On the Impact of Sparsification on Quantitative Argumentative Explanations in Neural Networks

Daniel Peacock[1], Mansi[1], Nico Potyka[2], Francesca Toni[1] and Xiang Yin[1]

[1]Imperial College London, UK
[2]Cardiff University, UK

Abstract

Neural Networks (NNs) are powerful decision-making tools, but their lack of explainability limits their use in high-stakes domains such as healthcare and criminal justice. The recent SpArX framework sparsifies NNs and maps them to (weighted) Quantitative Bipolar Argumentation Frameworks (QBAFs) to provide an argumentative understanding of their mechanics. QBAFs can be explained by various quantitative argumentative explanation methods such as Argument Attribution Explanations (AAEs), Relation Attribution Explanations (RAEs), and Contestability Explanations (CEs) - which assign numerical scores to arguments or relations to quantify their influence on the dialectical strength of an argument to be explained. However, it remains unexplored how sparsification of NNs impacts the explanations derived from the corresponding (weighted) QBAFs. In this paper we explore two directions for impact. First, we empirically investigate how varying the sparsification levels of NNs affects the preservation of these explanations: using four datasets (Iris, Diabetes, Cancer, and COMPAS), we find that AAEs are generally well preserved, whereas RAEs are not. Then, for CEs, we find that sparsification can improve computational efficiency in several cases. Overall, this study offers a preliminary investigation into the potential synergy between sparsification and explanation methods, opening up new avenues for future research.

Keywords

Explainability, Neural Networks, Argumentative Explanations

1. Introduction

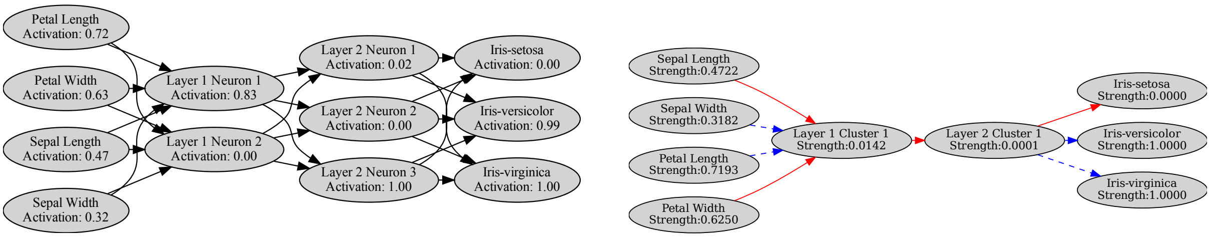
Improving the explainability of Neural Networks (NNs) has become a key concern in the development of trustworthy AI systems, particularly in high-stakes domains such as healthcare and criminal justice. Amongst various explanation methods (e.g., [1, 2, 3, 4, 5, 6]), *SpArX* [7] proposes an argumentative understanding of Multi-Layer Perceptrons (MLPs), a popular family of NNs. At a high level, SpArX first sparsifies a trained MLP, and then translates it into an equivalent weighted Quantitative Bipolar Argumentation Framework (weighted QBAF [8]), following [9]. A QBAF [10] models the reasoning process over conflicting or supporting information as a graph of weighted *arguments* connected via *support* and *attack* relations. In a weighted QBAF supports and attacks are also weighted. The dialectical strength of arguments in (weighted) QBAFs is given by numerical scores determined with gradual semantics (e.g. [11, 12, 13, 9])

The weighted QBAF obtained with SpArX offers a sparse and structured explanation for the decision-making process of the original MLPs. Figure 1 shows an example for a simple MLP. By sparsifying (Figure 1b), we reduce the size, making it simpler and easier to understand. Often information for subtle concepts resides within a subset of the neurons of the MLP [14, 15], making sparsification a useful tool.

Several explanation methods have been proposed for (weighted) QBAFs. For example, to explain the dialectical strength of an argument of interest (the *topic argument*), Argument Attribution Explanations (AAEs) [16] measure the influence of each argument on the topic argument by assigning numerical scores; Relation Attribution Explanations (RAEs) [17] follow a similar intuition but instead quantify the influence of individual attack or support relations. While these methods ignore weights on support

ArgXAI 2025: 3rd International Workshop on Argumentation for eXplainable AI

daniel.peacock20@imperial.ac.uk (D. Peacock); m24@ic.ac.uk (Mansi); potykan@cardiff.ac.uk (N. Potyka); ft@imperial.ac.uk (F. Toni); x.yin20@imperial.ac.uk (X. Yin)
0009-0000-9835-4584 (Mansi); 0000-0003-1749-5233 (N. Potyka); 0000-0001-8194-1459 (F. Toni); 0000-0002-6096-9943 (X. Yin)



(a) Original MLP.

(b) Sparse QBAF.

Figure 1: An illustration of SpArX. The original MLP (a) is sparsified to create a smaller version with similar outputs. The smaller MLP can be converted to an equivalent QBAF which provides an argumentative understanding. Each neuron (node) is modelled as an argument, red, solid edges indicate attacks and blue, dotted edges indicate supports (we ignore edge (attack and support) weights here). Note that in real-world usage for classification tasks, softmax can be applied to the final layer to ensure deterministic predictions.

and attack relations, Contestability Explanations (CEs) [18] determine how the edge weights can be modified to reach a desired dialectical strength for the topic argument. Together, these methods offer a fine-grained and quantitative understanding of the reasoning process within (weighted) QBAFs.

