Exploring Comparative Visual Approaches for Understanding Model Trade-offs in Adversarial Machine Learning

Yuzhe You
School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
y28you@uwaterloo.ca

Jian Zhao
School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
jianzhao@uwaterloo.ca

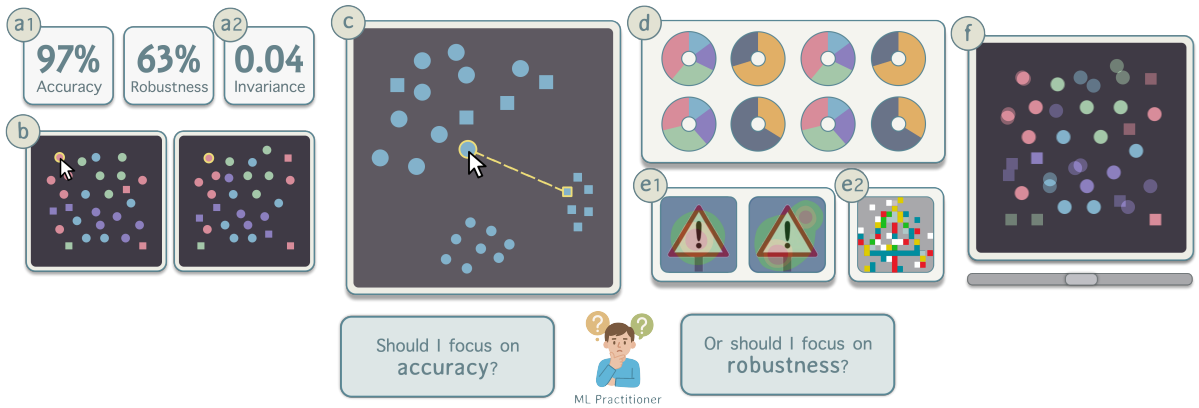

Figure 1: Effective comparative visual designs we identified for trade-off analysis: (a1) Juxtaposition of metrics with (a2) explicitly encoded invariance for high-level metric comparison; (b) Juxtaposed views with linked highlighting for global embedding comparison; (c) Superimposed projections with explicitly encoded trajectories and animations for class embedding comparison; (d) Juxtaposed doughnut charts for class metric comparison; (e1) Juxtaposed images with superimposed heatmaps and (e2) explicitly computed perturbation for instance comparison; (f) "shine-through" interaction for model comparison.

Abstract

Despite the effectiveness of adversarial training (AT) in enhancing model robustness, it suffers from the accuracy-robustness trade-off and the "robust fairness" problem. To strategize effectively, practitioners have the need to explore and compare model performance in both standard and adversarial settings concurrently. This work presents a design study with 11 experts to explore effective comparative visual techniques for multi-level trade-off analysis. We first collaborated with five adversarial machine learning (AML) experts in an iterative design process, based on which we developed a visual analytics design probe, VATRA, that employs an augmented hybrid comparative design to support concurrent accuracy and robustness evaluations for assessing model trade-offs. Further, we conducted user studies with six domain experts and derived two in-depth use cases of VATRA, providing empirical knowledge about how ML practitioners can leverage comparative visualizations for AML trade-off analysis.

CCS Concepts

• **Human-centered computing** → **Visualization**; • **Computing methodologies** → *Machine learning*.

Keywords

adversarial machine learning, information visualization, explainable AI, evasion attack

ACM Reference Format:
Yuzhe You and Jian Zhao. 2025. Exploring Comparative Visual Approaches for Understanding Model Trade-offs in Adversarial Machine Learning. In *Proceedings of Graphics Interface (GI '25), May 26-29, 2025, Kelowna, British Columbia, Canada.* ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
GI '25, May 26-29, 2025, Kelowna, British Columbia, Canada
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Rapid advancements in Computer Vision (CV) [28, 34, 37] have resulted in increasing deployment of vision-based classifiers in applications such as autonomous driving [27], facial recognition [36], and healthcare diagnostics [18]. Nonetheless, it has been observed that state-of-the-art (SOTA) classifiers like deep neural networks (DNNs) are extremely brittle to *adversarial examples* [24, 56], i.e., input data crafted maliciously with minor perturbations to produce

model mistakes, known as *evasion attacks*. These adversarial examples, though often indistinguishable from the original dataset, can significantly undermine model accuracy, raising safety concerns [11, 24, 41, 42, 62, 62]. As such, building adversarially robust models has become a key design goal and a frequently studied topic of the ML community [14, 41]. Currently, the standard and most effective defense is *adversarial training* (AT), which trains a classifier with adversarial examples close to the training data, such as adding adversarial examples to the training set [41] or integrating adversarial methods into regularization [65].

However, AT is known to come with a two-fold trade-off, which introduces critical decision-making challenges for ML practitioners. First, while AT improves a model's robustness on adversarial examples, it lowers the model's clean accuracy on natural examples—the *accuracy-robustness trade-off* problem [59]. Second, AT exacerbates the inter-class discrepancy for both accuracy and robustness, meaning that an AT model may significantly underperform in certain classes compared to a natural model [5]. This is defined as the *robust fairness* problem. Recent papers suggest that these two costs are unavoidable consequences of existing AT methods [5, 7, 21, 59]. Therefore, ML practitioners are faced with a dilemma: *should they pursue greater accuracy, aim for robustness, or achieve a balance between the two for optimal outcomes?*

Presented with this dilemma, ML practitioners need to thoroughly explore and evaluate their models in both *natural* and *adversarial* settings, and investigate existing attack strategies to identify potential trade-offs. A comparative visual paradigm is well-suited for this high-level need, as one can utilize it to 1) identify and link object differences (e.g., clean vs. perturbed datasets) to performance outcomes, 2) compare individuals to dissect how similar patterns (e.g., model characteristics) lead to positive performance (e.g., increased accuracy/robustness), and 3) assess overall patterns to determine if the performance meets expectations (e.g., if the model is satisfactory in both natural and adversarial conditions) [19, 22, 23]. We thus aim to explore comparative visual techniques for model trade-off analysis across levels and develop a visual analytics design probe to provide empirical insights on using hybrid comparative visualizations for real-world model trade-off assessment. Although several works have explored visual analytics for adversarial attacks, they either do not meet our objective or have a completely different focus. For example, Cao et al. [8] focused on using visual analytics to explain the causes of misclassifications rather than trade-offs, VATLD [25] is non-comparative and restricted to traffic light detectors, and Ma et al. [40] focused solely on data poisoning attacks in binary classifications.

To fill in this gap, we carried out a design study involving a total of 11 experts to investigate visual comparative techniques for trade-off analysis, consisting of three main phases. First, we collaborated with five experts in *adversarial machine learning* (AML, i.e., the study of adversarial attacks and their defenses) in an iterative co-design process to explore user tasks, levels of comparisons, and design goals for developing effective visual comparative techniques. Second, based on our findings, we created VATRA, a visual analytics design probe that employs an augmented hybrid comparative approach to support concurrent *accuracy* and *robustness* evaluations for assessing model *trade-offs*. Finally, we conducted a user study with six domain experts, deriving two use cases of VATRA that provide further empirical insights into the application of comparative visualizations for trade-off analysis.

In summary, we make the following contributions:

- An **iterative co-design** with five AML experts that explored various comparative approaches, providing insights into user goals, tasks, and effective comparative visual techniques for model trade-off analysis at different levels.
- A visual analytics **design probe**, VATRA, developed through our iterative design process, that employs an augmented hybrid comparative approach to support trade-off analysis.
- A **user study** with six experts from two application domains that derived two use cases of VATRA, as well as further empirical insights on how ML practitioners can leverage comparative visualizations for trade-off evaluation in AML.

2 Related Work

2.1 Adversarial Attacks & Defenses

Recently, a growing body of research has been done on adversarial attacks and their defenses. A white-box attack assumes that attacker has full access to model internals, which includes the Fast Gradient Sign Method (FGSM) [24], Basic Iterative Method (BIM) [35], and Projected Gradient Descent (PGD) [41]. A black-box attack assumes the attacker only has access to the inputs and outputs of a targeted model. Query-based black-box attacks include ZOO [11] and HopSkipJump [10], while transfer-based attacks include substitute model [45] and ensemble attacks [39].

Presently, adversarial training (AT) [24, 41, 56] is the standard and most effective approach for building robust models that can withstand strong attacks, which involves training a model with adversarial examples. The most well-known AT is by Madry et al. [41], which formulates the task as a saddle point (min-max) optimization problem. More advanced methods also exist, such as TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization (TRADES) [65], which captures the observed accuracy-robustness trade-off through a regularized surrogate loss, robust self-training (RST) [48], a robust variant of self-training that leverages extra unlabeled data, and "free" adversarial training (Free-AT) [52], which recycles the model's gradient information for producing adversarial examples.

The abundance of existing research on adversarial attacks (e.g., [2, 41, 48, 59, 64, 67, 68]) and defenses (e.g., [47, 48, 61, 64, 65]) attests to their relevance as AML methodologies. In this paper, we primarily explored comparative visual approaches with attacks including PGD [41] and SimBA [26], and defenses including TRADES [65] and RST [48], though our design probe supports analysis of different AT methods under any evasion attack. We selected these methods to examine how practitioners can leverage comparative visual analytics for both white-box and black-box attacks, focusing on SOTA ATs to ensure our insights reflect the most effective AT approaches available.

2.2 Exploring Properties in Adversarial Training

Despite AT's success in model defense, it comes with a two-fold cost, including 1) *accuracy-robustness trade-off*, a reduction in the model's

natural accuracy, and 2) the *robust fairness* problem, a reduction in the model's class-wise fairness. Raghunathan et al. [49] showed that while both standard and AT CIFAR-10 [33] models achieved 100% training accuracy, the latter had a notable drop in testing accuracy despite its improvements in robustness. Xu et al. [63] applied natural and adversarial training on a CIFAR-10 PreAct-ResNet18 model [29], and noticed that a standard model's class performance was more consistent while an AT model displayed a severe class-wise performance discrepancy.

