Benchmark Report - Vector Stitching with Persistence Images in Image Classification

April 8, 2025

Abstract

This report presents the results of benchmarks conducted for the image classification task on three different datasets: mushroom species in laboratory study, magnetic resonance images of brain tumors, and military vehicles. For each dataset, sample images, accuracy charts, and loss function graphs for different dataset sizes are presented.

1 Adopted methodology and assumptions

During experiments, we used the Optuna hyperparameter optimization library. This gave us the ability to verify how models behave for different combinations of network hyperparameters. We selected the number of neurons in three separate convolutional layers as hyperparameters to be optimized. Optuna modified these hyperparameters in each successive individual training, optimizing the accuracy result. We deliberately used a small model (from 35,112 to 943,944 parameters depending on the adopted layer sizes in blocks and the dataset used) to be able to repeat experiments multiple times in order to capture statistically significant effects. We also decided to focus on the number of neurons in the aforementioned layers because during manual tests, these parameters proved to most significantly affect the effectiveness of methods based on persistence images. The charts presented below show sorted (or in a few cases unsorted) results of many separate training sessions and accuracy and loss tests for different datasets (samples of datasets are included in each section). The visible results are values averaged using cross-validation. Images have standardized resolutions and, except for resonance images, have 3 RGB channels. We will denote "Acc RAW" as the accuracy of networks trained on raw data, and "Acc VS" as the accuracy of networks trained on data to which persistence images (calculated for each sample) have been added.

1.1 Used architecture

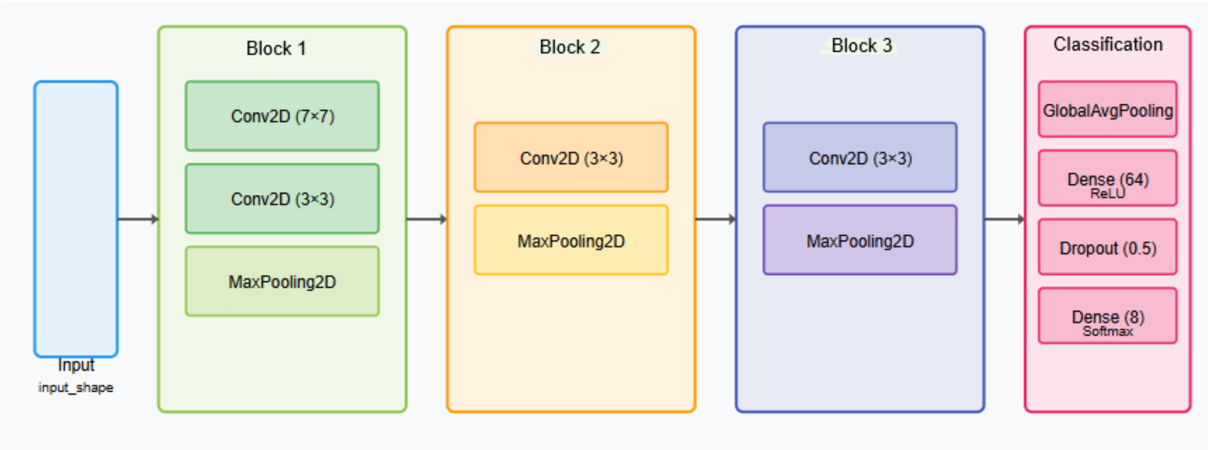

Figure 1: Diagram showing the architecture of the optimized model.

2 Dataset: Mushroom species in laboratory study

2.1 Dataset description

The dataset contains microscopic images of different mushroom species obtained in laboratory conditions. It consists of over 9,000 images belonging to 5 different classes representing individual species. Images have a resolution of 500x500.

2.2 Sample specimen

Figure 2: Sample image from the mushroom dataset observed under a microscope.

2.3 Results

2.4 Comparison of 100 experiments without sorting

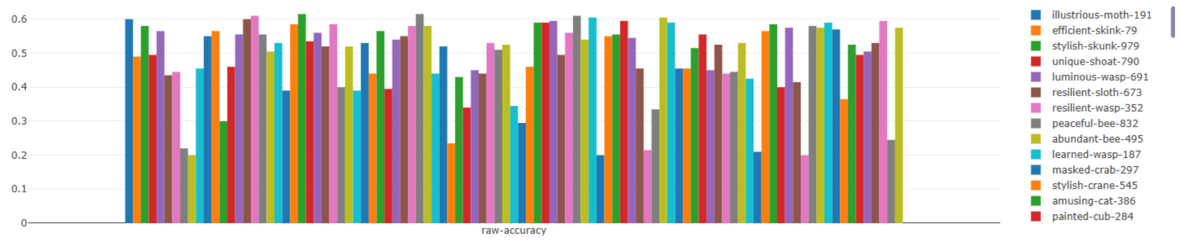

Figure 3: Chart showing 100 experiments on a dataset of 1000 samples (800 training/200 test) without sorting for accuracy.