Both sparsification and quantitative argumentative explanations (i.e. AAEs, RAEs and CEs) advance the interpretability of NNs, but there has been little investigation into how the former impacts the latter. This gap is particularly concerning because sparsification simplifies the structure of MLPs, which may alter or distort the quantitative explanations derived from the resulting (weighted) QBAF[1], potentially misleading users, and even resulting in ethical or legal risks. For example, in a healthcare setting, misleading explanations may lead incorrect treatments being given to patients.

To address this gap, we focus on two core research questions (as illustrated in Figure 2): *(1) To what extent does sparsification preserve AAEs and RAEs? (2) Can sparsification improve CEs' computational efficiency?* We distinguish these two questions because the nature of the explanations differs: AAEs and RAEs quantify how arguments and relations contribute to a fixed outcome (as opposed to identifying changes leading to a different outcome, as in CEs), so preservation under sparsification is crucial to assess whether interpretability can be maintained. In contrast, generating CEs typically involves heuristic or optimization-based search procedures, rather than direct computation as in AAEs and RAEs. Therefore, the primary concern for CEs is whether sparsification can accelerate this search. We empirically investigate these questions and make the following contributions:

1. We analyse the impact of sparsification on AAEs, and find that AAEs are generally well-preserved across varying sparsity levels.
2. We analyse the impact of sparsification on RAEs, and find that RAEs are not as well-preserved under sparsification as AAEs.
3. We propose a method that leverages the sparsification to improve the runtime of computing CEs.

The code is available at https://github.com/DanielPeacock/ArguingWithNeuralNetworksPublic.

2. Preliminaries

In this paper we focus on NNs in the form of MLPs. These are directed, acyclic graphs as illustrated in Figure 1a, processing inputs in an input layer (on the left in Figure 1a) through hidden layers (layer 1 and 2 in Figure 1a) to obtain a prediction in the last layer (on the right in Figure 1a). Nodes in all layers amount to neurons, whose activation is determined by an activation function applied to the (edge) weighted sum of the neuron's incoming connections plus a bias value assigned to each neuron. Throughout this paper we use the logistic activation function.

[1] Note that AAEs and RAEs use the unweighted QBAFs, while CEs use the weighted QBAFs. With a slight abuse of notation, we use "QBAF" to refer to both throughout the paper.

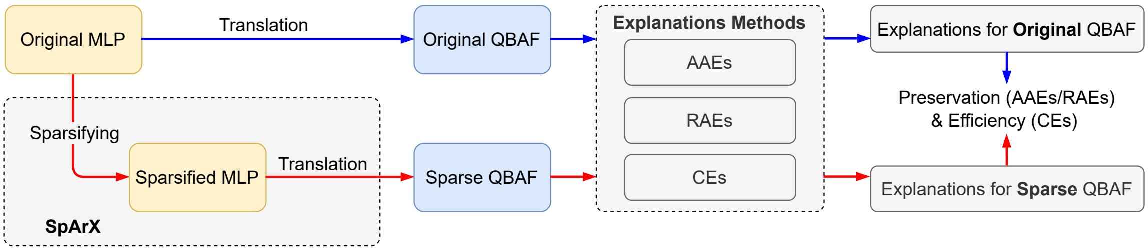
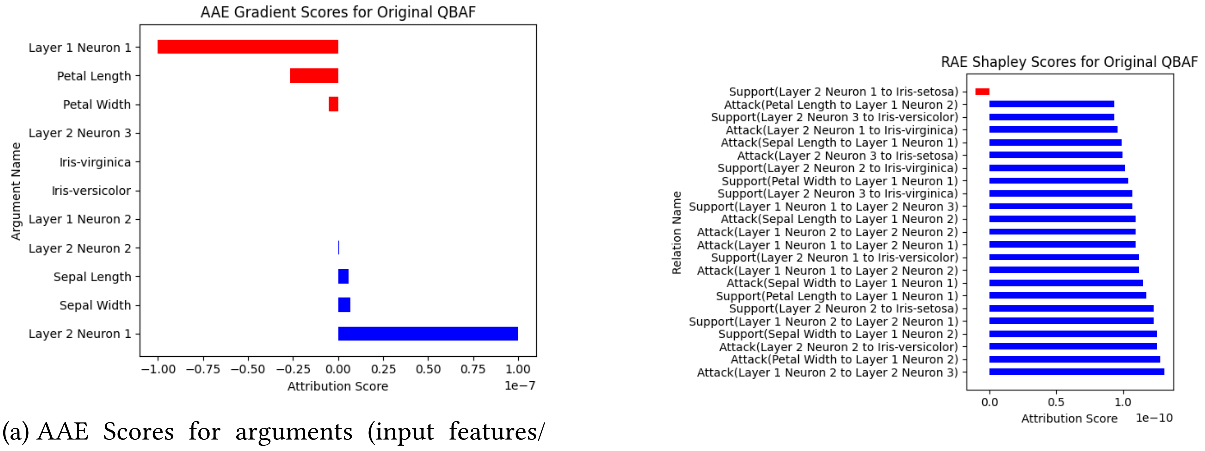


Figure 2: Impact of sparsification on quantitative argumentative explanations (AAEs, RAEs, and CEs). Red and blue arrows indicate sparsified and original flows, respectively.



