Adversarial Attacks on Machine Learning-Aided Visualizations Supplementary Materials

Takanori Fujiwara, Kostiantyn Kucher, Junpeng Wang, Rafael M. Martins, Andreas Kerren, and Anders Ynnerman

A Supplementary Experiments for Sec. 4.1

We show the issues seen in Sec. 4.1 can happen even when using parametric UMAP (PUMAP) with a smaller NN. While the default PUMAP employs an MLP with three 100-neuron hidden layers, here we use an MLP with only one hidden 4-neuron layer. As in Sec. 4.1, Fig. A.1 shows the results after applying the PUMAP to the Wine dataset and performing adversarial attacks. In these results, we can see similar patterns to Fig. 3, indicating that the close-to-linear mapping exists even when using a small NN.

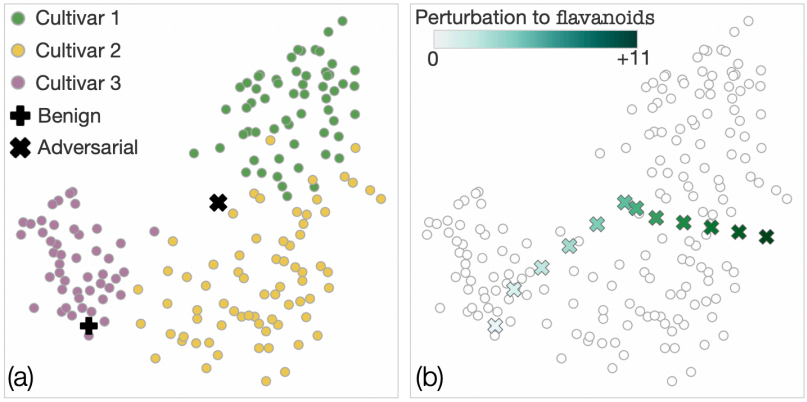

Figure A.1: The investigation of the one-attribute attack on the PUMAP using a **small MLP** (one hidden 4-neuron layer) and trained with **the Wine dataset**: (a) a scatterplot obtained by applying the PUMAP and (b) the input coordinate migration in response to the perturbations to flavanoids.

We further demonstrate and analyze adversarial attacks on two additional datasets: the breast cancer dataset[1] and the handwritten digits dataset.[2] The breast cancer dataset consists of 569 instances/masses, 30 attributes, and 2 labels (malignant or benign mass). The handwritten digits dataset consists of 1797 instances/handwritten digits, 64 attributes/pixels, and 10 labels corresponding to different digits, 0–9. For each dataset, we apply PUMAP with the default setting, perform one-attribute attacks, and analyze attack results as in Sec. 4.1.

Fig. A.2 shows the results for the breast cancer dataset. In Fig. A.2-a, one arbitrary benign mass is selected as a benign input and the adversarial counterpart is then crafted by adding a value of 15 into the benign input's mean cocavity. We can see that the adversarial input is projected near malignant masses. Similar to the analyses in Sec. 4.1, we can observe that, even for this dataset, the trained PUMAP shows a close-to-linear response to the perturbation to mean cocavity (see Fig. A.2-b), which has a clearly different distribution for each label (see Fig. A.2-c). Also, from the result in Fig. A.2-d, where we perturb both mean cocavity and mean texture, we can expect that, by adjusting mean cocavity and mean texture, adversaries can place an adversarial input to their desired coordinate as in the attack using a substitute model in Sec. 4.1.

