Panda or not Panda? Understanding Adversarial Attacks with Interactive Visualization

YUZHE YOU, School of Computer Science, University of Waterloo, Canada

JARVIS TSE, School of Computer Science, University of Waterloo, Canada

JIAN ZHAO, School of Computer Science, University of Waterloo, Canada

Adversarial machine learning (AML) studies attacks that can fool machine learning algorithms into generating incorrect outcomes as well as the defenses against worst-case attacks to strengthen model robustness. Specifically for image classification, it is challenging to understand adversarial attacks due to their use of subtle perturbations that are not human-interpretable, as well as the variability of attack impacts influenced by diverse methodologies, instance differences, and model architectures. Through a design study with AML learners and teachers, we introduce AdvEx, a multi-level interactive visualization system that comprehensively presents the properties and impacts of evasion attacks on different image classifiers for novice AML learners. We quantitatively and qualitatively assessed AdvEx in a two-part evaluation including user studies and expert interviews. Our results show that AdvEx is not only highly effective as a visualization tool for understanding AML mechanisms, but also provides an engaging and enjoyable learning experience, thus demonstrating its overall benefits for AML learners.

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

Additional Key Words and Phrases: adversarial machine learning, information visualization, explainable AI, evasion attack

ACM Reference Format:

Yuzhe You, Jarvis Tse, and Jian Zhao. 2025. Panda or not Panda? Understanding Adversarial Attacks with Interactive Visualization. 1, 1 (March 2025), 30 pages. https://doi.org/

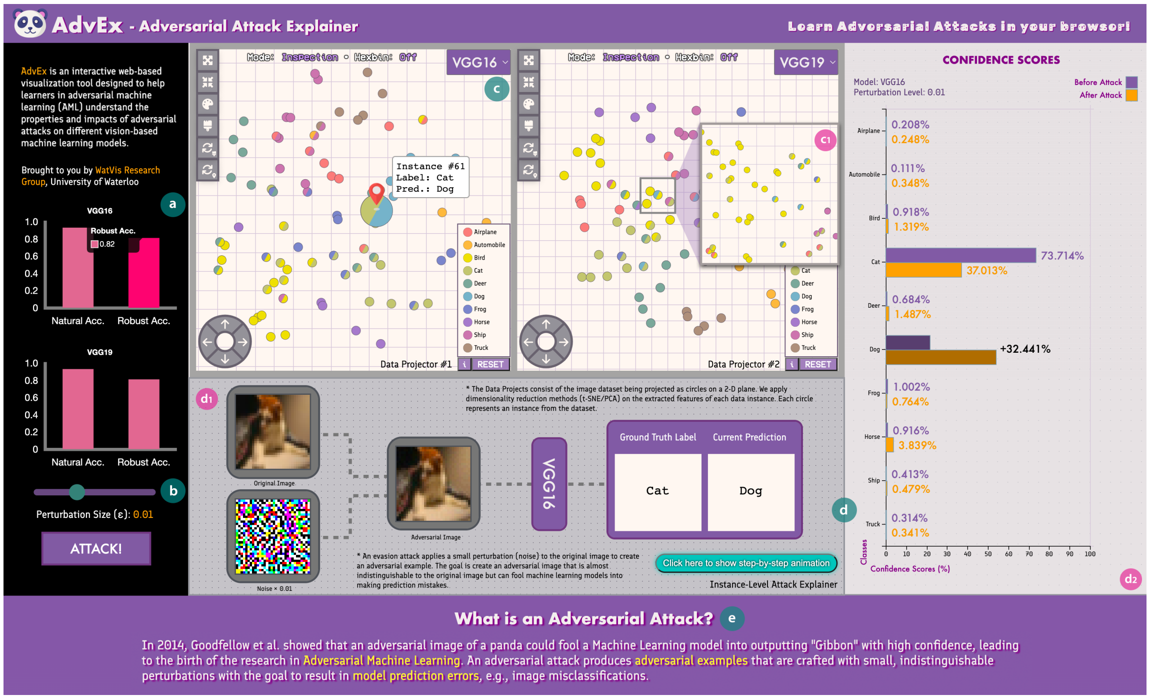
1 Introduction

Adversarial *evasion attacks* produce deceptive inputs (e.g., adversarial images) that are subtly altered with human-imperceptible perturbations to fool machine learning (ML) models into making prediction mistakes. In 2014, Goodfellow et al. [15] showed that an adversarial image of a panda could easily fool GoogLeNet [48] into labeling it as a gibbon with high confidence, resulting in the birth of *adversarial machine learning* (AML) research. Similar attack methods have been shown to achieve high misclassification rates in road sign classifiers [12] and evade automated surveillance cameras [49]. Though more and more people are studying and applying ML, many remain uninformed about the dangers of adversarial attacks to their models due to a lack of knowledge in AML. As a result, the models developed often achieve good natural accuracy but are highly susceptible to attack-perturbed inputs [47]. For these users (e.g., students, novice ML developers) to design or calibrate models to be adversarially robust for real-world applications, it is essential to educate them about the concepts and impacts of adversarial attacks.

Many studies have shown that visualizations serve as effective educational tools for teaching complex ML concepts to non-experts interactively, augmenting passive learning experience (e.g., textbooks and videos) [21, 24, 51]. Specifically, we aim to design an educational visualization tool to benefit learners who have an ML background but are unfamiliar with

Authors' Contact Information: Yuzhe You, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, y28you@uwaterloo.ca; Jarvis Tse, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, jarvis.tse@uwaterloo.ca; Jian Zhao, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, jianzhao@uwaterloo.ca.

Manuscript submitted to ACM



AdvEx is an interactive web-based visualization tool designed to help learners in adversarial machine learning (AML) understand the properties and impacts of adversarial attacks on different vision-based machine learning models.

Brought to you by WatVis Research Group, University of Waterloo

* The Data Projects consist of the image dataset being projected as circles on a 2-D plane. We apply dimensionality reduction methods (t-SNE/PCA) on the extracted features of each data instance. Each circle represents an instance from the dataset.

* An evasion attack applies a small perturbation (noise) to the original image to create an adversarial example. The goal is create an adversarial image that is almost indistinguishable to the original image but can fool machine learning models into making prediction mistakes.

In 2014, Goodfellow et al. showed that an adversarial image of a panda could fool a Machine Learning model into outputting "Gibbon" with high confidence, leading to the birth of the research in Adversarial Machine Learning. An adversarial attack produces adversarial examples that are crafted with small, indistinguishable perturbations with the goal to result in model prediction errors, e.g., image misclassifications.

Fig. 1. AdvEx user interface: (a) Robustness Analyzers that display the models' prediction accuracy pre- and post-attack; (b) Perturbation Adjuster that initiates the attack sequence with specified magnitude; (c) Data Projectors that visualize data embeddings in a 2-D latent space; (d) Instance-level Attack Explainer that displays in-depth information of the highlighted instance; (e) General Information Provider that provides more background on AdvEx and AML.

the risks of adversarial attacks, and are interested in learning AML to seek to build safer models for their applications. For this work, we focus on evasion attacks in image classification, a highly active AML research path that most existing work [15, 22, 57] focuses on since such models are frequently used in safety-critical applications [16, 27]. Compared to adversarial attacks in certain other ML tasks such as NLP [58] and recommender systems [7], the perturbations applied to images also tend to be more human-imperceptible, making them even more challenging to understand and thus increasing the value of visualizing them for education.

Understanding adversarial attacks for image classification involves certain key challenges. First, the attack process is often non-intuitive, as adversarial attacks leverage subtle perturbations that exploit data features beyond human perception [22]. These modifications are imperceptible to human observers, making adversarial images appear almost identical to their clean counterparts. Second, adversarial attacks exhibit high variability depending on several factors, including instance differences [1], model architectures and training methods [22, 31], and attack strategies [5, 57]. This variability requires a multi-level inspection: instance-level analysis reveals localized attack behaviors, dataset-level analysis uncovers broader patterns and trends in attack impacts across an entire dataset, and comparisons across different models help uncover how attack impacts vary across classifiers. Addressing these challenges requires visualizations that effectively illustrate attack behavior and variability, support model comparisons, and accommodate diverse attack strategies.

However, existing visualization-based educational tools for AML fail to address these challenges, lacking comprehensiveness and generalizability in presenting various attack properties. For instance, Adversarial-Playground [34] is limited by its simplistic approach that displays an adversarial image beside its original, a method that is ineffective when the two images look identical from subtle perturbations. Bluff [10] relies on visualizing the internal neuron logic on benign and adversarial examples, sacrificing model generalizability. Both tools, limited to specific models/attacks, insufficiently represent evasion attacks by neglecting the influence of varying model architectures, training methods, and instance differences, leaving learners with an incomplete understanding. While advanced AML visual analytic tools exist (e.g., AEVis [2] and Ma et al.'s [30]), they are designed for experts to perform model analysis with complex visualizations that are challenging for novices to understand, thus not suitable for educational purposes. Further, AEVis lacks model comparisons and dataset-level visualizations, while Ma et al.'s is limited to data poisoning attacks in binary classification, lacking support for evasion attacks in multiclass classification.

As such, designing an AML visualization targeted at novices for educational scenarios presents its unique challenges. These include effectively presenting advanced ML concepts like adversarial attacks in a non-overwhelming manner without losing essential details (e.g., the "imperceptible" attack property), along with creating an interactive learning environment that is comprehensive yet intuitive to understand. Therefore, to better augment learners' experience and address the limitations of existing tools, we carried out a study to design an interactive educational visualization to help learners understand evasion attacks at multiple levels, while allowing observation of their impacts on different models. Our primary objective is to help novice *learners* gain a comprehensive understanding of the properties and risks of adversarial attacks from multiple lenses, thus enabling them to make more informed decisions during model development to mitigate the risks posed by adversarial attacks. Through this work, we have made the following contributions:

- We conducted a **design study** on employing interactive visualization for educational purposes to augment the learning experience for AML. Our study involved both literature reviews and user interviews with AML learners (N=3) and teachers (N=3) for design guideline formulation, followed by system development and an extensive evaluation. Our findings provide insights and design lessons on how visualizations can support learning and engagement with concepts related to adversarial attacks.
- We designed and implemented **AdvEx**, an interactive educational visualization for novice learners to gain a comprehensive understanding of adversarial attacks. To the best of our knowledge, AdvEx is the first multi-faceted visualization designed specifically to support comprehensive learning of evasion attacks on both instance and population levels. Additionally, it supports model comparison and can readily adapt to different image classifiers and evasion attacks, addressing the generalizability gap in existing works (e.g., [10, 34]).
- We performed a **two-part evaluation** with 24 novice learners and 7 AML experts/teachers to quantitatively and qualitatively evaluate the learning aspects and usability of AdvEx. Our results show that AdvEx not only is highly effective in facilitating understanding of adversarial attacks, but also offers an engaging and enjoyable learning experience, thus amplifying its educational impact. The strengths and limitations of AdvEx are discussed, providing in-depth insights on how such a tool can be effectively utilized in an educational setting.

2 Related Work

2.1 Adversarial Machine Learning

Many adversarial attacks have been proposed to work under different threat models, namely white-box and black-box attacks. A white-box attack has full access to the model's internals, while a black-box attack can only access model inputs and outputs. Fast Gradient Sign Method (FGSM) [15], Basic Iterative Method (BIM) [26], and Projected Gradient

Descent (PGD) [31] are some of the well-known white-box attacks. Advanced black-box attacks include Zeroth Order Optimization (ZOO) [5], HopSkipJump Attack [4], and Substitute Model Attack [36]. To counter adversarial attacks, various defense methods have been proposed to fortify model robustness against adversarial inputs. The most effective defense is *adversarial training*, which trains classifiers with adversarial examples by adding them to the training set [15, 31] or through regularizations [41, 57]. TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization (TRADES) [57] is a state-of-the-art adversarial training method that leverages a regularized surrogate loss from the observed trade-off between robustness and accuracy. Other examples of adversarial defenses include standard adversarial training [31], robust self-training (RST) [42], local linear regularization (LLR) [41], etc.