(a) AAE Scores for arguments (input features/ neurons). Red/blue scores indicate a negative/positive influence on the topic argument (output).

(b) RAE Scores for relations. Red/blue scores indicate a negative/positive influence on the topic argument (output).

Figure 3: An illustration of AAEs (a) and RAEs (b) for the QBAF corresponding to the MLP in Figure 1a, using the topic argument `Iris-setosa`. A user can understand which arguments and relations are the most influential/important in determining the output of the MLP (i.e. the MLP prediction of `Iris-setosa`).

Each MLP can be represented by an *equivalent* Quantitative Bipolar Argumentation Framework (QBAFs). Here, we view each neuron as an argument, and each edge as a relation. Edges with negative weights are attacks, and edges with positive weights are supports. Each argument has a base score corresponding to its initial strength and relations are weighted. For more details on the translation process see [9]. This provides a new argumentative interpretation of MLPs with dialectical strength values for each argument (mathematically equivalent to the activations of neurons in the MLP).

We use several existing argumentation-based explanation methods, overviewed here (see original papers for more details).

SpArX [7] The QBAF interpretation of MLPs does not necessarily improve explainability since QBAFs are of the same size and density as the MLPs, which can be very large. SpArX provides explanations by reducing the size of the given MLPs first. The neurons in the hidden layers are clustered based on their activations, and then merged by averaging their biases and edge weights. The sparse MLPs are then converted to equivalent QBAFs (Figure 1b) from which *qualitative* explanations can be found, for example by creating word clouds of the most important input features or examining the dialectical relationships between the arguments. In this paper we consider instead quantitative argumentative explanations drawn from the sparsified QBAFs.

AAEs [16] AAEs attempt to explain QBAFs by examining the contribution of other arguments to a topic argument. Throughout this paper, we use *topic argument* to refer to the argument we are trying to explain (usually this is an argument corresponding to one of the output neurons in the equivalent MLP).

We focus on Gradient-based AAEs (although other types exist such as Shapley-based [19] and Removal-based AAEs [20]). Gradient methods work by computing a score which represents the sensitivity of the topic argument to changes in the base score of other arguments. An example is shown in Figure 3a.

RAEs [17] RAEs attempt to understand the role of the relations in contributing to the strength of a topic argument. In this paper, we focus on Shapley-based RAEs (although other types such as Gradient-based RAEs [18] also exist). These are based on Shapley values [21], and look at every subset of the attacks and supports to understand the influence of each one on the topic argument. Due to the complexity in computing these scores, an approximation is used. Figure 3b shows an example.

CEs [18] CEs calculate how the weights of each relation in the QBAF must be modified in order to reach a certain dialectical strength in the topic argument (called the *desired strength*). This is similar to the counterfactual problem in AI, where methods are used to try and explain how a model's outputs would change with modifications to the inputs [22, p. 847 - 848]. CEs are computed by iteratively updating the weights using the gradient-based RAE (G-RAE) to guide the search until the desired strength is reached. Table 1 shows an example.

Table 1
An illustration, showing (a selection of) the original edge weights and CE edge weights for Figure 1a. A user can observe how the edge weights must be changed in order to reach a desired strength for the topic argument.

Relation (Edge)	Original Weight	CE Edge Weight
(Sepal Width, Layer 1 Neuron 2)	6.96	0.00
(Sepal Length, Layer 1 Neuron 2)	6.33	0.97
(Petal Width, Layer 1 Neuron 2)	12.30	0.07
(Petal Length, Layer 1 Neuron 2)	12.63	0.57
(Layer 2 Neuron 3, Iris-virginica)	11.91	0.94

3. Methodology

In order to ascertain whether AAEs and RAEs are preserved after sparsification, we train MLPs of various sizes and compare an aggregation of the scores for the original MLPs to the scores for the MLP after sparsification. The aggregation is needed to allow a comparison of scores since there are significantly more scores for the original MLPs due to their larger size. Here, we define these aggregations.

Let $C = \{c_1, \ldots, c_n\}$ be a cluster of interest after sparsification in hidden layer l, containing neurons c_i for $i = 1, \ldots, n$. Similarly, let $C' = \{c'_1, \ldots, c'_m\}$ be another cluster of interest in the next layer $l + 1$ containing neurons c'_j for $j = 1, \ldots, m$.

Aggregation of AAEs We aggregate the AAE scores by averaging the score for each neuron in the cluster of interest. Formally, the aggregated score for cluster C is

$$\text{Agg\_aae\_score}(C) = \frac{1}{n} \sum_{c_i \in C} \text{aae\_score}(c_i).$$

A simple example of this process is shown in Figure 4.