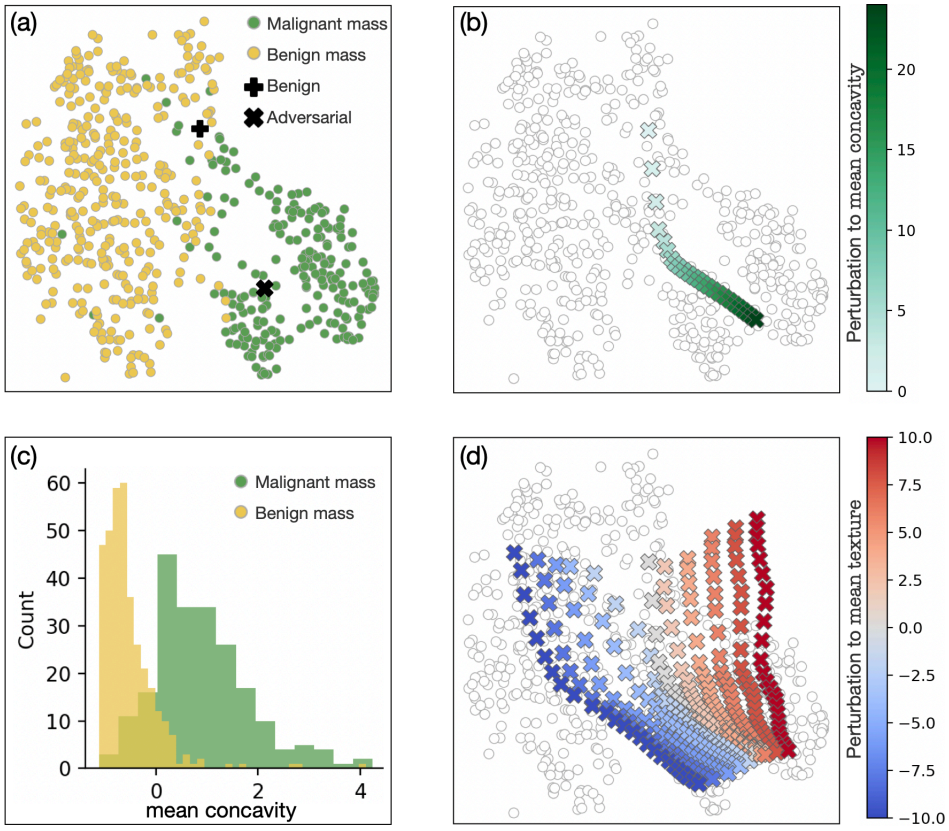

Figure A.2: The investigation of the one-attribute attack on the default PUMAP trained with **the breast cancer dataset**: (a) a scatterplot obtained by applying the PUMAP; (b) the input coordinate migration in response to the perturbations to mean concavity; (c) the value distribution of mean concavity for each label; and (d) the input coordinate migration when perturbing mean concavity (from 0 to 25) and mean texture (from -10 to 10).

[1] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)
[2] https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits

Fig. A.3 shows the results for the handwritten digits dataset. In Fig. A.3-a, one 0 digit is selected as a benign input and then an adversarial input is crafted by adding a value of 20 into the benign input's pixel_3_4. We can see that the adversarial input is projected near 8 digits. As in the other analyses, the trained PUMAP shows a close-to-linear response to the perturbation to pixel_3_4 (see Fig. A.3-b). As shown in Fig. A.3-c, pixel_3_4's distribution is clearly different between digits $\{0, 6\}$ and $\{1, 2, 3, 7, 8, 9\}$. Also, from the result in Fig. A.3-d, we can expect that, by adjusting pixel_3_4 and pixel_6_6, adversaries can place an adversarial input to their desired coordinate.