While AdvEx can be employed with any evasion attack algorithm of the user's choice, to showcase the system's adaptability to different types of attacks, in this paper, we demonstrate AdvEx using two attack examples recommended by the AML instructors we consulted, including FGSM, one of the earliest and most well-known white-box attacks [15], and ZOO, a highly effective black-box query-based attack [5]. Several prior studies have tried to understand the characteristics of these two attacks. For example, Zhang et al. [59] discovered that FGSM may create not only 2-D adversarial images but also 3-D adversarial examples by applying the attack methodology to PointNet [40], a DNN designed for 3-D point cloud data. Ye et al. [54] applied adversarial attacks to a DL-based multiuser OFDM detector [53] and showed that ZOO achieved the best performance among black-box methods. Additionally, both attacks are frequently used in existing works to evaluate the effectiveness of adversarial defenses [17, 31, 45, 55, 56] or as comparisons to other attacks [28, 32, 46]. The abundance of existing works on both methods shows that they are well-known attacks and hence good introductory examples for those new to AML. As AML is a relatively new area of ML, it is crucial to raise awareness on attacks like FGSM and ZOO to encourage users to build safer AI applications, especially those that are safety-critical.

2.2 Visualizations of Adversarial Attacks

In general, interactive tools designed for visualizing adversarial attacks are relatively under-explored. A few tools with educational purposes have been proposed in past studies. Adversarial-Playground [34] is a simple web application that demonstrates the efficacy of three attack algorithms against a small CNN on the MNIST dataset [11]. The tool allows users to choose from a set of pre-defined inputs and displays the adversarial image next to its original alongside classification likelihoods to illustrate the attack. Bluff [10] visualizes attacks on a vision-based network, but focuses on model internals instead by highlighting the neurons and connections that an attack exploits to confuse the model.

However, these tools lack comprehensiveness and multi-faceted approaches in visualizing adversarial attacks. For instance, Adversarial-Playground [34] offers a simple image comparison approach that becomes ineffective if used to visualize attacks that generate "imperceptible" inputs, a common characteristic among adversarial attacks. Its applied perturbations on the black and white MNIST dataset are also highly visible, which could create a false sense of security among learners about their abilities to discern adversarial images from clean ones. Bluff [10], while offering a dataset-level view of internal neurons, does so in a way that abstracts away from individual instances and loses information due to the use of median values and neuron filtering. Both tools are constrained to specific attacks/models and a single level of analysis: Adversarial-Playground focuses solely on instance-level visualization, while Bluff is limited to abstracted dataset-level attack insights. We define dataset-level attack visualization as the ability to reveal patterns, trends, or distributions of attack impacts across an entire dataset, rather than focusing on individual examples.

In addition, a few advanced AML visual analytics tools have been developed as well. AEVis [2] uses a river-based visual metaphor to show how the datapaths of clean and adversarial examples merge or diverge within the network.

However, it suffers from the same limitation of lacking model comparisons and dataset-level information, and cannot be used to visualize attacks with varying perturbation sizes. Ma et al. [30] proposed a framework that employs a multi-level visualization scheme to support the analysis of data poisoning attacks in binary classification tasks. While comprehensive, it is designed specifically for data poisoning attacks in binary classifications, thus diverging in focus from this work and lacking support for evasion attacks in multiclass classifications. Moreover, both tools are designed primarily for experienced practitioners to perform visual analytics on models under adversarial attacks, featuring complex visualizations and interfaces that may be overwhelming for novice learners.

Hence, current educational tools lack comprehensiveness, often visualizing a few instances and limited to specific attacks and models; current advanced visual analytics tools are overly complex for our intended audience or have a different focus from this work. In contrast, with AdvEx, we aim to enable users who have little or no knowledge of AML to learn about adversarial attacks at both dataset and instance levels, while making it easy to be generalized to different evasion attacks and vision-based classifiers.

In addition, to visualize the shift in how models perceive the dataset before and after an attack, we incorporated a dimensionality reduction overview depicting each model's feature space in AdvEx. Dimensionality reduction has been used frequently to understand and visualize adversarial attacks. For instance, Ma et al. [30] and Park et al. [37] utilize t-SNE for data embedding views to visualize the impacts of data augmentations including adversarial attacks. Panda and Roy [35] introduced a Noise-based Learning (NoL) approach for training robust DNNs and provided simplistic PCA-based visualizations for adversarial dimensionality and loss surface visual analysis. Hendrycks and Gimpel [18] incorporated PCA into adversarial image detection and visualized how adversarial images abnormally emphasize coefficients for low-ranked principal components. Inspired by these works, in AdvEx, we apply similar methods to project the data embeddings onto a 2-D plane, and use animated transitions and colors of circular glyphs to visualize how the attacks alter the models' perception of the images.

2.3 Visualizations for Learning ML

Outside of AML, several visualization tools specifically designed for learning ML have been proposed as well. GAN Lab [24] is designed for non-experts to learn and experiment with generative adversarial networks (GANs) by visualizing GANs' dynamic training processes on a simple dataset. CNN Explainer [51] enables learners to inspect the interplay between CNNs' low-level mathematical operations and their high-level model structures. Summit [21] provides higher-level explanations of DNNs by visualizing image features detected by the networks and how those features interact to make predictions. More recently, TransforLearn [14] provides interactive visual tutorials for learners to understand transformer models by supporting architecture-driven and task-driven exploration. While these tools are effective for demonstrating basic ML concepts, they are not suitable for our study's design objective in the context of AML learning. For example, GAN Lab [24] is only for exploring generative models on low-dimensional training datasets and significantly diverges from the focus of this work. Similarly, TransforLearn [14] is specifically designed for learning transformer models' layer operation and mathematical details. While CNN Explainer [51] and Summit [21] could potentially be extended to explore a model's internal datapaths on adversarial examples, they would still share the limitations of lacking model generalizability and dataset-level attack information like Bluff [10] and AEVis [2]. As such, it is important to have a designated educational tool for AML that addresses the gaps of existing works. We selected AML as our target domain as it is crucial to educate novice practitioners who apply ML across diverse domains but do not understand their models' vulnerability due to their gap in AML knowledge. We believe by helping them understand adversarial attacks in a hands-on manner, they would be equipped with the necessary knowledge to design models that are robust for real-world applications in the future.

Despite focusing on visualizing common DNNs instead of adversarial attacks, all aforementioned studies have provided us with inspirations for AdvEx's design. Specifically, similar to GAN Lab [24] and CNN Explainer [51], AdvEx is accessible to any user with a modern browser without the need to install specialized hardware for deep learning. Motivated by GAN Lab [24]'s step-by-step training visualization, AdvEx provides step-by-step executions of the attack methodology to visualize the detailed attack process. Like CNN Explainer [51] and Summit [21], AdvEx also adopts smooth transitions across different levels of abstraction to facilitate visual exploration and to serve as the link that connects different views of the visualization tool. Inspired from TransforLearn [14], AdvEx uses task-driven exploration to help users gain a deeper understanding of model robustness with actual image classification tasks. Based on existing work, we aim to develop AdvEx as a tool with comprehensive visualizations and animations that can enable intuitive exploration of attack properties across multiple levels.

3 Design Goals

To formulate the design guidelines for AdvEx, we conducted user interviews with six participants, including three interviewees (S1, S2, S3) who have AML learning experience and three AML teachers (E1, E2, E3). Our goal was to understand learners' needs in understanding adversarial attacks and to have experienced AML teachers envision how such a tool can be utilized in an educational setting. The learners involved come from computer science and data science backgrounds, and their employed learning methods varied from enrolling in AML courses to reading academic papers or online blog posts. The teaching experience of the interviewed educators ranged from leading graduate-level AML seminars to overseeing AML components within undergraduate ML courses. The semi-structured interviews lasted between 60 to 90 minutes and covered the following topics: 1) the participants' background and experience in AML learning/teaching, 2) existing content or tools used to understand/teach AML, 3) the challenges in understanding/teaching adversarial attacks, 4) features and functionalities to include in an educational visual tool for adversarial attacks, and 5) how participants envision using such a tool in an educational setting. The participants were compensated $20/hour for the interview.

While none of the interviewees had previously used any visualization tool for adversarial attacks, all recognized the value of introducing a multi-level visualization tool to demonstrate evasion attacks to learners. Specifically, they believed that an interactive visualization tool would have multiple educational benefits, including *"providing an accessible way to demonstrate attacks in practical applications"* -E2, *"making the learning experience more engaging"* -E3, and *"accommodating learners with different backgrounds"* -E1. The interviewees also thought that the tool could be used either in a self-learning scenario for exploration or incorporated into AML courses to demonstrate concepts and better augment students' learning experience. These comments confirm the need for a visualization tool like AdvEx in both independent and guided AML learning contexts.

We transcribed our interviews and employed a hybrid method of open and closed coding, informed by an extensive literature review (Section 2), to analyze the gathered qualitative data. Using an affinity diagram, we identified recurring themes and requirements of such a visualization tool for AML learning. This process involved collaboratively organizing observations and insights from literature and transcriptions into sticky notes, which we grouped on a large canvas based on their similarities and common themes. We started off by using open coding to freely identify themes and patterns in our data. Then, as the themes became clearer, we organized them into a more structured framework. Through iterative sessions of discussion and reorganization, clusters of notes representing common themes/requirements for our AML visualization were formed. As a result, we derived the following design goals to guide the development of AdvEx:

G1 **Present visual abstraction of the attack impact at multiple levels.** Many existing tools (e.g.,[10, 34]) only display instance-level attack information, such as how a specific image is modified by the attack. These instance details are insufficient to illustrate the reason behind misclassifications or the overall attack impact on a larger dataset. E3 mentioned, *"When simply comparing the images, we can observe the differences from a human perspective, but it remains unclear why the model misclassifies them."* E2 agreed that *"Examining images prone to misclassification is vital, but seeing the broader impact is equally important to fully grasp the risks."* Therefore, visual abstractions at multiple levels should be included to provide both dataset-level overviews of the attack and the options to conduct more in-depth investigations on specific instances.

G2 **Design a visualization framework that can be generalized to different evasion attacks and image classifiers.** Generalizability is crucial as it enables learners to grasp the variability of attack methods, assess different kinds of models under attacks, and connect theoretical knowledge with practical applications. E3 confirmed that *"A key learning objective should be the various methods to generate adversarial examples, which is essential for understanding how to defend against these diverse attack strategies."* S2 and S3 agreed that generalization can help learners gain practical insights into the variability of attack impact by exploring different attacks in actions and visualizing models/attacks that align with their backgrounds, rather than being restricted to a predetermined set of models/attacks. For AdvEx, we aim to address the gap of existing works [10, 34] being constrained to specific attacks/models by designing a general framework in these aspects to enable a more holistic and practical understanding of the attacks. Specifically, we aim to adapt a "plug-and-play" approach to allow users to easily swap out the attack algorithms and models based on their interests and learning goals.

G3 **Enable comparative analysis of different models' robustness under attack.** Models with different architectures and training methods vary in their robustness against the same attack [15, 22, 57], but most learning tools [10, 34] demonstrate attacks with a single, arbitrary model. Enabling visual analysis of multiple models is important to facilitate understandings of the variability in attack impact and to highlight the rationales for why certain models would fail. E3 stated, *"Comparing the model differences provides insights into why certain attacks succeed or fail, going beyond just seeing changes in model accuracy."* E2, S2, & S3 agreed that model comparisons *"highlight the models' varying defense abilities"* -S3 and in turn *"helps learners defend and improve their own in their future applications."* -E2. Thus, we aim to provide side-by-side comparisons of different models under various attack scenarios.

G4 **Facilitate dynamic experimentation with fluid transition between different perturbation sizes.** As the perturbation size increases, the attack becomes more effective and the applied noise also becomes more visible. Allowing users to dynamically experiment with the perturbation size and observe the changes in real time *"facilitates a better understanding of this correlation"* -E2 and *"creates a more game-like, engaging learning process"* -S1. E1 pointed out that such experimentation *"allows learners to observe how the model's perception of an instance changes, and identify the threshold at which misclassification occurs."* S3 stated that the approach *"helps learners understand when exactly the image starts to look different for humans."* Therefore, interfaces are included to allow easy manipulation of the perturbation size and visualize the changes in real time.