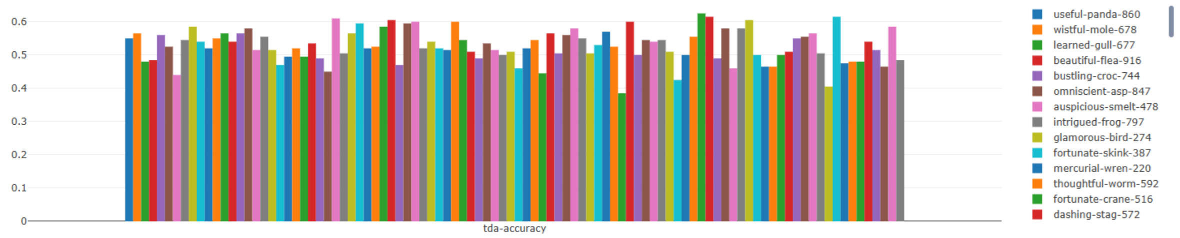

Figure 4: Chart showing 100 experiments on a dataset of 1000 samples (800 training/200 test) without sorting for accuracy.

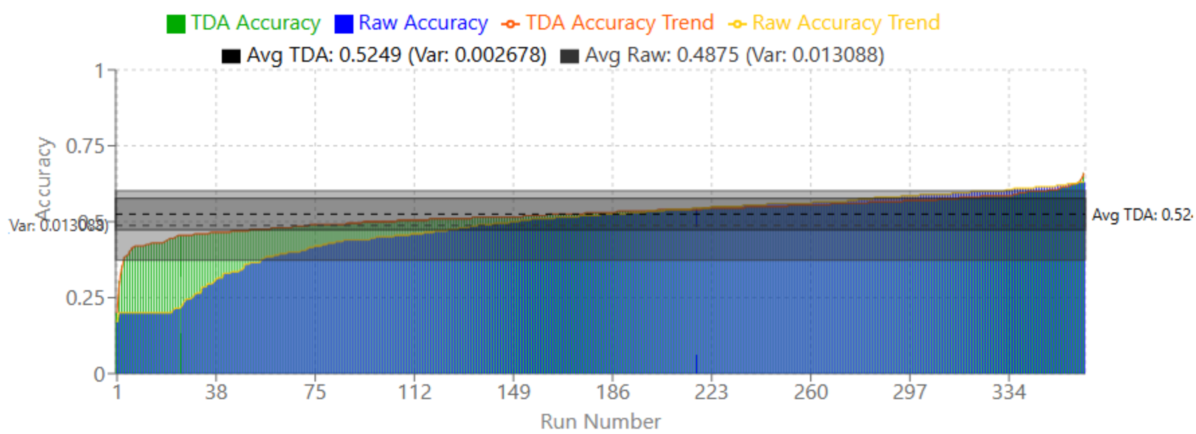

Figure 5: Chart showing 360 experiments on a dataset of 1000 samples (800 training/200 test) with sorting for accuracy.