Existing literature have explored these properties of AT, but the trade-offs remain largely ill-understood. Several pessimistic papers suggest the trade-off may be intrinsic. For instance, there may exist an inherent tension between the goal of adversarial robustness and that of standard generalization that provably manifests even in simple settings [59]. Hu et al. [30] conducted a theoretical analysis of the impact of robustness from AT, and attributed the decline in accuracy to an inevitable change in the model's decision boundary. More optimistic works also exist. Yang et al. [64] presented an alternative perspective, claiming that the trade-off is not inherent but a consequence of the current training methods due to a large gap between theory and practice.

In light of these observations and theoretical analyses, the dual trade-offs seem inevitable. ML practitioners thus have the need to navigate these trade-offs by concurrently exploring model performance on clean and adversarial datasets, a task well-suited for a comparative visual paradigm. Our work seeks to address this need by exploring various comparative visual techniques, identifying those effective for trade-off analysis at different granularities, and examining how ML practitioners can utilize our design probe.

2.3 Visualizations in Adversarial Machine Learning

In general, AML visual analytics remains relatively under-explored. While some visual analytics tools have been proposed, they either lack a comparative approach designed for trade-off explorations or completely diverge in focus. Cao et al. [8] used a river-based metaphor to visualize datapaths of clean and adversarial examples but lacks support for trade-off analysis. VATLD [25] is designed to assess the accuracy and semantic robustness of traffic light detectors but also overlooks the consequential trade-offs that may arise. The tool also focuses on evaluating "semantic robustness" that is human-interpertable, which differs from the traditional definition of adversarial robustness targeted in our work. Sietzen et al. [53] developed a tool for exploring CNN activations under 3D scene alterations but focus solely on adversarial attacks in 3D and instance-level evaluation without addressing trade-offs. Ma et al. [40] proposed a framework for exploring model vulnerabilities, but it is specialized for analyzing attacks that poison training data instead of producing adversarial examples, in binary classification tasks.

Additionally, though some of these works include simple comparative elements [8, 25, 53], such as basic juxtaposition, none investigated effective comparative designs that can be integrated into trade-off analysis. These tools are largely limited to juxtaposed instance-level comparisons, such as placing natural and adversarial images or feature heatmaps side by side, which, based on insights from our later design study, are insufficient for effective trade-off analysis. Furthermore, to better support the workflow, our study revealed that analyzing models at multiple levels of detail is essential. Other AML visualizations that target non-experts include Bluff [16], which highlights the abstracted activation pathways of InceptionV1 [55] exploited by attacks, and Adversarial Playground [44], which juxtaposes a natural MNIST image and its adversarial counterpart alongside classification probabilities. While both visualizations effectively explain attack logic to learners, our work focuses on exploring comparative visual approaches that support experienced ML practitioners in trade-off analysis at multiple levels, and thus diverges in focus.

2.4 Comparative Visualization Approaches

As a common task in data analysis, comparison involves looking for differences and similarities between objects, and identifying trends or patterns that shed light on their relationships [22]. Visualizations have been shown to be highly effective in assisting users with comparison [3, 22, 23, 31]. The visual designs for comparison can be divided into three categories, including 1) *juxtaposition*, 2) *superposition*, and 3) *explicit encoding* [22]. *Juxtaposition* involves placing items separately in different spaces, often next to each other. Some examples include Sequence Surveyor [1], which juxtaposes large-scale multiple genome sequences as rows. *Superposition* involves placing items in the same space, often on top of each other. For instance, ContraNA [20] utilizes a contrastive representation view that compares target and background networks in the same space. Lastly, *explicit encoding* involves explicitly visualizing the relationships between the compared objects. For example, Mauve [15] visualizes connections between aligned genome blocks by drawing lines that logically connect the homologous collinear blocks from each genome. The three designs may also be combined to create hybrid designs [6, 13, 22, 23, 50]. In addition, Tominski et al. [57] showed that the three designs can also be augmented to form interaction techniques, including the side-by-side arrangement (i.e., juxtaposition), shine-through (i.e., superposition), and folding methods (i.e., explict encoding). An interactive prototype integrating all three interaction techniques was implemented to support comparisons of table and matrix visualizations [57].

Inspired by these works, we aim to investigate effective comparative visual designs, augmented by interaction, for model trade-offs through an iterative design process. We believe that identifying these approaches will help ML practitioners more effectively perform AML comparison tasks, such as assessing differences in model performance between datasets, recognizing similarities that enhance accuracy and robustness, and verifying overall patterns to ensure the model meets expectations [22].

3 Design

We employed an iterative user-centered design process with AML experts to explore various comparative designs. This section details our co-design process, design requirements, and empirical insights into the effective comparative approaches for trade-off analysis.

3.1 Design Process

Our co-design process involves five experts (E1 ~ E5; all men), all of whom are experienced AML practitioners skilled in designing adversarial defenses or training models adversarially. All participants are knowledgeable about the trade-offs, and four have conducted research specifically related to them. Specifically, our design process consisted of four stages as detailed below.

3.1.1 Design Requirement & Task Formulation. To first understand the design requirements, we conducted an extensive literature survey (Section 2) and interviews with E1 ~ E5 on trade-off analysis. The 90–120 minute semi-structured interviews covered topics including experts' trade-off analysis workflows, tools, challenges, desired visualization tasks, and visualization integration preferences. Each interviewee was compensated $20/hour. We transcribed and analyzed the interviews using a combined approach of open and closed coding. Through affinity diagramming, we confirmed the need for a multi-level comparative approach and identified key recurring themes. These insights informed our high-level domain tasks (Section 3.2) and design goals (Section 3.4) for our comparative visualizations and design probe.

3.1.2 Comparative Approach Exploration. Based on the workflows described by our experts, we identified a set of multi-level comparison tasks that can help them achieve the established high-level goals. Guided by our design guidelines and expert input, we explored various combinations of comparative designs (visual + interaction techniques) to support these comparison tasks. We created low-fidelity prototypes (e.g., sketches, mock-ups) and presented them to E1 ~ E5 during our co-design sessions, discussed the pros and cons of each comparative approach, and collected their feedback. Our insights on effective comparative visual designs for trade-off analysis, including findings from the high-fidelity prototype, are detailed in Section 3.3. From our preliminary design exploration, we selected the initial design for our design probe, VATRA, and defined the view components for each comparison task.

3.1.3 Design Probe Development. After selecting an initial design, we engaged E1–E5 via email, sharing updates and gathering feedback to refine our design probe and comparative visualizations. Incorporating expert insights, we subsequently implemented HF versions. Then, we hosted remote sessions to engage experts in HF prototype walkthroughs, during which we asked them to perform AML domain tasks (**T1** ~ **T4**; see Section 3.2) on their preferred datasets. After the walkthrough, we conducted a semi-structured interview to gather expert feedback on the design probe and suggestions for further improvement. Each session lasted between 60 to 90 minutes. Experts were compensated $20/hour, and their provided insights were incorporated into the final design.

3.1.4 User Studies with Experts. To explore how VATRA's comparative visual analytics fit into the workflows of ML practitioners across various application domains, we conducted 90-minute interviews with six domain-specific ML experts. Based on the collected insights, we created two detailed use cases to offer empirical knowledge into how ML practitioners can leverage comparative visualizations to perform trade-off evaluation in AML. Details about the studies will be described in Section 5.

3.2 Tasks to Support in Trade-off Analysis

From our interviews (Section 3.1.1), we found that none of the AML experts use existing visual analytics for model trade-offs due to limited comparative capabilities. E1, E2, and E5 rely on simple coding libraries for basic visualizations like tables and bar charts. However, these methods are often cumbersome, tedious, and insufficient. All interviewees recognized the value of adopting comparative visual approaches in their workflows and expressed interest in exploring them. While some AML visual tools exist (e.g., [8, 25, 53]), none are designed for trade-off analysis or integrate well into their workflows. Experts noted these tools focus on low-level analyses, like instance features or neuron pathways, lacking support for dataset- or class-level comparisons needed for trade-off identification.

Based on expert insights, we identified four essential high-level domain tasks to guide the design of our visual comparative approach. These tasks included:

- **T1** Exploring models' behaviors in standard and adversarial conditions to understand their trade-offs.
- **T2** Locating interesting insights related to model trade-offs or the underlying causes of these trade-offs.
- **T3** Identifying the optimal model for their specific dataset, or alternatively,
- **T4** Determining directions to improve the existing models or address their trade-offs.

T1 is the first step to identify "how much" of a trade-off exists, which typically involves directly comparing how the model's performance differs across metrics. **T2** involves examining the model's characteristics at each level to analyze how they differ under each conditions to generate insights. E2 and E3 mentioned examining feature space proximity to assess instance vulnerability, while E4 used feature heatmaps to determine if the model focuses on objects or noise. **T3** involves comparing models with different defenses, and identifying those that are most aligned with the target dataset, as *"different ATs vary in their effectiveness across different types of data"* -E2. **T4** involves devising potential strategies to improve model trade-offs based on insights gathered from comparative assessments.

3.3 Effective Visual Comparative Techniques for Trade-off Analysis

During our interview, our experts emphasized the need to perform comparison tasks at multiple granularities to gain a comprehensive understanding of the trade-offs. Therefore, the visual techniques should be designed to support comparison at different levels for users to perform **T1** ~ **T4**. To achieve this, we explored how each domain task could be supported with multi-level comparisons in our design exploration and prototype walkthroughs. This led to the identification of six levels of comparisons that align with the high-level objectives and our experts' workflows (**C1** ~ **C6**). Through our iterative design process, we explored various comparative visual techniques with our experts, and based on their feedback, identified those effective for supporting these comparison tasks. Below we present a report of our empirical insights into the effective comparative visual methods for trade-off analysis.

