# A Technical Appendix

This supplementary material serves as technical appendix with sections given detailed information about 1) how difficult the different datasets are; 2) the distribution of class levels within each difficulty range across all datasets; 3) an illustrative selection of images in each difficulty bin and 4) performance comparison of $M_{\text{orig}}$ and $M_{\text{SAdv}}$ on a per-instance basis for each difficulty level.

## *A.1 Class difficulty*

Figure 6 shows the difficulty distribution per dataset and class, sorted by difficulty. Of course, the difficulty level of the classes for the different datasets can vary depending on the population of systems used to solve the classification tasks. Still, in our case, we are using a large population of systems (1000 for MNIST, 449 for FashionMNIST and 156 for CIFAR10), using the data from [19], which in turn comes from the OpenML platform[4].

Starting with MNIST, this is a dataset of handwritten digits from 0 to 9. The images in the MNIST dataset are normalised and centred, which makes the dataset less challenging (for image classification tasks) compared to other datasets such as CIFAR-10. In general, all the classes in the MNIST dataset are considered relatively easy to classify, as the images are well-segmented and have a clear contrast. However, some digits such as 5, 8 seem to be slightly more difficult to classify (probably because they are more similar to other digits like 6 or 9 respectively).

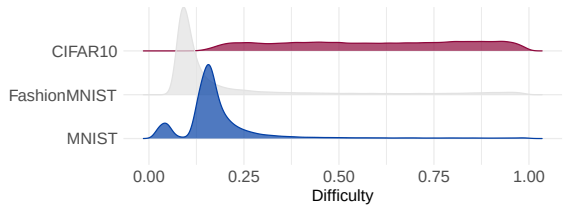


**Figure 5**: Difficulty distribution per dataset. Sorted by average difficulty.

Newer datasets such as FashionMNIST (a dataset of images of clothing and accessories) are considered to be more challenging than MNIST in terms of complexity and diversity. However, we see that FashionMNIST does not make a great difference in terms of the aggregated distribution of difficulty compared to MNIST (see Figure 5). In general, some classes in the FashionMNIST dataset are more difficult to classify than others. For example, the trousers, bags, sneakers or sandals classes are relatively easy to classify, while the shirt, pullover, coat or t-shirt are considered more difficult (see the tail to the right of their difficulty distributions). This is probably due to the similarities in the images of these classes and the variations in the texture, shape, and color.

Finally, we see higher average difficulties and differences between classes for CIFAR10. The images in this dataset are relatively complex, with objects that are often occluded, partially visible or in non-frontal poses. These factors make the dataset more challenging than the above datasets. What we see in Figure 6 is that, in general, the vehicle-related images for CIFAR10 are considered relatively easy, while the animal-related images are considered more difficult. This is due to the similarities in the images of these classes, which can make them hard to distinguish. Some images of deer and horse are similar and can cause confusion to a model.

[4] https://www.openml.org/

## *A.2 Class distribution*

Figure 7 shows the class distribution for each difficulty range for the original and all adversarial datasets.

For the original dataset $D_{\text{Orig}}$, we see that some classes are easy, and have a high proportion of very easy instances: class 1 (digit 1) for MNIST, classes 1 (trouser), 8 (bag) and 7 (sneaker) for Fashion-MNIST, and class 8 (ship) for CIFAR10. Other difficult classes are more dominated by difficult instances: class 5 (digit 5) for MNIST, class 6 (shirt) for FashionMNIST, and class 3 (cat) for CIFAR10.

For the Simple Adversarial dataset ($D_{\text{SAdv}}$), the number of difficult instances has increased, but in an uneven way for some classes over others (note that the $y$-axis changes across plots).

For the Balanced Adversarial dataset ($D_{\text{BAdv}}$), is constructed in the same way as $D_{\text{SAdv}}$, we see how they have increased the number of difficult instances for all classes, but now the number of instances for each class is the same (which does not mean of course that the distribution of difficulty is the same for all classes).

For Double Adversarial dataset ($D_{\text{DAdv}}$), as we undersample the easiest instances, there are no instances in the lower bins for all classes.