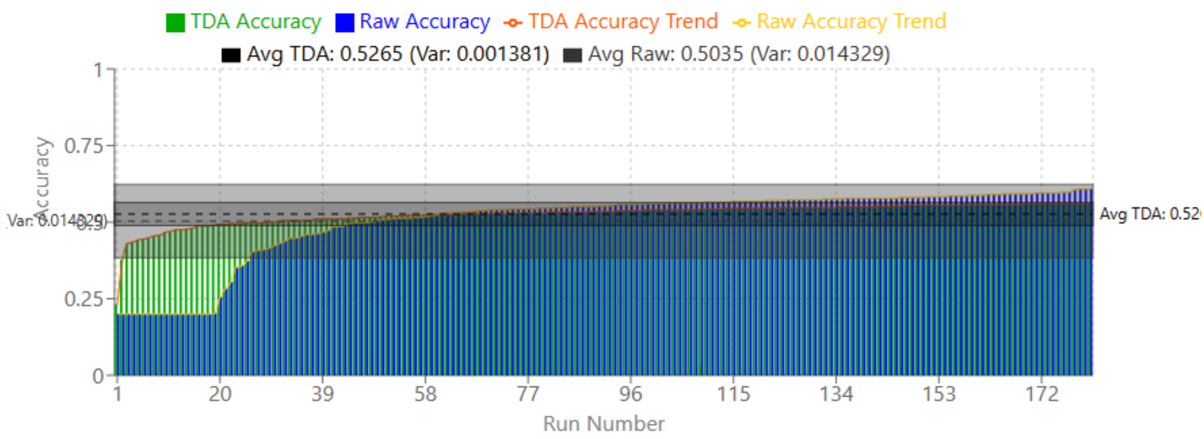

Figure 6: Chart showing 180 experiments on a dataset of 2000 samples (1600 training/400 test) with sorting for accuracy.

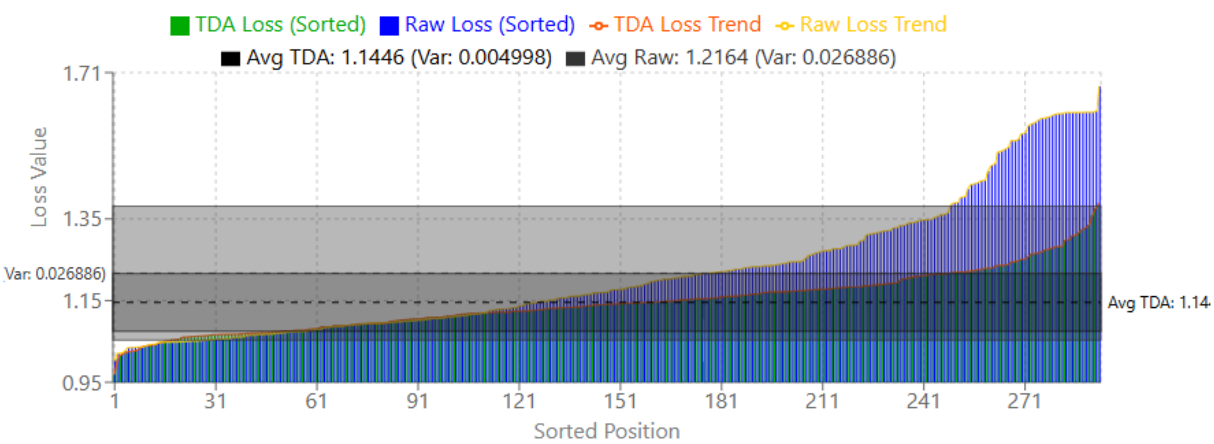

Figure 7: Chart showing 300 experiments on a dataset of 1000 samples (800 training/200 test) with sorting for loss.

Dataset size	Acc RAW [%]	Acc VS [%]	Variance RAW	Variance VS
1000	48.75	52.49	0.013	0.002
2000	50.35	52.65	0.014	0.001

Table 1: Averaged results in the above comparison.

At this stage, we noticed that Vector Stitching methods don't necessarily significantly impact the expected accuracy value, but there is a noticeable difference in the variance of this metric.