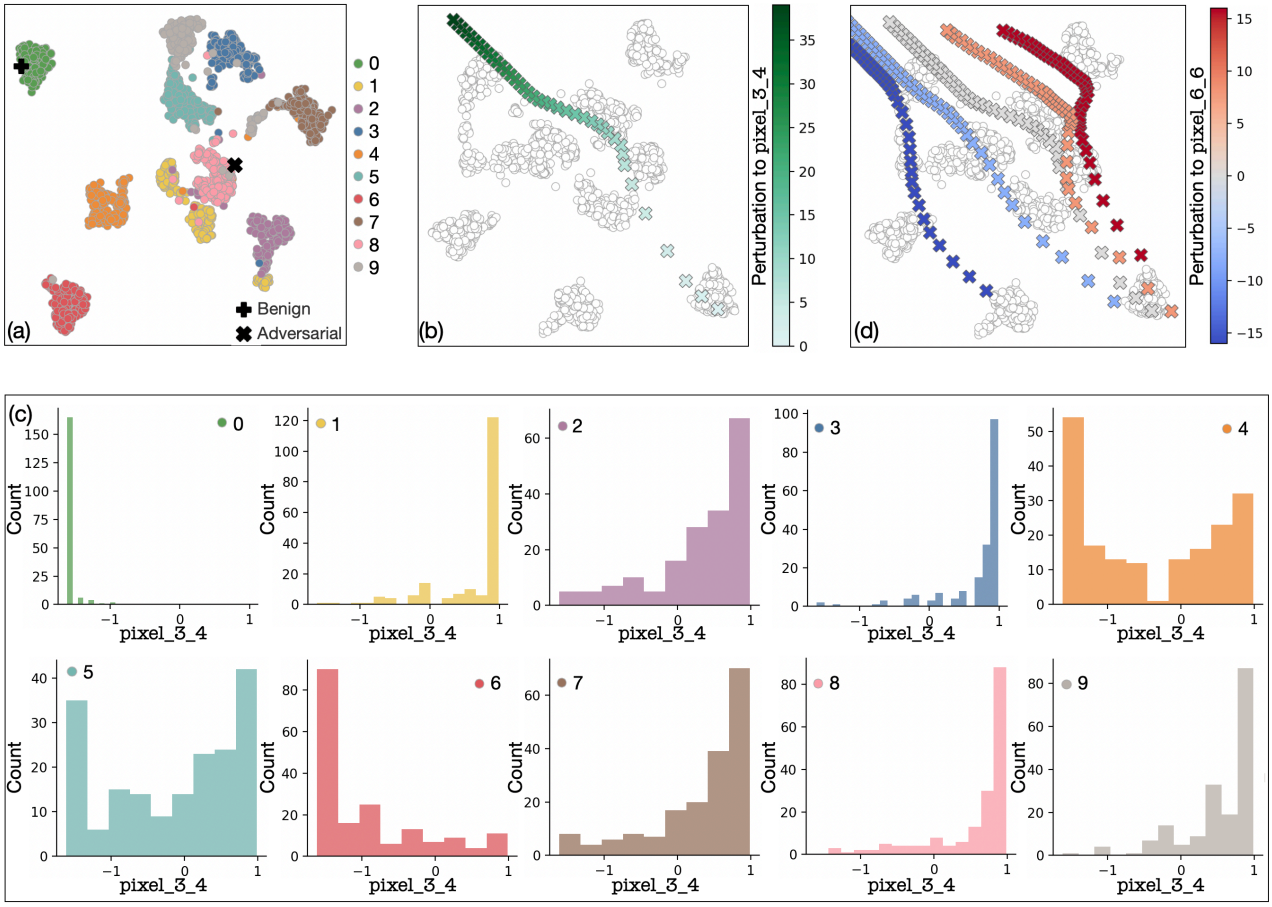

Figure A.3: The investigation of the one-attribute attack on the default PUMAP trained with **the handwritten digits dataset**: (a) a scatterplot obtained by applying the PUMAP; (b) the input coordinate migration in response to the perturbations to pixel_3_4; (c) the value distribution of pixel_3_4 for each label; and (d) the input coordinate migration when perturbing pixel_3_4 (from 0 to 40) and pixel_6_6 (from -16 to 16).

Furthermore, as shown in Fig. A.4, we perform attacks on a different parametric DR method—parametric t-SNE implemented by Lai et al. (`https://github.com/a07458666/parametric_dr`). The results in Fig. A.4 indicate that even for parametric t-SNE, adversaries can perform effective attacks. However, as Lai et al.'s implementation uses the sigmoid activation function (instead of rectified linear activation functions used in PUMAP), parametric t-SNE shows less linearity when compared with PUMAP. Note that we also applied Lai et al.'s parametric t-SNE to the Wine and breast cancer datasets; however, it did not produce reasonable low-dimensional representations for these two datasets (e.g., every instance is aligned along one line).

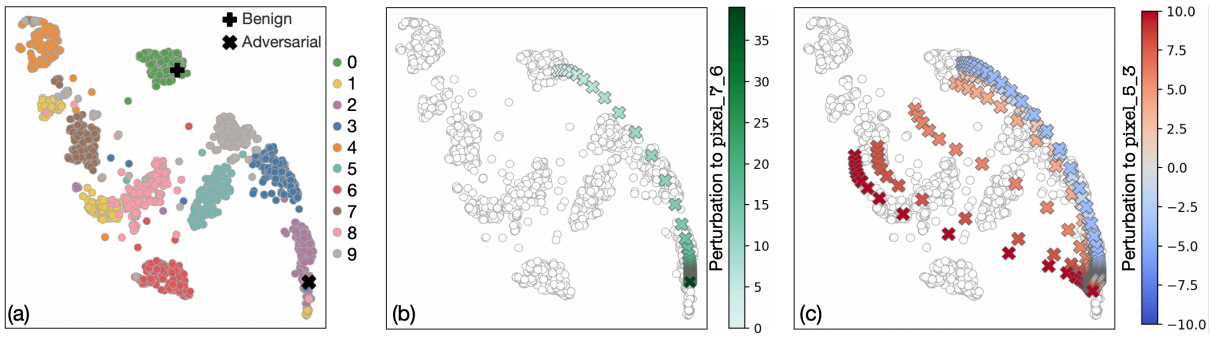

Figure A.4: The results of attacks on Lai et al.'s **parametric t-SNE** implementation using the **sigmoid activation functions** [43]: (a) a scatterplot obtained by applying the trained parametric t-SNE to the handwritten digits dataset; (b) the input coordinate migration in response to the perturbations to pixel_7_6; and (d) the input coordinate migration when perturbing pixel_7_6 (from 0 to 40) and pixel_5_3 (from -4 to 10).