C1: High-level metric comparison. For comparing high-level performance metrics (e.g., overall accuracy/robustness), *juxtaposition* is a simple but effective strategy (Figure 1a1). Since these

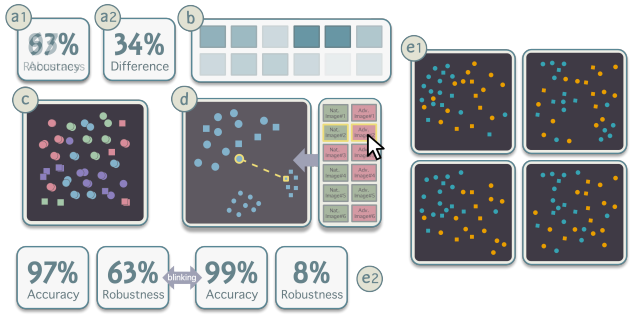

Figure 2: Examples of other comparative visual designs we explored during our iterative co-design process.

metrics are numerical data, superposition is less ideal for lacking a shared coordinate system for overlaying them (Figure 2a1). While explicit encoding can visualize trade-offs more explicitly (Figure 2a2), the experts preferred juxtaposition for its better context when assessing individual models. Our interviews and prototype walkthroughs also revealed that the experts were generally not interested in knowing the precise accuracy/robustness difference from explicit encoding, as simple juxtaposition already helps them compare and identify key trade-off patterns and differences [23]. Placing metrics close together also avoids the issue of object separation in juxtaposition, since accuracy and robustness are simple metrics that are easy to compare [17, 23]. However, when comparing conditions within a single model, *explicit encoding* offers another layer of representation for understanding object relationships (Figure 1a2). For instance, displaying the average model invariance between two datasets emphasizes the trade-off to viewers by visualizing it as a computed relationship. For class precision and recall (considered **C1** as it compares overall class metrics), *juxtaposition* with color-coding supports identification of class unevenness through comparison of high-level visual features such as color differences (Figure 2b). E3 noted that this comparison is particularly helpful for quickly finding fairness irregularities across classes.

C2: Global-level embedding comparison. We found that a hybrid of *juxtaposition + explicit encoding* is highly effective for this comparision (Figure 1b). Juxtaposing the global embeddings of natural and adversarial datasets, with interactive highlighting of corresponding selections—a form of explicit encoding [23]—ensures that each dataset is fully visible and comparable. The experts found this hybrid approach particularly valuable in this context, because they were comparing complex objects (e.g., scatterplots with many data points) and each comparative technique effectively compensates for the other's limitations. Specifically, juxtaposition maintains original data by displaying them independently in separate spaces, but we found that users struggled to make connections between them due to the complexity of embedding projections. Explicit encoding directly visualizes object differences but loses the context of the original objects. By combining these two approaches, the hybrid design leverages juxtaposition to maintain context and explicit encoding to highlight object differences [23]. Superposition creates visual clutter and difficulty in interpretation due to a large number of instances, making it more suitable for class-level embedding comparison with fewer data (Figure 2c).

C3: Class-level embedding comparison. From our exploration, we found that a hybrid design of *superposition + explicit encoding* is particularly effective for embedding comparison at a class level (Figure 1c). In global embeddings, where a large number of data points can lead to visual clutter, juxtaposition was preferred as it separates datasets into different spaces for clearer comparison. However, with class embeddings, we found that the smaller number of instances makes superposition more advantageous. Specifically, by superimposing natural and adversarial instances in the same space, users could directly compare and identify trade-offs based on spatial proximity, reducing the cognitive load of shifting attention between separate views [23]. During walkthroughs, experts confirmed that superposition helps them quickly determine data relationship: overlapping indicates similarity, while spatial differences reveal dissimilarity. Furthermore, when explicit encoding is added thoughtfully, it reveals additional embedding relationships. Specifically, we observed that superposition reveals general distribution differences, while explicit representations, like instance trajectories or embedding movement animations, highlight finer trade-offs between individual instances from the two datasets. In one walkthrough example, proximity revealed that a class cluster broke into smaller clusters after an attack, but explicit encoding helped users trace which parts of the original cluster shifted into these smaller clusters.

However, *juxtaposition* proved useful when used to address navigation challenges. Specifically, it can organize class instances from both datasets into side-by-side columns in a separate linked view, allowing for *"quick comparison of instance prediction correctness before and after an attack."* -E5 (Figure 2d). This immediate visual feedback reduces user burden by supporting efficient instance navigation based on model correctness [22, 23].

C4: Class-level metric comparison. For metric comparison at a class level (e.g., false positives, false negatives), *juxtaposition* by itself proves to be an intuitive design (Figure 1d). By placing class performance side by side in a consistent spatial layout, our experts noted that they can efficiently compare metric trade-offs without cognitive overload [23]. Visual elements from simple charts (e.g., doughnut charts) are sufficient to help them retain context and rely less on memory to make comparisons.

C5: Instance-level image comparison. For this task, we found the hybrid design of *juxtaposition + superposition* and *explicit encoding* particularly useful. Juxtaposition and superposition complement each other well in this context to address the common imperceptibility of adversarial attacks (Figure 1e1). Juxtaposition allows users to directly compare input images from both datasets close together, making it easy to spot any visible semantic differences. When the images appear visually indistinguishable, superposition becomes useful to highlight model perception trade-offs by overlaying feature-based visual explanations. One effective example, which we found through our design exploration, is juxtaposing natural and adversarial versions of an image with the option to interactively toggle heatmap overlays to highlight model perception differences. Many heatmap visualizations (e.g., saliency maps) are already widely used by AML practitioners for instance-level analysis [9, 32, 66], and these visual explanations themselves act as a form of superposition. When mixed with juxtaposition, this

creates "comparisons of comparisons" [23], visualizing how significant features perceived by models differ before and after the attack. E5 found this design valuable as it allows for first comparing raw images and then overlaying heatmaps to observe model perception trade-offs. Explicit encoding is less effective for heatmaps as it obscures original images, but can be effectively used to visualize the image differences as the applied perturbation, which highlights the image features the attack specifically targets (Figure 1e2).

C6: Model-level comparison. For model comparison, an effective design is a hybrid approach augmented by interaction, using the "blink" and "shine-through" techniques. We consider these techniques to incorporate elements of all three designs: *juxtaposition*, *superposition*, and *explicit encoding*. During the initial design exploration, both our team and experts originally leaned toward using a small multiples design, where models are compared side by side using juxtaposition, with either superposition or explicit encoding to compare the two conditions within each model (Figure 2e1). However, prototype walkthroughs revealed that this approach led to cognitive overload, especially when combined with comparisons from other levels (**C1** ∼ **C5**), making it too visually overwhelming. This led us to explore more dynamic designs [23, 57].

Through our exploration, we found that the "blink" interaction [23] allows users to toggle between models while retaining context, such as maintaining focus on a specific class or instance when switching models (Figure 2e2). This method effectively combines elements of superposition and juxtaposition by allowing comparison within the same coordinate system, but sequentially in time rather than space. We found that it worked well for simpler comparisons like **C1**, **C4**, and **C5**, as it still predominantly relies on users' memory to make comparisons. However, our experts noted that with this technique, comparing complex objects (**C2** & **C3**) was more challenging due to the mental effort required to track shifts across models. From further exploration, we found that the "shine-through" interaction can effectively address this problem (Figure 1f). Proposed in [57], it allows users to overlay embeddings from different models with the ability to control the transparency of each model's embeddings. This flexibility lets users retain as much context from each model as needed, and fine-tune comparisons to reveal subtle trade-offs between models without overwhelming the display. Though the original work [57] largely views this as an augmented form of superposition, we consider it to incorporate elements of all three visual comparison designs: superposition by overlaying embeddings in the same space, juxtaposition in time through the ability to transition between models, and transparency control as explicit encoding to emphasize model differences. As such, from our exploration of **C6**, we verified that, beyond providing comparison assistance (e.g., in **C1** ∼ **C5**), interaction can address inherent issues in basic visual comparative designs. An interaction-heavy design effectively mitigates visual clutter and cognitive overload that arise in other hybrid approaches.

3.4 System Design Goals

To support the designs of a complete system, we performed additional qualitative analysis on our expert interviews (Section 3.1.1) to extract further design requirements. Here, we present the complete set of design goals for VATRA:

G1: Utilize a comparative visual approach. Given the known trade-offs [49, 59, 65], all interviewees confirmed that a comparative visual approach is needed to assess models on both datasets concurrently. E3 commented, *"Focusing on a single metric can be misleading, since many models perform well only in either standard or adversarial condition."* E2 agreed, *"Visually comparing is important to guide us in turning model parameters and seeing how much we can trade accuracy for robustness."*

G2: Visualize trade-offs at multiple levels of details. Many ATs demonstrate effectiveness by reporting their overall accuracy and robustness [47, 52, 65]. However, to assess real-world viability, trade-offs should be visualized at different granularities, as *"models with strong overall performance can still reveal trade-offs at specific instances or classes"* -E3. Additionally, all interviewees agreed that class-wise comparison should be specifically supported to assess fairness trade-offs [4, 5, 63]. E5 stated, *"When examining safety-critical classes like stop signs, if the visualization shows a fairness trade-off in misclassifying them under poor lighting, I would prioritize addressing it by training on poorly-lit signs."*

G3: Facilitate interactive exploration of trade-off causes. The experts believed that the visualizations should be augmented with interaction to support dynamic exploration. E4 explained, *"If users can interactively explore the embedding space pre- and post-attack, they can navigate to the most affected areas for a close-up to investigate trade-off reasons and be more informed."* E2 agreed, *"If users can interactively drill down into class overlaps between datasets, it can reveal subtle patterns and help users explore why certain class performances decline sharply compared to benchmark models."*

G4: Adapt to various evasion attacks, ATs, and image classification tasks. As many attacks [10, 11, 24, 35, 41] and ATs [24, 41, 48, 56, 65] exist, the experts emphasized having a visual framework that can be generalized to different attacks and defenses. E1 stated, *"To find the best method in our workflow, we compare various AT methods to see how each holds against perturbed data."* E1 and E2 also stressed that the framework should be generalizable to domain-specific image classification tasks. E2 explained, *"Certain images, such as medical ones, are much susceptible to attacks than others. Users should be able to explore their own datasets to better assess models' reliability in their use cases."*

4 VATRA

Here, we describe the details of our design probe, VATRA, guided by our empirical insights into effective comparative approaches (Section 3.3) and developed through the aforementioned design process (Section 3.1).

