utility. However, the degree of trade-offs varies. As can be seen in Table 2, Adult Income demonstrates a small accuracy reduction from 0.861 to 0.767 when improving the group disparity from 0.167 to 0.0, whilst Support2 shows significant accuracy reductions from 0.977 to 0.255 when improving the group disparity from 0.0259 to 0.0. Similarly, we can train a model with lower Carbon emissions by trading Accuracy and Fairness. Compared to Accuracy, the trade-offs between Carbon Emission and Fairness are more significant across all five datasets. For example, when applying the model with the lowest Carbon Emission model to the Adult income dataset, the Accuracy and Group Disparity were reduced by 0.135 and increased by 0.279, respectively, compared to the optimal for Accuracy and Fairness.

Multiple objectives can be improved during model tuning. Although trade-offs (disregarding their degree) are seen across all datasets tested, as shown in Table 2, we also observed that multiple objectives could still be improved simultaneously. First, the models selected by optimising the Equal Importance (Eq. (7)) demonstrate balanced performances between all three metrics - Accuracy, Fairness and Carbon Emission, indicating that it is possible to find a suitable solution across all datasets. For example, we find a model in which the Accuracy and Fairness are reduced by only 0.057 and 0.007, respectively, with the COMPAS dataset. Second, from Figure 2 we observed that the trade-offs between objectives varies as per dataset. For example, Accuracy and Group Disparity are positively correlated (i.e., higher accuracy and lower fairness) in Adult Income and COMPAS datasets, whilst the same correlation in the LSAC and Support 2 datasets is negative. Although it is possible to improve multiple objectives simultaneously, the trajectory toward optimal (concerning the surrogate objective, such as the Equal Importance in Eq. (7)) can vary as per dataset.

Model Inputs	Dataset				
	Adult	**COMPAS**	**Loan**	**LSAC**	**Support2**
Logistic Regression	0.518	0.439	-0.348	0.300	-0.202
Decision Tree	0.472	0.383	-0.608	0.253	-0.321
Random Forest	-0.072	0.526	-0.612	0.137	-0.054
XGBoost	0.519	0.871	-0.283	-0.093	-0.188
Neural Network	0.485	0.449	0.493	-0.570	0.028

Table 3: The correlation between Accuracy and Group Disparity.

The models yielding lower Group Disparity are usually more deferentially private. Heuristically, we know that protected attributes (e.g., gender in our experiments) could potentially be utilised to identify individuals. Therefore, differential privacy (DP) could also be improved when building fair models for those protected attributes. Our experiment results shown in Table 2 confirm the above hypothesis. The best model concerning fairness always provides better DP (fulfilling smaller privacy budget ϵ) compared to other scenarios.

Simpler models are usually more explainable and carbon-friendly. This observation reinforces our heuristic regarding the trade-offs between model complexity and explainability. As seen in Table 2, the best models for Carbon Emission adopt the decision tree architecture (i.e., more explainable and easy to compute) across four datasets. In contrast, Neural Network (NN) and XGBoost, having extensive capability to approximate any continuous function with lower Explainability, achieved the best Accuracy across four datasets. Further observations regarding the impact of model architecture will be presented in the following subsection.

4.5 The Impact of Model Architecture

The choice of model architecture also plays a significant role when searching for better solutions under the FPIG framework. The best model architecture varies significantly across datasets. We further see this in Figure 3, where we compare accuracy and fairness for the three datasets - Loan Default, COMPAS and LSAC. We plot a line that best fits each model type and dataset. As can be seen, the trade-offs not only vary between datasets but also differ between model architectures. The correlations between Accuracy and Group Disparity when applying different model architectures across the five datasets shown in Table 3 also support this observation. In the COMPAS dataset, we see similar behaviour across all model types. As we increase Accuracy, this must come at the expense of Fairness, as Group Disparity also increases. This is shown by the positive gradient of the lines of best fit in Figure 3 and the positive correlations in Table 3. In contrast, the correlation becomes negative when applying NNs to the LSAC dataset, which means that the NN models could improve Accuracy and Group Disparity simultaneously compared to other model architectures. This difference is less obvious in Loan Default and COMPAS datasets.