G5 **Allow step-by-step execution for learning the attack process in detail.** Mentioned by E1, E2, & S2, navigating complex mathematical steps in papers to understand attack logic is a daunting task for learners. A step-by-step attack execution *"provides a more structured understanding of attack strategies"* -E2. This approach allows learners to *"grasp*

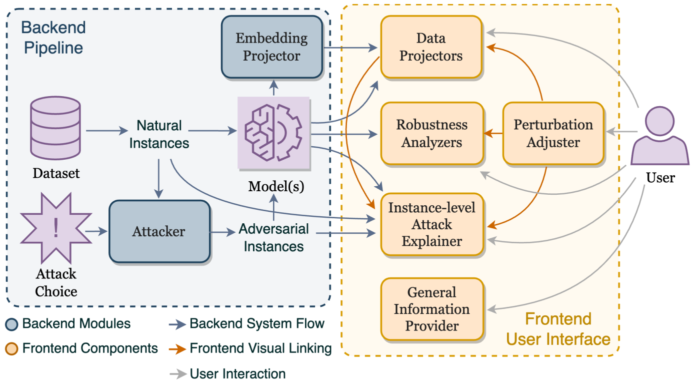


Fig. 2. A schematic diagram depicting the system architecture of AdvEx. In the backend pipeline, an Attacker module performs users' choice of attacks on the image dataset, targeting models specified by users (**G2**). Once processed, the backend outputs are passed to the frontend interface for user interaction.

not just the impact but also the design and rationale behind the attacks" -E3. E2 confirmed, *"A step-by-step approach simplifies the attack process and reduces learners' burden compared to interpreting steps directly from papers."* In AdvEx, we aim to incorporate a step-by-step view to clarify the underlying attack logic for learners, guiding them through the complexities of various attack strategies.

4 AdvEx

Based on our design guidelines, we developed AdvEx. Here, we begin with an overview of AdvEx's system, followed by detailed descriptions of its backend modules and frontend components.

4.1 System Overview

As depicted in Figure 2, AdvEx is a web application with two main system components: A) a *backend pipeline* (Section 4.3) and B) a *frontend user interface* (Section 4.4).

In the backend pipeline, an *Attacker* module begins by normalizing the image data and performing users' chosen attack methods to generate adversarial examples. Both the original and adversarial examples are fed into the models to obtain information such as image embeddings, confidence scores, and prediction accuracy. An *Embedding Projector* is employed to extract each model's embedding vectors by removing the final output layer and applying dimensionality reduction methods (e.g., t-SNE [50], PCA [38]) to prepare the projection coordinates of the data representations. The processed outputs are relayed to the frontend components to be presented visually for user interaction.

The frontend interface comprises five key components: 1) *Data Projectors* (Figure 1c), 2) *Instance-level Attack Explainer* (Figure 1d), 3) *Robustness Analyzers* (Figure 1a), 4) *Perturbation Adjuster* (Figure 1b), and 5) *General Information Provider* (Figure 1e) + interactive tutorials. The Robustness Analyzers feature two interactive bar charts that assess the models' overall robustness under a specified attack (**G1**) and offer a comparative view of this robustness to natural accuracy (**G3**). The Data Projectors utilize coordinates from the Embedding Projector to visualize data representations as two interactive, side-by-side scatterplots. These scatterplots enable exploration of attack-induced embedding changes (**G1**) and offer comparisons of embeddings between different models (**G2, G3**). The Instance-Level Attack Explainer provides detailed insights into a specific instance (**G1**), complemented by a confidence score view and a step-by-step guide to the instance's attack process (**G5**). The Perturbation Adjuster allows users to select their desired perturbation size and initiates animations within the three aforementioned components to simulate the attack in real time (**G4**). Finally, along with interactive tutorials, the General Information Provider guides users through the navigation of the interface and offers further context on AML.

4.2 Dataset and Models

In this paper, we use the CIFAR-10 dataset [25] to demonstrate AdvEx, but our system can be employed with any image dataset with ≤ 12 classes due to color distinguishability [33] or a subset of a dataset with more classes. The CIFAR-10 dataset consists of 60,000 32 × 32 colored images from 10 different classes (50,000 training data and 10,000 testing data), with 6,000 images per class. We chose this dataset due its popularity of being used in ML research to evaluate the accuracy and robustness of image classifiers [9, 19, 57].

In addition, AdvEx supports a variety of image classifiers and allows the user to compare two models side by side (**G2, G3**). For example, users could compare CNNs with the same architecture but different numbers of convolutional layers, or investigate how a classifier trained adversarially may outperform a standard model in an attack. For this paper, we loaded two pairs of models for our studies: 1) VGG-16 vs. VGG-19, and 2) ResNet-34 trained naturally vs. trained adversarially with TRADES [57].

4.3 Backend Pipeline

In this section, we describe how the backend processes and analyzes the data in AdvEx, including how it generates the adversarial examples and prepares the data instances and model outputs for frontend display.

4.3.1 Attacker Module. The "Attacker" module produces adversarial examples of the original dataset by conducting adversarial attacks on the targeted models. It first feeds the natural images into the targeted models (or surrogate models) to obtain information relevant to the attack, then adjusts the pixel values of the input image based on the information. While users can swap out the attack algorithm in the Attacker module with any evasion attack they wish to learn about (**G2**), here we use one white-box attack and one black-box attack, FGSM [15] and ZOO [5], as examples for demonstrating our system's adaptability to different attacks.

We chose the FGSM attack due to its notoriety for creating the very first adversarial panda image [15] that is well-known among AML researchers. It is commonly used as a baseline for evaluating model robustness and defense effectiveness [31, 43, 59]. The attack was also recommended by the consulted AML instructors as it is relatively simple in logic and used as the introductory attack in AML courses and tutorials. However, the attack has been proven to be extremely effective [15]:

$$\mathbf{x}' = \mathbf{x} + \epsilon \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y)). \tag{1}$$

It modifies image $\mathbf{x}$ by maximizing the loss $J(\theta, \mathbf{x}, y)$ towards the gradients' sign to produce the adversarial image $\mathbf{x}'$. Here, y is the true label, θ is model parameters, and ϵ scales the perturbation. We used L^{∞} norm to restrict the maximum pixel change to create bounded examples.

Additionally, we employed the ZOO attack [5], an advanced black-box attack, as the second demonstration method. The AML instructors highlighted that ZOO is often used as a representative example of black-box attacks in AML courses. It is also frequently used to evaluate defenses [17, 55] or as a benchmark for other attacks [28, 53]. ZOO only has access to model inputs (e.g., images) and outputs (e.g., confidence scores), and finds $\mathbf{x}'$ by solving the following optimization problem:

$$\begin{aligned} \underset{\mathbf{x}'}{\text{minimize}} \quad & \|\mathbf{x}' - \mathbf{x}\|_2^2 + c \cdot f(\mathbf{x}') \\ \text{subject to} \quad & \mathbf{x}' \in [0,1]^p \end{aligned} \tag{2}$$

The first term $\|\mathbf{x}' - \mathbf{x}\|_2^2$ applies L^2 norm regularization to enforce similarity between $\mathbf{x}'$ and $\mathbf{x}$. The loss $c \cdot f(\mathbf{x}')$ represents the level of unsuccessful attacks, with $c > 0$ as the regularization parameter. The attack approximates the gradient with a finite difference method and solves the optimization problem via zeroth order optimization.

Employed with users' selected attack, the module performs attacks respectively with the selected perturbation sizes ϵ. We enable this flexibility through a "plug-and-play" framework, allowing users to "plug in" an attack and dataset and "play them" without needing to modify the foundational elements of the module. Users can add new attacks/datasets by creating functions that follow a predefined interface, specifying parameters such as the model, input data, perturbation magnitude, and other relevant settings. Based on *RobustBench*'s suggested limits [8], we selected ϵ of 0.00, 0.01, 0.02, and 0.03 for FGSM with L^∞ norm, and ϵ of 0.0, 0.1, 0.3, and 0.5 for ZOO with L^2 norm for our demonstration. The resulting adversarial examples, along with the original data, are then inputted in the models for classification and embedding extraction.

4.3.2 Embedding Projector. The Embedding Projector is tasked with 1) processing the models' produced embeddings and 2) analyzing the information of the extracted features and preserving it in a low-dimensional representation. The goal is to unveil important patterns in the embeddings and transform them into a format easily fetched by frontend. The module temporarily detaches the final output layer to obtain the embeddings and reduces their dimensions by applying users' choice of dimensionality reduction for later 2-D visualizations. For instance, in the case of t-SNE, the module analyzes instance features by constructing a lower-dimensional probability distribution that represents the similarities between the objects in the high-dimensional space. If PCA is used, the module preserves the most significant variability in the embeddings while reducing the number of features. The resulting outputs are scaled to be used as the x- and y-coordinates of the instances in scatterplots and are stored as tabular data easily accessed by the frontend Data Projectors.

4.4 Frontend User Interface

Here, we detail the frontend components of AdvEx. We demonstrate our approach using FGSM on VGG-16 and VGG-19 models pre-trained with CIFAR-10 [39].

4.4.1 Data Projectors. The Data Projectors (Figure 1c) represent dimensionality reduction overviews of the dataset and consist of two scatterplots where the image embeddings are projected as circles on a 2-D plane. They present dataset-level attack information by illustrating how the overall dataset distribution shifts in the embedding space under adversarial perturbations, enabling users to observe broader patterns and trends in attack impacts across the dataset. Each circle corresponds to a data instance and is sliced into two halves: the color of the left half represents the instance's ground truth label, while the color of the right half represents its current prediction. The spatial positions of the circles encode the relationships between them in the original high-dimensional space (e.g., similarities, variance, local and global structure).

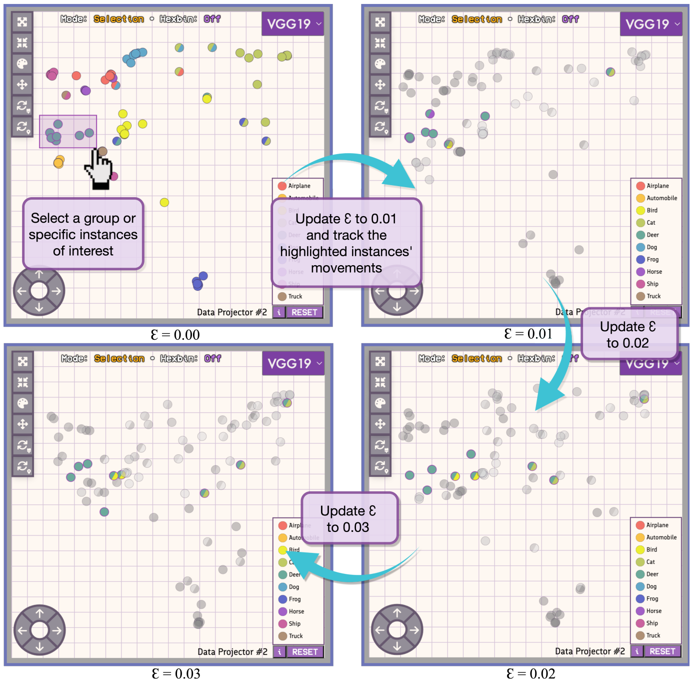
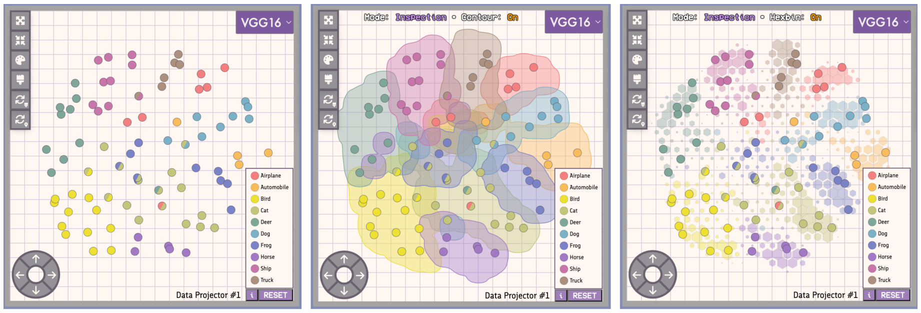


Fig. 3. A user highlights and tracks a specific class from the dataset with selection mode. Under this mode, one can evaluate model performance on a dataset subset.



(a) Scatterplot (b) Scatterplot + Density Contours (c) Scatterplot + Hexbin Map

Fig. 4. We explored a variety of visual encodings and aggregating features for the Data Projectors. We chose binned aggregation with multiple zoom levels, with an optional hexbin toggle to display the overall distribution (Fig. c). This preserves data scalability and displays global data structure without the need for high-performance devices.