4.1 Design Probe Overview

VATRA is a visual analytics design probe divided into two system components: 1) a backend analytic pipeline (Section 4.2) and 2) a frontend user interface (Section 4.3).

The *backend analytic pipeline* generates adversarial examples and evaluates models for trade-off analysis. The *Perturber* module applies user-selected attacks and processes images for metric evaluation and embedding analysis (**G4**). The *Feature Analyzer* extracts embeddings from both inputs, using a dual DR approach combining *independent* and *conjoint* methods (Section 4.2.2). It also employs

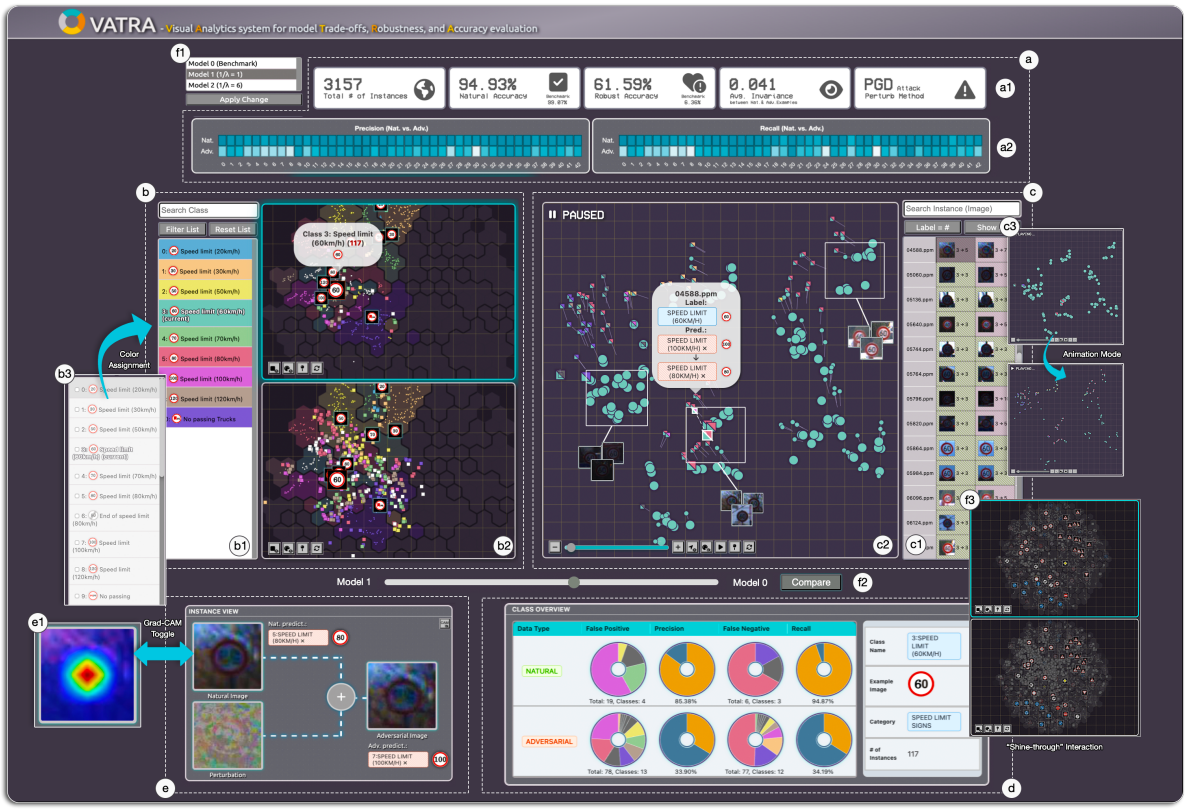

Figure 3: VATRA's interface, consisting of: (a) Summary View, a high-level metric comparison of model performance; (b) Global Dual Projection View, enabling the comparison of global embeddings between natural and adversarial datasets; (c) Class CoProjection View, allowing for class-wise comparison of local embedding relationships; (d) Class Overview Display, offering a statistical comparison of the current class; (e) Instance View, highlighting specific data instances; and (f) Model Comparison feature, which supports model comparison through interactive techniques.

Grad-CAM [51] to highlight key image regions and quantifies invariance to assess accuracy-robustness balance.

The *frontend user interface* includes the followings: 1) a *Summary View* (Figure 3a), 2) a *Global Dual Projection View* (Figure 3b), 3) a *Class CoProjection View* (Figure 3c), 4) a *Class Overview Display* (Figure 3d, 5) an *Instance View* (Figure 3e), and a 6) *Model Comparison feature* (Figure 3f). An overview of the interface components, along with their comparison levels, visual comparative designs, and a summary of their functionalities, is provided in Table 1.

4.2 Backend Analytic Pipeline

We first describe the backend's role in data preparation and analysis, including adversarial example generation, standard/adversarial evaluations, and embedding analysis for trade-off assessment.

4.2.1 Perturber. The Perturber generates adversarial data by conducting the chosen attack with user-defined parameters. The module first analyzes model predictions using standard inputs to extract perturbation-relevant information (e.g., gradients). It then adjusts pixels within ϵ bounds to generate adversarial examples aimed at deceiving the model. We adopted an attack-agnostic approach, allowing for generalization across different attacks (**G4**). Users can add new attacks by defining functions that adhere to the required interface, such as model parameters and input data.

While the Perturber can be swapped with different attack methodologies, here we use PGD [41] and SimBA [26], two well-known attacks (one white-box and one black-box), to demonstrate the design probe. We chose PGD because it is considered the strongest attack using local first order information [41], which is known to be effective against many classifiers [41, 59, 67] and frequently used to evaluate other attacks and defenses [48, 64, 65]. The second method, SimBA [26], is one of the most efficient black-box attacks and commonly taught in AML classes, as mentioned by our experts.

4.2.2 Feature Analyzer. This module evaluates models in the latent space by extracting, reducing, and comparing models' perceived features of standard and adversarial examples, and transforming them into interpretable representations. It maps each input (i.e., x, x_{adv}) to a latent vector by temporarily detaching the final output layer to extract embeddings, then applies users' choice of DR method (e.g., t-SNE [60], PCA [46]) to prepare them for 2-D visual interpretation. We apply reduction in a two-fold approach to support multi-level embedding comparison (**G2**): 1) *Independent reduction*, in which natural and adversarial datasets are processed separately

Interface Component	Level	Comparative Designs	Summary
Summary View	**C1**	Juxtaposition, Explicit Encoding	A comparison overview of model metrics w.r.t the benchmark; key dataset and attack information; class-wise precision + recall.
Global Dual Projection View	**C2**	Juxtaposition + Explicit Encoding	Reveals global embedding patterns; direct comparisons of the structural similarities and differences between nat. and adv. embeddings against the decision boundaries.
Class CoProjection View	**C3**	Juxtaposition, Superposition + Explicit Encoding	Comparison of dynamics and trajectories between nat. and adv. embeddings within the same class in a unified projection space.
Class Overview Display	**C4**	Juxtaposition	Statistical metric comparison of the class currently investigated by the user between nat. and adv. datasets.
Instance View	**C5**	Juxtaposition + Superposition, Explicit Encoding	Provides comparison of raw images and visual feature explanations on the selected instance.
Model Comparison	**C6**	Juxtaposition, Superposition, Explicit Encoding	Model comparison of performance metrics via "blinking" [57]; comparison of dataset- and class-level embeddings between models via "shine-through" interaction [57].

Table 1: An overview of VATRA's interface components.

to analyze global structural patterns, and 2) *Conjoint reduction*, in which both datasets' instances within a class are reduced together in a combined feature space.

To quantify how closely adversarial examples mimic the natural inputs, the analyzer also calculates the average invariance between natural and adversarial examples, i.e., $\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{v}_i - \mathbf{v}_{\text{adv},i}\right\|^2$. Here, $\mathbf{v}_i$ represents the latent space vector of the i-th natural input, and $\mathbf{v}_{\text{adv},i}$ represents the latent space vector of the corresponding adversarial example. This provides a metric for the model's degree of balance between accuracy and robustness. A *Grad-CAM* component computes gradients of the output class w.r.t key feature layers, producing a localization map highlighting important image regions for its prediction.

4.3 Frontend User Interface

Here, we detail the frontend interface, which includes several interactive components designed to support exploration and comparison of model trade-offs through multi-level comparative visualizations.

*4.3.1 Summary View (Figure 3a) - **C1**.* This view utilizes *juxtaposition* and *explicit encoding* to compare the current model's high-level performance statistics w.r.t the benchmark, and presents key dataset and attack information. From left to right in Figure 3a1, the view presents total data instances, juxtaposed model accuracy and robustness, explicitly encoded embedding invariance, and the attack method. Through attention shifting across metrics, users can compare between models to understand the high-level accuracy-robustness trade-off patterns (**G1, G2**).