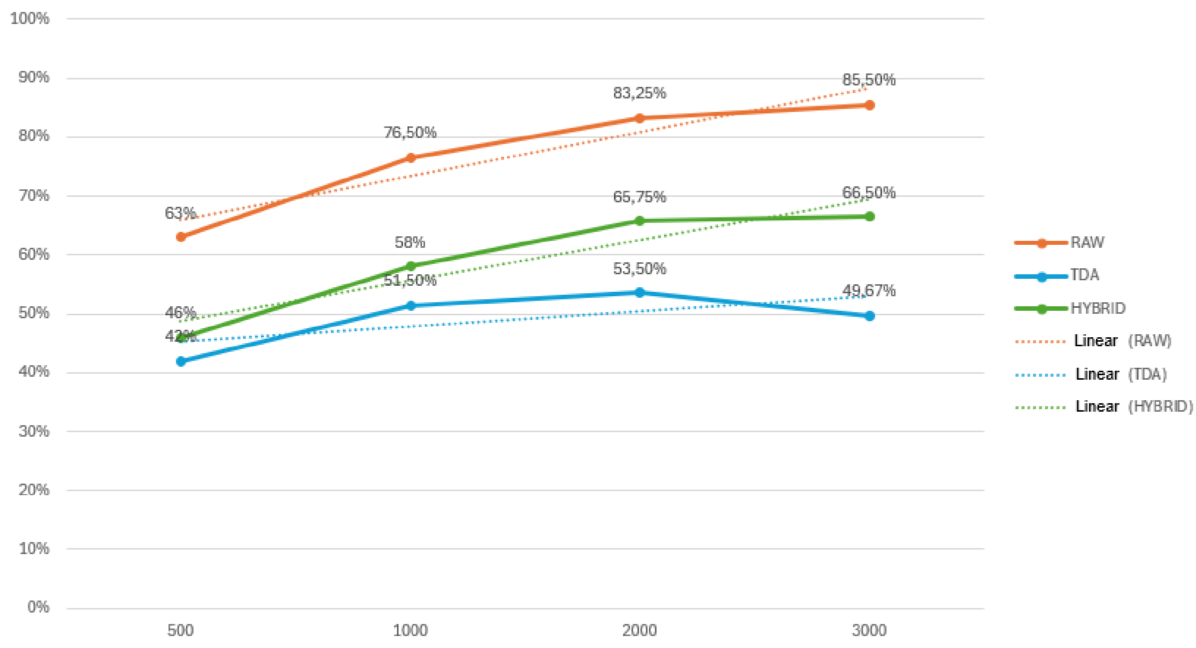

Figure 8: The chart shows the effectiveness of a network which, on the full raw dataset (9100 samples), should achieve about 92% effectiveness. We used 3 methods: RAW - raw data, TDA - pure persistence images, HYBRID - Our vector stitching method.

Above are experiments conducted with different training set size thresholds (500, 1000, 2000, 3000). It's easy to notice that our method is more effective for each threshold than using persistence images alone for mushroom classification. Compared to using raw data, we can observe a significant decrease in effectiveness for both methods.

3 Dataset: Magnetic resonance images of brain tumors

3.1 Dataset description

The dataset contains magnetic resonance imaging (MRI) images showing different types of brain tumors. It consists of 1728 images belonging to 4 different classes (one of which is brain resonances without neoplastic changes) representing individual tumor types. Images have a resolution of 256x256.

3.2 Sample specimen

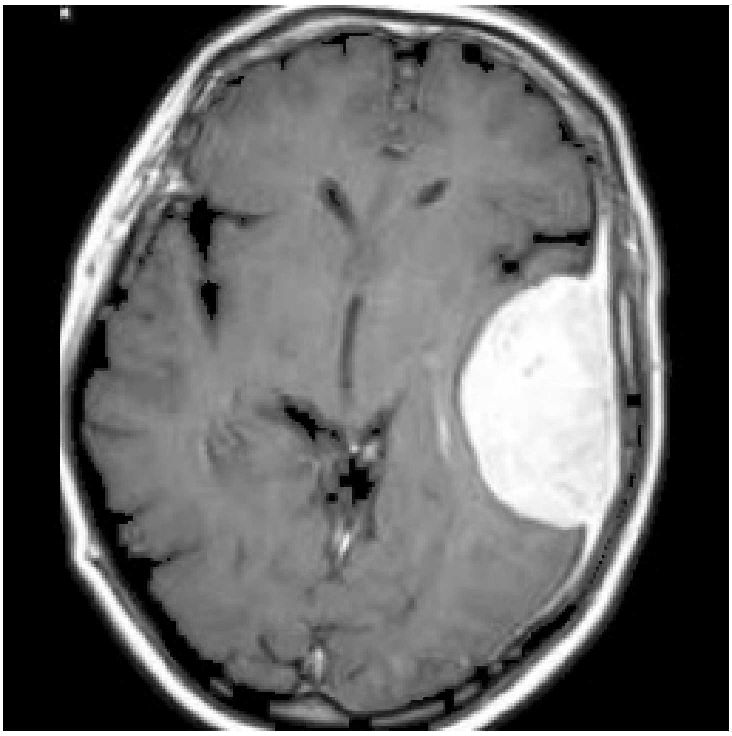

Figure 9: Sample MRI image of a brain tumor from the dataset.