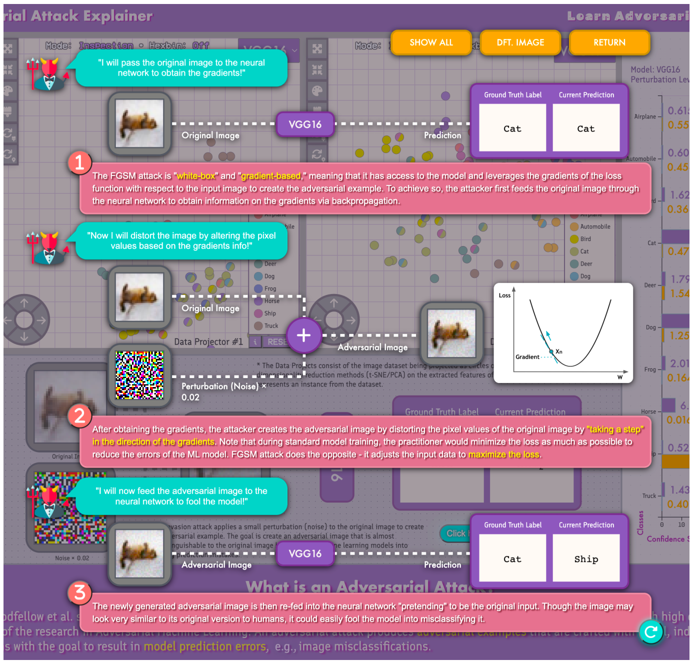
Inspired by *nanocubes* [29], we use a combination of binned aggregation and hierarchical clustering with multiple zoom levels to preserve data scalability and organize embeddings into a multi-level structure (Figure 1c1). Each level of data cubes represents varying granularity of data aggregation, with the reduced space split up into a grid of equally shaped and sized bins. The number of bins at the highest level is scaled based on the data range in reduced dimensions, and follows a fixed multiplying pattern for subsequent levels (e.g., 10 × 10, 20 × 20, 30 × 30, ...). Data points are allocated to bins based on their positions in the scatterplot, and a representative instance, predicted as the most frequently predicted class within that bin, is sampled for display at that specific level.

To assist users in identifying data patterns at higher zoom levels with fewer sampled instances, the size of the hexagons in the background indicates the density of embeddings in each bin (Figure 4c), visualizing the overall clustering patterns. During exploration, the projectors dynamically sample and display instances from each level, allowing for an iterative and detailed examination of larger datasets without getting messy from too many moving dots. Our approach allows interactive exploration of data sources with large numbers of instances while maintaining the global data structures without high-performance devices.

When an attack is conducted with newly specified magnitude, the Data Projectors visualize the attack with an animated sequence that emphasizes each circle's change in position and color (**G4**) (Figure 3). For example, if a circle transitions to a different coordinate, this indicates that the model's perception of the instance's features has been altered by the attack. To mitigate potential artifacts produced by projection methods' randomness, multiple runs of the projection are conducted and the results are averaged. This approach allows us to ensure that the transitions observed in the projectors are predominantly indicative of the models' changes in feature perceptions. Moreover, if the class "airplane" is represented by the color red and the class "automobile" is represented by the color orange, then a red circle transitioning into a half-red, half-orange circle means that this is an airplane image incorrectly classified as an automobile due to the attack. To improve the usability of the projectors, the following functionalities are also incorporated:

- **Inspection mode.** ✥ Under this mode, users can zoom and drag freely within the scatterplots to explore the models' embedding distributions. To avoid overlap when instances share similar features, AdvEx dynamically adjusts the radius of the projected circles at different zoom levels, allowing users to precisely examine each individual instance. Clicking a circle highlights the instance by enlarging its radius and pinning it, then panning the entire plot to recenter that circle within the 2-D plane.
- **Selection mode (Figure 3).** 📌 In this mode, users can highlight a subset of the dataset, including a single item, by specifying a selected region with a pointing gesture. As a result, only colors of the selected circles within the region remain visible, while all other instances are grayed out. This feature allows users to track the movements of specific subgroups or instances across different perturbation sizes, adding a subpopulation-level display (**G1**). When a group/instance is highlighted in one Data Projector, the same group/instance is simultaneously highlighted in the other projector for comparison (**G3**).
- **Hexagonal binning toggle.** 🎨 While navigating, users can toggle a hexagonal binning map (Figure 4c) for each projector to track the global data structure. The hexbin map displays the general trends in instance clustering based on model predictions, allowing quick identification of decision boundaries and similarly classified image groups (**G1**). Moreover, this approach preserves visibility of the whole dataset's distribution even when the projectors are only displaying a subset at higher zoom levels.

In summary, the Data Projectors provide interactive visualizations of image embeddings, illustrating instance relationships via spatiality and revealing population-level attack impacts though animated transitions (**G1, G4**). Given that the set of generated adversarial examples varies depending on the chosen attack method and the model targeted, we provide side-by-side visualizations of two distinct models' embeddings to illustrate the differential impacts of the attack (**G3**). We design the Data Projectors to not only be technically robust to handle varying data scales and complexities, but also intuitively understandable through its animated transitions to visualize changes in model perception. Compared to works like [34], our approach illustrates attacks' impacts on models while preserving their imperceptibility at the instance level (see Section 4.4.2). By interacting with the Data Projectors, users can intuitively observe how changes



"I will pass the original image to the neural network to obtain the gradients!"

1 The FGSM attack is "white-box" and "gradient-based," meaning that it has access to the model and leverages the gradients of the loss function with respect to the input image to create the adversarial example. To achieve so, the attacker first feeds the original image through the neural network to obtain information on the gradients via backpropagation.

"Now I will distort the image by altering the pixel values based on the gradients info!"

2 After obtaining the gradients, the attacker creates the adversarial image by distorting the pixel values of the original image by "taking a step" in the direction of the gradients. Note that during standard model training, the practitioner would minimize the loss as much as possible to reduce the errors of the ML model. FGSM attack does the opposite - it adjusts the input data to maximize the loss.

"I will now feed the adversarial image to the neural network to fool the model!"

3 The newly generated adversarial image is then re-fed into the neural network "pretending" to be the original input. Though the image may look very similar to its original version to humans, it could easily fool the model into misclassifying it.

Fig. 5. An example of the final state of the step-by-step execution view for explaining the FGSM attack. The view progressively reveals attack elements and explanations, animated one by one to illustrate the flow of the attack process.

in the perturbation size influence both the models' data representations and their resulting image predictions, thus gaining a deeper understanding of the attack's impact on model performance.

4.4.2 *Instance-level Attack Explainer.*

While the Data Projectors visualize attacks on dataset or subpopulation levels, the Instance-level Attack Explainer (Figure 1d) provides in-depth information for each perturbed input. It outlines the attack process for each image, detailing instance-level information such as the original image, applied noise, and confidence scores (**G1**). To examine an instance, users click on the corresponding circle in the Data Projectors to update the panels associated with the attack explainer. Specifically, the Instance-level Attack Explainer consists of the following components:

- **General view.** The general view (Figure 1d1) displays key information of the selected instance, including its original image, applied noise, adversarial image, targeted model, true label, and current prediction (**G1**). To help users contextualize these instance-level details, subtle animations are used to link the information together to depict the high-level attack flow. For instance, a repeated animated sequence shows the original image and the generated perturbation progressively moving towards each other with reduced transparency and stacking on top of each other, then gradually fading into the final perturbed image. Dashed lines connecting the images are animated to continuously move from the clean image + noise to the resulting adversarial image. By presenting a visual attack narrative, these animations are designed to help users better interpret the instance-specific information within the context of the current attack.
- **Side-by-side image inspection.** For a closer inspection of the images, users can click on the image thumbnails in the general view to view enlarged versions. A comparison mode is available to examine the clean and adversarial images side by side and observe the exact pixel differences.

- **Confidence score view.** The interactive bar chart panel (Figure 1d2) showcases the model's pre- and post-attack confidence scores across all classes for the selected instance. These scores are grouped together in pairs to provide comparison between the model's confidence before and after an attack. Hovering over each pair of them reveals their exact difference in percentage, allowing users to quantitatively assess the attack impact on class-wise classification probabilities (**G1**).
- **Step-by-step execution view.** The step-by-step execution view (Figure 5) provides detailed explanations of the underlying attack logic. Clicking the button at the bottom-right of the general view activates this feature, which initiates a series of step-by-step animated sequences with accompanying explanations (**G5**). Once activated, these explanations unfold sequentially. For instance, Explanation #2 (Figure 5-2) does not appear until users click the play button next to Explanation #1 (Figure 5-1), which becomes visible only after Explanation #1 has finished playing. Within each explanation, individual elements are animated to present the flow of the attack. This includes animating the appearance of components one by one to show how the input image is fed into the model to obtain relevant information, or how the generated noise is added to the image to create the adversarial input. A toggle allows users to replace the default image with their selected instance for the view's demonstration, allowing them to apply step-by-step explanations to an actual adversarial example they are examining. This feature provides users with a more tangible and personalized understanding of how adversarial attacks manifest and operate on real-world examples (**G2**).

In short, the Instance-level Attack Explainer offers a focused, in-depth look at adversarial attacks on individual instances (**G1**). To achieve a balance between technical detail and novice engagement, the view translates complex attack processes into intuitive visual narratives, and adopts a step-by-step approach to guide novices through the underlying attack logic (**G5**). Also, its confidence score view enables users to quantitatively explore and assess how the attack impacts the model's confidence for the given instance (**G1**). Together, these features offer a detailed perspective of the instance-specific properties and consequences of adversarial attacks.

4.4.3 Robustness Analyzers. The Robustness Analyzers (Figure 1a) in the leftmost panel feature two compact, interactive bar charts, each containing two bars. These charts evaluate the model's robustness under the given attack and compare it to the pre-attack accuracy (**G1, G3**). The left bars represent natural accuracy, indicating the model's prediction accuracy on the clean dataset, while the right bars represent robust accuracy, reflecting the model's performance on the adversarial dataset. As users adjust the perturbation size (Figure 1b), the right bars dynamically adjust their heights to visualize the corresponding changes in the model's robust accuracy (**G4**). With the Robustness Analyzers, users can compare 1) a model's robustness to its baseline performance and 2) the relative performance of different models under standard and adversarial conditions (**G2, G3**). Consequently, users can gain insights into the attack's varying impact across models, identify which models are resistant or vulnerable to the current attack, and quantify the degree of performance degradation from adversarial inputs.

4.4.4 Perturbation Adjuster. The Perturbation Adjuster (Figure 1b), situated below the Robustness Analyzers, features a slider and an attack button. The slider allows users to choose a perturbation size from a range they have pre-set in the backend, which they can adjust horizontally to visualize the desired attack strength. We chose the perturbation size as AdvEx's primary control parameter as in the context of AML, it is standard to evaluate attacks under a defined upper bound on the perturbation size. While different attacks vary in their logic, all have a perturbation size that can be defined when applied to input images, which is what we focus in AdvEx. Upon selecting a perturbation size, users initiate an animated attack sequence by clicking the attack button, which triggers changes in other components of the interface. For example, the circles of the Data Projectors' (Figure 1c) may shift to new coordinates alongside prediction color changes, while the right bars of the Robustness Analyzers (Figure 1b) adjust their heights up or downward based on the model's accuracy with the new adversarial dataset. With the Perturbation Adjuster, users can dynamically modify the perturbation size and observe the increased attack strength with larger perturbation sizes, as well as the growing visibility of applied image noise (**G4**). The integration of dynamic perturbation control and real-time visual feedback enables users to intuitively understand the interplay between perturbation size, attack intensity, and resulting image distortions across a variety of attack methods.

4.4.5 Interactive Tutorials + General Information Provider. To help users pick up AdvEx more easily, an interactive tutorial system is also integrated. Upon launching the application, users encounter an overlay tutorial that introduces every component of AdvEx's interface, highlighting its key features. During interaction, hovering over any Data Projectors' button displays a tooltip explaining its function. If users have not engaged with certain key features (e.g., hexbin map, step-by-step execution) within 10 minutes, an animated arrow prompts them to explore these features.

Furthermore, if users wish to learn more about AdvEx and AML research, they may read the information placed beneath the interactive components (Figure 1e), which provides more in-depth explanations for both. By including interactive tutorials and reading materials, users will not only learn our tool faster, but also gain detailed and accurate knowledge of adversarial attacks in addition to perceiving them through interactive visualizations. As such, we designed AdvEx to place balanced emphasis on both visualizations and text, reinforcing novices' learning by presenting content in different formats, allowing learners to quickly grasp complex topics through multiple forms of interpretations.

4.5 Example Scenario

Here we provide an example scenario to show how AdvEx can help learners understand adversarial attacks and gain insights. Zoey, a learner, explores FGSM and ZOO attacks on ResNet-34 models using the CIFAR-10 dataset [25]. She aims to understand the attacks' properties and how their impact differs between standard and adversarially trained models. For simplicity, we refer to the standard ResNet as *ResNet*, and the adversarially trained version as *ResNet**. For a demonstrated workflow, see the video which is available along with this article in the ACM Digital Library.