B Supplementary Experiments and Information for Sec. 4.2

As an additional example of the attack on MultiVision, we use the Penguins dataset,[3] which consists of 344 instances and 7 attributes with several missing values. As shown in Fig. B.1-a, MultiVision recommends useful charts for understanding this data. The multiple stacked histograms (a1) depict value distributions organized by three nominal attributes, Sex, Species, and Island. The two bar charts (a2, a3) inform Beak Depth and Beak Length differences by Sex. Similar to Sec. 4.2, to generate an adversarial input, we add an empty column into the data table and a blank space for a randomly selected row of the empty column. Note that our attack succeeded even without adding a blank space. However, in this case, MultiVision caused an execution error due to the zero division during the feature extraction and did not generate any charts. Fig. B.1-b shows the recommended charts for the adversarial input. While the top-recommended chart is still the same as Fig. B.1-a, the second (b2) and third (b3) charts do not convey meaningful information. We also investigate the cause of the issues by checking MultiVision's intermediate outputs. We notice that the y-axes in Fig. B.1-b2 and b3 represent the empty column we added, while this information is not visible from the charts due to the Vega-Lite specifications (i.e., Data-VIS Mapping process helps adversaries hide the cause of this issue). We further observe that the empty column is categorized as a nominal attribute and assigned high importance; consequently, the empty column is selected for chart generation.

As in Sec. 4.2, based on the gradient information indicating the high influence of column_idx_normed, we shuffle the order of columns, resulting in the recommendations shown in Fig. B.1-c. While Fig. B.1-c1 is still a similar visualization to Fig. B.1-a1, the other recommended charts do not show useful information due to the overplotting.

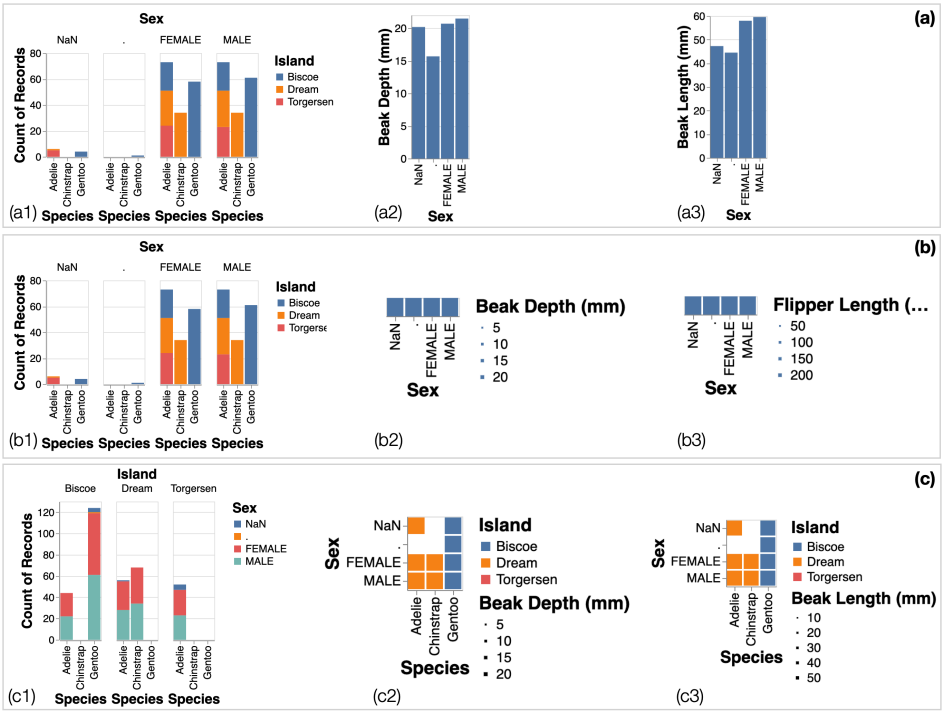

Figure B.1: The top-3 recommended charts by MultiVision for the Penguins dataset; (a) before and after (b) adding one empty column and (c) shuffling the column order. The charts are placed in order of the recommendation ranks (i.e., a1: the first, a2: the second, a3: the third). Note that we show the generated charts as they are, even though the legends located in b2, b3, c2, c3 do not match with the employed visual encodings.

[3] https://github.com/mwaskom/seaborn-data/blob/master/penguins.csv

We here provide the full-size charts corresponding to Fig. 8-b1, b2, and Fig. 9.