The two matrices (Figure 3a2) utilize *juxtaposition* with color-coding to display class-wise *precision* and *recall* in both conditions (**G2**). The rows are juxtaposed, with the first showing natural dataset metrics and the second reflecting adversarial dataset metrics (**G1**). The value at ith column represents the cell's class value, and the color intensity witin each cell indicates the magnitude. This visual comparison offers an at-a-glance view of variations in color distribution, enabling quick identification of vulnerable classes for further inspection (**G1, G2**).

*4.3.2 Global Dual Projection View (Figure 3b) - **C2**.* This view has a selection panel and two *juxtaposed* scatterplots mixed with *explicit encoding* to support comparison of global embeddings (**G1**). It is designed to support exploration of trade-off causes in separate latent coordinate systems (**G2, G3**).

Class Selection Panel (Figure 3b1). This panel lets users select specific classes for analysis. Clicking "Filter List" reveals available classes, displaying all embedding points in neutral gray. Users can assign distinct colors to up to 12 classes [43], and consistent colors are ensured across all views for unified comparison.

Dual Scatterplot View (Figure 3b2). This view includes two side-by-side scatterplots with linked highlighting (i.e., *juxtaposition* + *explicit encoding*) to support comparison of global embeddings from both datasets (**G1**). The upper scatterplot displays the global structure of natural data, while the lower plot shows the global adversarial structure. When users highlight a cluster in one scatterplot, the same instances are automatically highlighted in the other. Correct predictions are rendered as circles and incorrect predictions as squares. Users can interactively zoom and pan to highlight specific points or clusters, maintaining mental models of each dataset's context while comparing their structural trade-offs (**G3**).

Users are provided with two navigation methods: 1) instance-by-instance by selecting any point, and 2) class-by-class by clicking on a class centroid (**G2**). A hexagonal binning map is integrated in the backgrounds, which depicts the model's estimated decision boundaries (i.e., abstraction as a form of *explicit encoding* [23]). We introduced it based on our experts' suggestions to help users infer the predicted class of instances by examining colors of surrounding hexagons. By maintaining a consistent hexbin map within both natural and adversarial projections, the view enables more explicit comparisons of how embedding distributions shift in response to an attack (**G1, G3**). This comparison reveals data migration between decision regions, informing trade-off mitigation such as prioritizing training for classes more susceptible to adversarial shifts.

*4.3.3 Class CoProjection View (Figure 3c) - **C3**.* This view leverages *juxtaposition* and a hybrid of *superposition* + *explicit encoding* to support exploration of models' class perception at a more fine-grained level (**G1, G2**).

Instance Selection Panel (Figure 3c1). This panel updates with instances from the selected class, displaying rows with three *juxtaposed* columns: image name, a natural image with its label and prediction, and an adversarial image with its updated prediction (**G1**). Colored hatching indicates prediction correctness—green for correct, red for incorrect. *Juxtaposition* with color-hatching enables quick comparison of instance classifications and supports efficient instance navigation based on model correctness.

CoProjection Scatterplot View (Figure 3c2). This view supports comparison of local relationships between natural and adversarial embeddings within a class (**G2**). A hybrid design of *superposition + explicit encoding* is employed to project both datasets into the same spatial context with visual cues and animations to explicitly highlight instance movements (**G1**). The view also differentiates correct and misclassified instances using circles and squares. Each square's left half indicates its label, while the right half shows its misclassification. Users can choose either dataset as the "foreground," where larger points highlight it as the primary investigation focus, while the "background" dataset is displayed with smaller points for contrast. Each background instance has an *explicitly encoded* tail tracing its movement from the foreground (**G1**). When a point is highlighted, an animated line connects it to its counterpart, visualizing embedding shifts between datasets to identify class-level trade-off causes (**G1, G3**). An animation highlights background instances moving into the foreground (a form of *explicit encoding*; Figure 3c3), showing class migration within the projection space during an attack (**G2, G3**).

*4.3.4 Class Overview Display (Figure 3d) - **C4**.* This view offers a *juxtaposed* metric summary of the class being investigated (**G2**). The right column displays the class name, an example image, class category, and the number of class instances. The left column features a series of juxtaposed doughnut charts that summarize the model's class metrics under both conditions (**G1**). This column has two rows displaying natural and adversarial class metrics. It displays false positives, precision, false negatives, and recall from left to right, enabling class comparison through linked visual elements without cognitive overload.

*4.3.5 Instance View (Figure 3e) - **C5**.* This view utilizes a hybrid design of *juxtaposition + superposition* and *explicit encoding* to provide detailed information on an instance (**G2**). It depicts the attack process by juxtaposing the followings: the natural image and the applied noise (i.e., explicitly encoded image difference), the model's original prediction and its correctness, the adversarial image, and the model's new prediction and its correctness. Beyond comparing pixel differences and model correctness, users can toggle *Grad-CAM* (Figure 3e1) to compare heatmaps of significant image features for the two predictions (**G1**), overlaid on the original images. By juxtaposing the model's decision-making for individual instances with superimposed heatmaps, this view helps assess trade-offs on a case-by-case basis, informing users' strategies to balance robustness, accuracy, and fairness across different instances.

*4.3.6 Model Comparison (Figure 3f) - **C6**.* To support trade-off comparisons between models across **C1**∼**C5**, we incorporated an interaction feature designed specifically for model comparison. Based on our design considerations for **C6** (Section 3.3), this feature combines all three comparative designs: *juxtaposition*, *superposition*, and *explicit encoding*. It includes two key comparative interactions. The first is the "blink" interaction [23], which allows users to toggle between models while retaining the selection focus of a specific class or instance. This allows for quick comparisons of simple metrics (e.g., overall accuracy/robustness, class precision/recall) by seamlessly alternating between models. For comparing more complex objects like embeddings, the second interaction, i.e., the "shine-through" feature [57], offers a transparency control mechanism. This lets user overlay the embedding projections and control the visibility of each model's embeddings with a slider, allowing them to retain context and reducing the cognitive load of attention shifting between models. Together, these interactions offer a flexible and intuitive way to compare models without leading to visual clutter.

5 User Studies

To explore how the comparative visualizations from VATRA can be integrated into the workflows from different ML application domains, we conducted 90-minute interviews with six experts (E6 ∼ E11). We developed two comprehensive use cases and provide empirical insights into how AML practitioners can leverage our comparative visual approaches for trade-off analysis. We aim to answer the following questions:

RQ1 How do AML practitioners utilize comparative visual designs to effectively gather trade-off insights?

RQ2 How do these comparative designs enhance existing workflows for AML model evaluation and improvement?

Each interview began with an introduction to the research and a tutorial on the design probe. Experts received access to VATRA along with tasks and scenarios in PDF format for free-form interaction. They explored classifiers with the same architecture but varying AT levels, assessing performance under both conditions to understand trade-offs (**T1**). Initial tasks familiarized them with the interface, but they were encouraged to identify as many trade-off insights and their causes as they could (**T2**). Based on findings, they either selected the best classifier for their application (**T3**) or explored ways to improve models and mitigate trade-offs (**T4**). A think-aloud protocol was used, with an experimenter taking notes and assisting with technical questions. Sessions concluded with a follow-up interview for additional feedback. Participants were compensated $20/hour.

5.1 RQ1: Insight Gathering

Entry model identification. AML practitioners begin their insight gathering by first identifying an entry model as their analysis starting point through high-level visual comparisons. Practitioners typically start with the Summary View (Figure 3a), which presents overall model accuracy and robustness w.r.t the benchmark model. Occasionally, they complement this by utilizing the Global Dual Projection View (Figure 3b), where they toggle between models and rely on their memory to compare differences in the global embedding distribution across models. Though VATRA offers a "shine-through" feature (Figure 3f2) for superimposing model distributions, we found that practitioners mostly use it to examine more fine-grained distribution patterns. When identifying an entry model, practitioners find model toggling and viewing juxtaposed scatterplots sufficient, as they only require a high-level overview of the distribution at this stage.

Back-and-forth insight refinement. After identifying an entry model, practitioners proceed to refine their insights iteratively by leveraging a combination of views. This phase involves transitioning between the Global Dual Projection View, Class CoProjection View (Figure 3c), and Instance View (Figure 3e), depending on

the nature of the insight being pursued. Practitioners often switch back and forth between these views, using the Global Dual Projection View to compare overall patterns or shifts in data distributions, the Class CoProjection View to analyze feature-specific relationships, and the Instance View to compare natural and adversarial instances. This back-and-forth comparison enables practitioners to balance high-level overviews with targeted deep dives to uncover meaningful insights into model performance. For example, practitioners may begin by comparing the overall distribution to identify clusters of commonly misclassified instances. Then, they transition to the Class CoProjection View to examine how close the instances are to each other within the shared feature space. From there, they use the Instance View to compare specific instances and investigate the correlation between spatiality and semantic features, such as lighting, colors, or textures, that may explain why the model struggles with these instances.

Guidance for next steps. Some comparative designs in VATRA are not primarily used for direct insight gathering, but instead serve as guidance for determining the next steps in the analysis workflow. For example, from our interviews, the experts agreed that the Class Overview display (Figure 3d) provides useful insights into the common classes that the current class tends to be confused with. However, in practice, we observed that they rely on this view the least for direct insight gathering. Instead, the practitioners use it to guide their strategy after completing a round of insight gathering with other views. For instance, once they identify an entry model and gather insights using the embedding and instance comparison views, the Class Overview helps them decide which class or instances to analyze next using the same comparative workflow. Similarly, the Instance Selection Panel (Figure 3c1) within the Class CoProjection View plays a guiding role. When loading a class into the Class CoProjection View, practitioners rely on the juxtaposed color differences in the Instance Selection Panel to inform their next steps, i.e., which instances to focus on next. These guiding views help practitioners navigate their analysis more effectively to ensure a more structured approach to gathering trade-off insights.