## *A.3 Sample images for each difficulty bin*

Figure 8 shows sample images from each difficulty bin for MNIST, FashionMNIST and CIFAR10. Images belonging to the first three bins are relatively easy instances, but the images of the two hardest bins (d and e) are sometimes really challenging even from a human perspective. A human will find it difficult or will misclassify the examples provided in this figure for the hardest bin (Figure 8 (e)).

## *A.4 $M_{orig}$ VS. $M_{SAdv}$ confusion matrix*

Figure 9 compares the performance of $M_{\text{Orig}}$ and $M_{\text{SAdv}}$ on a per-instance basis for each difficulty level on the dataset constructed from MNIST, FashionMNIST and CIFAR10. We can observe that $M_{\text{Orig}}$ performs equally good or in most cases better than $M_{\text{SAdv}}$ in the first 4 bins when tested on MNIST and FashionMNIST and only performs worse in the last bin. We see almost the same trend for CIFAR10 with the difference of $M_{\text{SAdv}}$ performing slightly better in the fourth bin, too.

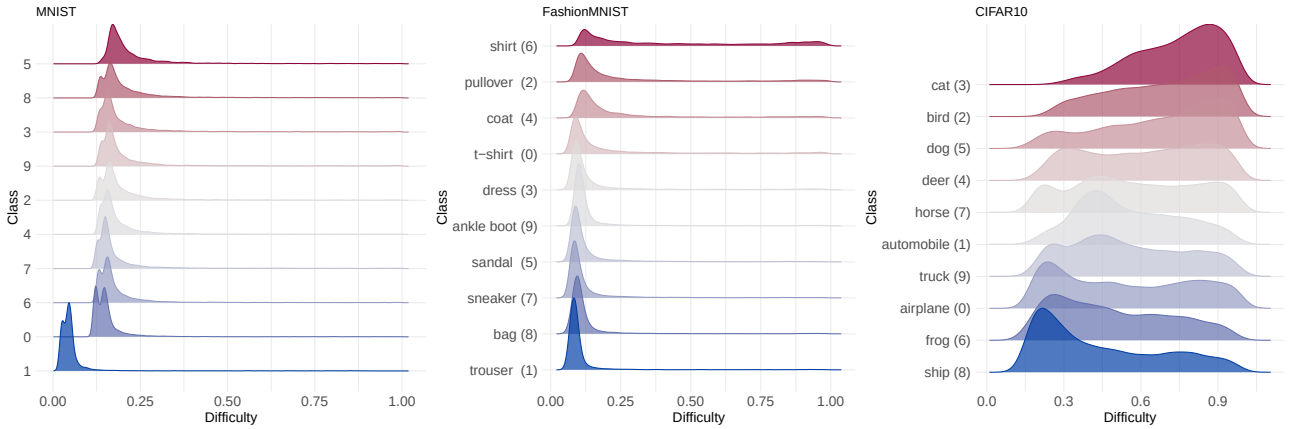


**Figure 6**: Difficulty distribution per dataset and class. Classes sorted by average difficulty.

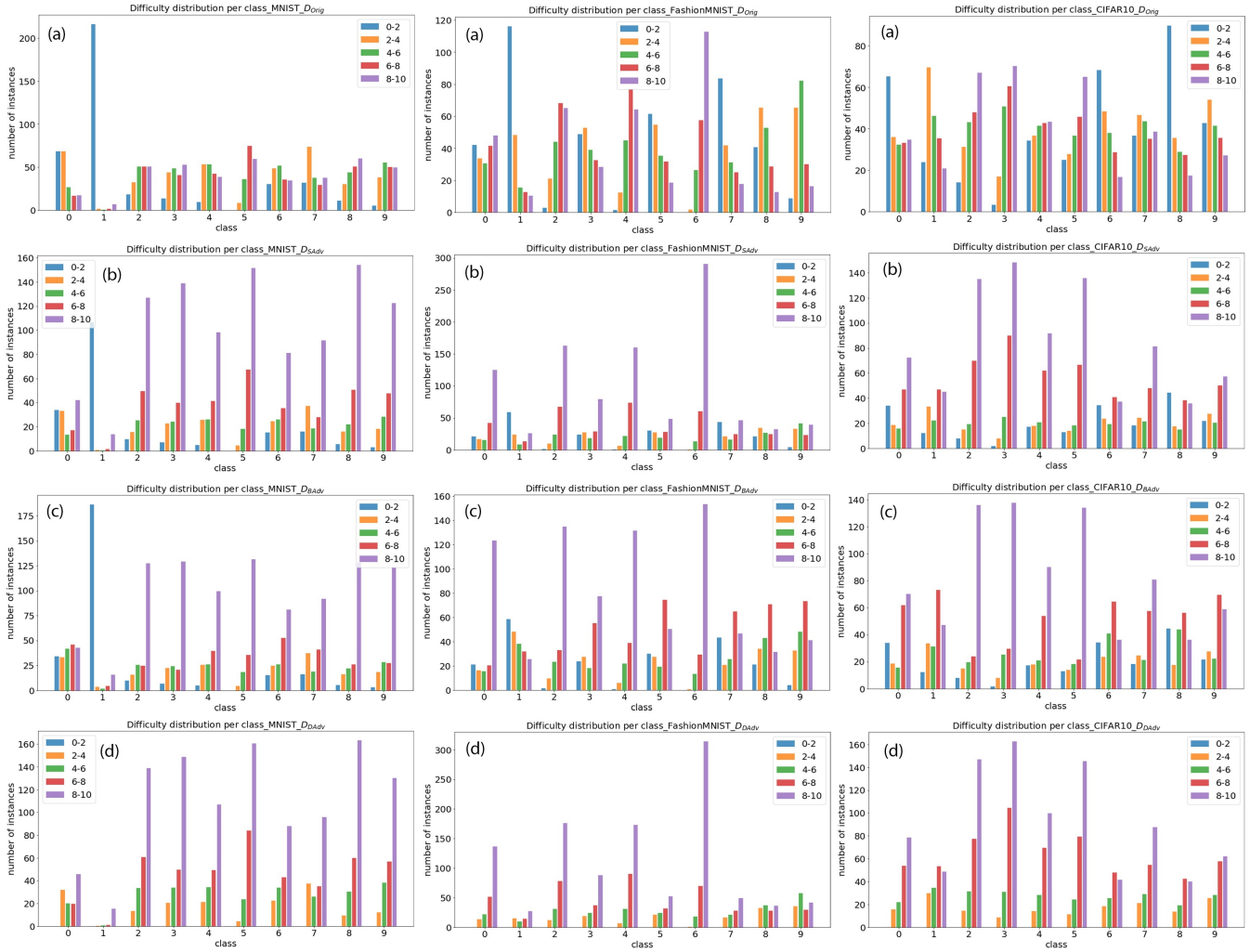


**Figure 7**: Proportion of each class for each difficulty range for all modified datasets constructed from MNIST (left), FashionMNIST (centre), and CIFAR10 (right), for $D_{\text{Orig}}$ (a), $D_{\text{SAdv}}$ (b), $D_{\text{BAdv}}$ (c), $D_{\text{DAdv}}$ (d). Class names for FashionMNIST and CIFAR10 in Figure 6.


**Figure 8**: The figure presents sample images from the MNIST (top), FashionMNIST (centre), and CIFAR10 (bottom) datasets organised by difficulty bin. The bins are designated (a) through (e); (a) represents the easiest bin, (b) the difficulty level between 2 and 4, (c) the difficulty level between 4 and 6, (d) the difficulty level between 6 and 8, and (e) the hardest bin which covers difficulty level between 8 and 10.