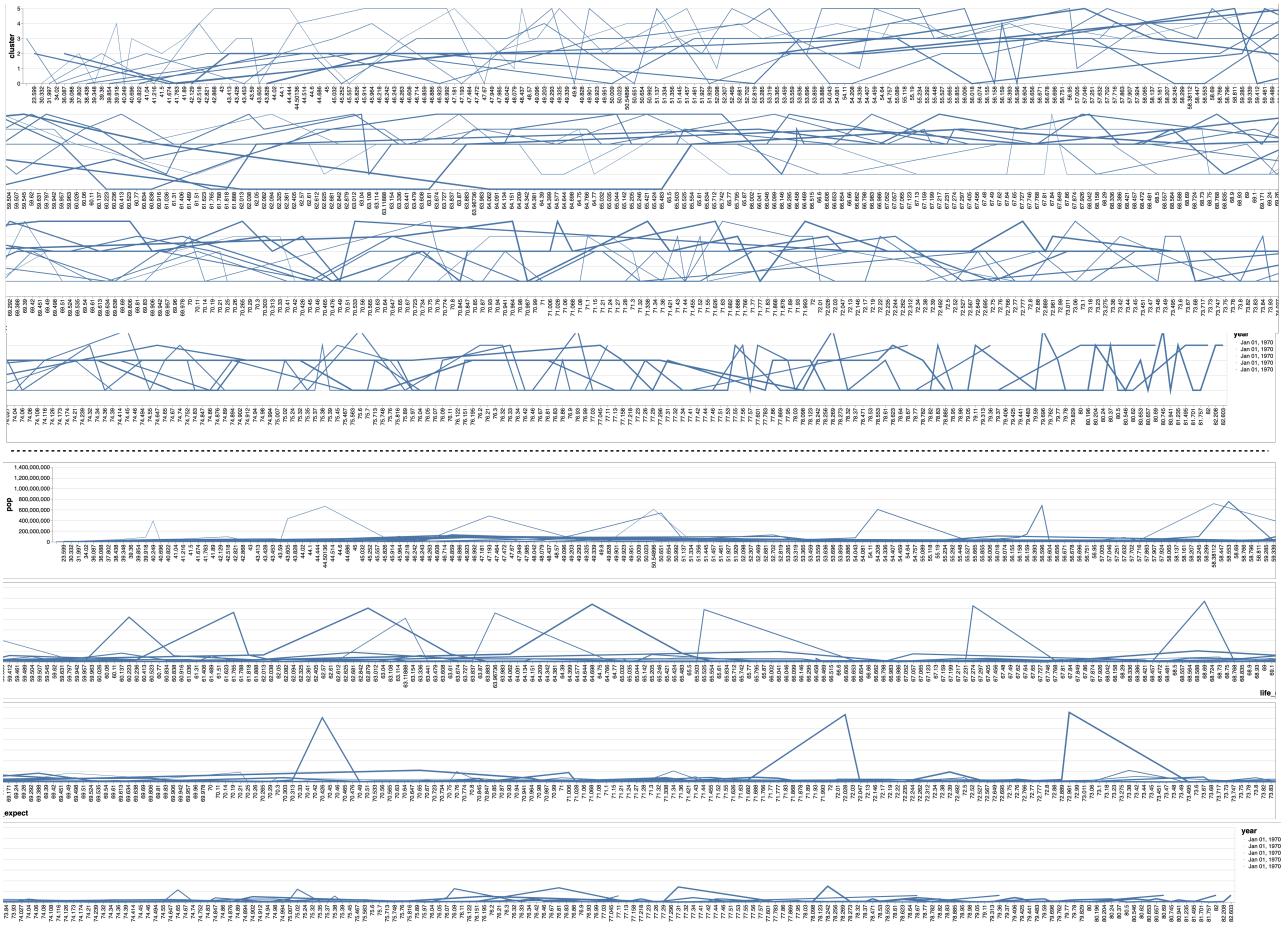

Figure B.2: The full-size charts corresponding to Fig. 8-b1 (top) and b2 (bottom).

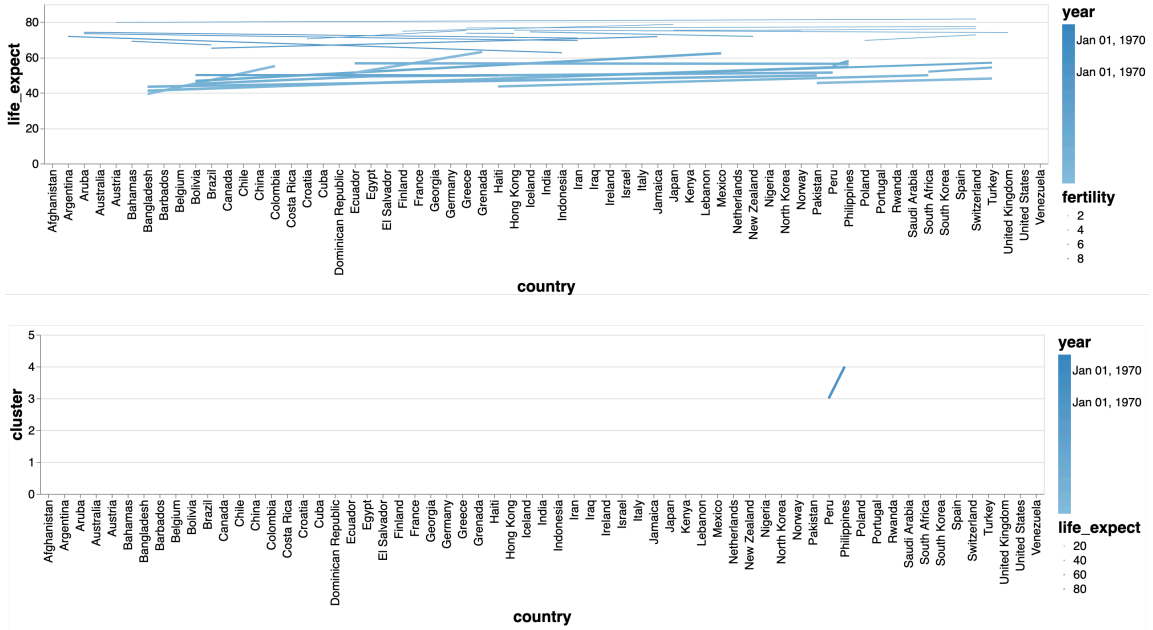

Figure B.3: The full-size charts corresponding to Fig. 9.