3.3 Results

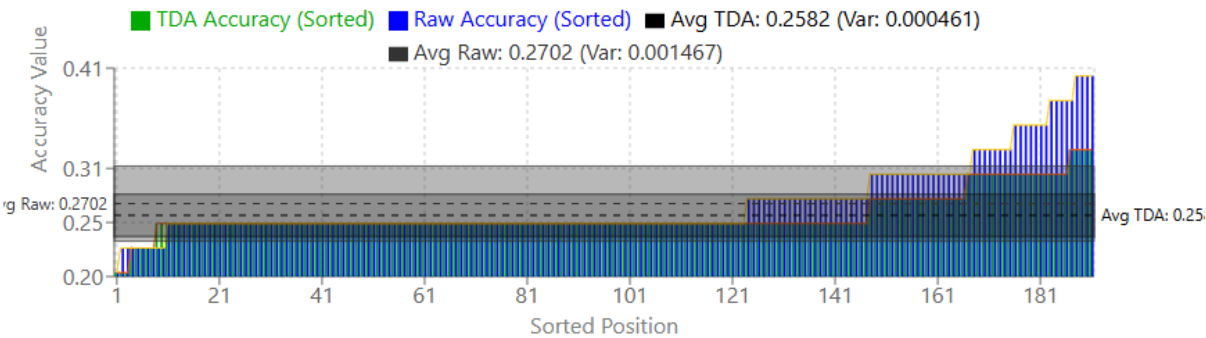

Figure 10: Chart showing 200 experiments on a dataset of 200 samples (160 training/40 test) with sorting for accuracy.

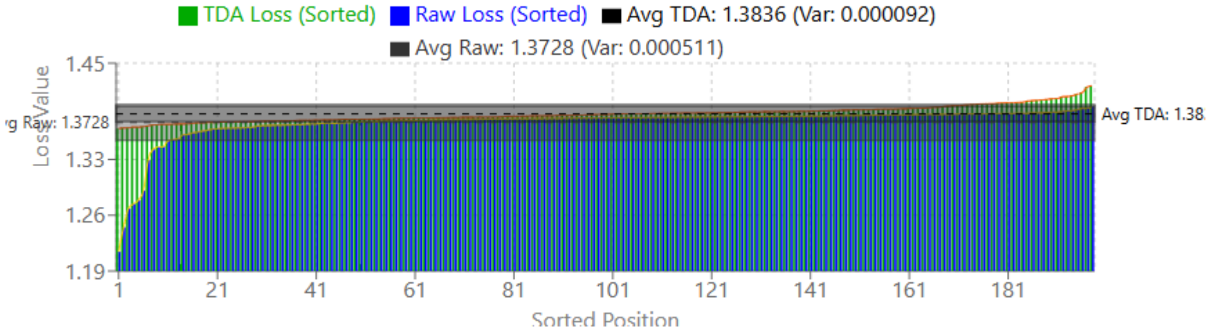

Figure 11: Chart showing 200 experiments on a dataset of 200 samples (160 training/40 test) with sorting for loss.

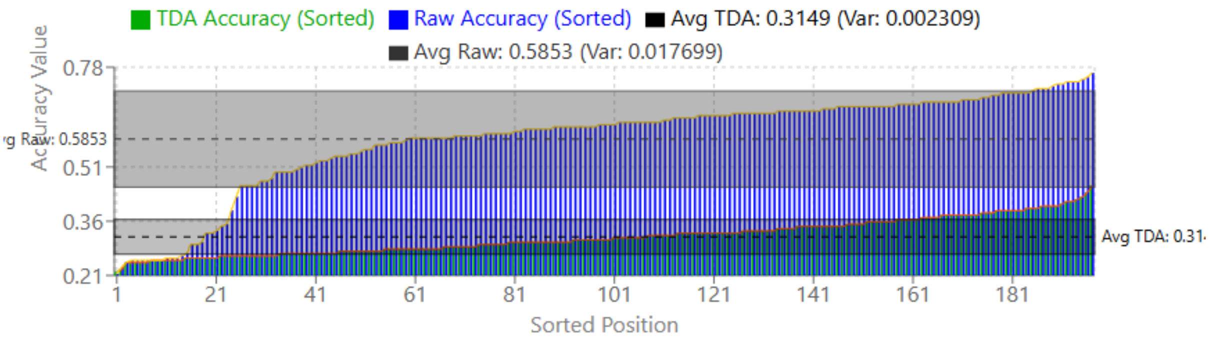

Figure 12: Chart showing 200 experiments on a dataset of 800 samples (640 training/160 test) with sorting for accuracy.