5.2 RQ2: Workflow Enhancement

Refining and structuring workflows. From our user interviews, we identified that the workflows of AML practitioners generally follow a high-level structure: first, assessing the extent of a trade-off (**T1**); next, gathering insights and understanding the causes behind these trade-offs (**T2**); and finally, deciding on actions such as selecting the best model for their scenario (**T3**) or identifying areas for improvement (**T4**). While VATRA's comparative visual designs align with this general workflow, they also help refine and structure practitioners' approaches, making their workflows more defined and goal-oriented. In traditional workflows, practitioners often lacked a fixed starting point or clear strategy, with insight gathering varying greatly depending on individual preferences. Typically, they would begin with high-level metrics like accuracy and robustness, but proceed in a more open-ended manner, often unsure about their specific goals. In contrast, with VATRA, practitioners developed workflows that were both more structured and flexible, adapting to their objectives while leveraging its comparative features.

For instance, E6 began their workflow by selecting traffic sign classes they deemed most critical based on their background knowledge, independently of other views. From this starting point, they gathered insights and worked their way back to compare additional classes relevant to these critical signs. Meanwhile, E7 used the Global Dual Projection View to identify classes that appeared more vulnerable and focused their analysis on these areas. Others leveraged embedding distributions to select classes that were visually or semantically similar for further exploration. These examples suggest that VATRA's comparative designs not only align with the high-level workflow logic of AML model evaluation, but also guide practitioners in creating workflows that are structured, goal-driven, and tailored to their specific needs.

Introducing an iterative workflow. Another way VATRA enhances existing workflows is by facilitating an iterative approach, which contrasts with the more linear workflows practitioners described in their prior processes. In our initial interviews, many practitioners mentioned that their existing workflows typically involve looking at high-level statistics, such as overall accuracy and robustness, followed by generating some graphs to gather trade-off insights. However, with VATRA, we observed that practitioners' workflows became more iterative, involving repeated transitions between views after completing the initial high-level comparison (**T1**). For example, practitioners would perform tasks **T2**, **T3**, and **T4** through a dynamic and iterative workflow. In the most straightforward scenario, this iterative process followed a high-level-to-low-level structure. Practitioners began with *"high-level metrics to identify trade-offs, then move down into lower-level views for embedding and instance comparisons"* -E8, and finally *"returned to higher levels to iteratively refine the analysis and repeat the process."* -E11. However, we also observed workflows that did not follow this high-to-low structure. In some cases, practitioners began their analysis with the global distribution to identify misclassified instances directly, then jumped straight to the instance view. From there, they selected classes to investigate further based on interesting patterns or outliers observed at the instance level. This flexibility suggests that VATRA's comparative designs could adapt to the unique needs and goals of practitioners, allowing them to iteratively explore trade-offs and uncover insights at multiple levels of detail.

5.3 Case Study #1: Traffic Sign Recognition

This case study involves three researchers (E6 ~ E8; all men) with ML/CV backgrounds in autonomous driving and traffic modeling. E6 and E7 have six years of ML experience, while E8 has five. E6 has two years of AML experience, E8 has one, and E7 is familiar with AML concepts. Domain-wise, E6 specializes in traffic sign recognition and integrating image-LiDAR models for pedestrian and vehicle detection. E7 researches CV and image processing for autonomous driving, and E8 specializes in ML-driven traffic modeling. We loaded VATRA with three ResNet-101 models trained with various degrees of TRADES [65], a SOTA AT method recommended by our AML experts, on the German traffic sign recognition benchmark (GTSRB) dataset [54], and used PGD [41] as our attack. Based on our observations and interviews, we distilled a use case to provide insights on how practitioners can leverage comparative visualizations to analyze model trade-offs for sign recognition.

Understanding model-level overview. The expert wants to compare three classifiers: *Model 0*, a standard model, *Model 2*, a model trained with TRADES ($1/\lambda = 1$), and *Model 3* trained with a higher degree of TRADES ($1/\lambda = 6$). *Model 0* is first selected to explore its trade-offs when no AT is applied. From the Summary View, juxtaposed metrics direct the expert's attention between two contrasting values, revealing that *Model 0* has a high accuracy of 99.07%, but low robustness of 6.36%. Concerned about its safety, the expert toggles the other models with the "blink" interaction. By focusing on the alternating metrics superimposed in the same view, the expert discerns with little memory load that *Model 1* achieves significantly higher robustness (61.59%), despite relatively lower accuracy (94.93%) (Figure 3a1). Meanwhile, *Model 2* exhibits marginally better robustness (62.79%) but much lower accuracy (88.65%). The expert recognizes that the accuracy-robustness trade-off persists even with SOTA methods like TRADES. Finding *Model 1* the most promising yet hoping to alleviate the trade-offs, the expert decides to take a closer look at its embeddings.

Exploring global-level embeddings. The expert navigates to the Global Dual Projection View (Figure 3b) to display the embedding distributions of all 43 signs. By carefully adjusting the "shine-through" (Figure 3f2) back and forth to compare *Model 0* and *Model 1*'s global embeddings (Figure 3f3), the expert manages to retain both data distributions in their mental model, making several observations. First, the juxtaposed embeddings of *Model 1* highlight a similar layout under both conditions. This suggests that *Model 1* perceives both datasets as alike, contributing to its robustness. Second, since juxtaposition preserves the context of the original embeddings, the expert knows that *Model 1* relies more on human-interpretable features for prediction, as *"speed limit signs cluster at the top left, triangular signs group on the right, and blue signs concentrate at the bottom."* -E8. However, by skimming across the class matrices (Figure 3a2), the significant variations in color intensity between juxtaposed squares show that *"Model 1 experiences a significant fairness trade-off among speed limit signs."* -E6. Linked highlighting reveals the issue in both adversarial and natural embeddings, leading practitioners to suspect it as the cause of the accuracy trade-off. The expert assigns each speed limit sign an unique color to focus the analysis more on them (Figure 3b3).

Investigating class-level local relationships. With non-speed limit signs hidden, the expert pans and zooms in on the scatterplots to focus solely on speed limit sign clusters. The expert begins with "Class 3: Speed Limit (60km/h)" by clicking on its natural class centroid. The Class Overview Display (Figure 3d) and Class CoProjection View (Figure 3c) updates, with natural embeddings as the "foreground" and adversarial embeddings as the "background." Superimposed projections enable comparison of structural shifts, revealing that adversarial data from Class 3 form distinct smaller sub-clusters. Observing the explicitly encoded trajectories, the expert also notices that "after" cluster data points were previously scattered throughout the "before" cluster. Aiming to identify the patterns behind the clustering, the expert selects individual data points to examine their actual natural and adversarial images.

Examining instance-level trade-off causes. From the Instance View, the expert notices that instances within the same "after" cluster tend to share similar lightings (Figure 3c2). For example, within the natural dataset, *"all misclassifications share a common bluish tint with no direct sunlight."* -E7. The expert therefore considers using data augmentations or generative AI to simulate various lighting conditions to enhance model robustness. The expert also notes that under adversarial conditions, *"PGD modifies the semantic visual features of sign numbers in Model 1"* -E7—a departure from its usual tactic of adding imperceptible noise. To address the severe fairness trade-off in speed limit signs, the expert considers using semantic integrity checks to look for inconsistencies in identifiable attributes such as shape, size, or colors. By examining the local similarities among instances and their actual appearances, the expert continues to gather new insights, aiding in the development of new solutions to enhance the model.

5.4 Case Study #2: Skin Lesion Recognition

This case study involves three researchers (E9 ~ E11; all women) in healthcare CV. E9 is knowledgeable about AML through past courses and projects; E10 and E11 is familiar with the concepts from reading AML literature. E9 collaborates with hospitals to apply vision models for tumor classification in medical videos. E10 researches CV and image processing in healthcare. E11 specializes in CV-based disease prediction and neuroscience data visualization. We loaded VATRA with three VGG-16s trained with different degrees of robust self-training (RST) [48], another SOTA AT method [64], and used the HAM10000 skin lesion dataset [58] with SimBA attack [26]. We devised another use case within the context of healthcare CV to further demonstrate the workflow of our augmented comparative approach.

Observing global shift across decision boundaries. An expert aims to investigate three VGG-16s: the standard *Model 0*, *Model 1* trained with RST ($\beta = 0.5$), and *Model 2* trained with RST ($\beta = 4$). They begin with *Model 0* to assess standard training, noting from the Summary View that it achieves 84.28% accuracy but only 5% robustness. Intrigued, they navigate to the Global Dual Projection View, and examine global embeddings to explore reasons behind the trade-off. By explicitly highlighting classes in the linked juxtaposed scatterplots, they note a pronounced shift: under attack, many instances of "Class 4: Melanocytic Nevi" move to the border of the prediction region, crossing the decision boundary and approaching other class clusters (Figure 4a). *"This shift results in the misclassification of most melanocytic nevi in adversarial conditions."* -E10.