As the supplementary information of Table 2, we here list the gradients (Table B.1) and the meanings (Table B.2) of all the column features.

Table B.1: All column features' gradients corresponding to Table 2.

	Column set			Chart type		
	year	life_expect	fertility	year	life_expect	fertility
column_idx_normed	-0.06	-0.63	-0.48	-0.06	-0.63	-0.48
dataType_normed	0.02	0.03	0.18	0.02	0.03	0.18
aggrPercentFormatted	-0.05	-0.10	-0.22	-0.05	-0.10	-0.22
aggr01Ranged	-0.04	-0.34	-0.19	-0.04	-0.34	-0.19
aggr0100Ranged	-0.02	-0.06	-0.04	-0.02	-0.06	-0.04
aggrIntegers	0.00	-0.09	-0.21	0.00	-0.09	-0.21
aggrNegative	-0.02	-0.25	-0.17	-0.02	-0.25	-0.17
aggrBayesLikeSum	0.00	0.03	0.02	0.00	0.03	0.02
dmBayesLikeDimension	-0.01	0.01	-0.02	-0.01	0.01	-0.02
commonPrefix	-0.03	0.07	0.04	-0.03	0.07	0.04
commonSuffix	-0.02	0.06	0.08	-0.02	0.06	0.08
keyEntropy	0.05	0.18	-0.05	0.05	0.18	-0.05
charEntropy	0.00	0.04	0.13	0.00	0.04	0.13
norm_range	0.00	-0.13	-0.03	0.00	-0.13	-0.03
changeRate	0.01	0.24	0.00	0.01	0.24	0.00
partialOrdered	-0.02	-0.40	-0.06	-0.02	-0.40	-0.06
norm_var	0.00	-0.03	0.00	0.00	-0.03	0.00
norm_cov	0.01	-0.11	-0.13	0.01	-0.11	-0.13
cardinality	0.01	-0.02	-0.08	0.01	-0.02	-0.08
spread	0.00	0.11	0.22	0.00	0.11	0.22
major	-0.06	-0.38	0.19	-0.06	-0.38	0.19
benford	0.01	0.02	0.06	0.01	0.02	0.06
orderedConfidence	0.01	-0.08	0.08	0.01	-0.08	0.08
equalProgressionConfidence	-0.02	-0.04	0.14	-0.02	-0.04	0.14
geometircProgressionConfidence	0.00	-0.03	0.06	0.00	-0.03	0.06
medianLength	-0.03	-0.02	0.06	-0.03	-0.02	0.06
lengthStdDev	-0.04	-0.01	0.13	-0.04	-0.01	0.13
sumIn01	-0.06	0.01	0.08	-0.06	0.01	0.08
sumIn0100	0.00	-0.10	-0.02	0.00	-0.10	-0.02
absoluteCardinality	0.03	0.03	-0.04	0.03	0.03	-0.04
skewness	0.01	-0.03	0.01	0.01	-0.03	0.01
kurtosis	0.01	-0.05	-0.05	0.01	-0.05	-0.05
gini	0.00	-0.05	-0.06	0.00	-0.05	-0.06
nRows	0.01	-0.08	0.00	0.01	-0.08	0.00
averageLogLength	0.00	-0.08	-0.04	0.00	-0.08	-0.04
dummy0	0.01	0.00	0.03	0.01	0.00	0.03
dummy1	0.00	0.02	0.01	0.00	0.02	0.01
dummy2	0.01	0.03	0.00	0.01	0.03	0.00
dummy3	0.00	0.03	0.02	0.00	0.03	0.02
dummy4	-0.01	0.02	0.01	-0.01	0.02	0.01
dummy5	0.00	0.00	0.01	0.00	0.00	0.01
dummy6	-0.01	-0.03	-0.01	-0.01	-0.03	-0.01
dummy7	0.00	0.01	-0.01	0.00	0.01	-0.01
dummy8	0.00	-0.01	0.01	0.00	-0.01	0.01
dummy9	-0.01	-0.01	0.01	-0.01	-0.01	0.01
dummy10	0.00	-0.02	0.03	0.00	-0.02	0.03