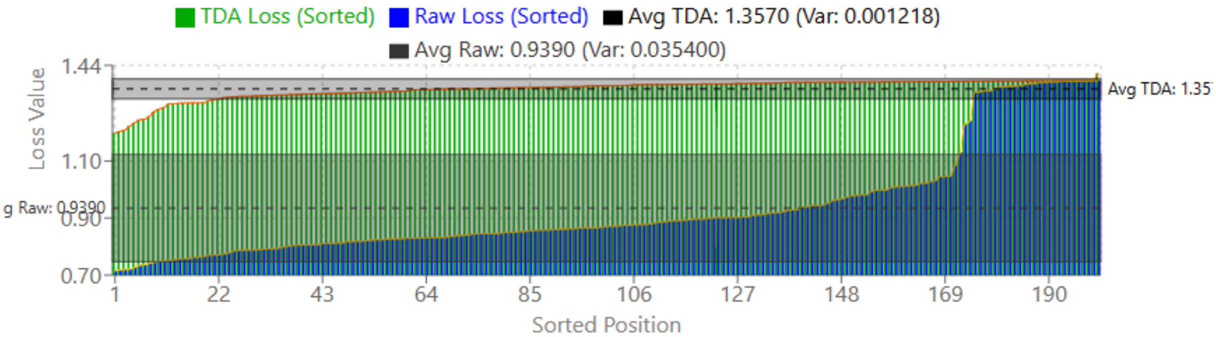

Figure 13: Chart showing 200 experiments on a dataset of 800 samples (640 training/160 test) with sorting for loss.

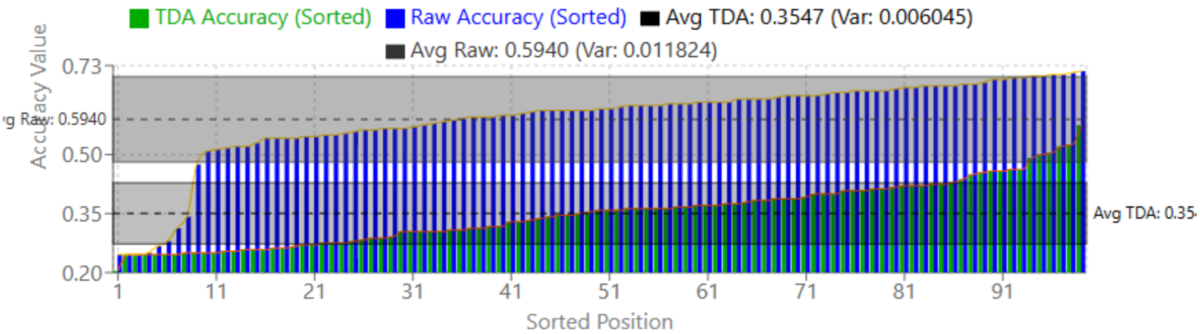

Figure 14: Chart showing 100 experiments on a dataset of 1200 samples (960 training/240 test) with sorting for accuracy.

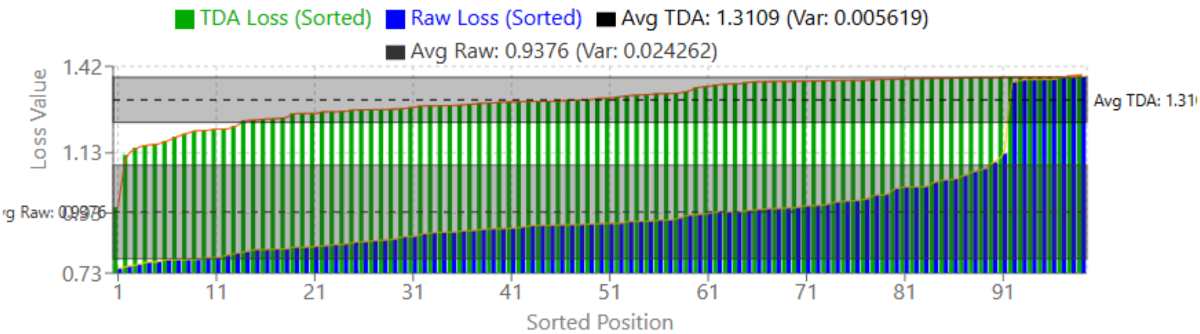

Figure 15: Chart showing 100 experiments on a dataset of 1200 samples (960 training/240 test) with sorting for loss.

