VTruST: Controllable value function based subset selection for Data-Centric Trustworthy AI

Soumi Das[*12] soumidas@mpi-sws.org
Shubhadip Nag[*2] shubhadipnag5555@gmail.com
Shreyyash Sharma[*] shrysh1701@gmail.com
Suparna Bhattacharya[†] suparna.bhattacharya@hpe.com
Sourangshu Bhattacharya[*] sourangshu@cse.iitkgp.ac.in

[*]*Dept. of Computer Science and Engg.*
Indian Institute of Technology, Kharagpur,
Kharagpur, WB, India -721302.

[†]*AI Research Lab,*
Hewlett Packard Labs,
Bengaluru, KA, India - 560048.

Reviewed on OpenReview: https://openreview.net/forum?id=cAc3xKxYX9

Abstract

Trustworthy AI is crucial to the widespread adoption of AI in high-stakes applications with *fairness, robustness, and accuracy* being some of the key trustworthiness metrics. In this work, we propose a controllable framework for data-centric trustworthy AI (DCTAI)- VTruST, that allows users to control the trade-offs between the different trustworthiness metrics of the constructed training datasets. A key challenge in implementing an efficient DCTAI framework is to design an online value-function-based training data subset selection algorithm. We pose the training data valuation and subset selection problem as an online sparse approximation formulation. We propose a novel online version of the Orthogonal Matching Pursuit (OMP) algorithm for solving this problem. Experimental results show that VTruST outperforms the state-of-the-art baselines on social, image, and scientific datasets. We also show that the data values generated by VTruST can provide effective data-centric explanations for different trustworthiness metrics.

1 Introduction

Trustworthiness (Kaur et al., 2022; Li et al., 2023a) of predictions made by Machine Learning models is crucial in many applications. In applications impacting society, e.g. loan eligibility prediction (Hardt et al., 2016), criminal recidivism risk prediction (Angwin et al., 2016), etc, fairness in prediction across different marginalized groups is as important as overall prediction accuracy. Similarly, the robustness of object detection systems for autonomous driving against perturbed input images(Song et al., 2024), or robustness against label corruption in phase transition prediction of sub-atomic particles (Benato et al., 2022) are important metrics compared to overall prediction accuracy. Tradeoffs between various notions of fairness with accuracy, e.g. individual fairness (demographic parity/equalized odds) (Roh et al., 2020; Romano et al., 2020) or group fairness (Accurate Fairness) (Li et al., 2023b) are being studied for different models and training mechanisms. Similar studies have also been reported on inherent tradeoffs between feature robustness in im-

1. Currently at: MPI-SWS, Saarbruecken, Germany.
2. Equal contribution.

ages (Tsipras et al., 2019; Hu et al., 2023) with accuracy, and adversarial label robustness (Pang et al., 2022; Madry et al., 2018). While inherent tradeoffs between the trustworthiness metrics such as accuracy vs fairness or robustness is generally accepted, the nature of additional bias introduced by algorithms e.g adversarial training (Madry et al., 2018) or FairBatch (Roh et al., 2020) is not clear. In this paper, we follow a data-centric approach to designing trustworthy AI techniques.

Data-centric approaches to AI model development (Zha et al., 2023) strive to design methods for creating high-quality training datasets, that when used with standard SGD-based training algorithm can lead to models with specific trustworthiness properties. This eliminates the algorithmic bias introduced by specific algorithms, while limiting the bias only to the newly created training dataset, which is easier to interpret. While many data valuation (Koh and Liang, 2017; Park et al., 2023; Ghorbani and Zou, 2019) techniques have been developed for the selection of high-quality training data subsets, most of them optimize only one property, e.g. validation set error rate. While well-known metrics for robustness and fairness exist, their use as a viable and efficient value function remains to be studied. A key research issue in designing a data-centric approach is the design of an appropriate "value function" that captures the notion of value of a training datapoint (toward trustworthiness metrics) while also being efficiently optimizable. Another important research question is: can the value functions corresponding to various trustworthiness metrics be combined into a single value function using user-defined weightage? In this paper we address these research questions, effectively leading to a general data-centric framework to achieve user-controlled tradeoffs between different trustworthiness metrics.

We propose additive value functions for accuracy, fairness, and robustness, which can be combined to form composite value functions. The additiveness of the value functions is a key property that allows us to pose the problem of training data valuation and selection as an *online sparse approximation problem*. We propose a novel *online orthogonal matching pursuit (OMP)* algorithm that greedily replaces features corresponding to a selected datapoint with those of a new datapoint, if there is a net improvement in the overall value of the selected set. Unlike the traditional OMP (Cai and Wang, 2011) which makes a pass through the entire training dataset to select an example, the proposed online OMP makes a pass through the selected datapoints (a much smaller set) at the time of training update to optionally replace an existing selected point. Experimental results on various applications demonstrate that models trained on subsets selected by VTruST can outperform all state-of-the-art baselines by $\sim 10 - 20\%$ and can also provide data-centric explanations behind its performance.

2 VTruST: Value-driven Trustworthy AI through Selection of Training Data

We propose a controllable value function-based framework for developing trustworthy models using a data-centric paradigm. Our system has two components: (1) A general value function-based framework that allows users to specify a combination of trustworthiness metrics (sections 2.1 and 2.2), and (2) a novel online subset selection algorithm for constructing high-quality training dataset based on the specified value function (section 2.3).

2.1 A Controllable Value Function-based Framework for DCTAI

Let $\mathcal{D} = \{d_i | i = 1, 2, .., N\}$ be the training dataset and $\mathcal{D}' = \{d'_j | j = 1, 2, .., M\}$ be the validation dataset. Every datapoint $d' \in \mathcal{D}'$ can be used to define the value function $\mathcal{V}(\theta, d')$ which is used for calculating the value of a model θ. Given a run of model training, we define the incremental value function $v_i^t(d'_j)$ as the decrease in loss incurred due to an SGD update (Pruthi et al., 2020) using the datapoint d_i: $v_i^t(d'_j) = l(\theta_t^{i-1}, d'_j) - l(\theta_t^i, d'_j)$, where θ_t^{i-1} and θ_t^i are the model parameters before and after the SGD update involving the training datapoint d_i in the t^{th} epoch. Hence the value of a model θ^T can be defined as: $\mathcal{V}(d'_j) = \sum_{t=1}^{T} \sum_{i=1}^{N} v_i^t(d'_j)\ \forall d' \in \mathcal{D}'$. We overload the notation to define the value function vector $\mathcal{V}(\mathcal{D}') = \sum_{t=1}^{T} \sum_{i=1}^{N} v_i^t(\mathcal{D}')$, where $v_i^t(\mathcal{D}') \in \mathbb{R}^M$ is the vector of incremental values over all validation set datapoints.