4.6 Sustainable Meta Learning

In the previous subsection, we studied the trade-offs between sustainable objectives and realised that the preferred hyperparame-

Model Inputs	Sustainability Feature Coefficients		
	Accuracy	Group Disp.	GHG
No DP Applied	1.118	0.843	-0.153
Classifier NN (y/n)	0.537	-0.200	1.902
DP 5.5 (y/n)	0.446	0.008	0.008
DP 7.0 (y/n)	0.402	0.215	-0.018
DP 10.0 (y/n)	0.326	0.180	0.091
Dataset Column Number	0.322	-0.416	0.006
DP 3.0 (y/n)	0.306	0.326	0.018
# of categorical features	0.069	-0.413	0.071
# of NN layers	-0.024	0.021	0.417
Size of NN layer	-0.067	-0.006	0.102
Feature cardinality	-0.094	0.317	-0.016
Variance of target	-0.461	-0.149	0.034
E-net R^2 On Test Set	0.6967	0.3799	0.6806

Table 4: Meta-learning model: Ridge Regression ($\alpha = 1$) of accuracy, fairness (group disparity) and GHG emissions on selected features of the combined datasets. Coefficients impose the importance of each feature on the metric of interest. Accuracy and fairness are sensitive to differential privacy (lower is more differentially private), whilst GHG emissions depend on the size of the original training set and the model used.

ters and model architectures vary as per dataset. A single solution does not exist that fits all scenarios. Following the methodology in [64], we trained regression models $\mathcal{M}_i$ for each of $i \in$ [accuracy, disparity, emissions] that learn the relationship between the key objectives (i.e. accuracy, group disparity, and GHG emission), based on features of the dataset d_X (e.g., number of features, number of entries) as well as features on the model and training d_m (e.g., hyperparameters of the model architecture, number of training epochs) and finally privacy level requirements, d_p. This is based on the FPIG framework's trained models using Optuna. We aim to offer users a framework to determine which architecture and "sustainable hyperparameters" to pick given requirements before training, which is typically time-consuming and leads to extensive energy consumption and GHG emissions. This is the first step towards developing broader frameworks and attempting to define a meta-learning approach that will allow for a more automated ML system.

Table 4 summarises the results by running separate regression models against each objective of interest and showing the learnt coefficients of selected inputs. We list the most essential features in the Table. Our experiment demonstrates the relationship between dataset, model hyperparameters and sustainability features. We observe that using a less differentially private dataset increases the accuracy, thus illustrating the trade-off between accuracy and privacy previously discussed. We also note that using neural networks tends to increase accuracy, whilst using a dataset with a significant variance in the target will decrease the model's accuracy. Further, group disparity tends to decrease with more privacy. Using no differential privacy is the most significant factor in having a more extensive group disparity, with a coefficient of 0.843. Lastly, we note that using Neural Networks has the most significant impact on GHG Emissions, thus increasing with model complexity. There doesn't seem to be any significant relationship between privacy requirements and carbon emissions.

Algorithm 1 demonstrates how the meta-learning algorithms can be used in practice. This takes in a dataset X, a set of candidate model architectures $d_m^{(k)}$ for $k = 1, \cdots, K$, along with user requirements τ on the minimum accuracy, maximum disparity and maximum emissions the user is willing to accept. The algorithm then returns only the model architectures whose estimated metrics sit within the thresholds based on the meta-learning models. From this, users can select their chosen model based on preference across the metrics.

Algorithm 1 Candidate Model Evaluation

```
Require: Dataset $X$, meta-learning models $\mathcal{M}_i$ for $i \in [\text{accuracy}, \text{disparity}, \text{emissions}]$, user requirements $\tau = [\tau_{\text{acc}}, \tau_{\text{disp}}, \tau_{\text{em}}]$, privacy requirement $d_p$, candidate model architectures $d_m^{(k)}$ for $k = 1, \cdots, K$.
    Compute $d_x$ for dataset $X$
    for $k = 1, \cdots, K$ do
        Compute
            $m_{\text{acc}}^{(k)} = \mathcal{M}_{\text{accuracy}}[d_x, d_p, d_m^{(k)}]$
            $m_{\text{disp}}^{(k)} = \mathcal{M}_{\text{disparity}}[d_x, d_p, d_m^{(k)}]$
            $m_{\text{em}}^{(k)} = \mathcal{M}_{\text{emissions}}[d_x, d_p, d_m^{(k)}]$
    end for
    return $\{d_m^{(k)}: m_{\text{acc}}^{(k)} \geq \tau_{\text{acc}}, m_{\text{disp}}^{(k)} \leq \tau_{\text{disp}}, m_{\text{em}}^{(k)} \leq \tau_{\text{em}}\}$
```

5 Conclusion

This paper introduces the FPIG framework for Sustainable Machine Learning, considering a multi-objective optimisation problem involving accuracy, fairness, privacy, explainability and carbon emissions. We demonstrate this approach on five datasets and show the trade-offs between these sustainable objectives and the possibility of finding models to balance these trade-offs. We also extended these observations by building a meta-learning approach to predict these critical metrics based on the dataset and model characteristics. The results further validate our above observations and show that fairness, as one of the sustainable objectives, is more data-dependent, meaning that it is more difficult to provide guarantees by selecting suitable model architectures and corresponding hyperparameters.