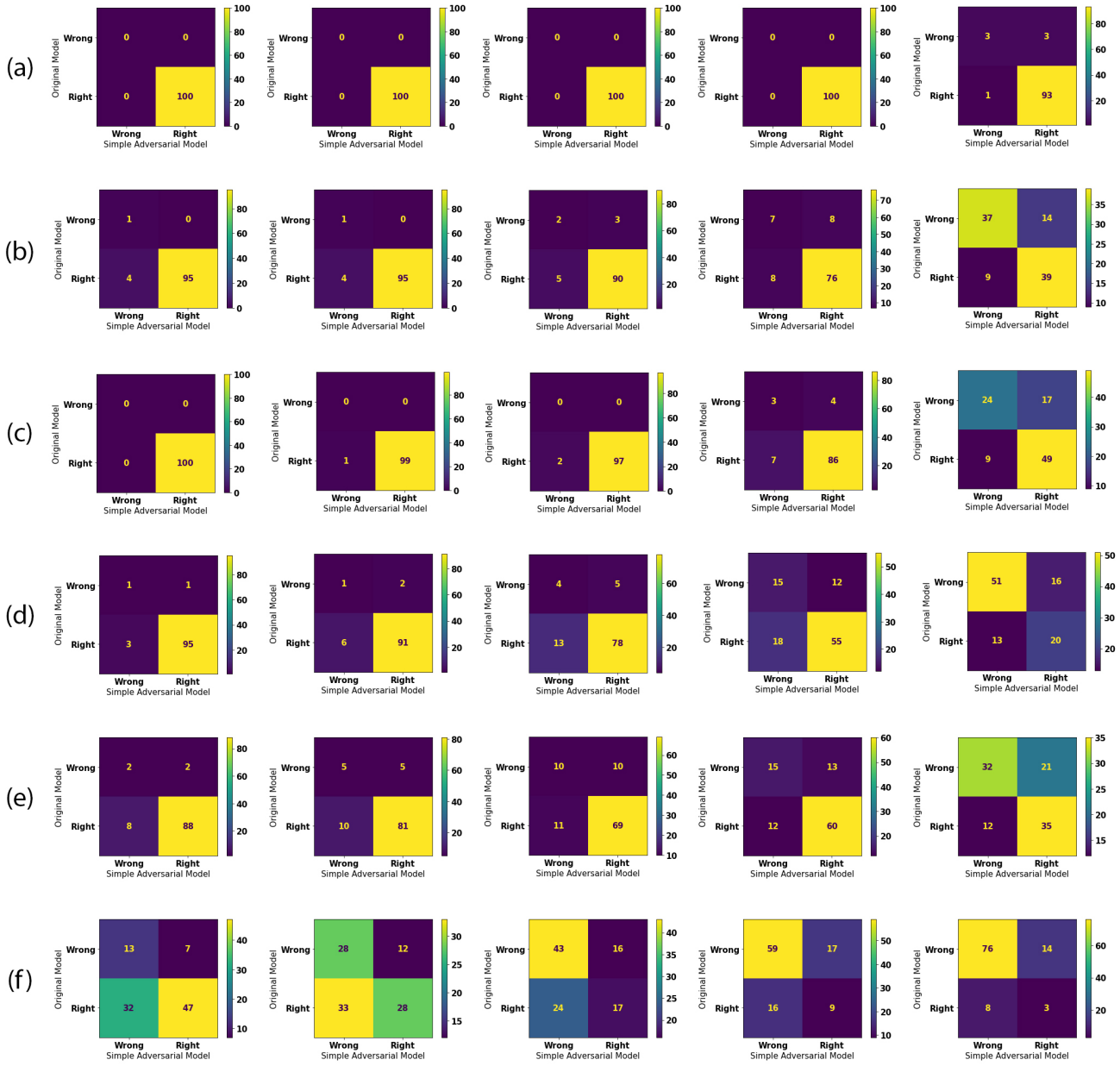


**Figure 9**: This graph compares the performance of $M_{\text{orig}}$ and $M_{\text{SAdv}}$ on a per-instance basis for each difficulty level on MNIST ((a) using CNN and (b) using NN), FashionMNIST ((c) using CNN and (d) using NN), and CIFAR10 ((e) using CNN and (f) using NN).