Our data-centric framework aims to find a subset of training datapoints $\mathcal{S} \in \mathcal{D}$ that leads to a high-value model θ^t after training for t-epochs. Let $\vec{y}_t = \sum_{k=1}^{t} \sum_{i=1}^{N} v_i^k(\mathcal{D}')$ be the cumulative value function till the t^{th} epoch. We formulate the training data subset selection problem as a sparse approximation: $\vec{y}_t \approx \sum_{d_i \in S \subseteq \mathcal{D}} \alpha_i [\sum_{k=1}^{t} v_i^k(\mathcal{D}')]$, where α_i are the weights for the selected training datapoint d_i. Next, using a second order Taylor series expansion of the change in loss function and plugging in the SGD update $\theta_t^i - \theta_t^{i-1} = \eta_t \nabla l(\theta_t^{i-1}, d_i)$, we obtain the following approximation for each term in the value function $l(\theta_t^i, \mathcal{D}') - l(\theta_t^{i-1}, \mathcal{D}') \approx \eta_t \nabla l(\theta_t^{i-1}, d_i)^T \nabla l(\theta_t^{i-1}, \mathcal{D}') + \mathcal{O}(||\theta_t^i - \theta_t^{i-1}||_2^2)$. We truncate the Taylor expansion till the second-order terms to arrive at the following sparse approximation problem: $\vec{y}_t \approx \sum_{d_i \in S} \alpha_i \left[\sum_{k=1}^{t} \vec{X}_i^k\right]\ \forall t = 1, ..., T$ where $\vec{X}_i^k = \nabla l(\theta_k^{i-1}, d_i)^T \nabla l(\theta_k^{i-1}, \mathcal{D}') + \frac{(\nabla l(\theta_k^{i-1}, d_i)^T \nabla l(\theta_k^{i-1}, \mathcal{D}'))^2}{2}$ are the features for the i^{th} training point calculated in epoch t. We use $\vec{y}_t$ and $\vec{X}_i^t$ to denote the predictor and predicted variables for valuating training datapoint d_i using the entire validation set $\mathcal{D}'$. The main challenge in solving this approximation problem is that we need to store all the features $\vec{X}_i^k$ for all training datapoints i over epochs $k = 1, ..., t$, in order to compute $\sum_{k=1}^{t} \vec{X}_i^k$. This becomes prohibitively expensive. Instead, we solve the following *online sparse approximation* (OSA) problem for each epoch t:

$$\vec{y}_t \approx \sum_{(p,q) \in S_t} \beta_p^q \vec{X}_p^q \tag{1}$$

Here, S_t is the set of selected training datapoints after epoch t. Note that the set S_t can contain datapoints indexed by p with features from any of the epochs $q = 1, ..., t$. We constrain the size of S_t to be less than a user-specified parameter ω. We describe an online algorithm for solving the above problem in Section 2.3. Note that the value function $\mathcal{V}(\mathcal{D}')$ only needs to be additive over the training datapoints and epochs for the above formulation to be valid. Hence, this framework applies to a composite value function $\mathcal{V}(\mathcal{D}') = \sum_f \lambda_f \mathcal{V}_f(\mathcal{D}')$, where each value function $\mathcal{V}_f(.)$ satisfies the additive property. This leads us to a general *controllable* framework for incorporating many trustworthiness value functions, controlled using the user-specified weights λ_f.

2.2 Value Functions for Trustworthy Data-centric AI

For the accuracy metric, we use the value function proposed in (Pruthi et al., 2020), which is defined as the decrease in loss incurred due to an SGD update using the datapoint d_i:

$v_i^t(d'_j) = l(\theta_t^{i-1}, d'_j) - l(\theta_t^i, d'_j)$ where θ_t^{i-1} and θ_t^i are the model parameters before and after the SGD update involving the training datapoint d_i in the t^{th} epoch. Hence, the **accuracy value function** vector in defined as $\mathcal{V}_a(\mathcal{D}') = \sum_{t=1}^{T}\sum_{i=1}^{N} v_i^t(\mathcal{D}')$.

Robustness Value Function: Training data augmentation using various perturbations has been observed to improve robust accuracy (Rebuffi et al., 2021; Addepalli et al., 2022). We use various perturbations to create the augmented training set $\mathcal{D}_a$ and validation set $\mathcal{D}'_a$ from $\mathcal{D}$ and $\mathcal{D}'$ respectively. The robustness value function is defined as $\mathcal{V}_r(\mathcal{D}'_a) = \sum_{t=1}^{T}\sum_{d_i \in \{\mathcal{D}\cup\mathcal{D}_a\}} l(\theta_t^i, \mathcal{D}'_a) - l(\theta_t^{i-1}, \mathcal{D}'_a)$. Since $\mathcal{V}_r$ is derived from the loss function (that is additive), it also follows the additive property.

Fairness Value Function: Existing literature in fairness (Roh et al., 2020; Romano et al., 2020) uses equalized odds (EO) disparity and demographic parity disparity for achieving fair models. Let $x \in \mathcal{X}$ be the input domain, $\{y_0, y_1\} \in \mathcal{Y}$ be the true binary labels, and $\{z_0, z_1\} \in \mathcal{Z}$ be the sensitive binary attributes. We define the fairness value function as the change in EO disparity: $\mathcal{V}_f(\mathcal{D}') = \sum_{t=1}^{T}\sum_{d_i \in \mathcal{D}} ed(\theta_t^i, \mathcal{D}') - ed(\theta_t^{i-1}, \mathcal{D}')$. Based on (Roh et al., 2021b), it is defined as the maximum difference in accuracy between the sensitive groups ($z \in \mathcal{Z}$) pre-conditioned on the true label ($y \in \mathcal{Y}$): $ed(\theta, \mathcal{D}') = max(\|l(\theta, \mathcal{D}'_{y_0,z_0}) - l(\theta, \mathcal{D}'_{y_0,z_1})\|, \|l(\theta, \mathcal{D}'_{y_1,z_0}) - l(\theta, \mathcal{D}'_{y_1,z_1})\|)$. Considering we have $ed(\theta, \mathcal{D}')$ defined for two validation sets $\mathcal{D}'_1$ and $\mathcal{D}'_2$ and the loss function is inherently additive, $ed(\theta, \mathcal{D}'_1) + ed(\theta, \mathcal{D}'_2) = ed(\theta, (\mathcal{D}'_1 + \mathcal{D}'_2))$ also holds true, thus $\mathcal{V}_f$ turning out to be additive.

Composite value functions: We can combine the value functions for accuracy ($\mathcal{V}_a(\mathcal{D}')$), robustness ($\mathcal{V}_r(\mathcal{D}'_a)$) and fairness ($\mathcal{V}_f(\mathcal{D}')$) to construct different composite value functions for observing tradeoffs between different trustworthiness metrics. We use the following combinations for our experiments: (a) Accuracy-Fairness : $\mathcal{V}_{af}(\mathcal{D}') = \lambda\mathcal{V}_a(\mathcal{D}') + (1-\lambda)\mathcal{V}_f(\mathcal{D}')$; (b) Accuracy-Robustness : $\mathcal{V}_{ar}(\mathcal{D}', \mathcal{D}'_a) = \lambda\mathcal{V}_a(\mathcal{D}') + (1-\lambda)\mathcal{V}_r(\mathcal{D}'_a)$; (c) Robustness-Fairness : $\mathcal{V}_{rf}(\mathcal{D}', \mathcal{D}'_a) = \lambda\mathcal{V}_r(\mathcal{D}'_a) + (1-\lambda)\mathcal{V}_f(\mathcal{D}')$. The user-defined parameter λ is used to control the tradeoff between the two objectives.