Zoey loads her ResNet models and dataset into AdvEx. As she enters the main interface, she notices two Data Projectors (Figure 1c) side by side, each representing the image embeddings of one model. The embeddings from ResNet form distinct, small clusters, while those from ResNet* are more spread out. Examining the Robustness Analyzers (Figure 1a), she discovers that, in the absence of any attack, ResNet* has lower accuracy compared to ResNet. To better understand the general attack process, she clicks on the step-by-step execution view (Figure 5) and follows the guided, animated explanations of the FGSM attack before proceeding to investigate further.

Zoey decides to experiment with the Perturbation Adjuster (Figure 1b), setting the perturbation size to values above 0. In the left projector, the circles start to travel rapidly, while in the right projector, the circles only shift slightly. At a perturbation size of 0.03, the accuracy of ResNet drops to just 57%, whereas ResNet* maintains at 81%. From this, Zoey understands that the two models rely on different sets of features for classification. The FGSM attack drastically changes the way ResNet perceives the dataset, but it struggles to alter how ResNet* perceives the same dataset.

When the perturbation size reaches 0.03, Zoey notices that instance #6 in ResNet's Data Projector, an automobile, is incorrectly predicted as a cat. Upon clicking the instance, the Instance-level Attack Explainer (Figure 1d) updates. Zoey clicks on the image and enters the side-by-side comparison mode. From this, Zoey understands why adversarial attacks are known to be human-imperceptible: even though the adversarial image still looks similar to the original, the changes are meaningful enough to ResNet to alter its prediction. Zoey then clicks on the same instance in ResNet*'s projector, which is classified correctly. She immediately notices that the noise visualized here has more defined shapes. Since the noise is based on the model's gradient information, Zoey realizes that ResNet* relies on more human-interpretable features for classification, making it more robust.

Next, Zoey decides to load the ZOO attack into AdvEx. As she experiments with the perturbation size, she makes a few interesting observations. Firstly, she observes from the Data Projectors that, unlike FGSM, which scatters instances across all classes, the ZOO attack generally pushes each instance toward its nearest adversary class. Combined with observations from the confidence score view (Figure 1d2) that ZOO slightly adjusts the adversary class's confidence just above the original, Zoey realizes that ZOO pushes instances just past its decision boundary. Because of this, ResNet*, which performs well against FGSM, does not show much robustness to the ZOO attack.

Secondly, she notices that as the perturbation size increases, fewer instances begin to move around. Through the step-by-step explanations, Zoey understands that this is because ZOO is an iterative attack, and thus finds the minimum perturbation needed to cause a misclassification. From this, Zoey learns that different strategies result in different attack properties, and a training method effective against one attack may not be so against another. Through her hands-on experience with AdvEx, Zoey gains a level of understanding that she could not achieve through textbooks or traditional tutorials alone.

5 User Study with Novice Learners

To assess how AdvEx can help novice AML learners, we conducted a user study with participants who knew basic ML but were unfamiliar with AML. We aimed to investigate two aspects of AdvEx as an educational tool: **(A1)** whether AdvEx is effective for helping learners understand the concepts and impacts of adversarial attacks, and **(A2)** whether AdvEx engages users in learning. Engagement is also crucial in learning as it is known to promote active participation, which has been shown to improve comprehension and retention of complex concepts [6, 13]. Given the abstract nature of AML, engagement can make these topics more accessible and prevent disengagement caused by cognitive overload. To evaluate these aspects, we incorporated Adversarial-Playground [34] as a baseline tool. Adversarial-Playground is an existing visualization tool also designed to help non-experts understand adversarial attacks. It does so by providing side-by-side comparisons of natural and adversarial images with classification likelihoods, making it a suitable point of comparison. By comparing it with AdvEx, we sought to evaluate the specific contributions of AdvEx to AML learning. We also aim to answer following research questions:

- **RQ1:** How does AdvEx support learners' understanding and engagement with adversarial attack?
- **RQ2:** What workflows do learners adopt when using AdvEx to facilitate their learning?

5.1 Study Setup

Participants and Apparatus. We recruited 24 participants (P1 ~ P24; 18 men, six women; aged 21~34) from a local university, with 12 participants assigned to each condition. We denote P1–12 as participants assigned to AdvEx and P13–24 as participants assigned to Adversarial-Playground. They came from different areas of study such as computer science, transportation engineering, and data science. All reported having a background in ML but were unfamiliar with AML. Specifically, on a 7-point Likert scale (self-rated; 1="Novice", 7="Expert"), we recruited participants that satisfied all the following constraints: ML experience ≥ 2, AML experience ≤ 2, completion of ≥ 1 ML project, completion of $\leq$

1 AML project. Their median ML experience was 4 (IQR = 2), and their median AML experience was 1 (IQR = 1). The median number of ML projects completed was 2.5 (IQR = 2), while the median number of AML projects completed was 0 (IQR = 0). Participants interacted with their assigned tool on provided laptops in-person.

Task and Procedure. We loaded AdvEx with CIFAR-10 testing data perturbed by FGSM in varying degrees to investigate the participants' learning of the properties and impacts of the attack. Similarly, we provided Adversarial-Playground pre-loaded with FGSM attack examples for comparison. We selected FGSM for our first evaluation study based on recommendations from all interviewed AML instructors/learners, citing it as ideal for introducing AML concepts. E2 mentioned, *"For learners, it is essential to start with foundational methods like FGSM given its straightforward and basic nature."* S2 agreed that *"When teaching learners, it is best to use simpler attacks like FGSM."* We measured the participants' learning through a pre-quiz before their interaction with their assigned tool, followed by a post-quiz afterwards to assess the amount of knowledge acquired through the use of the tool. The quizzes were collaboratively designed with a renowned AML researcher/instructor who co-authored TRADES, the state-of-the-art adversarial training method against evasion attacks that won first place in the robust model track of NeurIPS 2018 Adversarial Vision Challenge [57]. Prior to tool interaction, we asked the participants to complete the pre-quiz that consisted of 9 questions to assess their ML background and knowledge in AML. These questions included 4 checker questions on basic ML and 5 questions that would be taught by the tools 4 Checker Qs 5 Taught Qs. The checker questions were to ensure participants' self-reported expertise aligned with their background and to assess their attention during the study. After the pre-quiz, we introduced either AdvEx or Adversarial-Playground to the participants and provided them with 5 minutes to go through their beginning tutorials. The tutorials contained basic background on adversarial attacks (e.g., what an adversarial attack is) and guidance on navigating through the different components of the interface. Participants were also given the freedom to revisit these tutorials at any point during their interaction with the assigned tool. Following the tutorials, we provided the participants with 30 minutes to interact with the tool freely. We instructed the participants to use their assigned tool to learn about the FGSM attack as much as they could, and informed them that there would be a follow-up post-quiz to assess how much they had learned. Next, we asked the participants to complete a Likert scale post-questionnaire, which collected their opinions on the learning and usability aspects of their assigned tool. We then asked them to complete the post-quiz (19 questions), which comprised of the 9 original questions from the pre-quiz, along with 10 new questions that were taught 4 Checker Qs 5 Old Taught Qs 10 New Taught Qs. We ended the study with a qualitative interview that further asked for their thoughts and opinions on their assigned system.

The user study took about one hour and the participants received $15 for their effort. They were informed that the top 3 performers of both the pre-quiz and post-quiz would be awarded an additional $10.

5.2 Results and Analysis: Task Performance

Completion Times and Average Quiz Scores. Out of 24 participants, we removed two from the AdvEx condition and one from the Adversarial-Playground condition, whose pre-quiz checker scores were below 50%. On average, the 10 remaining AdvEx participants spent 3.97 minutes ($\sigma = 0.07$) on the pre-quiz, 16.17 minutes ($\sigma = 0.21$) on their interaction with AdvEx, and 5.17 minutes ($\sigma = 0.10$) on the post-quiz. For the 11 remaining Adversarial-Playground participants, they spent on average 4.38 minutes ($\sigma = 1.66$) on the pre-quiz, 16.26 minutes ($\sigma = 7.09$) on their interaction, and 7.52 minutes ($\sigma = 2.28$) on the post-quiz. Before interacting with AdvEx, the participants had an average pre-quiz score of 65.56% ($\sigma = 16.93\%$), and 50% ($\sigma = 17\%$) if excluding the checker questions. After, the participants earned an

Table 1. The results of the paired t-tests and the quiz averages of our participants (filtered & all). Our results show that AdvEx has a strong learning effect on both filtered and all participants, outperforming Adversarial-Playground. *OQ (“old questions”): 9 questions from the pre-quiz that are also included in the post-quiz. †NQ (“new questions”): 10 questions that are newly added in the post-quiz. ‡Average of total quiz (checkers + taught) scores.

Results of Adversarial-Playground Participants

	Paired T-Tests					**Quiz Averages**			
	Pre-quiz vs. Post-quiz	Pre-quiz vs. Post-quiz OQ*	Pre-quiz vs. Post-quiz NQ†	Pre-quiz Taught vs. Post-quiz Taught	Pre-quiz Checkers vs. Post-quiz Checkers	Pre-Quiz Checkers	Pre-Quiz Taught	Post-Quiz Checkers	Post-Quiz Taught
Filtered (11 Participants)	t = -2.432, p = 0.035	t = -2.283, p = 0.045	t = -1.785, p = 0.105	t = -3.131, p = 0.011	t = 0.289, p = 0.779	79.55% (σ = 18.77%)	58.18% (σ = 28.92%)	77.27% (σ = 23.6%)	80% (σ = 16.87%)
						67.68% (σ = 20.76%)‡		79.42% (σ = 16.04%)‡	
All (12 Participants)	t = -2.448, p = 0.035	t = -1.995, p = 0.0714	t = -2.031, p = 0.0671	t = -2.697, p = 0.0208	t = 0.29, p = 0.777	75% (σ = 23.84%)	60% (σ = 28.28%)	72.92% (σ = 27.09%)	78.89% (σ = 16.54%)
						66.67% (σ = 20.1%)‡		77.63% (σ = 16.51%)‡	

Results of AdvEx Participants

	Paired T-Tests					**Quiz Averages**			
	Pre-quiz vs. Post-quiz	Pre-quiz vs. Post-quiz OQ*	Pre-quiz vs. Post-quiz NQ†	Pre-quiz Taught vs. Post-quiz Taught	Pre-quiz Checkers vs. Post-quiz Checkers	Pre-Quiz Checkers	Pre-Quiz Taught	Post-Quiz Checkers	Post-Quiz Taught
Filtered (10 Participants)	t = -5.264, p = 0.00052	t = -6.128, p = 0.00017	t = -4.229, p = 0.00221	t = -6.482, p = 0.00011	t = 1.0, p = 0.34344	85% (σ = 21.08%)	50% (σ = 17%)	82.5% (σ = 26.48%)	93.33% (σ = 6.28%)
						65.56% (σ = 16.93%)‡		91.05% (σ = 7.46%)‡	
All (12 Participants)	t = -6.225, p = 0.00006	t = -6.661, p = 0.00004	t = -5.197, p = 0.0003	t = -5.88, p = 0.00011	t = -0.561, p = 0.5863	75% (σ = 30.15%)	51.67% (σ = 15.86%)	77.08% (σ = 27.09%)	90% (σ = 10.05%)
						62.04% (σ = 17.38%)‡		87.28% (σ = 11.32%)‡	

average post-quiz score of 91.05% (σ = 7.46%), and 93.33% (σ = 6.28%) if excluding the checker questions. For Adversarial-Playground, participants had an average pre-quiz score of 67.68% (σ = 20.76%), and 58.18% (σ = 28.92%) if excluding the checker questions. After their interaction, participants earned an average post-quiz score of 79.42% (σ = 16.04%), and 80.00% (σ = 16.87%) if excluding the checkers. These results show that AdvEx enables more significant learning improvements, with a ~25% increase in post-quiz scores compared to ~12% for Adversarial-Playground. Furthermore, AdvEx participants demonstrated more consistent learning outcomes, shown by lower standard deviations in post-quiz. While the difference between the mean pre-quiz and post-quiz scores clearly indicates AdvEx's stronger effectiveness in enabling learning, we further answered **A1** by performing several paired t-tests on our collected quantitative data.