Identifying causes of distribution change. To investigate the reason for the shift, the expert clicks on the centroid of melanocytic nevi, which populates the Class CoProjection View. Upon examining the distance of explicitly encoded links between the foreground and background data, and comparing the overlaid heatmaps in the Instance View, they note that *Model 0*'s attention to image features dramatically shifts to different areas post-perturbation. Yet, even without perturbation, the superimposed heatmaps show that the model's focus does not consistently align with the actual locations of skin lesions in the image (Figure 4b). The expert realizes that *Model 0* relies on non-human-interpretable features rather than actual visual characteristics of skin lesions—*"a mistake that could be deemed critical from a medical professional's perspective."* -E9

Analyzing trade-offs from data imbalance. Discontent with *Model 0*, the expert shifts focus to *Model 1*. They note from juxtaposed views that *Model 1* exhibits lower accuracy (79.28%) but much

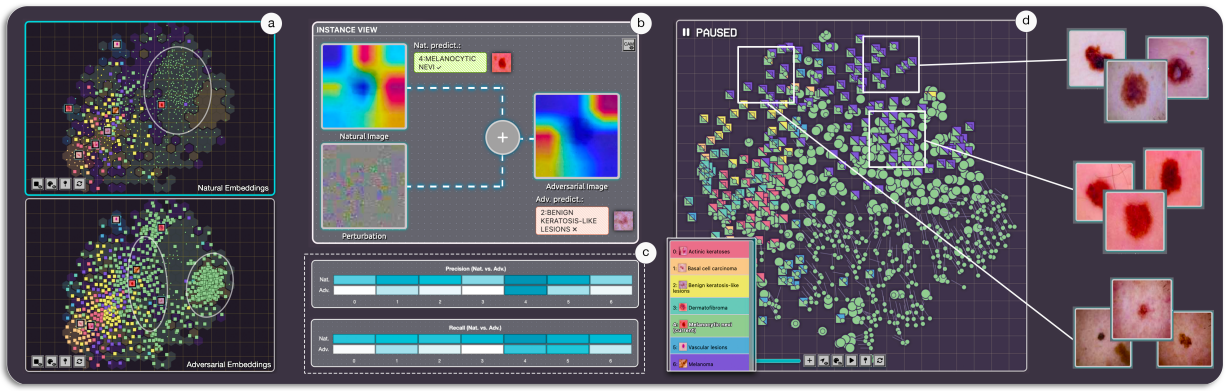

Figure 4: A healthcare CV expert leverages the comparative visualizations from VATRA to explore and compare trade-offs of three VGG-16s trained with varying degrees of RST.

improved robustness (40.57%). Nonetheless, the inconsistent color distribution in the class matrices indicates a pronounced fairness trade-off (Figure 4c). Specifically, the model demonstrates much higher precision/recall for melanocytic nevi compared to other classes. The expert returns to the Global Dual Projection View to explore these trade-offs. From the scatterplots, the expert observes a unique property of this class: a clear imbalance, with melanocytic nevi having significantly more instances than other classes. The test set, sampled from the whole dataset, reveals class imbalance in training, causing the model to *"excel in classifying melanocytic nevi but struggle with other classes due to insufficient data."* -E9. The expert realizes that to reduce the fairness trade-off of *Model 1*, several approaches could be adopted. These include cost-sensitive learning to penalize minority class misclassifications and data augmentation to expand underrepresented classes.

Investigating trade-off causes case by case. To explore the accuracy-robustness trade-off in *Models 1 & 2*, the expert decides to examine the local embeddings more closely, starting with melanocytic nevi. As the expert alternates between *Models 1 & 2* with the "blink" interaction, they notice from the superimposed Class Overview Display that the more robust the model becomes, more standard instances are misclassified as "Class 6: Melanoma." Reviewing images often misclassified as melanoma, the expert notes their similar lesion shapes and colors (Figure 4d). Thus, beyond generic data augmentation for class imbalance, targeted augmentations can generate diverse melanocytic nevi and melanoma examples, emphasizing shape and color differences. The expert continues their investigation by analyzing the distributions of various classes, gaining new insights to address the model's trade-offs during the process.

6 Discussion

Here, We discuss emerging themes from our study, providing additional insights and design implications, and outline study limitations with future research directions.

Interaction in comparative visual design. From prototype walkthroughs and user studies, we show that interaction plays a dual role in visual comparison: it can assist users to better perform existing comparison tasks (**C1**~**C5**), but also addresses issues inherent in comparative visual designs (**C6**). One example of using interaction to support existing comparative designs is linked highlighting, a form of explicit encoding powered by interaction. On the other hand, interaction techniques like "blinking" and "shine-through" [23, 57] help mitigate the cognitive overload and visual clutter in traditional comparative designs. While most visual analytics [1, 3, 12, 23] focus on spatial comparisons, we argue that time-based comparisons remain underexplored and can be effectively achieved through interaction. By incorporating temporal elements like blinking or shining through objects in a sequential order, this addresses limitations in space-based design and adds another layer of visual comparison without cluttering the interface. Further, though the original taxonomy of comparative visual design [23] stated that a hybrid of all three designs (i.e., *juxtaposition*, *superposition*, and *explicit encoding*) was possible but not encountered, our findings suggest that this can be realized through the integration of interaction. By leveraging interactions, users can combine elements of comparative strategies within one design to achieve a more comprehensive comparison, as shown in VATRA's Model Comparison feature (Figure 3f).

Key design learnings beyond AML. The development and evaluation of VATRA revealed several design principles that can be applied to comparative workflows for AML and beyond. First, VATRA demonstrates the importance of balancing structure and flexibility. While its design provides a structured framework for practitioners to follow, it allows them to adapt their workflows to specific goals and insights. Second, the iterative nature of VATRA's workflow proved valuable for insight gathering, as it enables users to transition easily between views and revisit earlier steps to refine their understanding. Third, we observed that some comparative designs, such as the Class Overview (Figure 3d) and Instance Selection Panel (Figure 3c1), were particularly effective as guidance, helping users determine their next steps rather than directly serving as views to collect knowledge. Lastly, VATRA's ability to integrate insights across multiple levels, from high-level summaries to instance-level comparisons, helped practitioners connect global patterns with specific examples. These design principles not only enhance AML workflows but also provide a foundation for creating effective visualization tools in other domains requiring iterative exploration and decision-making.

Limitations and future work. One limitation of our study is the sample size. A larger sample of ML experts from diverse domains would provide deeper insights into how our comparative approach could integrate into diverse workflows. Another potential concern is scalability. While our system has proven effective with large datasets, recent deep-learning models trained on billions of images pose a challenge. To visualize such large-scale data, we might need additional approaches like binned aggregation and hierarchical clustering [38]. Additionally, our study focuses primarily on evasion attacks, without fully exploring other attack types, such as data poisoning. Future exploration of comparative designs for trade-off analysis under these attack types could provide design implications for visual analytics in other AML sub-domains. While we explored various interaction techniques to enhance comparative designs, the system's multi-layered design and complex interactions could present a steep learning curve for less experienced users. Additional user guidance or simplified modes for non-experts might improve accessibility. While we explored various interaction techniques to enhance comparative designs, other methods, such as analytical and statistical *automatic comparisons*, could further augment visual comparison [23]. Future work could examine how integrating these techniques might better support or improve comparative trade-off exploration. Finally, although this work centers on comparative approaches for trade-off exploration, other aspects of visual design, such as color schemes and layout configurations, could also be explored to enhance AML visual analytics. Future research could investigate how these visual elements contribute to or hinder tasks in trade-off analysis.

7 Conclusion

We have explored effective comparative visual designs for trade-off analysis through a design study with 11 experts. First, an iterative design process with five AML experts generated insights for our development of a visual analytics design probe, VATRA, which employs an augmented hybrid comparative design to support multi-level comparison of model trade-offs. Moreover, a user study with six ML experts from two application domains derived two in-depth use cases of VATRA, providing empirical insights into how ML practitioners can effectively leverage comparative visualizations to analyze model trade-offs, informing design considerations for future AML visual analytics.

Acknowledgments

This work is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant #RGPIN-2020-03966, the Canada Foundation for Innovation (CFI) John R. Evans Leaders Fund (JELF) #42371, and the University of Waterloo Interdisciplinary Trailblazer Grant. We acknowledge that much of our work takes place on the traditional territory of the Neutral, Anishinaabeg, and Haudenosaunee peoples. Our main campus is situated on the Haldimand Tract, the land granted to the Six Nations that includes six miles on each side of the Grand River.

References

[1] Danielle Albers, Colin Dewey, and Michael Gleicher. 2011. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2392–2401. https://doi.org/10.1109/TVCG.2011.232

[2] Motasem Alfarra, Juan C Pérez, Ali Thabet, Adel Bibi, Philip HS Torr, and Bernard Ghanem. 2022. Combating adversaries with anti-adversaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5992–6000. https://doi.org/10.1609/aaai.v36i6.20545

[3] Basak Alper, Benjamin Bach, Nathalie Henry Riche, Tobias Isenberg, and Jean-Daniel Fekete. 2013. Weighted graph comparison techniques for brain connectivity analysis. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 483–492. https://doi.org/10.1145/2470654.2470724

[4] Philipp Benz, Chaoning Zhang, Soomin Ham, Adil Karjauv, Gyusang Cho, and In So Kweon. 2021. Trade-off between accuracy, robustness, and fairness of deep classifiers. In *Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges*.