2.3 An online-OMP algorithm for online sparse approximation

Algorithm 1 : **VTruST**

Input:
i. ω : Total number of datapoints to be selected
ii. $\vec{y}$: Targeted value function
iii. $\vec{X}_i$: Features of all training points $d_i \in \mathcal{D}$
iv. S : Set of selected datapoint indices
v. $\vec{\beta} \in \mathbb{R}^{|S|}$: Weight of selected datapoints
Initialize:
$S \longleftarrow \phi$ //Indices of selected datapoints
for each epoch $t \in \{1, 2, ..., T\}$ **do**
  for each datapoint $d_i \in \mathcal{D}$ **do**
    Input: $\vec{y}_t$, $X_i^t \quad \forall i \in \{1, 2, .., N\}, ||X_i^t||_2 = 1$
    Process:
    if $|S_{t-1}| = \omega$ **then**
      $S_t \leftarrow$ **DataReplace**$(\vec{y}_t, \vec{\xi}_{t-1}, S_{t-1}, \vec{\beta}_{t-1}, \vec{X}_i^t)$
    else
      $S_t \longleftarrow S_{t-1} \cup \{i\}$ // Add datapoints till the cardinality of S_t reaches ω
    end if
    Update $\vec{\beta}_t = \text{argmin}_\beta \|\vec{y}_t - \sum_{p,q\in S_t}(\beta_q^p \vec{X}_q^p)\|_2$
    Update $\vec{\xi}_t = \sum_{p,q\in S_t} \beta_q^p \vec{X}_q^p$
  end for
end for
Output:Final set of selected datapoint indices S_T, learned coefficients $\{\beta_q^p | p, q \in S_T\}$

Algorithm 2 :**DataReplace**($\vec{y}_t, \vec{\xi}_{t-1}, S_{t-1}, \vec{\beta}_{t-1}, \vec{X}_i^t$) - *Replace an existing datapoint.*

$\vec{\rho}_t = \vec{y}_t - \vec{\xi}_{t-1}$
π_{max} = -∞
$(a, b) = \phi$
$\pi \longleftarrow \text{abs}(\vec{X}_i^t.\vec{\rho}_t)$
for each index $p, q \in S_{t-1}$ **do**
  $\pi' \longleftarrow \text{abs}(\vec{X}_q^p.\vec{\rho}_t)$
  $\gamma \longleftarrow \beta_q^p$
  if $\pi > \pi'$ & $\gamma \leq 0$ & $(\pi' + \gamma) > \pi_{max}$ **then**
    $\pi_{max} \longleftarrow \pi' + \gamma$
    $a, b \longleftarrow p, q$
  end if
end for
if $(a, b) \neq \phi$ **then**
  $S_t \longleftarrow S_{t-1} \setminus \{a, b\} \cup \{t, i\}$
end if
return S_t

In this section, we describe a novel online-OMP-based algorithm for the *online sparse approximation problem*(OSA) in Algorithm 1. The key difference between OSA and standard sparse approximation setting is that in OSA, new columns $\vec{X_p^q}$ are added and the target value $\vec{y_t}$ is updated at each epoch t. Line 10 in Algorithm 1 adds new datapoints till the cardinality of S_t reaches ω. Once the buffer is saturated, the *DataReplace* module is invoked in line 8 to replace an existing selected datapoint with the incoming datapoint. The criteria for replacement is to select the datapoints in S_t that contribute to a better approximation of the current value function $\vec{y_t}$. Hence a new datapoint with features $\vec{X_i^t}$ gets selected if the current approximation error reduces after the replacement. We compute the projection of the incoming datapoint features, $\vec{X_i^t}$ and that of the features of the selected datapoints $\vec{X_q^p}$ $\forall p, q \in S_t$ on the existing residual vector $\vec{\rho_t}$, measured by π and π' respectively. We also denote by γ, the contribution of datapoint $p, q \in S_t$ obtained through β_q^p. The datapoints with indices (p, q) in S_t whose additive impact $(\pi' + \gamma)$ is smaller than that of incoming datapoint (i, t), but larger than the current feature for replacement $(\vec{X_q^p})$ (line 8), gets substituted with the incoming point in line 14 of Algorithm 2. In terms of complexity, the per-epoch time complexity of OMP is $\mathcal{O}(\omega MN)$ and that of *VTruST* is $\mathcal{O}(\omega M(N - \omega))$.

Hyperparameter selection: The proposed framework has two user-controlled hyperparameters, the tradeoff λ and the subset-size ω. Since the metrics are not monotone in ω, we perform a grid search with various selection fractions between 10 - 90%. Exploiting the monotonicity of metrics w.r.t. λ, one can fix a threshold on the first metric, say accuracy in case of $\mathcal{V}_{af}$ and perform a binary search to arrive at an optimal point w.r.t. the second metric (fairness in $\mathcal{V}_{af}$), once the threshold w.r.t. the first metric has been satisfied.

3 Experimental Evaluation

In this section, we describe the datasets, models, and evaluation metrics used for the trustworthiness metrics - Accuracy, Fairness and Robustness. We analyze the performance of *VTruST* (VTruST-F with V_{af}, VTruST-R with V_{ar}, VTruST-FR with V_{rf}) over various applications. All our experiments have been executed on a single Tesla V100 GPU.

3.1 Error rate, Fairness and Robustness on Social Data

We evaluate the ability of VTruST to achieve a tradeoff between pairs of the three important social trustworthiness metrics: error rate (ER), fairness and robustness. Our *baselines* are: Wholedata standard training (ST), Random, SSFR (Roh et al., 2021a), FairMixup (Mroueh et al., 2021) and FairDummies (Romano et al., 2020). We report results on three benchmark datasets: COMPAS (Angwin et al., 2016) , Adult Census (Kohavi et al., 1996) and MEPS-20 (mep). We use a 2-layer neural network for all the datasets. We report two fairness metrics: equalised odds (EO) Disparity (Hardt et al., 2016) and demographic parity(DP) Disparity (Feldman et al., 2015) following (Roh et al., 2021a).

Fairness and Error Rate comparison (VTruST-F) with baselines: We compare the performance metrics of VTruST-F with the baselines in Table 1. The better the model is, the lower its ER as well as its fairness measures. We can observe in Table 1 that VTruST-F with 60% selected subset outperforms all the other methods in terms of fairness measures by a margin of $\sim 0.01 - 0.10$, and performs close to Wholedata-ST that yields the lowest

ER. This denotes that it is able to condemn the error-fairness tradeoff emerging out to be the best performing method. We report these results with standard deviation across 3 runs.