Aggregation of RAEs We aggregate the RAE scores by averaging the RAE scores of edges between all pairs of neurons contained in two clusters of interest. Formally, the aggregated score for the edge between clusters C and C' is

$$\text{Agg\_rae\_score}(C, C') = \frac{1}{n} \sum_{c_i \in C} \frac{1}{m} \sum_{c_j \in C'} \text{rae\_score}(c_i, c_j).$$

A simple example of this process is shown in Figure 5.

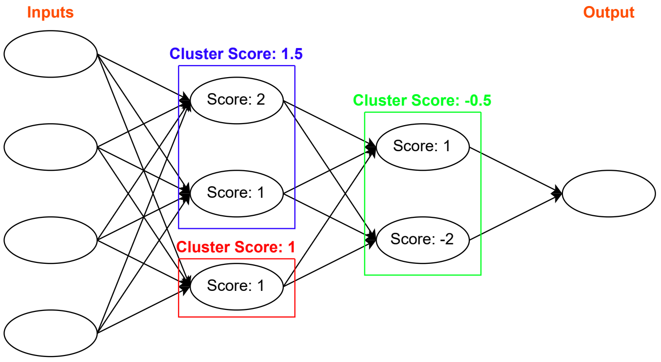


Figure 4: An example of aggregation of AAE scores. The boxes indicate neurons clustered together by SpArX. We create an aggregated score by averaging the scores of the neurons within each box. Note that the input and output layers are not modified by SpArX, so no aggregation is necessary for these layers.

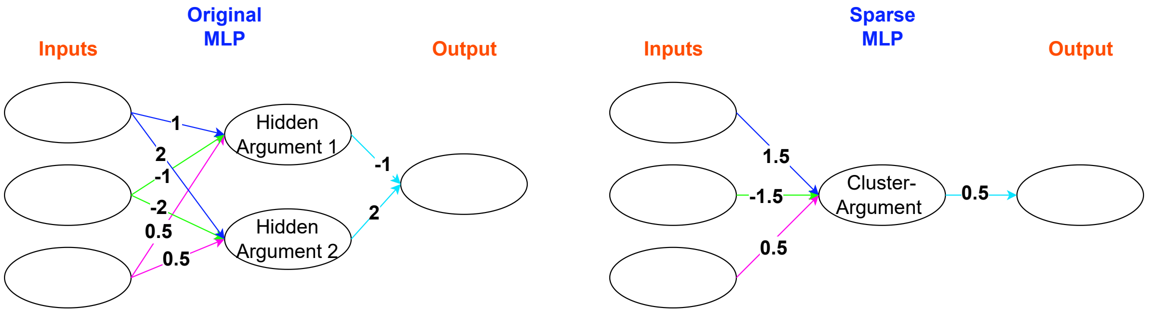


Figure 5: An example of RAE aggregation. The aggregated RAE score (right) is computed by averaging all pairs of edges between arguments in the two clusters, indicated by the same colour (left). Note that the input and output layers are not sparsified by SpArX.

De-aggregation of CEs We do not attempt to directly understand if CEs are preserved with sparsification since this question is ill-defined. Indeed, CEs do not give a fixed score to each component in the same way as AAEs and RAEs and so it is challenging to define what preservation means in this setting. Instead, we look at CEs in the opposite direction: that is, we attempt to de-aggregate the CEs for the sparse MLPs to approximate/recover the CE for the original MLPs and improve computational efficiency. The sparse CE assigns weights to each edge in the sparse QBAF. We de-aggregate by assuming the weights are *equally distributed* amongst edges merged together by SpArX. Every edge merged together is assigned the same weight in our approximate CE. Formally, consider the edge between two clusters C and C', assigned weight w in the sparse CE. There are a total of mn edges between every pair of neurons in these clusters. So every edge between these pairs is assigned weight $\frac{w}{mn}$ in the approximate CE. A simple example of this process is shown in Figure 6.

4. Results and Analysis

To compare the aggregated AAE and RAE scores with the sparse MLP scores, we look at two approaches: the overall pattern in the scores and the highest scoring arguments/ edges. All of our analysis is for the Iris [23], Diabetes [24], Cancer [25] and COMPAS [26] datasets and we average our results over the test set for each dataset. These datasets are commonly used and of varying levels of complexity (number of input features and dataset size). For AAEs we train MLPs with 1 - 2 hidden layers each with 10 - 100 neurons. For RAEs we train MLPs with 1 - 2 hidden layers each with 2 - 10 neurons. These are significantly smaller than the MLPs used for AAEs due to the time complexity involved in computing Shapley-based RAEs making it impractical to use large MLPs.