	Column set			Chart type		
	year	life_expect	fertility	year	life_expect	fertility
wordEmb0	0.02	0.17	-0.11	0.02	0.17	-0.11
wordEmb1	-0.03	-0.27	-0.41	-0.03	-0.27	-0.41
wordEmb2	-0.01	0.03	0.28	-0.01	0.03	0.28
wordEmb3	-0.02	-0.32	-0.47	-0.02	-0.32	-0.47
wordEmb4	-0.04	-0.08	0.13	-0.04	-0.08	0.13
wordEmb5	-0.01	0.37	0.17	-0.01	0.37	0.17
wordEmb6	-0.01	0.07	0.01	-0.01	0.07	0.01
wordEmb7	-0.01	0.34	0.16	-0.01	0.34	0.16
wordEmb8	0.06	0.32	-0.25	0.06	0.32	-0.25
wordEmb9	0.00	0.15	0.34	0.00	0.15	0.34
wordEmb10	-0.02	0.59	0.60	-0.02	0.59	0.60
wordEmb11	0.03	-0.14	-0.39	0.03	-0.14	-0.39
wordEmb12	-0.02	-0.37	-0.02	-0.02	-0.37	-0.02
wordEmb13	0.01	-0.05	-0.11	0.01	-0.05	-0.11
wordEmb14	-0.03	-0.65	0.39	-0.03	-0.65	0.39
wordEmb15	0.00	-0.34	-0.30	0.00	-0.34	-0.30
wordEmb16	-0.01	-0.17	-0.13	-0.01	-0.17	-0.13
wordEmb17	0.02	0.09	-0.43	0.02	0.09	-0.43
wordEmb18	0.05	0.01	-0.09	0.05	0.01	-0.09
wordEmb19	-0.01	0.33	0.32	-0.01	0.33	0.32
wordEmb20	0.02	0.49	0.07	0.02	0.49	0.07
wordEmb21	-0.01	-0.46	-0.31	-0.01	-0.46	-0.31
wordEmb22	0.02	-0.34	-0.04	0.02	-0.34	-0.04
wordEmb23	-0.05	0.26	0.05	-0.05	0.26	0.05
wordEmb24	-0.04	0.77	-0.11	-0.04	0.77	-0.11
wordEmb25	-0.01	0.27	-0.29	-0.01	0.27	-0.29
wordEmb26	-0.04	0.18	-0.62	-0.04	0.18	-0.62
wordEmb27	0.00	-0.42	-0.18	0.00	-0.42	-0.18
wordEmb28	0.03	0.68	-0.06	0.03	0.68	-0.06
wordEmb29	0.02	0.24	0.16	0.02	0.24	0.16
wordEmb30	0.01	-0.60	-0.17	0.01	-0.60	-0.17
wordEmb31	0.03	0.00	0.13	0.03	0.00	0.13
wordEmb32	0.06	0.40	1.00	0.06	0.40	1.00
wordEmb33	0.00	-0.09	-0.05	0.00	-0.09	-0.05
wordEmb34	-0.02	-0.03	0.01	-0.02	-0.03	0.01
wordEmb35	-0.01	-0.17	-0.16	-0.01	-0.17	-0.16
wordEmb36	0.03	-0.30	0.82	0.03	-0.30	0.82
wordEmb37	0.00	-0.50	-0.03	0.00	-0.50	-0.03
wordEmb38	-0.03	0.18	0.27	-0.03	0.18	0.27
wordEmb39	-0.06	0.90	-0.28	-0.06	0.90	-0.28
wordEmb40	0.02	-0.11	0.27	0.02	-0.11	0.27
wordEmb41	-0.02	-0.08	-0.03	-0.02	-0.08	-0.03
wordEmb42	-0.04	-0.59	0.42	-0.04	-0.59	0.42
wordEmb43	-0.06	-0.07	0.04	-0.06	-0.07	0.04
wordEmb44	0.06	0.26	0.30	0.06	0.26	0.30
wordEmb45	-0.04	-0.17	0.58	-0.04	-0.17	0.58
wordEmb46	0.01	-0.17	-0.32	0.01	-0.17	-0.32
wordEmb47	0.01	-0.69	-0.03	0.01	-0.69	-0.03
wordEmb48	-0.01	0.02	0.10	-0.01	0.02	0.10
wordEmb49	0.04	0.07	0.00	0.04	0.07	0.00

Table B.2: The meanings of column features.