Table 1: **Comparison of VTruST-F with baselines over 60% subset for fairness evaluation.**

Methods	COMPAS			AdultCensus			MEPS20		
	ER ±std	**EO Disp ±std**	**DP Disp ±std**	**ER ±std**	**EO Disp ±std**	**DP Disp ±std**	**ER ±std**	**EO Disp ±std**	**DP Disp ±std**
Wholedata-ST	0.34 ±0.001	0.31 ±0.05	0.24 ±0.03	0.16 ±0.002	0.19 ±0.06	0.13 ±0.06	0.09 ± 0.001	0.09 ± 0.007	0.08 ± 0.0008
Random	0.35 ±0.002	0.20 ±0.10	0.23 ±0.09	0.19 ±0.002	0.16 ±0.05	0.13 ±0.05	0.12 ±0.017	0.06 ±0.02	0.08 ±0.005
SSFR	0.35 ±0.002	0.26 ±0.03	0.17 ±0.02	0.21 ±0.001	0.18 ±0.03	0.12 ±0.01	0.14 ±0.003	0.10 ±0.011	0.06 ±0.005
Fair-Dummies	0.35 ±0.002	0.24 ±0.02	0.17 ±0.01	0.16 ±0.002	0.14 ±0.01	0.10 ±0.01	0.12 ±0.001	0.13 ±0.005	0.08 ±0.003
Fair-Mixup	0.35 ±0.03	0.15 ±0.03	0.13 ±0.04	0.24 ±0.04	0.11 ±0.05	0.1 ±0.02	0.89 ±0.02	0.02 ±0.04	0.05 ±0.03
VTruST-F	0.34 ±0.002	**0.15 ±0.01**	**0.13 ±0.01**	0.18 ±0.001	**0.11 ±0.03**	**0.05 ±0.01**	0.09 ±0.003	**0.01 ±0.001**	**0.05 ±0.0008**

Tradeoffs between Error rate, Fairness and Robustness (VTrust-F , VTruST-FR): We observe the tradeoffs between error rate vs fairness (VTruST-F: Figure 1a) and fairness vs robustness (VTruST-FR: Figure 1b) through pareto frontal curve by varying $\lambda \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}$. The error rate is measured on the clean test sets while robust error rates are measured on the label flipped test sets. We can observe that Wholedata-ST has a lower error rate but high disparity and robust error values. The other baselines continue to have a higher error rate and disparity compared to VTruST. We report the results on other datasets in the Appendix.

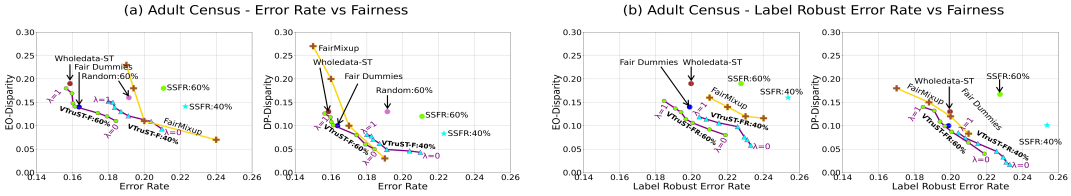


Figure 1: **Controlling tradeoffs in trustworthiness metrics for social data - Adult Census.**

3.2 Accuracy and Robustness on Image and Scientific Datasets

We evaluate VTruST on three image datasets: CIFAR10 (Krizhevsky et al., 2009) , MNIST (Deng, 2012) and Tinyimagenet (Le and Yang, 2015) using ResNet-18 (He et al., 2016). For evaluation, we use the standard accuracy (SA) computed on the clean test sets and the robust accuracy (RA) computed on the corrupted test sets, CIFAR-10-C, Tiny ImageNet-C (Hendrycks and Dietterich, 2019) and MNIST-C (Mu and Gilmer, 2019). While augmentation leads to robustness (Rebuffi et al., 2021), it also leads to a large dataset with redundancy. We use VTruST-R to select high-quality subset from augmented data. Empirically we find that creation of augmented data by sampling images based on how difficult an augmentation is (*Sampled Augmentation (SAug)*) leads to better performance compared to uniform selection across augmentations (*Uniform Augmentation (UAug)*). We describe SAug (Algorithm 3) and its comparison with UAug in Table 5 in the Appendix. Next, we define the baselines. (i) *Clean-ST*: Unaugmented training dataset. (ii) *Uniform Augmentation (UAug)* (iii) *Sampled Augmentation (SAug)* (iv) *SSR* (Roh et al., 2021a): Training subset using the robustness objective function. (v) *AugMax* (Wang et al., 2021).

Table 2: **Comparison of VTruST-R over varying subset sizes for robustness evaluation. The numbers in brackets indicate the difference with the second best among baselines.**

Methods	MNIST			CIFAR10			TinyImagenet		
	#Data points	SA	RA	#Data points	SA	RA	#Data points	SA	RA
Clean-ST	60K	99.35	87.00	50K	95.64	83.95	100K	63.98	23.36
AugMax	240K	97.62	88.79	200K	94.74	86.44	400K	54.82	40.98
SAug	260K	**99.36** (1.74)	**97.31** (8.52)	200K	**94.9** (0.16)	**90.13** (3.69)	300K	**62.04** (7.22)	**42.04** (1.06)
After subset selection from SAug									
SSR:40%	104K	98.98	94.96	80K	93.3	85.73	120K	32.82	24.42
VTruST-R:40%	104K	**99.04** (0.06)	**96.29** (1.33)	80K	**94.74**	**88.23** (1.79)	120K	**57.3** (2.48)	39.69
SSR:60%	156K	99.07	96.53	120K	93.77	88.0	180K	41.94	30.07
VTruST-R:60%	156K	**99.12** (0.05)	**97.09** (0.56)	120K	**94.77** (0.03)	**89.21** (1.21)	180K	**60.88** (6.03)	**41.50** (0.52)

Table 3: **Performance comparison on scientific datasets**

Metrics	Spinodal							EOSL			
	Whole data	Rand 40%	SSFR 40%	VTruST -R 40%	Random 60%	SSFR 60%	VTruST -R 60%	Whole data	Rand 40%	SSFR 40%	VTruST -R 40%
SA	**83.08**	73.05	74.84	**80.33**	77.06	78.94	**81.93**	70.01	63.74	62.40	**66.10**
RA	76.89	61.11	62.32	**78.36**	75.11	75.18	**80.41**	66.72	60.04	56.90	**65.27**

Robustness and Accuracy comparison (VTruST-R) with baselines: We compare VTruST-R with the baselines in Table 2 where it can be seen that model trained on clean datasets (*Clean-ST*) performs abysmally in terms of RA, indicating the need of data augmentations. VTruST-R is seen to outperform AugMax in most of the scenarios, thus indicating that data-centric approaches help in creating quality training datasets.

Scientific datasets : We analyzed the performance of VTruST-R on binary class scientific datasets - Spinodal and EOSL (Benato et al., 2022) that have 29,000 samples with 400 features and 180,000 samples with 576 features respectively. We used the experimental setup as (Benato et al., 2022) for evaluation. Table 3 shows that VTruST-R (using label flipping for robustness) performs close to the wholedata in standard accuracy (SA) and better in terms of robust accuracy (RA). The remaining results can be found in the Appendix.

3.3 Data-centric analysis: Post hoc explanation

In this section, we explore the characteristics of the selected samples to justify their quality.

Explanation for fairness: We use the metric *Counterfactual Token Fairness Gap (CF-Gap)*(Garg et al., 2019) for our evaluation. Given a selected instance x, we generate a counterfactual instance x' by altering its sensitive attribute and define CF-Gap(x) as $\|f(x_i) - f(x_i')|$ where $f(x)$ corresponds to the model confidence on the target label. We plot the distribution of CF-Gap in Figure 2. It can be observed that VTruST-F acquires the least value, justifying its retainment of fair subsets leading to fair models. We show 10 anecdotal samples from the Adult Census dataset in Table 4 on the basis of high *CF-Gap* and we can observe that SSFR has a large number of redundant samples with similar attribute values (highlighted) while VTruST-F which anyway has relatively lower CF-gap contains a diverse set of samples.