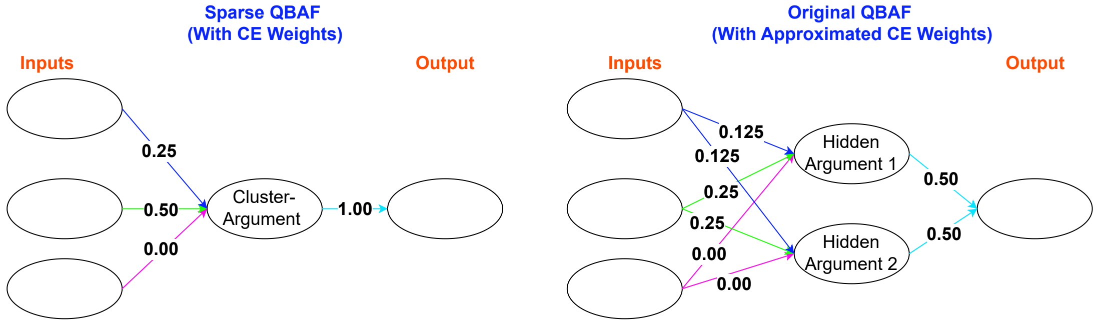


Figure 6: The methodology for reverse engineering/approximation of CEs for the original QBAF. For each edge in the CE for the sparse QBAF, divide it by the number of edges which were averaged together to create the sparse edge. Every edge averaged together in the original QBAF will be assigned the same weight in the approximate CE. For example, the light blue edge in the sparse QBAF has a CE weight of 1.0. We divide this by the number of edges which were averaged together to create this sparse edge (in this case, 2), resulting in an approximate CE weight of 0.5 for both these edges in the original QBAF.

Overall Pattern To check if the overall pattern was preserved we check the Spearman Rank [27] and Kendall-τ coefficient [28]. These provide a measure of the strength of the relationship between two variables. We use these measures to examine the strength of the correlation between the aggregated scores and the sparse scores. A rank/coefficient close to 1 means a strong correlation, indicating the pattern in both sets of scores is similar and hence the pattern in the scores is preserved with sparsification. We also rank the arguments and edges based on their scores. We then look at the percentage difference in ranking between each aggregated argument/edge and the corresponding argument/edge in the sparse QBAF. A small difference in rankings would indicate a similar pattern after sparsifying.

Highest Scores We also look specifically at the highest scoring arguments/edges. These are important since they are the most influential components of the MLPs so it is important these are preserved. Firstly, we look at the top-ten scoring arguments/edges and check what percentage of arguments/edges in the aggregated scores are also in the top-ten of the sparse MLP scores i.e. how many of the highest scoring arguments/edges stay the same after sparsification. In addition, for RAEs we also look at the top-scoring aggregated edge and check whether this edge is in the top-ten of the sparse scores i.e. checking that the most important edge remains important after sparsification. High percentages would indicate that the highest scoring arguments/edges are preserved after sparsifying.

4.1. Preservation of AAEs

Overall our results are positive, showing that AAEs are preserved well by sparsification.

Overall Pattern The results can be found in Table 2. We find that the pattern/distribution of scores matches closely before and after sparsification. We can see that the correlation between the scores is very strong. The coefficients are always higher than 0.7 and in most cases at least 0.9. Since the coefficients/ranks computed are all close to 1, we can conclude that the pattern in scores is preserved well with sparsification. We should note that the coefficients do reduce slightly towards higher levels of sparsification (around 90%), but this is very little and the correlation still remains strong. Considering the rankings differences (Δ in Table 2), we can see that the rankings are similar. Towards the lower levels of sparsification, there is only around a 10% difference in rankings, and at high levels of sparsification this goes up to 30%. However, this is still relatively low, and only appears with high levels of sparsification (90%). This is also to be expected, since high levels of sparsification should result in greater loss of information.

Table 2
Results for the preservation of AAE scores (↑: higher is better; ↓: lower is better), including Spearman ρ(↑), Kendall τ(↑), and the percentage rankings differences (%) Δ(↓) between the aggregated and sparse AAE scores.

Dataset		Preservation of Overall Pattern of AAE Scores								
		Sparsification Amount								
		10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
Iris	ρ	0.907	0.894	0.882	0.871	0.867	0.857	0.854	0.859	0.859
	τ	0.904	0.893	0.882	0.872	0.868	0.858	0.853	0.855	0.848
	Δ	7.89	10.06	11.66	12.31	12.71	12.87	13.16	13.70	16.35
Diabetes	ρ	0.995	0.992	0.989	0.988	0.984	0.978	0.972	0.968	0.816
	τ	0.994	0.992	0.990	0.989	0.987	0.982	0.978	0.976	0.847
	Δ	5.89	8.05	10.22	11.80	13.60	15.53	18.50	22.84	31.32
Cancer	ρ	0.965	0.963	0.958	0.952	0.942	0.921	0.901	0.874	0.824
	τ	0.967	0.965	0.960	0.955	0.945	0.918	0.891	0.858	0.783
	Δ	6.65	8.91	11.18	13.06	14.84	16.84	19.75	22.76	27.03
COMPAS	ρ	0.994	0.994	0.993	0.992	0.991	0.991	0.987	0.985	0.926
	τ	0.995	0.995	0.994	0.993	0.992	0.992	0.989	0.987	0.915
	Δ	3.56	5.79	8.63	10.89	14.29	17.23	20.98	25.34	31.33

Table 3
Results for the preservation of the top ten scoring arguments, checking what percentage of the top ten scoring arguments stay the same before and after sparsification.