Learning Improvements Within Conditions. For AdvEx, our first paired t-test shows a significant difference between the participants' overall pre-quiz and post-quiz performance ($t = -5.264, p = 0.00052$); the difference is also significant in the second paired t-test when the checker questions are excluded ($t = -6.482, p = 0.00011$). Both results indicate a strong performance improvement after interaction with AdvEx. A third paired t-test shows a significant difference between their performance on the same 9 questions in the pre-quiz and post-quiz ($t = -6.128, p = 0.00017$). This indicates that the participants have successfully learned the answers to the questions that were originally included in the pre-quiz. Similarly, a significant difference can be observed between the participants' pre-quiz performance and their performance on the 10 newly added questions in the post-quiz ($t = -4.229, p = 0.00221$). This shows that the participants have picked up additional knowledge that was not mentioned in the pre-quiz during their interaction with AdvEx. Lastly, another paired t-test was performed between their performance on the same checker questions in the pre-quiz and post-quiz and no significant difference was found ($t = 1.0, p = 0.34344$). In conjunction with the fact that all

10 qualified participants scored a minimum of 50% on the pre-quiz checker questions, this suggests that our participants maintained consistency in their checker responses and did not select answers randomly.

For the Adversarial-Playground condition, participants also showed significant differences in overall pre-quiz to post-quiz performance ($t = -2.432, p = 0.035$), pre-quiz to post-quiz without checkers ($t = -3.131, p = 0.011$), and on questions present in both quizzes ($t = -2.283, p = 0.045$). However, no significant improvement was observed on newly added questions (t = -1.785, p = 0.105), indicating that participants struggled to acquire entirely new insights. This suggests that participants were able to consolidate and apply knowledge from Adversarial-Playground to questions they had already encountered, but gaining entirely new insights requires more support. Compared to AdvEx, which showed statistically significant improvement across both familiar and newly added questions, Adversarial-Playground appears less effective in facilitating the acquisition of new AML knowledge.

We repeated our statistical tests on all 24 participants for both conditions, including the three participants who were originally excluded. Our results demonstrate that AdvEx still has a strong learning effect on all its participants (Table 1). This finding suggests that while AdvEx is primarily designed for learners with a basic ML background who are new to AML, it benefits not only the intended users but also proves effective for those without fundamental ML knowledge seeking to understand adversarial attacks. On the other hand, for Adversarial-Playground, while the overall quiz improvement was statistically significant ($t = -2.448, p = 0.035$), this significance does not hold when isolating the seen and newly added questions separately ($t = -1.995, p = 0.0714$; $t = -2.031, p = 0.0671$). This suggests that for those without foundational ML knowledge, Adversarial-Playground only supports broad engagement, while for those with prior knowledge, it can deepen understanding of familiar content but require additional scaffolding for new concepts. The ability of AdvEx to accommodate a wider audience further emphasizes its value as an educational tool, extending its potential impact by making complex AML concepts more approachable even to those just beginning to explore the field of ML. The full results of all our paired t-tests and the quiz averages of the participants are shown in Table 1.

Comparative Learning Outcomes Across Conditions. To complement the paired t-tests, we performed independent samples t-tests to compare learning gains between AdvEx and Adversarial-Playground directly with effect sizes (Cohen's d), providing a broader perspective on how each tool supports AML learning. We defined five measures to assess learning differences: Δ_1 **(Post-quiz − Pre-quiz)**, Δ_2 **(Post-quiz OQ − Pre-quiz)**, Δ_3 **(Post-quiz NQ − Pre-quiz)**, Δ_4 **(Post-quiz Taught − Pre-quiz Taught)**, and Δ_5 **(Post-quiz Checkers − Pre-quiz Checkers)**. Our results show that AdvEx demonstrated significantly larger overall learning gains (Δ_1) compared to Adversarial-Playground ($t = -2.364, p = 0.027$, $d = -0.97$). For newly added questions ($\Delta 3$), AdvEx participants showed statistically significant greater gains ($t = -2.387, p = 0.026$, $d = -0.97$), which confirms its strength in facilitating novel knowledge acquisition. Gains on non-checker questions (Δ_4) revealed a marginally non-significant trend favoring AdvEx ($t = -2.032, p = 0.054$, $d = -0.83$), suggesting potential advantages in explicitly taught material. For foundational checker questions ($\Delta 5$), no significant difference was found ($t = -0.515, p = 0.614$, $d = -0.21$), which indicates comparable support for basic ML understanding between the tools. For familiar questions (Δ_2), the unpaired t-test showed no significant difference between conditions ($t = -1.596, p = 0.128$, $d = -0.65$). However, since previous within-condition paired t-tests revealed that AdvEx participants showed significant differences on seen questions compared to Adversarial-Playground participants, this disparity suggests that variability across participants may reduce the observable difference in direct comparisons. These results reinforce AdvEx's strong effectiveness as a learning tool for AML understanding, with particular strengths in promoting general and advanced knowledge acquisition. Adversarial-Playground, while less consistent, supports modest gains in familiar content consolidation.

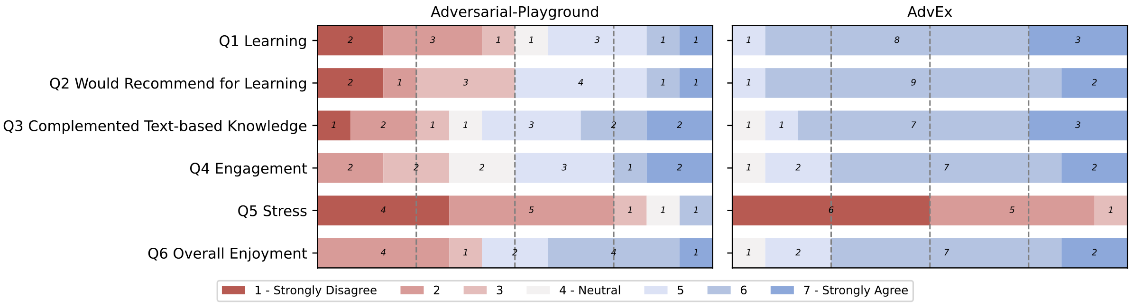


Fig. 6. Participants' questionnaire ratings (1 = "strongly disagree"; 7 = "strongly agree") on the learning and usability aspects of Adversarial-Playground and AdvEx.

5.3 Results and Analysis: Participants' Feedback

To further investigate **A1** and **A2**, we analyzed the participants' post-questionnaire responses on a 7-point Likert scale (Figure 6) and their qualitative feedback from the interviews on the learning and usability of AdvEx. The questionnaire asked users whether they: Q1) learned about adversarial attacks through AdvEx, Q2) would recommend AdvEx for learning, Q3) found AdvEx complemented text-based knowledge, Q4) felt engaged, Q5) felt stressed, and Q6) enjoyed overall interaction with AdvEx. We also asked them to complete a 10-item System Usability Scale (SUS) survey rated on a 5-point scale.

For Q1, all AdvEx participants agreed that they had learned about adversarial attacks through interacting with AdvEx (MD = 6, IQR = 0.25) and gave a positive rating (≥ 5) on AdvEx's learning effect. Specifically, the participants felt that AdvEx offered comprehensive visualizations and found the explanations very easy to understand. P3 stated that *"AdvEx teaches all aspects of adversarial attacks very thoroughly,"* and P8 commented that *"The clear explanations made the learning process much easier."* The participants also thought that AdvEx's visualizations were highly informative for them to understand the key attack properties and the underlying attack process. For Adversarial-Playground, participants felt it gave a basic idea of adversarial attacks but lacked depth (MD = 3.5, IQR = 3). They found it unclear on attack behavior or logic, offering only a surface-level overview. P19 mentioned, *"I tried experimenting to find the pattern of noise generation. But by the end of the interaction, it still just feels entirely random to me."*

Similarly, for Q2, all participants stated that they would recommend AdvEx to others for learning AML (MD = 6, IQR = 0). They believed that AdvEx would be highly beneficial for beginners and can serve as an effective entry point to those interested in learning about adversarial attacks. P2 thought that *"AdvEx is a valuable educational tool for illustrating the attacks,"* and P5 believed that *"AdvEx is great for beginners, it can teach them a lot about the attack process."* To further strengthen AdvEx as a learning tool, P5 suggested visualizing the internal attack process in more detail, such as how gradients are modified. For Adversarial-Playground, participants thought its lack of sufficient explanations and reliance on prior knowledge limits its practical learning value (MD = 4, IQR = 2.25). P22 commented, *"Though its visualization is simple, you will need AML background to understand both the text-based tutorial and what's happening."*

Q3's ratings show that AdvEx complemented the provided text-based knowledge well for learning (MD = 6, IQR = 0.25). From this, some participants believed that AdvEx could be used in conjunction with text-based documents, such as textbooks, to *"improve learning experience in traditional classroom settings"* -P1. Other participants felt that AdvEx was sufficient on its own for explaining adversarial attacks. *"I don't think AdvEx needs any additional complementary materials. Its visualizations are enough to thoroughly explain the attack logic."* -P4. For Adversarial-Playground, some found the text-based knowledge to be a helpful introduction to the visualization (MD = 5, IQR = 3.25). However, others criticized the visualization for failing to effectively illustrate the concepts. P19 and P16 tried to recreate the concepts described with the visualization but could not match the text's explanation, leading to frustration over the disconnection.

Eleven out of 12 participants gave a rating ≥ 5 on AdvEx's engagement in Q4 (MD = 6, IQR = 0.25). They applauded AdvEx for its highly interactive interfaces and enjoyed dynamically experimenting with the perturbation size to see all the real-time changes. *"It is highly engaging to change the noise level and observe how the resulting image differs."* -P1 Similarly, P5 stated, *"It is fun to see all the points move around in the Data Projectors as I adjust the slider."* However, one participant, P12, rated AdvEx's engagement a 4 and explained, *"In general, the application is good. But as a programmer, I feel like I should be able to get more involved and write custom code directly."* Participants found Adversarial-Playground to be a bit less engaging (MD = 4.5, IQR = 2.25). Many initially showed interest but became frustrated as the process felt random, with no discernible patterns or meaningful feedback to guide their learning of the attack logic.

For Q5, all participants agreed that it was not stressful to interact with AdvEx (MD = 1.5, IQR = 1). This was likely because AdvEx had an interactive tutorial system that provided guidance on AdvEx's functionalities, along with the General Information Provider that offered further assistance. Moreover, everything AdvEx visualized (e.g., 2-D latent space, confidence scores) were familiar to learners who knew ML, thus making AdvEx intuitive to learn with. *"Using AdvEx is very simple, I didn't encounter any difficulties. The visualizations are all quite straightforward and intuitive."* -P7 For Adversarial-Playground, participants also found it generally not stressful to use (MD = 2, IQR = 1.25). This was attributed to its very simple and intuitive visualizations, which made it easy to pick up and interact with. However, P23 rated their stress level a six, and explained that despite the tool's simplicity, they felt their exploration did not answer key questions about the attack or provide meaningful learning, leading to frustration.

In general, the participants rated AdvEx's enjoyment positively in Q6 (MD = 6, IQR = 0.25). They offered different reasons for why they enjoyed AdvEx. P9 and P10 claimed that AdvEx's visually appealing interfaces and animations made their interactions entertaining. P3, P8 & P10 emphasized the amount of knowledge they gained from AdvEx and found the learning experience fruitful. P4, P6 & P11 applauded AdvEx for its high level of interactivity. *"I enjoy AdvEx because I can do a lot with it. I can investigate different examples, try out different noise levels, and observe how the embedding distribution changes."* -P4. Similarly, participants in general enjoyed interacting with Adversarial-Playground, but with more varied opinions (MD = 5, IQR = 4). Overall, they appreciated Adversarial-Playground as an introductory tool to adversarial attacks, but hoped it would offer more detailed explanations or visuals to help them better understand how adversarial attacks work behind the scenes.

To further assess the usability of AdvEx, we analyzed and computed the SUS scores. AdvEx achieved a significantly higher overall SUS score (MD = 80 > 60, $p < 0.002$) compared to Adversarial-Playground. For effectiveness in aiding understanding, AdvEx scored higher (MD = 4 > 3, $p < 0.002$), again showing that it better supports learning. AdvEx also showed significantly better integration of features (MD = 4 > 2, $p < 0.001$), highlighting its more cohesive design compared to Adversarial-Playground. Additionally, neither system was found unnecessarily complex (MD = 2 = 2, $p < 0.610$), and both systems scored low difficulty for onboarding (MD = 1.0 < 1.5, $p < 0.151$). These results collectively demonstrate that AdvEx, despite its comprehensive features, is not overwhelming and supports intuitive interaction for learners, outperforming Adversarial-Playground in key usability metrics.