Figure 2: **Box plot representation of CF-gap.**

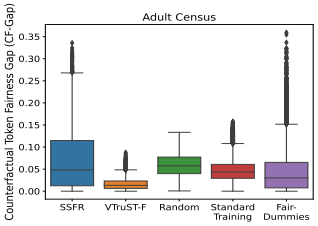


Table 4: **Sample instances with High Counterfactual Token Fairness Gap**

VTruST-F					SSFR				
Feat	Rel	Race	Sex	NC	*Feat*	Rel	Race	Sex	NC
D_1	ORel	B	F	JM	D_1	Husb	W	M	US
D_2	NIF	W	M	US	D_2	Husb	W	M	US
D_3	NIF	W	M	US	D_3	Husb	W	M	US
D_4	OC	API	F	TW	D_4	Husb	W	M	US
D_5	Husb	W	M	US	D_5	Husb	W	M	US
D_6	UnM	W	F	US	D_6	Husb	W	M	US
D_7	Wife	W	F	US	D_7	NIF	W	M	US
D_8	OC	W	M	US	D_8	OC	W	M	US
D_9	NIF	AIE	F	Col	D_9	Husb	W	M	DE
D_{10}	UnM	W	F	DE	D_{10}	OC	W	M	US

Explanation for robustness: Delving into the literature (Swayamdipta et al., 2020; Huang et al., 2018), we pick two measures - *uncertainty* and *distinctiveness*. Having a set of hard-to-learn and distinguishable samples in the subsets makes the model more generalizable and robust. We quantify uncertainty of an instance x as predictive entropy $(-f(x)logf(x))$ and distinctiveness as $\mathbb{E}_{e \in \mathcal{S}} dist(fv(x), fv(e))$ where $dist(,)$ is the euclidean distance and $fv(.)$ is the feature from the model's penultimate layer. Based on the data maps in Figure 9 in Appendix, we show anecdotal samples in Figure 3 having High Distinctiveness-High Uncertainty (HD-HU). The anecdotal samples and the histogram visualization show that VTruST-R selects diverse samples with difficult augmentations like impulse noise and glass blur, while similar(mostly white-background) and no-noise or easier augmentation-based samples like brightness are more observed in SSR samples, thus justifying the robust selection across augmentations using VTruST-R.

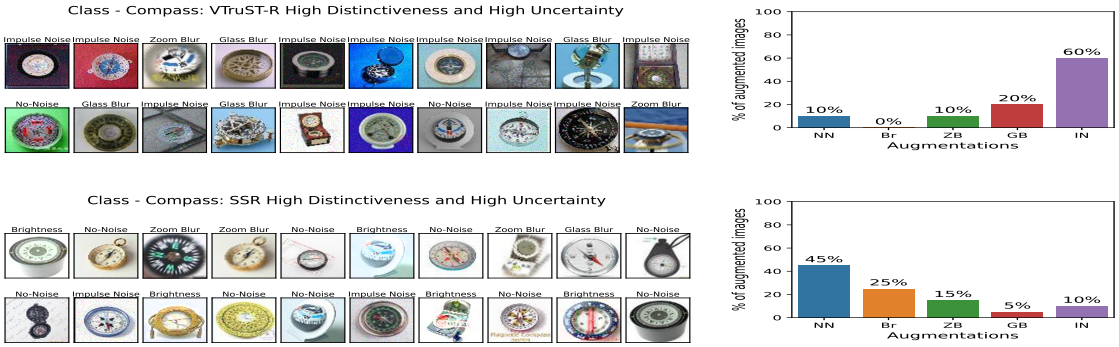


Figure 3: **Anecdotal samples from VTruST-R & SSR with High Distinctiveness and Uncertainty from TinyImagenet for class Compass.**

4 Discussion and Related works

Existing works on trustworthy AI (Liang et al., 2022) have focussed on designing fair (Zemel et al., 2013; Romano et al., 2020; Sattigeri et al., 2022; Chuang and Mroueh, 2021) and robust (Wang et al., 2021; Chen et al., 2022) models. Several works have also been done to inevstigate the tradeoffs between pairs of trustworthiness metrics - fairness vs accuracy (Roh et al., 2020); robustness vs accuracy (Pang et al., 2022; Hu et al., 2023) ; fairness vs robustness (Roh et al., 2021a). The closest approach to our method is that of (Roh et al., 2021a) which selects fair and robust samples by enforcing a constraint on the number of selected samples per class. However, none of the above methods have the flexibility of a user-controllable tradeoff between trustworthiness metrics. Besides, they impose an additional constraint on the training objective that may lead to a potential bias. Hence,

there arises a need for a paradigm shift from model-centric to data centric approaches that would look at the input space and sample the quality datapoints with potentially less bias.

The existing works in data-centric AI (DCAI) have explored data valuation approaches for obtaining quality data. (Swayamdipta et al., 2020; Ethayarajh et al., 2022; Seedat et al., 2022a,b) work on data quality measures to determine hard and easy samples. The other category of valuation methods are mostly based on Shapley values (Ghorbani and Zou, 2019; Wang and Jia, 2022), influence functions (Koh and Liang, 2017; Park et al., 2023), reinforcement learning (Yoon et al., 2020), gradient-based approximations (Yang et al., 2020; Paul et al., 2021; Killamsetty et al., 2021; Das et al., 2021) and training free scores (Just et al., 2023; Nohyun et al., 2022; Wu et al., 2022). However, all the methods only account for accuracy and none of the other trustworthiness metrics. Our proposed method, VTruST, lies in an intersectional area between trustworthy AI and data valuation. To the best of our knowledge, ours is one of the first works in DCAI that develops a controllable framework to provide a tradeoff across different trustworthiness metrics (fairness, robustness and accuracy) leading to desired subsets in an online training paradigm.

Reproducibility Statement

We run all our experiments on publicly available datasets and thus all our results can be seamlessly reproduced. The code is available at `https://github.com/dmlr-vtrust/VTruST/`. Details on model architectures and datasets are provided in the main paper. The remaining details for obtaining reproducible results can be found in the Appendix.

References

Medical expenditure panel survey. URL `https://meps.ahrq.gov/data_stats/data_use.jsp`.

S. Addepalli, S. Jain, et al. Efficient and effective augmentation strategy for adversarial training. *Advances in Neural Information Processing Systems*, 35:1488–1501, 2022.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

L. Benato, E. Buhmann, M. Erdmann, P. Fackeldey, J. Glombitza, N. Hartmann, G. Kasieczka, W. Korcari, T. Kuhr, J. Steinheimer, H. Stöcker, T. Plehn, and K. Zhou. Shared data and algorithms for deep learning in fundamental physics. *Computing and Software for Big Science*, 6(1), May 2022. ISSN 2510-2044. doi: 10.1007/s41781-022-00082-6. URL `http://dx.doi.org/10.1007/s41781-022-00082-6`.

T. T. Cai and L. Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688, 2011. doi: 10.1109/TIT.2011.2146090.

T. Chen, P. Wang, Z. Fan, and Z. Wang. Aug-nerf: Training stronger neural radiance fields with triple-level physically-grounded augmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15191–15202, June 2022.

C.-Y. Chuang and Y. Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=DNl5s5BXeBn`.