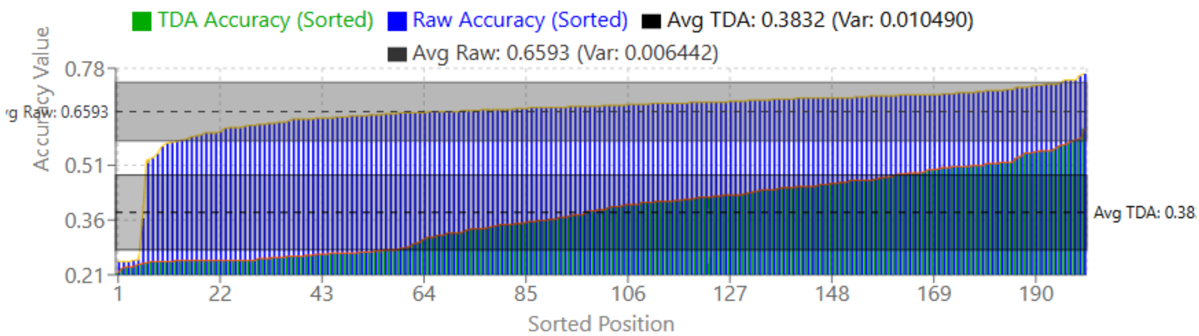

Figure 16: Chart showing 200 experiments on a dataset of 1728 samples (1382 training/346 test) with sorting for accuracy.

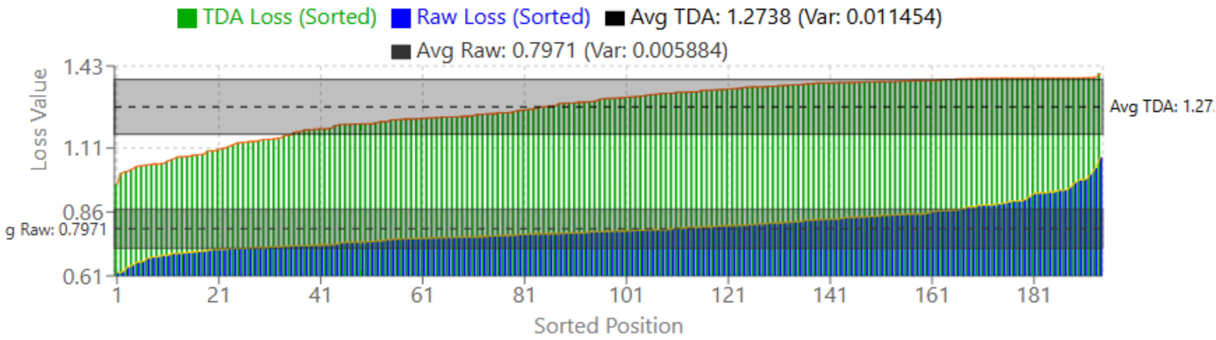

Figure 17: Chart showing 200 experiments on a dataset of 1728 samples (1382 training/346 test) with sorting for loss.

Dataset size	Acc RAW [%]	Acc VS [%]	Variance RAW	Variance VS
200	27.02	25.82	0.001	0.0004
800	58.53	31.49	0.017	0.002
1200	59.40	35.47	0.011	0.006
1728	65.93	38.32	0.006	0.010

Table 2: Averaged results in the above comparison.

4 Dataset: Military vehicles

4.1 Dataset description

The dataset contains images of various military vehicle types. It consists of 2508 images belonging to 7 different classes representing individual vehicle types. Images have a resolution of 280x180 and include, besides professional photographs, images of military objects from games and films.

4.2 Sample specimen

Figure 18: Sample image of a military vehicle from the dataset.