[5] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. 2021. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*. PMLR, 325–342. https://proceedings.mlr.press/v148/benz21a.html

[6] Ulrik Brandes, Tim Dwyer, and Falk Schreiber. 2003. Visualizing related metabolic pathways in two and a half dimensions. In *International Symposium on Graph Drawing*. Springer, 111–122. https://doi.org/10.1007/978-3-540-24595-7_10

[7] Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. 2019. Adversarial examples from computational constraints. In *International Conference on Machine Learning*. PMLR, 831–840. https://proceedings.mlr.press/v97/bubeck19a

[8] Kelei Cao, Mengchen Liu, Hang Su, Jing Wu, Jun Zhu, and Shixia Liu. 2020. Analyzing the noise robustness of deep neural networks. *IEEE transactions on visualization and computer graphics* 27, 7 (2020), 3289–3304. https://doi.org/10.1109/TVCG.2020.2969185

[9] Tanmay Chakraborty, Utkarsh Trehan, Khawla Mallat, and Jean-Luc Dugelay. 2022. Generalizing adversarial explanations with Grad-CAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 187–193. https://doi.org/10.1109/CVPRW56347.2022.00031

[10] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. 2020. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*. IEEE, 1277–1294. https://doi.org/10.1109/SP40000.2020.00045

[11] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 15–26. https://doi.org/10.1145/3128572.3140448

[12] Michael Correll, Adam L Bailey, Alper Sarikaya, David H O'Connor, and Michael Gleicher. 2015. LayerCake: a tool for the visual comparison of viral deep sequencing data. *Bioinformatics* 31, 21 (2015), 3522–3528. https://doi.org/10.1093/bioinformatics/btv407

[13] Pier Francesco Cortese, Giuseppe Di Battista, Antonello Moneta, Maurizio Patrignani, and Maurizio Pizzonia. 2006. Topographic visualization of prefix propagation in the internet. *IEEE transactions on visualization and computer graphics* 12, 5 (2006), 725–732. https://doi.org/10.1109/TVCG.2006.185

[14] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2020. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670* (2020). https://doi.org/10.48550/arXiv.2010.09670

[15] Aaron CE Darling, Bob Mau, Frederick R Blattner, and Nicole T Perna. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research* 14, 7 (2004), 1394–1403. https://doi.org/10.1101/gr.2289704

[16] Nilaksh Das, Haekyu Park, Zijie J Wang, Fred Hohman, Robert Firstman, Emily Rogers, and Duen Horng Polo Chau. 2020. Bluff: Interactively deciphering adversarial attacks on deep neural networks. In *2020 IEEE Visualization Conference (VIS)*. IEEE, 271–275. https://doi.org/10.1109/VIS47514.2020.00061

[17] Mark C Detweiler. 1991. Envisioning Information. *Cartographic Perspectives* 10 (1991), 22–24. https://doi.org/10.14714/CP10.1055

[18] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118. https://doi.org/10.1038/nature21056

[19] Takanori Fujiwara, Xinhai Wei, Jian Zhao, and Kwan-Liu Ma. 2021. Interactive dimensionality reduction for comparative analysis. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 758–768. https://doi.org/10.1109/TVCG.2021.3114807

[20] Takanori Fujiwara, Jian Zhao, Francine Chen, and Kwan-Liu Ma. 2020. A visual analytics framework for contrastive network analysis. In *2020 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 48–59. https://doi.org/10.1109/VAST50239.2020.00010

[21] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. 2018. Adversarial spheres. *arXiv preprint arXiv:1801.02774* (2018). https://doi.org/10.48550/arXiv.1801.02774

[22] Michael Gleicher. 2017. Considerations for visualizing comparison. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 413–423. https://doi.org/10.1109/TVCG.2017.2744199

[23] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D Hansen, and Jonathan C Roberts. 2011. Visual comparison for information visualization. *Information Visualization* 10, 4 (2011), 289–309. https://doi.org/10.1177/1473871611416549

[24] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014). https://doi.org/10.48550/arXiv.1412.6572

[25] Liang Gou, Lincan Zou, Nanxiang Li, Michael Hofmann, Arvind Kumar Shekar, Axel Wendt, and Liu Ren. 2020. VATLD: A visual analytics system to assess, understand and improve traffic light detection. *IEEE transactions on visualization and computer graphics* 27, 2 (2020), 261–271. https://doi.org/10.1109/TVCG.2020.3030350

[26] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. 2019. Simple black-box adversarial attacks. In *International Conference on Machine Learning*. PMLR, 2484–2493.

[27] Guodong Guo and Na Zhang. 2019. A survey on deep learning based face recognition. *Computer vision and image understanding* 189 (2019), 102805. https://doi.org/10.1016/j.cviu.2019.102805

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034. https://doi.org/10.1109/ICCV.2015.123

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[30] Yuzheng Hu, Fan Wu, Hongyang Zhang, and Han Zhao. 2023. Understanding the Impact of Adversarial Robustness on Accuracy Disparity. In *International Conference on Machine Learning*. PMLR, 13679–13709. https://proceedings.mlr.press/v202/hu23j.html

[31] Yansong Huang, Zherui Zhang, Ao Jiao, Yuxin Ma, and Ran Cheng. 2023. A Comparative Visual Analytics Framework for Evaluating Evolutionary Processes in Multi-objective Optimization. *IEEE Transactions on Visualization and Computer Graphics* (2023). https://doi.org/10.1109/TVCG.2023.3326921

[32] Hyoeun Kang, Howon Kim, et al. 2021. Robust adversarial attack against explainable deep classification models based on adversarial images with different patch sizes and perturbation ratios. *IEEE Access* 9 (2021), 133049–133061. https://doi.org/10.1109/ACCESS.2021.3115764

[33] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90. https://doi.org/10.1145/3065386

[35] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*. Chapman and Hall/CRC, 99–112.

[36] Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. 2020. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems* 22, 2 (2020), 712–733. https://doi.org/10.1109/TITS.2019.2962338

[37] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[38] Lauro Lins, James T Klosowski, and Carlos Scheidegger. 2013. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2456–2465. https://doi.org/10.1109/TVCG.2013.179

[39] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770* (2016). https://doi.org/10.48550/arXiv.1611.02770

[40] Yuxin Ma, Tiankai Xie, Jundong Li, and Ross Maciejewski. 2019. Explaining vulnerabilities to adversarial machine learning through visual analytics. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1075–1085. https://doi.org/10.1109/TVCG.2019.2934631

[41] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017). https://doi.org/10.48550/arXiv.1706.06083

[42] Jian-Xun Mi, Xu-Dong Wang, Li-Fang Zhou, and Kun Cheng. 2023. Adversarial examples based on object detection tasks: A survey. *Neurocomputing* 519 (2023), 114–126. https://doi.org/10.1016/j.neucom.2022.10.046

[43] Tamara Munzner. 2014. *Visualization analysis and design*. CRC press.

[44] Andrew P Norton and Yanjun Qi. 2017. Adversarial-Playground: A visualization suite showing how adversarial examples fool deep learning. In *2017 IEEE symposium on visualization for cyber security (VizSec)*. IEEE, 1–4. https://doi.org/10.1109/VIZSEC.2017.8062202

[45] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. 506–519. https://doi.org/10.1145/3052973.3053009

[46] Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2, 11 (1901), 559–572. https://doi.org/10.1080/14786440109462720

[47] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. 2019. Adversarial robustness through local linearization. *Advances in Neural Information Processing Systems* 32 (2019). https://proceedings.neurips.cc/paper/2019/hash/0defd533d51ed0a10c5c9dbf93ee78a5-Abstract.html

[48] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. 2020. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716* (2020). https://doi.org/10.48550/arXiv.2002.10716

[49] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. 2019. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032* (2019). https://doi.org/10.48550/arXiv.1906.06032

[50] Natascha Sauber, Holger Theisel, and Hans-Peter Seidel. 2006. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE Transactions on Visualization and Computer Graphics* 12, 5 (2006), 917–924. https://doi.org/10.1109/TVCG.2006.165

[51] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626. https://doi.org/10.1109/ICCV.2017.74

[52] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *Advances in Neural Information Processing Systems* 32 (2019). https://proceedings.neurips.cc/paper_files/paper/2019/hash/7503cfacd12053d309b6bed5c89de212-Abstract.html

[53] Stefan Sietzen, Mathias Lechner, Judy Borowski, Ramin Hasani, and Manuela Waldner. 2021. Interactive analysis of cnn robustness. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 253–264. https://doi.org/10.1111/cgf.14418

[54] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks* 0 (2012), –. https://doi.org/10.1016/j.neunet.2012.02.016

[55] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[56] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013). https://doi.org/10.48550/arXiv.1312.6199

[57] Christian Tominski, Camilla Forsell, and Jimmy Johansson. 2012. Interaction support for visual comparison inspired by natural behavior. *IEEE Transactions on visualization and computer graphics* 18, 12 (2012), 2719–2728. https://doi.org/10.1109/TVCG.2012.237

[58] Philipp Tschandl. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. https://doi.org/10.7910/DVN/DBW86T

[59] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2018. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152* (2018). https://doi.org/10.48550/arXiv.1805.12152

[60] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008). https://jmlr.org/papers/v9/vandermaaten08a.html

[61] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems* 33 (2020), 2958–2969. https://proceedings.neurips.cc/paper/2020/hash/1ef91c212e30e14bf125e9374262401f-Abstract.html?ref=https://githubhelp.com

[62] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. 2017. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*. 1369–1378. https://doi.org/10.1109/ICCV.2017.153

[63] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. 2021. To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*. PMLR, 11492–11501. https://proceedings.mlr.press/v139/xu21b.html

[64] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. 2020. A closer look at accuracy vs. robustness. *Advances in neural information processing systems* 33 (2020), 8588–8601. https://proceedings.neurips.cc/paper_files/paper/2020/hash/61d77652c97ef636343742fc3dcf3ba9-Abstract.html

[65] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*. PMLR, 7472–7482. https://proceedings.mlr.press/v97/zhang19p.html

[66] Zhun Zhang, Qihe Liu, and Shijie Zhou. 2021. GGCAD: A Novel Method of Adversarial Detection by Guided Grad-CAM. In *International Conference on Wireless Algorithms, Systems, and Applications*. Springer, 172–182. https://doi.

org/10.1007/978-3-030-86137-7_19

[67] Tianhang Zheng, Changyou Chen, and Kui Ren. 2019. Distributionally adversarial attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2253–2260. https://doi.org/10.1609/aaai.v33i01.33012253

[68] Linjun Zhou, Peng Cui, Xingxuan Zhang, Yinan Jiang, and Shiqiang Yang. 2022. Adversarial eigen attack on black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15254–15262. https://doi.org/10.1109/CVPR52688.2022.01482