S. Das, A. Singh, S. Chatterjee, S. Bhattacharya, and S. Bhattacharya. Finding high-value training data subset through differentiable convex programming. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 666–681. Springer, 2021.

L. Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

K. Ethayarajh, Y. Choi, and S. Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$ -usable information. In *International Conference on Machine Learning*, pages 5988–6008. PMLR, 2022.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019.

A. Ghorbani and J. Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251. PMLR, 2019.

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

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Y. Hu, F. Wu, H. Zhang, and H. Zhao. Understanding the impact of adversarial robustness on accuracy disparity. In *International Conference on Machine Learning*, pages 13679–13709. PMLR, 2023.

S.-J. Huang, J.-W. Zhao, and Z.-Y. Liu. Cost-effective training of deep cnns with active model adaptation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1580–1588, 2018.

H. A. Just, F. Kang, J. T. Wang, Y. Zeng, M. Ko, M. Jin, and R. Jia. Lava: Data valuation without pre-specified learning algorithms. *arXiv preprint arXiv:2305.00054*, 2023.

D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, 55(2):1–38, 2022.

K. Killamsetty, D. Sivasubramanian, B. Mirzasoleiman, G. Ramakrishnan, A. De, and R. Iyer. Grad-match: A gradient matching based data subset selection for efficient learning. *arXiv preprint arXiv:2103.00123*, 2021.

P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

R. Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Y. Le and X. Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023a.

X. Li, P. Wu, and J. Su. Accurate fairness: Improving individual fairness without trading accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14312–14320, 2023b.

W. Liang, G. A. Tadesse, D. Ho, L. Fei-Fei, M. Zaharia, C. Zhang, and J. Zou. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4(8):669–677, 2022.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

Y. Mroueh et al. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021.

N. Mu and J. Gilmer. MNIST-C: A robustness benchmark for computer vision. *CoRR*, abs/1906.02337, 2019.

K. Nohyun, H. Choi, and H. W. Chung. Data valuation without training of a model. In *The Eleventh International Conference on Learning Representations*, 2022.

T. Pang, M. Lin, X. Yang, J. Zhu, and S. Yan. Robustness and accuracy could be reconcilable by (Proper) definition. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17258–17277. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/pang22a.html`.

S. M. Park, K. Georgiev, A. Ilyas, G. Leclerc, and A. Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

M. Paul, S. Ganguli, and G. K. Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34: 20596–20607, 2021.

G. Pruthi, F. Liu, S. Kale, and M. Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, 2020.

S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

Y. Roh, K. Lee, S. E. Whang, and C. Suh. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*, 2020.

Y. Roh, K. Lee, S. Whang, and C. Suh. Sample selection for fair and robust training. *Advances in Neural Information Processing Systems*, 34:815–827, 2021a.

Y. Roh, K. Lee, S. E. Whang, and C. Suh. Fairbatch: Batch selection for model fairness. In *ICLR*, 2021b. URL https://openreview.net/forum?id=YNnpaAKeCfx.

Y. Romano, S. Bates, and E. Candes. Achieving equalized odds by resampling sensitive attributes. *Advances in Neural Information Processing Systems*, 33:361–371, 2020.

P. Sattigeri, S. Ghosh, I. Padhi, P. Dognin, and K. R. Varshney. Fair infinitesimal jackknife: Mitigating the influence of biased training data points without refitting. *arXiv preprint arXiv:2212.06803*, 2022.

N. Seedat, J. Crabbé, I. Bica, and M. van der Schaar. Data-iq: Characterizing subgroups with heterogeneous outcomes in tabular data. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23660–23674. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/95b6e2ff961580e03c0a662a63a71812-Paper-Conference.pdf.

N. Seedat, J. Crabbe, and M. van der Schaar. Data-suite: Data-centric identification of in-distribution incongruous examples. *arXiv preprint arXiv:2202.08836*, 2022b.

Z. Song, L. Liu, F. Jia, Y. Luo, G. Zhang, L. Yang, L. Wang, and C. Jia. Robustness-aware 3d object detection in autonomous driving: A review and outlook. *arXiv preprint arXiv:2401.06542*, 2024.

S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of EMNLP*, 2020. URL https://arxiv.org/abs/2009.10795.

D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

H. Wang, C. Xiao, J. Kossaifi, Z. Yu, A. Anandkumar, and Z. Wang. Augmax: Adversarial composition of random augmentations for robust training. *Advances in neural information processing systems*, 34:237–250, 2021.

T. Wang and R. Jia. Data banzhaf: A data valuation framework with maximal robustness to learning stochasticity. *arXiv preprint arXiv:2205.15466*, 2022.

Z. Wu, Y. Shu, and B. K. H. Low. Davinz: Data valuation using deep neural networks at initialization. In *International Conference on Machine Learning*, pages 24150–24176. PMLR, 2022.

Y.-Y. Yang, C. Rashtchian, H. Zhang, R. R. Salakhutdinov, and K. Chaudhuri. A closer look at accuracy vs. robustness. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8588–8601. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/61d77652c97ef636343742fc3dcf3ba9-Paper.pdf.

J. Yoon, S. Arik, and T. Pfister. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, pages 10842–10851. PMLR, 2020.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, Z. Jiang, S. Zhong, and X. Hu. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

Table 5: **Comparison of Uniform Augmentation with Sampled Augmentation.**

Metrics	UAug (260K) MNIST	SAug (260K) MNIST	UAug (200K) CIFAR10	SAug (200K) CIFAR10	UAug (300K) TinyImagenet	SAug (300K) TinyImagenet
Standard Accuracy	99.34	99.37 (0.14)	94.84	94.9 (0.06)	60.92	62.04 (1.12)
Robust Accuracy	97.12	97.31 (0.19)	89.06	90.13 (1.07)	26.87	42.04 (15.87)

Appendix A. Appendix

1. Empirical evaluation on Social Data

In this section, we report the empirical results for the fairness value function. We show the variation in performance measures with varying subset fractions in Figure 4. It can be clearly observed that VTruST-F outperforms SSFR and that VTruST-F has the lowest Error Rate (ER) and disparity measures across all the considered fractions. We show the pareto-frontal curve for both clean and noisy datasets from MEPS20 and COMPAS in Figure 5 where we can observe that VTruST-F has the lowest disparity for $\lambda = 0$ and lowest error rate for $\lambda = 1$. It lies relatively in the bottom-left region compared to other baselines.