References

[1] Bank of england model risk management framework. https://www.bankofengland.co.uk/prudential-regulation/publication/2023/may/model-risk-management-principles-for-banks-ss/. Accessed: 2023-07-24.

[2] Codecarbon. https://codecarbon.io/. Accessed: 2023-07-24.

[3] Fca consumer duty. https://www.fca.org.uk/firms/consumer-duty. Accessed: 2023-07-24.

[4] Environmental, social, and corporate governance. https://en.wikipedia.org/wiki/Environmental,_social,_and_corporate_governance. Accessed: 2023-07-24.

[5] Eu ai act. https://artificialintelligenceact.eu/the-act/. Accessed: 2023-07-24.

[6] United nation's 17 goals for sustainable development. https://sdgs.un.org/goals. Accessed: 2023-07-24.

[7] The eu aritificial intelligence act. https://artificialintelligenceact.eu/. Accessed: 2023-07-24.

[8] Information commissioner's office. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/. Accessed: 2023-07-24.

[9] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[10] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proc. SIGKDD 2019*.

[11] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Compas. Kaggle, 2016. URL: https://www.kaggle.com/datasets/danofer/compass.

[12] J. Bader and E. Zitzler. Hype: An algorithm for fast hypervolume-based many-objective optimization. *Evolutionary computation*, 19(1):45–76, 2011.

[13] B. Becker and R. Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[14] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

[15] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, Aug. 2018. doi: 10.1177/0049124118782533.

[16] H. Bossel. *Modeling and Simulation*. A K Peters/CRC Press, Oct. 2018. doi: 10.1201/9781315275574.

[17] F. Buet-Golfouse and I. Utyagulov. Towards fair unsupervised learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1399–1409, 2022.

[18] F. Buet-Golfouse and I. Utyagulov. Fairness trade-offs and partial debiasing. In *Asian Conference on Machine Learning*, pages 112–136. PMLR, 2023.

[19] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010. doi: 10.1007/s10618-010-0190-x.

[20] D. Chen, T. Orekondy, and M. Fritz. Gs-wgan: A gradient-sanitized approach for learning differentially private generators. *Advances in Neural Information Processing Systems*, 33:12673–12684, 2020.

[21] J.-W. Chen, C.-M. Yu, C.-C. Kao, T.-W. Pang, and C.-S. Lu. Dpgen: Differentially private generative energy-guided network for natural image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8387–8396, 2022.

[22] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani. A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 53:5113–5155, 2020.

[23] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[24] A. Clair. LSAC. Kaggle, 2022. URL: https://www.kaggle.com/code/eds8531/lsac-dataset-eda-and-predictions.

[25] M. Coeckelbergh. AI for climate: freedom, justice, and other ethical and political challenges. *AI and Ethics*, 1(1):67–72, Feb. 2021. ISSN 2730-5961. doi: 10.1007/s43681-020-00007-2. URL https://doi.org/10.1007/s43681-020-00007-2.

[26] K. Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

[27] B. J. De Vries. *Sustainability science*. Cambridge University Press, 2023.

[28] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

[29] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[30] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9 (3–4):211–407, 2014.

[31] C. J. A. F. et al. A controlled trial to improve care for seriously iii hospitalized patients. *Management Review Quarterly*, 1995.

[32] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI, Jan. 2020. URL https://papers.ssrn.com/abstract=3518482.

[33] M. Graziani, L. Dutkiewicz, D. Calvaresi, J. P. Amorim, K. Yordanova, M. Vered, R. Nair, P. H. Abreu, T. Blanke, V. Pulignano, et al. A global taxonomy of interpretable ai: unifying the terminology for the technical and social sciences. *Artificial intelligence review*, 56(4):3473–3504, 2023.

[34] T. Hagendorff. Blind spots in ai ethics. *AI and Ethics*, 2(4):851–867, 2022.