Dataset	Top-Ten Arguments Preservation (%)								
	Sparsification Amount								
	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
Iris	63.03	57.72	59.23	61.20	63.94	66.84	68.92	72.56	74.53
Diabetes	86.59	82.00	78.17	74.92	70.83	67.30	63.27	59.05	48.83
Cancer	85.58	77.82	71.38	67.93	67.64	67.27	65.62	60.36	47.19
COMPAS	92.75	88.12	83.80	78.85	73.02	68.83	64.55	60.03	52.63

Highest Scoring Arguments The results can be found in Table 3. We again see that in general the highest scoring arguments are preserved. With moderate levels of sparsification (10% - 60%), around 70% of the top-scoring arguments are preserved. This does reduce as we increase the amount of sparsification, but again this is to be expected as the amount of information lost should increase as we sparsify more. There is a balance between sparsification and loss of information, but this depends on the type of dataset and its complexity e.g. the COMPAS dataset (the most complex) loses more information with high sparsification levels, but this does not happen to the same extent with less complex datasets such as Iris. However, in general, most of the top scoring arguments do remain within the top 10, and so we can conclude that the highest scoring arguments with AAEs are preserved well.

4.2. Preservation of RAEs

Overall, the results are mixed and we find that RAEs are not preserved well in the way seen for AAEs.

Overall Pattern The results can be found in Tables 4a and 4b. Considering first the Spearman Rank and Kendall τ coefficients (Table 4a), there is a relatively strong correlation between the aggregated scores and sparse scores. The ranks/coefficients are around 0.8, although this reduces as the sparsification level is increased. For example, the Spearman Rank of the Diabetes dataset decreases from 0.888 at 20% sparsification to only 0.696 at 80%. This is a similar pattern to what we saw for AAEs, and largely what is to be expected; the sparser an MLP, the more information is lost. This indicates that the overall pattern in scores is relatively well preserved. However, compared to the equivalent AAEs analysis (Table 2), the correlation is significantly lower (around 0.9 for AAEs). Therefore, although the pattern looks to be preserved, we do lose more information about RAEs with sparsification compared to AAEs.

Table 4
Results for the preservation of the overall pattern of RAEs. Averaged over whole test set.

(a) Results for the Spearman Rank (ρ) and Kendall τ Coefficient (τ) between the aggregated RAE scores and sparse RAE scores.

Dataset		Average Rank/Coefficient			
		Sparsification Amount			
		20%	40%	60%	80%
Iris	ρ	0.844	0.813	0.810	0.857
	τ	0.885	0.864	0.861	0.892
Diabetes	ρ	0.888	0.832	0.793	0.696
	τ	0.912	0.873	0.840	0.761
Cancer	ρ	0.695	0.716	0.719	0.724
	τ	0.786	0.801	0.803	0.807
COMPAS	ρ	0.838	0.805	0.786	0.779
	τ	0.877	0.853	0.838	0.835

(b) Results for the average difference in rankings between the aggregated RAE scores and the sparse RAE scores.

Dataset	Average Rankings Difference (%)			
	Sparsification Amount			
	20%	40%	60%	80%
Iris	30.40	30.71	30.83	32.98
Diabetes	31.89	32.23	32.04	31.94
Cancer	33.27	33.99	33.63	33.46
COMPAS	32.75	29.99	30.08	28.67

Table 5
Results for the preservation of the highest scoring edges for RAEs.

(a) Results for the preservation of the top ten scoring edges (percentage of edges which stay the same before and after sparsification).

Dataset	Top-Ten Edges Preservation (%)			
	Sparsification Amount			
	20%	40%	60%	80%
Iris	27.53	31.94	39.57	38.63
Diabetes	24.69	30.43	39.84	49.30
Cancer	26.73	23.17	26.71	28.29
COMPAS	31.02	41.57	52.10	62.38

(b) Results for the average percentage of RAEs where the top scoring edge in the aggregated scores is within the top-ten of the sparse scores.

Dataset	Most Important Edge In Top 10(%)			
	Sparsification Amount			
	20%	40%	60%	80%
Iris	39.21	41.36	52.19	41.15
Diabetes	23.53	33.14	40.81	48.08
Cancer	40.50	31.87	30.70	29.23
COMPAS	27.93	53.62	64.04	82.65

Looking at Table 4b, compared to AAEs (Table 2), the rankings of the edges change significantly more for RAEs. The rankings on average change by around 30% (over all levels of sparsification), compared to only around 5 - 15% for AAEs. The rankings only change by around 30% for AAEs with very high levels of sparsification (90%), but this is always the case for RAEs, even at very low levels of sparsification. This indicates that significantly more information is lost through sparsification for RAEs, and the overall pattern in scores is not preserved well for RAEs. This does seem to contradict the correlation coefficients/ranks seen previously, but this only measures the correlation in the scores and does not look at the individual scores themselves.

Looking at both sets of results, we can conclude that although the scores before and after sparsification are highly correlated, the individual scores and their rankings *are* affected by sparsification. The overall pattern is to some extent preserved (strong correlation), but lots of information is lost as a result especially in the individual scores.