5.4 RQ1: Support of Understanding and Engagement

Based on statistical analyses of quiz performance and self-reported learning effects, AdvEx demonstrates strong learning outcomes and engagement. To investigate how AdvEx's visualization design effectively supports learning and uncover key design lessons, we further analyzed qualitative data from user interactions with the system and their follow-up interviews, addressing **RQ1**.

Linking between text-based explanations and visualizations. AdvEx establishes a meaningful connection between its text-based explanations and visualization components, ensuring they complement each other rather than functioning as standalone features. While many tools like Adversarial-Playground include both text and visualizations, they often feel disconnected, as noted by P16 and P19, making it challenging for learners to draw connections. AdvEx addresses this by tightly linking its textual and non-textual components to create a cohesive system. For example, P6, P7, and P8 started by using the step-by-step execution view (Figure 5) to gain a general understanding of the attack logic through animated textual sequences. They then switched to exploring visualizations, such as the Data Projectors (Figure 1c), to identify specific data instances or patterns of interest. Once an instance was identified, they returned to the step-by-step execution, using its built-in feature to substitute the selected instance as a practical example for the textual explanations, reinforcing their understanding of attack behavior and logic with real-world data. This iterative interplay between the textual and visual components encourages learners to actively engage with the two types of material rather than passively absorb information.

Multi-level visualizations to uncover unique attack properties. AdvEx offers visualizations at multiple levels, which not only allows learners to examine attack at varying granularities, but also extract unique attack properties specific to each level. For instance, the Data Projectors present dataset-level embeddings that help learners identify general attack behaviors. By observing adversarial perturbations, learners noted that *"attacks often push instances from one class cluster toward another cluster."* -P4. This helps them conceptualize adversarial attacks as deliberate manipulations of data distributions rather than random alterations. We observed that the Instance-level Attack Explainer (Figure 1d) allows learners to identify finer-grained properties, such as attack imperceptibility and the trade-offs between noise strength and success. Learners noted that *"incrementally increasing attack strength makes attacks more effective but also results in noisier, more perceptible images."* -P6. This illustrates the balance attackers must achieve between subtlety and effectiveness, giving learners a hands-on understanding of these critical dynamics. In contrast, Adversarial-Playground offers only instance-level side-by-side input images before and after attacks, along with a noise slider for exploration. While this offers basic interactivity, learners often adjusted the slider randomly, leading to misconceptions that *"adversarial attacks are easily detectable due to visual noise."* -P20.

Comparative visual exploration through space and time. AdvEx incorporates comparative visual designs, allowing learners to examine adversarial attacks through both spatial and temporal comparisons. Spatial comparisons include the Robustness Analyzers (Figure 1a) for accuracy versus robustness, Data Projectors for embedding comparisons, Instance-level Attack Explainer (Figure 1d) for analyzing before-and-after images, and the confidence score view (Figure 1d2) for score comparisons. Learners often start by identifying a misclassified data instance, shown as a split-colored circle in the Data Projectors, and compare it side by side across models to understand *"why one worked while the other failed."* -P4. This process involves comparing between the two instances and analyzing properties like embedding distributions or confidence scores. These spatial comparisons guide learners to seek information systematically by examining differences across models and attack levels. In addition to spatial comparisons, AdvEx supports temporal comparisons through the Perturbation Adjuster (Figure 1b). By adjusting the slider, learners visualize how noise changes data instances from correctly classified to misclassified, showing attack progression dynamically. Temporal comparisons reveal how attack properties change with increasing perturbation strength, such as shifts in overall accuracy, embedding movements, or the perceptibility of image noise. By comparing data across models, levels, and time, learners better connect theoretical concepts to observable phenomena, which promotes deeper insights into attack behavior and their impact on model robustness.

5.5 RQ2: Learning Workflow

To investigate what workflows AdvEx's visual design supports to effectively facilitate learning, we observed how learners interacted with the system and performed further qualitative analysis to address **RQ2**.

Establishing foundational understanding before experimentation. The first stage of learners' workflows with AdvEx involves building a foundational understanding of the general attack logic and the data distribution before proceeding to experimentation. In contrast to Adversarial-Playground, where learners typically dove straight into experimenting by randomly adjusting the noise slider to observe its effects, AdvEx learners took the time to develop a concrete understanding of the attack and dataset first. During the study, most learners began by spending considerable time with the step-by-step execution view (Figure 5) to understand the underlying attack logic. Many also actively explored data points in the Data Projectors (Figure 1c) to familiarize themselves with the dataset and examine the embedding distribution prior to conducting any attacks. This initial stage of structured exploration helped learners gained a clear understanding of *"how the dataset was distributed and how adversarial attacks worked conceptually"* -P12 before attempting to gain deeper insights. Unlike Adversarial-Playground, where learners might lack context due to their exploratory yet unstructured workflows, AdvEx provides learners with visualizations and text-based explanations to make sense of the dataset and attack mechanics, ensuring that subsequent insights are built upon a clear conceptual framework, making advanced exploration more meaningful and effective.

High-level iterative perturbation experimentation. After establishing a foundational attack understanding, learners transitioned into an iterative process of adjusting the Perturbation Adjuster (Figure 1b) to experiment with different attack strengths. At this stage, they focus on higher-level views, such as the Robustness Analyzers (Figure 1a) and the Data Projectors, to observe the broader effects of increasing perturbation size. Through the Robustness Analyzers, learners quickly grasped how the model's robustness degraded as attack strength increased. Simultaneously, the Data Projectors allowed them to observe the dynamic movement in the embedding space, with color changes highlighting the growing number of misclassifications. In contrast, while Adversarial-Playground also supported an iterative workflow using its noise slider, the intent and outcomes differed. AdvEx learners used iterative adjustments with a clear purpose: to *"understand how attacks altered the model's overall robustness and feature representations."* -P3. Adversarial-Playground learners, on the other hand, often adjusted the slider without a specific objective, relying instead on trial-and-error to observe changes. Therefore, providing high-level visualizations that connect theoretical concepts with practical observations enables more purposeful and structured experimentation to understand attack dynamics.

Multi-level exploration of specific instances. After obtaining a high-level view of attack properties through iterative experimentation, learners transitioned into a detailed multi-level workflow. This stage involved navigating between high-level and low-level views to investigate specific data points. Learners started by identifying an instance of interest, often focusing on misclassified data points. They selected these points to explore lower-level visualizations, examining pixel changes and confidence scores. Learners adjusted the slider to observe prediction changes with attack strength or compared the instance across models using side-by-side Data Projectors. For example, when one model correctly classified the instance while another misclassified it, learners analyzed differences in how the attack affected predictions. In contrast, while learners attempted similar workflows with Adversarial-Playground, the tool's limitations hindered their ability to effectively analyze attacks. They often began by adjusting the noise slider and comparing side-by-side images to identify patterns. However, they found it difficult to extract meaningful insights from the applied noise alone. As a result, learners shifted their focus to the classification likelihoods, incrementally adjusting the noise level to detect patterns. Despite these efforts, learners often reported in interviews that they could not identify significant or consistent behaviors from these visualizations. This lack of actionable feedback limited their ability to progress in their workflow, highlighting the importance of AdvEx's multi-level design, which integrates views that complement one another to support both broad and granular analysis of adversarial attacks.

6 Interview Study with Experienced Experts/Teachers

To collect more in-depth qualitative feedback on AdvEx, we conducted an interview study with AML experts/teachers, who possess profound knowledge of the key aspects and requirements for understanding adversarial attacks. These interviews provided additional insights into how AdvEx can be utilized in an educational setting.

6.1 Study Setup

In this study, AML experts/teachers were prompted to use AdvEx to explore one white-box attack and one black-box attack, FGSM and ZOO, on four different models (VGG-16, VGG-19, standard ResNet-34, & adversarially trained ResNet-34) with the CIFAR-10 data in a free-form analysis session. We recruited seven AML experts (E1, E2, E4 ∼ E8; all men), six of whom have teaching experience that spans from leading AML seminars to teaching ML courses with AML components. Each study session began with a 5-minute introduction to the study background and AdvEx's key features. Next, we presented a task scenario where participants were asked to use AdvEx to explore and understand *"how the FGSM and ZOO attacks alter the input images to affect the models' performance,"* and *"how models display varying robustness against the attacks."* Participants had 30 minutes to explore each attack, and a task list was provided to guide their interaction. They were also informed that they could explore the tool freely without following these tasks as long as insights were gathered. We employed the think-aloud protocol, requiring participants to provide feedback from both the perspectives of *experts/teachers* and *learners*. An experimenter was responsible for providing help and answering questions regarding the interface, who also observed the experts' interactions and took notes. After the interaction, a semi-structured interview (≈30 minutes) was conducted to gain a better understanding of the participants' thoughts on AdvEx in light of the think-aloud feedback and observation gathered previously. The participants were compensated $20/hour for the study.

6.2 Results and Analysis

All seven experts successfully used AdvEx to gain insights into the attacks and expressed a positive sentiment toward it. We conducted a thematic analysis on the unstructured feedback gathered during the free-form analysis and the qualitative data provided to us during the semi-structured interviews. We came up with five systematic themes aligned with our design goals and an additional theme focused on usability, and adopted a deductive approach to identify patterns of them in our data.

Visualizations of attack impacts. All experts agreed that AdvEx can help learners quickly grasp the attack impacts. E4 and E6 liked the Robustness Analyzers for illustrating *"the overall trend of accuracy changes."* E1, E6, and E8 valued the Data Projectors for allowing learners to *"see how embeddings are drifted from their original positions."* They also found the lower-level visualizations highly useful. E7 noted, *"The confidence score view can show learners that ZOO [...] pushes instances just past the decision boundary."* This confirms that AdvEx effectively visualizes the attack impact at multiple levels (**G1**). A noted limitation is the occasional difficulty in distinguishing between points misclassified before and due to the attack. E1 suggested displaying both the original and current predictions in the attack explainer.

Generalizability. The experts praised AdvEx for its generalizability to different attacks and image classifiers. E2 explained, *"AdvEx's ability to adapt to different attacks and models is vital for learners to truly understand the risks by evaluating against diverse techniques."* This confirms that it can effectively help learners assess the variability of models and attack methods (**G2**). Moreover, the experts highlighted that such design simplifies the exploration by providing a more accessible way to investigate different attacks. Both E1 and E2 pointed out how AdvEx saves learners' time by eliminating the need to code from scratch when exploring different attack strategies on their own models.

Evaluation of model robustness. The experts believed that AdvEx can help learners easily discern their models' strengths and weaknesses. The model comparison feature was frequently highlighted, with E7 noting that it can reveal that *"deeper models do not necessarily excel under attacks."* Similarly, E5 and E6 commented that it shows that a robust model has *"embeddings that barely differ under standard or adversarial conditions."* These comments affirm AdvEx's capability for detailed visual analysis and model comparison (**G3**). While the experts valued how comprehensive the model visualizations are, E1 and E4 suggested adding comparison of the same model under attacks with different perturbation sizes side by side.

Dynamic experimentation with real-time changes. The experts enjoyed dynamically experimenting with the perturbation size and found the real-time visual feedback valuable. E2 commented, *"AdvEx answers questions that papers and tutorials may not cover, such as the effects of varying perturbation sizes on model embeddings."* Furthermore, they believed the integration of dynamic perturbation adjustment and real-time visual feedback offers an engaging learning experience. E2 explained that learners could play with AdvEx for self-learning, while E1 thought teachers could use the tool to *"demonstrate attacks in a fun, interactive, and engaging way."* These observations suggest that AdvEx provides a highly interactive learning experience with its perturbation experimentation and real-time feedback (**G4**).

Overall benefits as an educational tool for learners. All experts agreed that AdvEx is highly beneficial as an educational tool. E6 stated, *"AdvEx bridges theory and practice, enhancing learners' understanding [...] and encouraging them to further explore the field."* They also thought the step-by-step execution would be very informative for AML learners, confirming AdvEx's capability to enable detailed learning of the attack process (**G5**). In addition, the experts believed that AdvEx's interfaces would make the learning experience highly enjoyable. E1 commented, *"AdvEx's game-like experience makes learning and evaluating models much easier for learners without too many tedious formulas."*

Usability & beginner-friendly design. All experts thought AdvEx was very intuitive to pick up. E5 liked how the beginning tutorial highlighted specific areas of the interface, which helped him easily understand the purpose of each component. E1 thought learners less experienced with AML could also pick up AdvEx easily. This confirms that AdvEx was successfully integrated with a beginner-friendly design. The experts also thought that AdvEx was very accessible. E5 highlighted the zoomable binned aggregation feature and commented, *"This feature effectively accommodates different users' available computational power and enable smooth exploration of large-scale data for everyone."*

7 Discussion

Here, we discuss the limitations of our current system and outline future directions to enhance our work. We also present the potential avenues for extending and generalizing our proposed design.