Feature name	Meaning
column_idx_normed	Column index divided by # of columns
dataType_normed	Data type ID divided by 5 (e.g., string, datetime, decimal)
aggrPercentFormatted	Proportion of values formatted with %
aggr01Ranged	Proportion of values within a range of [0, 1]
aggr0100Ranged	Proportion of values within a range of [0, 100]
aggrIntegers	Proportion of integer values
aggrNegative	Proportion of negative values
aggrBayesLikeSum	(not computed based on Wu et al's implementation)
dmBayesLikeDimension	(not computed based on Wu et al's implementation)
commonPrefix	Proportion of the most common prefix digit
commonSuffix	Proportion of the most common suffix digit
keyEntropy	Entropy by values
charEntropy	Entropy by digits/chars
norm_range	Range of values
changeRate	Proportion of different adjacent values
partialOrdered	Maximum proportion of continuously increasing/decreasing values
norm_var	Normalized standard deviation
norm_cov	Normalized covariance
cardinality	Cardinality

Feature name	Meaning
spread	Cardinality divided by the range of values
major	Proportion of the most frequent value
benford	Skewness measured by Benford's law
orderedConfidence	Indicator of sequentiality
equalProgressionConfidence	Confidence for a sequence to be equal progression
geometircProgressionConfidence	Confidence for a sequence to be geometric progression
medianLength	Normalized median length of records
lengthStdDev	Standard deviation of lengths of records
sumIn01	Sum of values within a range of [0, 1]
sumIn0100	Sum of values within a range of [0, 100]
absoluteCardinality	Absolute cardinalitu
skewness	Skweness
kurtosis	Kurtosis
gini	Gini coeffcient
nRows	# of raws
averageLogLength	Average length of records in log scale
dummy0-10	-
wordEmb0-49	Word embeddig vector

C List of Publications with Publicly Available Source Code

Table C.1 lists the ML4VIS works that incorporate neural networks and provide source codes for replicating their ML models. We created this list by manually checking the contents of all full-length papers presented at EuroVis, PacificVis, and VIS from 2017 to 2022. Through this process, we found 69 ML4VIS-related papers and identified 22 papers providing source codes. Among these, only several source codes and pre-trained models were executable (after small modifications) with currently available programming libraries.

Table C.1: ML4VIS papers providing source code to replicate their neural network (NN) model, used datasets, and pretrained NN model.

Authors	Title	Year	Presented venue	Source code	Training data	Trained model
Poco and Heer	Reverse-Engineering Visualizations: Recovering Visual Encodings from Chart Images	2017	EuroVis	✓	✓	
Berger et al.	A Generative Model for Volume Rendering	2018	VIS	✓	✓	✓
Haehn et al.	Evaluating 'Graphical Perception' with CNNs	2018	VIS	✓	✓	✓
Chen et al.	Towards Automated Infographic Design: Deep Learning-based Auto-Extraction of Extensible Timeline	2019	VIS	✓	✓	
Chen et al.	LassoNet: Deep Lasso-Selection of 3D Point Clouds	2019	VIS	✓	✓	
He et al.	InSituNet: Deep Image Synthesis for Parameter Space Exploration of Ensemble Simulations	2019	VIS	✓	✓	
Wang et al.	DeepDrawing: A Deep Learning Approach to Graph Drawing	2019	VIS	✓	✓	
Lekschas et al.	PEAX: Interactive Visual Pattern Search in Sequential Data Using Unsupervised Deep Representation Learning	2020	EuroVis	✓	✓	✓
Fujiwara et al.	A Visual Analytics Framework for Contrastive Network Analysis	2020	VIS	✓	✓	
Engel and Ropinski	Deep Volumetric Ambient Occlusion	2020	VIS	✓	✓	✓
Lu et al.	Compressive Neural Representations of Volumetric Scalar Fields	2021	EuroVis	✓	✓	
Luo et al.	Texture Browser: Feature-based Texture Exploration	2021	EuroVis	✓	✓	✓
Madan et al.	Parsing and Summarizing Infographics with Synthetically Trained Icon Detection	2021	PacificVis	✓	✓	
Qin et al.	A Domain-Oblivious Approach for Learning Concise Representations of Filtered Topological Spaces for Clustering	2021	VIS	✓	✓	
Luo et al.	Natural Language to Visualization by Neural Machine Translation	2021	VIS	✓	✓	✓
Wu et al.	MultiVision: Designing Analytical Dashboards with Deep Learning Based Recommendation	2021	VIS	✓	✓	✓
Zhao et al.	ChartSeer: Interactive Steering Exploratory Visual Analysis With Machine Intelligence	2021	VIS	✓	✓	✓
Huesmann and Linsen	SimilarityNet: A Deep Neural Network for Similarity Analysis Within Spatio-temporal Ensembles	2022	EuroVis	✓	✓	
Shi et al.	GNN-Surrogate: A Hierarchical and Adaptive Graph Neural Network for Parameter Space Exploration of Unstructured-Mesh Ocean Simulation	2022	PacificVis	✓	✓	
Shi et al.	VDL-Surrogate: A View-Dependent Latent-based Model for Parameter Space Exploration of Ensemble Simulations	2022	VIS	✓	✓	
Wang et al.	Towards Natural Language-Based Visualization Authoring	2022	VIS	✓	✓	
Xia et al.	Interactive Visual Cluster Analysis by Contrastive Dimensionality Reduction	2022	VIS	✓	✓	✓