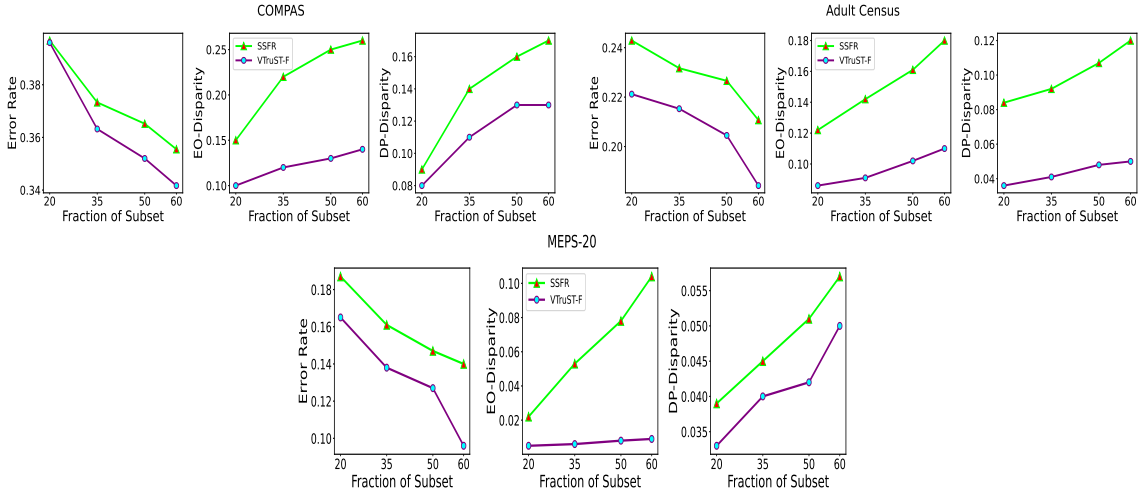


Figure 4: **Varying fraction of subsets: We report the ER and disparities for different subset sizes selected by the proposed method VTruST-F and SSFR. It can be observed that the proposed method always stays below the baselines in terms of error rate and disparity measures.**

2. Sampled Augmentation - SAug

Firstly, we look at the performance of the models across different augmentations that worked as an intuition for the sampling algorithm. Figure 6 depicts the difference in performance across difference augmentations. The cells (i, j) corresponds to performance of a model trained on augmentation i and tested on augmentation j. The diagonal element correspond to the self trained augmentation accuracies that turn out to be the best for any augmentation. Based on the intuition developed from the above heatmap, we present the

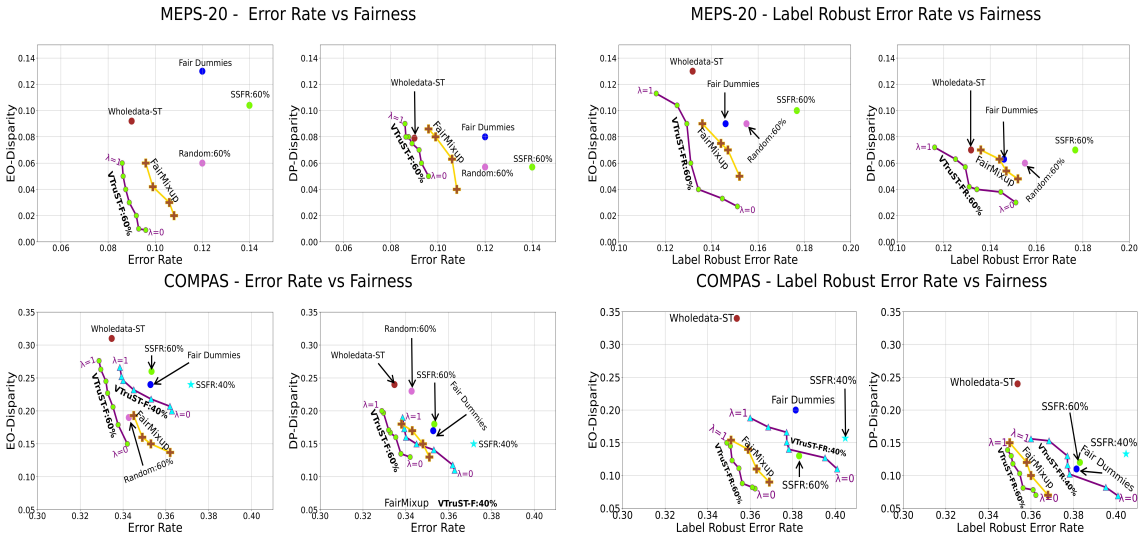
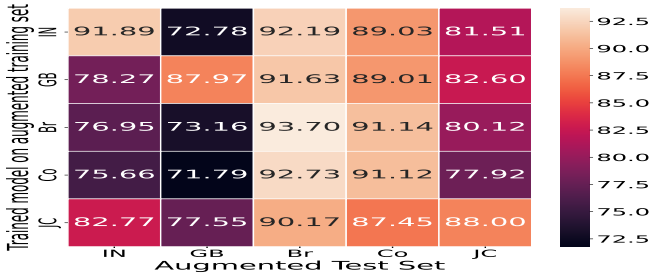


Figure 5: **Error Rate-Fairness and Robustness-Fairness tradeoff in clean and augmented data setup : We show the performance of the methods w.r.t the two dimensions - *Performance and Disparity* and can observe that the proposed method VTruST lies relatively on the bottom left region (*low error rate or robust error rate-low disparity*) with disparity being the lowest for $\lambda = 0$. Higher weightage to λ leads to a low error rate or robust error rate for the same fraction and increasing disparity.**

Figure 6: **Performance of self-trained augmentation models on augmented test sets.**



pseudocode of our sampling augmentation in Algorithm 3. We define Sampling Number (SN) for augmentation j as a normalized difference between the average RA for aug j (RA_j) and the self-trained accuracy. We execute the algorithm and plot the standard and robust accuracies across the different rounds in Figure 7. $R0$ corresponds to the round when we use the model trained on clean/non-augmented data. We can observe in Figure 7 that the similar pattern is observed across all the datasets where the standard accuracy gets compromised marginally with a significant improvement in robust accuracy, thus justifying the use of augmentations for robustness.

Algorithm 3 : **Sampling augmentations**

1: **Input**: $Mat_{M\times|A|}$ // Matrix[i,j] = Robust accuracy on test set augmented with corruption $j \in A$ using model $m \in M$ trained with data augmented with augmentation $i \in A$. $M = |A| + 1$ where we also test using model trained on clean data. ; Iteration: $t = 0$; Clean dataset: D^t ;
2: $SN_j^t = \frac{Mat[j,j] - \frac{\sum_{i\neq j} Mat[i,j]}{M-1}}{Mat[j,j]}$
3: Normalise $SN_j^t \forall j \in A$ and sample that fraction of images for the respective augmentations from all classes uniformly and form D^{t+1}.
4: Train on D^{t+1} and obtain model f.
5: Compute robust accuracy for each j using trained model f.
6: $SN_j^{t+1} = Mat[j,j] - RA_j^{t+1}$
7: $t = t + 1$
8: Repeat from line 3 till $\frac{\sum_j RA_j^t}{|A|} - \frac{\sum_j RA_j^{t-1}}{|A|} < \epsilon$
9: **Output**: D^t

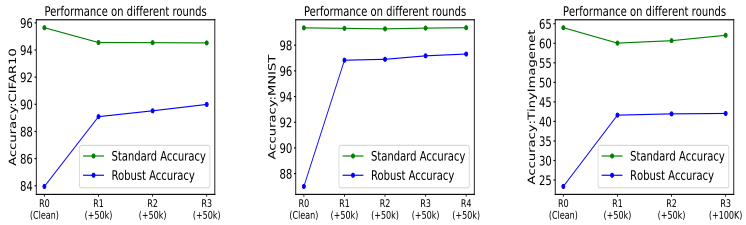


Figure 7: **Mean accuracies across all augmentations over each round. R0 represents the accuracy obtained using model trained on clean data. The subsequent rounds are using augmented images obtained from our proposed sampling algorithm. The drop in Standard Accuracy is marginal, while the increase in Robust Accuracy is significant.**

3. Empirical evaluation on scientific data

We report the results on EOSL dataset for the 60% subset in Table 6. It can be observed that VTruST-R performs better than the other baselines.