[35] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[36] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *The Journal of Machine Learning Research*, 21(1): 10039–10081, 2020.

[37] Y. Hilpisch. *Artificial Intelligence in Finance*. O'Reilly Media, 2020.

[38] H. Horigome, H. Kikuchi, and C.-M. Yu. Local differential privacy protocol for making key–value data robust against poisoning attacks. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 241–252. Springer, 2023.

[39] A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), Sept. 2019. doi: 10.1038/s42256-019-0088-2.

[40] Y. Katz. *Artificial Whiteness: Politics and Ideology in Artificial Intelligence*. Columbia University Press, Nov. 2020.

[41] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017. doi: 10.4230/LIPIcs.ITCS.2017.43.

[42] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[43] J.-P. Lai, Y.-M. Chang, C.-H. Chen, and P.-F. Pai. A survey of machine learning models in renewable energy predictions. *Applied Sciences*, 10 (17):5975, 2020.

[44] J. Lanier and G. Weyl. AI is An Ideology, Not A Technology. *Wired*, Mar. 2015. ISSN 1059-1028. URL https://www.wired.com/story/opinion-ai-is-an-ideology-not-a-technology/. Section: tags.

[45] C.-H. Lin, C.-M. Yu, and C.-Y. Huang. Dpview: Differentially private data synthesis through domain size information. *IEEE Internet of Things Journal*, 9(17):15886–15900, 2022.

[46] K. Lin and Y. Gao. Model interpretability of financial fraud detection by group shap. *Expert Systems with Applications*, 210:118354, 2022.

[47] Z. Lin, S. Gopi, J. Kulkarni, H. Nori, and S. Yekhanin. Differentially private synthetic data via foundation model APIs 1: Images. In *Proc. ICLR 2024*, 2024.

[48] K. Lottick, S. Susai, S. A. Friedler, and J. P. Wilson. Energy usage reports: Environmental awareness as part of algorithmic accountability. *arXiv preprint arXiv:1911.08354*, 2019.

[49] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[50] A. Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, FAT* 18, New York, USA, 2018.

[51] G. Novakovsky, N. Dexter, M. W. Libbrecht, W. W. Wasserman, and S. Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, 2023.

[52] Y. Ozaki, Y. Tanigaki, S. Watanabe, and M. Onishi. Multiobjective tree-structured parzen estimator for computationally expensive optimization problems. In *Proceedings of the 2020 genetic and evolutionary computation conference*, pages 533–541, 2020.

[53] H. Pallathadka, E. H. Ramirez-Asis, T. P. Loli-Poma, K. Kaliyaperumal, R. J. M. Ventayen, and M. Naved. Applications of artificial intelligence in business management, e-commerce and finance. *Materials Today: Proceedings*, 80:2610–2613, 2023.

[54] H. S. Sætra. Ai in context and the sustainable development goals: Factoring in the unsustainability of the sociotechnical system. *Sustainability*, 13(4):1738, 2021.

[55] D. J. Solove. Understanding privacy. 2008.

[56] E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

[57] S. Umamaheswari, A. Valarmathi, et al. Role of artificial intelligence in the banking sector. *Journal of Survey in Fisheries Sciences*, 10(4S): 2841–2849, 2023.

[58] A. Van Wynsberghe. Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, 1(3):213–218, 2021.

[59] S. Verma and J. Rubin. Fairness definitions explained. In *Proceedings of the International Conference on Software Engineering*, pages 1–7, New York, NY, USA, 2018. ACM. doi: 10.1145/3194770.3194776.

[60] G. Vietri, C. Archambeau, S. Aydore, W. Brown, M. Kearns, A. Roth, A. Siva, S. Tang, and S. Z. Wu. Private synthetic data for multitask learning and marginal queries. *Proc. NeurIPS 2022*.

[61] P. Weber, K. V. Carl, and O. Hinz. Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. *Management Review Quarterly*, pages 1–41, 2023.

[62] C.-J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. Aga, J. Huang, C. Bai, et al. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813, 2022.

[63] K. Wu, E. Wu, and G. Kreiman. Learning scene gist with convolutional neural networks to improve object recognition. In *Proc. CISS 2018*.

[64] C. Yang, Y. Akimoto, D. W. Kim, and M. Udell. Oboe: Collaborative filtering for automl model selection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1173–1183, 2019.

[65] M. H. Yasser. Loan default dataset. Kaggle, 2020. URL: https://www.kaggle.com/datasets/yasserh/loan-default-dataset.