7.1 Limitations and Future Work

While our study confirms that AdvEx is highly effective in helping learners understand adversarial attacks, it still has several limitations. Firstly, as commented by our participants, the current Data Projectors (Figure 1c) allow comparisons of two different models under the same perturbation level, but do not support comparing the same models side by side at different perturbation levels. Future extensions should enable this type of comparison without adjusting the slider back and forth. A simple solution is to add additional toggles to the projectors for switching between the different comparison modes. Similarly, AdvEx currently lacks historical tracking, making it less intuitive to revisit prior states. A lightweight timeline or step-based log could also help improve exploration continuity.

Secondly, when the perturbation size exceeds 0, distinguishing instances misclassified before the attack from those misclassified due to the attack becomes less intuitive. This can be easily mitigated by implementing additional visual encodings such as using different shapes (triangles and crosses) to represent the two types of misclassifications. However, this approach may increase the cognitive load of users. But an optional filtering feature can be added to allow users to focus only on either type of misclassification.

Thirdly, due to color distinguishability [33], AdvEx is currently limited to handling datasets with ≤ 12 classes or subsets of larger datasets. While our tool could theoretically support more colors, increasing the number beyond twelve would reduce the effectiveness of exploration due to colors appearing too similar. Nonetheless, Our approach aligns with common encoding methods used in existing visualization tools [23, 44], and AdvEx accommodates datasets with a larger number of classes by allowing learners to focus on exploring a smaller subset of more critical classes. In future work, alternative encoding methods (e.g., using a combination of color with other visual cues like shapes or patterns) could be explored to potentially expand the number of distinguishable classes while preserving the clarity and usability of the visualization.

Furthermore, the evaluation of AdvEx can be further enhanced. A larger sample size should be obtained to better evaluate AdvEx's effectiveness. Also, the current study was designed with a few selected models, and only the FGSM and ZOO attacks with the CIFAR-10 data were used to assess the learning effect and usability of AdvEx. In the future, deployment studies with other types of attacks and datasets should also be conducted to investigate how AdvEx can be used in various real-world domains. This will thoroughly examine the strengths and weaknesses of AdvEx, and help us understand how AdvEx can be potentially incorporated into learners' model development workflows.

Finally, this work primarily explored how multi-level visualization designs complement traditional teaching materials, particularly by enabling novices to gain hands-on experience with adversarial attacks. AdvEx allows learners to explore models and datasets tailored to their interests, bridging theoretical concepts with practical understanding. Participants in our study found AdvEx intuitive, well-integrated, and effective for learning, without being overwhelming. However, future work could enhance AdvEx's accessibility further by integrating visualization with traditional materials through approaches like scrollytelling. Such designs could provide more structured guidance for learners, transitioning them from foundational AML concepts to understanding the impacts of various attack algorithms and defense mechanisms.

7.2 Generalization and Extension

Generalization to other applications. We designed AdvEx as a system for visualizing adversarial attacks, but the tool is flexible enough to be adapted to visualize other data augmentations in image classification. For example, AdvEx can be extended to visualize noise applications (e.g., Gaussian noise, salt-and-pepper noise) or other image degradations (e.g., motion blur, JPEG compression). Learners may use AdvEx to understand how visual quality impacts the performance of models in those scenarios. While our focus is primarily adversarial attacks in image classification, we acknowledge that there are similar attacks in other prevalent ML tasks, such as object detection [52], audio processing [3], and NLP [58]. Although visualizing attacks in these domains is outside our current scope, we believe there is potential for our approach to be generalized to these domains. Components like Robustness Analyzers, Data Projectors, and confidence score view are already well-generalized and provide an intuitive way to explore and assess models' accuracy, feature representations, and output probabilities across different domains, making such designs easily extendable to other ML tasks. Learners can use these views to observe embedding movements and output changes when subtle background noises are applied to audio inputs, or after letters in text are maliciously replaced, deleted, or swapped. For other components like Instance-level Attack Explainer, future enhancement may be made to generalize them to different types of data instances. For example, further implementations could be made to detect input data type and adjust the layout of the attack explainer accordingly to better accommodate the data. For example, for NLP, the view can be adjusted to highlight which letters in the text instances have been modified. For object detection, the view can display the before and after of bounding box manipulation that causes models to miss the targets. The extension of our tool to these areas could further heighten its educational value and provide additional knowledge to learners into adversarial attacks across a spectrum of ML applications.

Extension to other ML concepts. AdvEx leverages a balanced combination of active visualizations and passive text-based information to help learners understand AML, and this design can be applied to visualization tools for learning other ML concepts. In fact, many existing tools (e.g., [10, 34]) only focus on their interactive visualizations and place little emphasis on text-based information, not providing enough guidance and background knowledge to the users. On the other hand, interactive articles [20] usually involve mainly text and provide insufficient visualizations. AdvEx places more balanced weights on both components, ensuring that the users may gain detailed and accurate AML knowledge from our General Information Provider (Figure 1e) in addition to exploration with the visualizations. Our design not only reinforces learning by presenting content in multiple formats, but also allows the learners to quickly grasp complex topics requiring visual interpretations, which could shed light on future research on the spectrum of modalities for teaching ML concepts. However, one potential addition is the visualizations of model learning dynamics, such as gradient and weight evolution. By doing so, AdvEx could generalize to other ML concepts by offering insights into how models adapt across different learning settings.

8 Conclusion

We have presented AdvEx, an interactive visualization tool for novice AML learners to explore and understand adversarial attacks. Based on the design guidelines derived from user interviews, we designed AdvEx to provide learners with detailed attack visualizations at multiple levels, highlighting attack's properties and effects on different image classifiers. Our design addresses the limitations of existing tools, which lack comprehensiveness and generalizability when visualizing the attacks. We quantitatively and qualitatively assessed AdvEx in a two-part evaluation, including a user study with 24 AML learners and an interview study with seven AML experts/teachers. Our results indicate that AdvEx is not only highly effective as an educational tool, but also provides an engaging and enjoyable learning experience, thus highlighting its overall benefits for AML learners. Additionally, we discuss the future directions to enhance our work and present potential avenues to extend and generalize AdvEx to other applications.

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] 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.

[2] 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

[3] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*. IEEE, 1–7.

[4] 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

[5] 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

[6] Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist* 49, 4 (2014), 219–243.

[7] Konstantina Christakopoulou and Arindam Banerjee. 2019. Adversarial attacks on an oblivious recommender. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 322–330.

[8] 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

[9] Francesco Croce and Matthias Hein. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*. PMLR, 2206–2216.

[10] 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

[11] Li Deng. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* 29, 6 (2012), 141–142. https://doi.org/10.1109/MSP.2012.2211477

[12] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1625–1634. https://doi.org/10.1109/CVPR.2018.00175

[13] Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences* 111, 23 (2014), 8410–8415.

[14] Lin Gao, Zekai Shao, Ziqin Luo, Haibo Hu, Cagatay Turkay, and Siming Chen. 2023. TransforLearn: Interactive Visual Tutorial for the Transformer Model. *IEEE Transactions on Visualization and Computer Graphics* (2023).

[15] 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

[16] 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

[17] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. 2019. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 588–597. https://doi.org/10.1109/CVPR.2019.00068

[18] Dan Hendrycks and Kevin Gimpel. 2016. Early methods for detecting adversarial images. *arXiv preprint arXiv:1608.00530* (2016). https://doi.org/10.48550/arXiv.1608.00530

[19] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. 2019. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems* 32 (2019).

[20] Fred Hohman, Matthew Conlen, Jeffrey Heer, and Duen Horng Polo Chau. 2020. Communicating with interactive articles. *Distill* 5, 9 (2020), e28. https://doi.org/10.23915/distill.00028

[21] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. 2019. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1096–1106. https://doi.org/10.1109/TVCG.2019.2934659

[22] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems* 32 (2019).

[23] Minsuk Kahng, Pierre Y Andrews, Aditya Kalro, and Duen Horng Chau. 2017. A cti v is: Visual exploration of industry-scale deep neural network models. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 88–97.

[24] Minsuk Kahng and Duen Horng Chau. 2019. How does visualization help people learn deep learning? Evaluation of GAN Lab. In *IEEE VIS 2019 Workshop on EValuation of Interactive VisuAl Machine Learning Systems*.

[25] A Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Tronto* (2009).

[26] 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.

[27] 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

[28] Da Lin, Yuan-Gen Wang, Weixuan Tang, and Xiangui Kang. 2022. Boosting Query Efficiency of Meta Attack With Dynamic Fine-Tuning. *IEEE Signal Processing Letters* 29 (2022), 2557–2561. https://doi.org/10.1109/LSP.2022.3229558

[29] 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

[30] 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

[31] 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

[32] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2574–2582. https://doi.org/10.1109/CVPR.2016.282

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

[34] 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

[35] Priyadarshini Panda and Kaushik Roy. 2021. Implicit adversarial data augmentation and robustness with Noise-based Learning. *Neural Networks* 141 (2021), 120–132. https://doi.org/10.1016/j.neunet.2021.04.008

[36] 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

[37] Cheonbok Park, Soyoung Yang, Inyoup Na, Sunghyo Chung, Sungbok Shin, Bum Chul Kwon, Deokgun Park, and Jaegul Choo. 2021. VATUN: Visual Analytics for Testing and Understanding Convolutional Neural Networks. In *Eurographics Conference on Visualization (EuroVis)*. The Eurographics Association. https://doi.org/10.2312/evs.20211047

[38] 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

[39] Huy Phan. 2021. huyvnphan/PyTorch_CIFAR10. (Jan 2021). https://doi.org/10.5281/zenodo.4431043

[40] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2017.16

[41] 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).

[42] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. 2020. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceedings of the 37th International Conference on Machine Learning*. 7909–7919.

[43] Pradeep Rathore, Arghya Basak, Sri Harsha Nistala, and Venkataramana Runkana. 2020. Untargeted, targeted and universal adversarial attacks and defenses on time series. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207272

[44] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D Williams. 2016. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 61–70.

[45] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. 2018. Is Robustness the Cost of Accuracy?–A Comprehensive Study on the Robustness of 18 Deep Image Classification Models. In *Proceedings of the European conference on computer vision (ECCV)*. 631–648. https://doi.org/10.1007/978-3-030-01258-8_39

[46] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation* 23, 5 (2019), 828–841. https://doi.org/10.1109/TEVC.2019.2890858

[47] Lu Sun, Mingtian Tan, and Zhe Zhou. 2018. A survey of practical adversarial example attacks. *Cybersecurity* 1 (2018), 1–9. https://doi.org/10.1186/s42400-018-0012-9

[48] 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

[49] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0. https://doi.org/10.1109/CVPRW.2019.00012

[50] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[51] Zijie J Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Polo Chau. 2020. Cnn explainer: Learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1396–1406. https://doi.org/10.1109/TVCG.2020.3030418

[52] 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.

[53] Hao Ye, Geoffrey Ye Li, and Biing-Hwang Juang. 2017. Power of deep learning for channel estimation and signal detection in OFDM systems. *IEEE Wireless Communications Letters* 7, 1 (2017), 114–117. https://doi.org/10.1109/LWC.2017.2757490

[54] Youjie Ye, Yunfei Chen, and Mingqian Liu. 2022. Multiuser Adversarial Attack on Deep Learning for OFDM Detection. *IEEE Wireless Communications Letters* 11, 12 (2022), 2527–2531. https://doi.org/10.1109/LWC.2022.3207348

[55] João G. Zago, Eric A. Antonelo, Fabio L. Baldissera, and Rodrigo T. Saad. 2020. It is double pleasure to deceive the deceiver: disturbing classifiers against adversarial attacks. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 160–165. https://doi.org/10.1109/SMC42975.2020.9282889

[56] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. 2017. Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. 39–49. https://doi.org/10.1145/3128572.3140449

[57] 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.

[58] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 3 (2020), 1–41.

[59] Yu Zhang, Gongbo Liang, Tawfiq Salem, and Nathan Jacobs. 2019. Defense-pointnet: Protecting pointnet against adversarial attacks. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 5654–5660. https://doi.org/10.1109/BigData47090.2019.9006307