4. Data centric explanation for Fairness

We report the CF-Gap for the COMPAS dataset in Figure 8 and show that the proposed algorithm has the lowest values of the measures (the lower, the fairer) compared to all other baselines.

Table 6: **Performance comparison on scientific datasets**

Metrics	**EOSL**			
	Whole data	Rand 60%	SSFR 60%	VTruST -R 60%
SA	70.01	64.94	64.64	**68.86**
RA	66.72	61.55	62.43	**67.67**

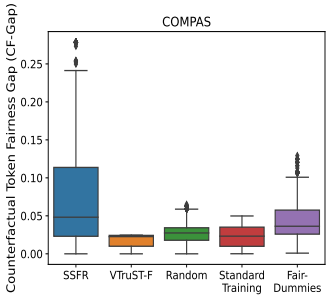


Figure 8: **Box plot representation of counterfactual token fairness gap on the selected subsets from VTruST-F and other baselines for COMPAS dataset.**

5. Data centric explanation for Robustness

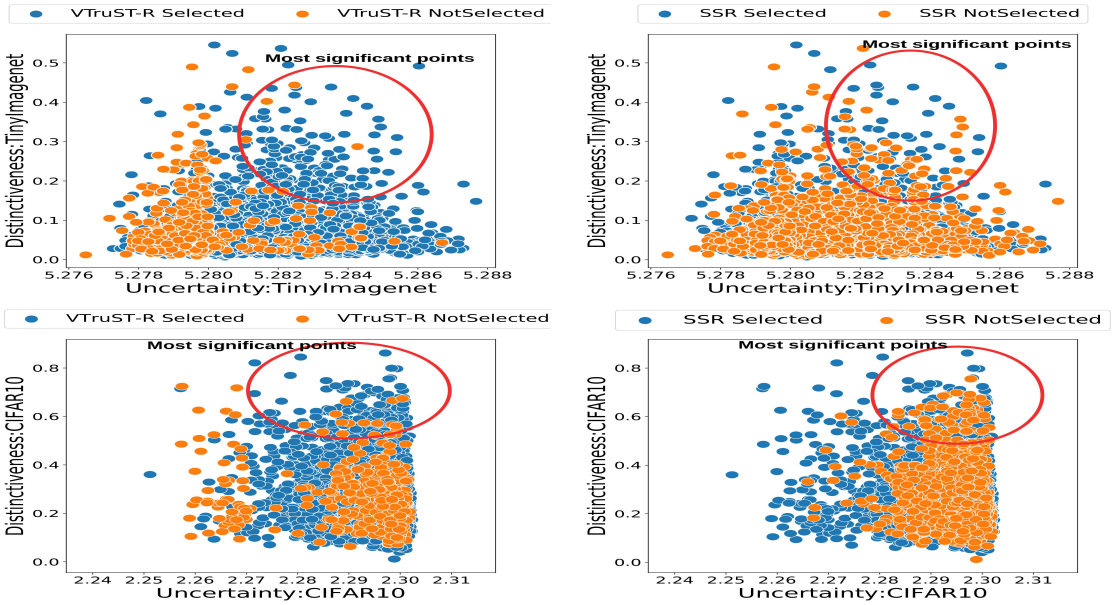


Figure 9: **Data Map for randomly taken 5000 samples from TinyImagenet and CIFAR10 augmented training dataset**

We visualize the datapoints in the two dimensions - Uncertainty and Distinctiveness (defined in the main paper) in Figure 9 where we choose a random set of 5000 points from CIFAR and TinyImagenet datasets, followed by marking them as *selected* and *not selected* by VTruST-R and SSR respectively. We can observe that points with relatively high uncertainty and high distinctiveness(HD-HU) values mostly belong to the *selected set of VTruST-R*, while the *unselected points from SSR* mostly cover the HD-HU region.

We show anecdotal samples for the class *Watertank* from TinyImagenet in Figure 10 and for the classes *Car* and *Truck* from CIFAR-10 in Figure 11 having high distinctiveness and high uncertainty. It can be noted that (a) VTruST-R selects diverse samples while SSR selects similar (mostly similar background) samples ; (b) VTruST-R mostly selects samples

from difficult augmentations like Impulse Noise and Glass Blur while SSR selects samples from unaugmented (No-Noise) or easier augmented samples like Brightness and Contrast. This justifies the outperforming capability in robustness from VTruST-R in comparison with SSR.

6.Details on training regime

Experiments using Social Data: For all the datasets, we use a 2-layer neural network and vary the learning rate on a grid search between 5^{-1} to 5^{-4}.
Experiments using Image Data: For all the datasets, we use ResNet-18 model and a learning rate of 10^{-1} with momentum of 0.9 and weight decay of 5^{-4}.
Experiments using Scientific Data: For all the datasets, we use convolutional neural networks as the experimental setup from (Benato et al., 2022) and vary the learning rate on a grid search between 10^{-2} to 10^{-4}.

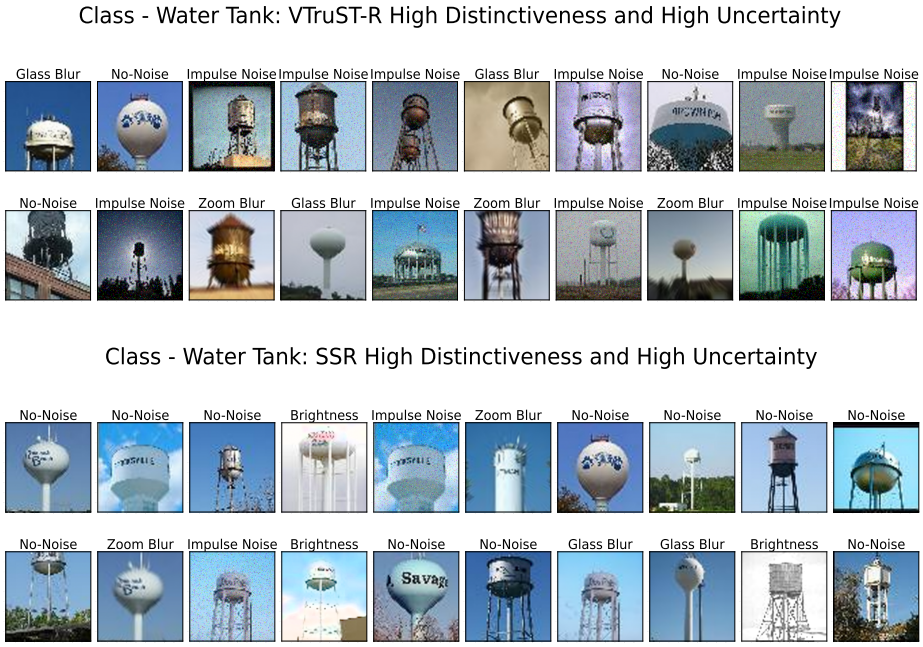


Figure 10: **Anecdotal samples from VTruST-R and SSR with High Distinctiveness-High Uncertainty from TinyImagenet for class Watertank.**

Figure 11: **Anecdotal samples from VTruST-R and SSR with High Distinctiveness-High Uncertainty from CIFAR10 for classes Car and Truck.**