Highest Scoring Edges The results can be found in Tables 5a and 5b. We can see that the most important edges are *not* well preserved by sparsification. First looking at Table 5a, we see that in all cases, a very low percentage of edges remain in the top ten. The results indicate that generally around 30% of the top-ten edges stay the same, but is as low as 23% in some cases. This fits with our previous analysis that the individual rankings of edges is not preserved well, and there is a large change in rankings. This tells us that in general the preservation of the highest scoring edges is poor, and information about RAEs is lost as a result of sparsification.

Looking at Table 5b, we again see poor preservation. In most cases the top scoring edge *does not*

Table 6
The percentage of cases observed where our approximated CE is valid (1)/ where the topic argument's strength is closer to the desired strength than before applying the CE (2).

	Percentage of Cases with Valid Approximated CEs (1)/ Lower Distance (2)								
Dataset	**Sparsification Amount**								
	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
Iris	2.38/ 93.90	2.06/ 94.11	2.53/ 94.43	3.12/ 94.53	3.54/ 94.85	3.91/ 95.70	5.82/ 96.08	9.74/ 96.71	11.53/ 99.51
Diabetes	5.47/ 82.93	7.02/ 83.27	5.74/ 83.82	6.73/ 84.60	8.77/ 85.91	9.66/ 86.52	10.38/ 88.40	6.28/ 90.41	10.22/ 92.27
Cancer	7.15/ 98.11	6.89/ 98.23	7.23/ 98.32	5.74/ 88.53	8.24/ 98.82	9.16/ 97.28	9.25/ 99.33	9.79/ 99.40	14.18/ 99.39
COMPAS	15.91/ 90.41	15.53/ 90.27	17.93/ 89.80	18.01/ 90.33	20.64/ 90.22	19.69/ 89.27	19.57/ 88.30	13.63/ 87.05	15.77/ 79.60

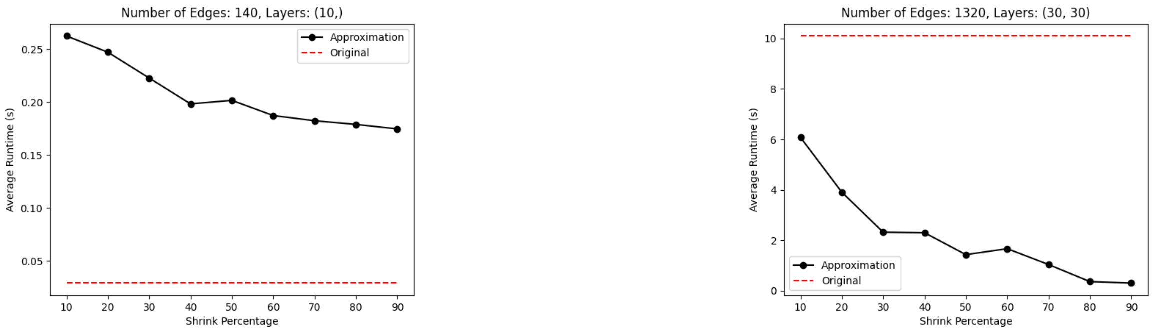
remain high scoring after sparsification. In general, only in around 40% of cases does the top scoring edge remain high-scoring and this is as low as 23% in some cases. We should note that for the COMPAS dataset, the highest scoring edges do look to be better preserved. Due to the size and the complexity of the dataset, significantly fewer MLPs were tested compared to the other datasets. This may have resulted in the slightly different results for COMPAS compared to the other datasets. However, the pattern across all datasets tested indicates that the highest scoring edges are not well preserved.

4.3. De-aggregated CEs

We analyse the de-aggregated CEs using a different methodology to that used for AAEs and RAEs. We look at the validity and distance to check the quality of the de-aggregated (approximate) CEs and use this methodology to improve runtime.

Validity A CE is *valid* if the topic argument attains the desired strength using the edge weights provided by the CE. We create an approximate CE for the original QBAF from the sparse QBAF using the de-aggregation method in Section 3. We check the validity of our approximate CE and the results are shown in Table 6 (results labelled 1). Clearly, we can see that the approximation does not successfully produce a valid CE in the majority of cases. In many cases, the percentage of valid CEs is less than 10% so our approximation clearly does not work effectively. For the COMPAS dataset, the percentage of valid CEs does increase, up to around 20% in some cases. This is positive since the COMPAS dataset is the most complex of the datasets. However, this is still a very low percentage, and therefore we can conclude that our approximation is not effective. It is likely that our assumption in the approximation that the edge weights were equally distributed is incorrect, causing this poor performance.

Distance To further understand the quality of the approximate CEs, we also look at the distance. We check if the topic argument's strength gets *closer* to the desired strength than before the CE weights are applied. The results can be found in Table 6 (results labelled 2). We can see that our approximate CE does bring the strength of the topic argument closer to the desired strength in the majority of cases. For the Iris and Cancer datasets, consistently in over 90% of cases, the approximate CE brings the strength closer. For the Diabetes dataset, this percentage does reduce to around 80%, but this is still high and in most cases the approximation does succeed. Finally, for the COMPAS dataset, around 85%-90% of cases generally are closer, except for the 90% sparsification case, where the percentage reduces to 73%. Overall, however, this is still a positive result in bringing the strength of the topic argument closer to the desired strength. Note also that the approximation does successfully get closer even at very high levels of sparsification. This implies that CEs are preserved well with sparsification.