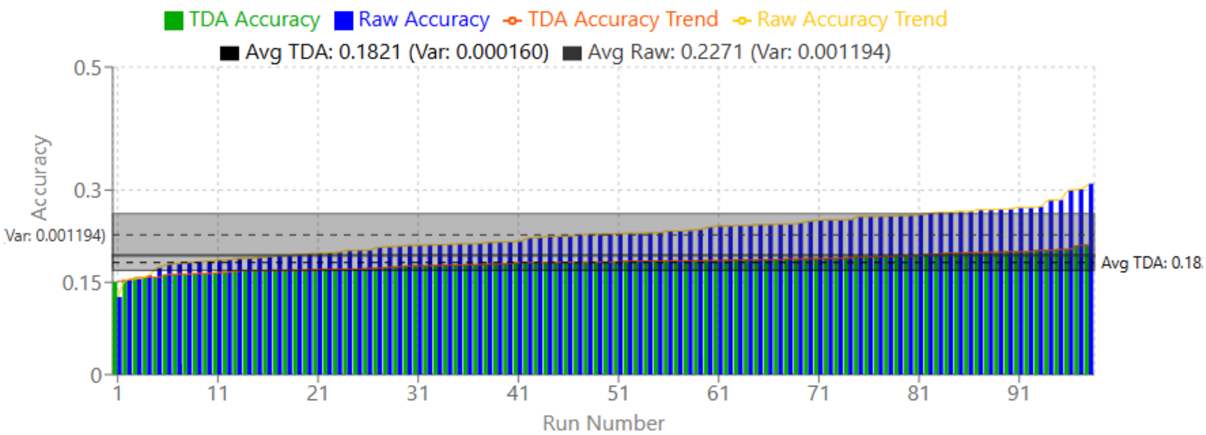

Figure 19: Chart showing 100 experiments on a dataset of 2508 samples (2006 training/502 test) with sorting for accuracy.

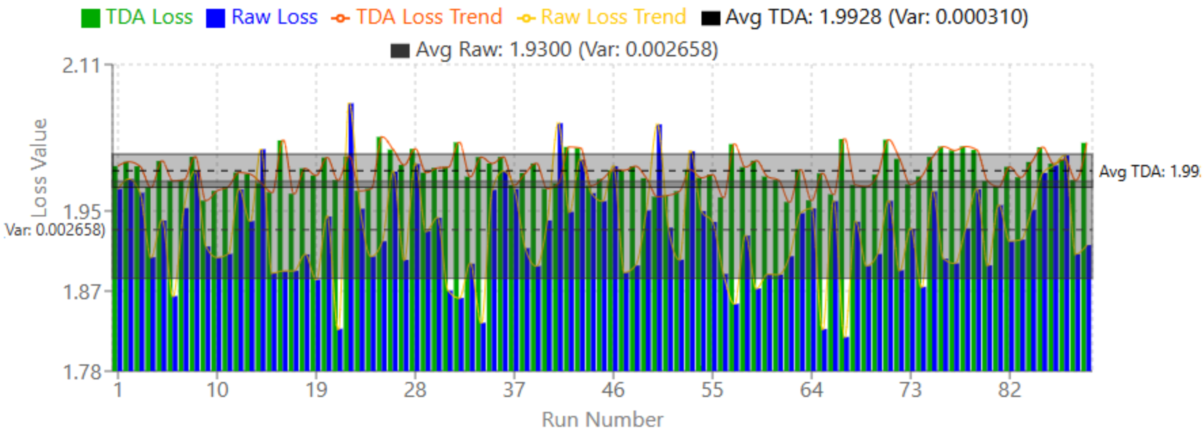

Figure 20: Chart showing 100 experiments on a dataset of 2508 samples (2006 training/502 test) without sorting for loss.

Dataset size	Acc RAW [%]	Acc VS [%]	Variance RAW	Variance VS
2508	22.72	18.21	0.001	0.0001

Table 3: Averaged results in the above comparison.

5 Comparison and analysis of results

Dataset	Acc RAW [%]	Acc VS [%]	Variance RAW	Variance VS
Microscopic mushrooms	49.55	52,57	0.0135	0.0015
Brain tumor MRI	52.72	32.78	0.0087	0.0046
Military vehicles	22.72	18.21	0.001	0.0001

Table 4: Averaged results in the above comparison.

6 Conclusions

To draw universal conclusions, the above tests would need to be conducted for even more complex and larger datasets. However, the above results "temper enthusiasm." What in the case of classifying microscopic mushroom images seemed to show potential for stabilizing network training and limiting the need for hyperparameter tuning, proves to be of little use, and sometimes even detrimental (for accuracy) in tasks such as classifying entire brain MRI images. We realize that the network architecture itself is simple and does not have the number of parameters that would be useful in commercial use. But if we observe such results in the case of moderately advanced classification tasks, then implementing a modified version of Vector Stitching methods in detection tasks and others related to Computer Vision will most likely (perhaps a naive assumption) be burdened with similar problems (in detection, we will encounter an identical classification mechanism at some stage). It is possible that the variability of image data and often their certain "topological poverty" make these methods not provide stable effects in a universal way (the situation may look much better for structures such as time series). Despite this, from our perspective, this approach may still be useful in some fields (even following the path of research on tissues under electron microscopes).