(a)

(b)

Figure 7: The average runtime for the COMPAS dataset using the original CE algorithm (red, dashed line) and our modified CE algorithm (black, solid line) for MLPs with one hidden layer of 10 neurons (a) and two hidden layers of 30 neurons each (b). The x-axis shows the sparsification percentage, and the y-axis the runtime (in s).

Table 7
Percentage reduction in the average runtime using the sparse CE methodology for MLPs with 500 or more edges.

	Percentage Reduction in Average Runtime								
Dataset	**Sparsification Amount**								
	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
Iris	38.06	44.44	54.17	56.11	59.03	60.97	62.64	62.08	63.89
Diabetes	-57.51	-212.38	-18.30	16.29	32.39	34.242	45.97	52.87	51.04
Cancer	-56.83	-36.09	-24.45	7.60	15.09	22.26	30.38	42.78	43.41
COMPAS	31.71	32.12	63.84	65.56	78.83	81.73	84.33	85.67	86.68

Runtime The CE algorithm in [18] works by iteratively updating the weights of the QBAF until the desired strength is reached. However, the initial weights are randomised and so the algorithm can take some time to converge to the desired strength (the algorithm may get stuck in a local minimum). Therefore, to improve convergence, we can initialise the weights of the QBAF with our approximated CE instead of randomised weights. We perform experiments to see if the runtime was improved using our approximation. We run all experiments on a Linux PC running 64-bit Ubuntu 24.04, Intel Core i7-8700 3.20GHz processor and 16GB memory. We compare the following:

(a) Apply the usual CE algorithm i.e. translate the MLP to the equivalent QBAF and apply the CE algorithm using randomised initial weights.
(b) Use our approximation method i.e. sparsify the MLP, translate to a sparse QBAF, apply the CE algorithm to the sparse QBAF with random initial weights, create an approximation for the original MLP and apply the CE algorithm to the original MLP using the approximation as the initial weights.

We plot graphs of the average runtime using the two methods for each dataset and MLP size. For the second method, this involves checking both how much time is spent on (1) computing the CE on the sparsified MLP and (2) computing the CE on the original MLP starting from the previously computed CE. We give two of these graphs in Figure 7 (one small MLP, and one larger MLP) for the COMPAS dataset only (the most complex analysed) for succinctness although the full set of graphs for each dataset can be found in Appendix A. From the figure, we can see that when the MLP is small (with a small number of edges), the runtime is longer using our method as more steps must be done to compute the CE, and due to the small number of edges, the CE algorithm converges quickly anyway. However, when the MLP is larger, our method does improve the runtime. For this reason, we calculate the percentage reduction in the average runtime for MLPs of more than 500 edges only. The results are given in Table 7. We see that for low levels of sparsification our method can still be slower (e.g. for Diabetes). However, in all cases, for high levels of sparsification there is a reduction in runtime, often by more than 50%. The much larger number of edges (over 1300 in the graph shown in Figure 7b) means that the CE algorithm takes longer to converge, so initialising the weights with the approximate CE does improve the runtime. We can also note that using our initial guess from a very high level of sparsification (e.g. 90%), our method performs well. This is positive, as even at a high level of sparsification, we can still recover a large amount of information. Although we cannot easily recover a valid CE, we can still make a good guess.

5. Conclusion

In this paper, we explored the impact of sparsification on quantitative argumentative explanations. Our investigations allow us to understand whether sparsification alters or distorts the argumentative explanations (AAEs, RAEs, CEs) produced from the resulting QBAF. Without this, users could be misled by explanations, leading to ethical or legal risks. Our findings showed that AAEs are well-preserved under sparsification, suggesting that AAEs can be reliably used alongside sparsification to enhance the interpretability of NNs. In contrast, RAEs appeared less robust, making them challenging to use alongside sparsification. Finally, for CEs, we saw that sparsification can improve the computational efficiency of explanation generation, which is particularly useful for large and dense MLPs.

There are a few avenues for future work. While we found promising empirical results for the preservation of AAEs, further work is needed to establish theoretical guarantees for such preservation. Additionally, while our findings do not support the preservation of RAEs, future work could explore whether gradient-based RAEs (G-RAEs) [18] exhibit better consistency, potentially enabling RAEs to contribute more effectively to explanations in sparsified settings. Future work could also explore using a weighted aggregation of the RAE scores, similar to the weighted averaging used by SpArX when merging the edge weights ([7, Def. 6]) to see if this results in better preservation. Further, for both AAEs and RAEs, other aggregation methods could be explored. Using other techniques instead of mean aggregation (e.g. min./max. aggregation) may result in sparsification having a lower impact.

Finally, although we observed that sparsification can improve the speed of computing CEs in large MLPs, our method is a heuristic. Further work is necessary to find guarantees as to when our method is faster, perhaps by finding a lower bound on MLP size for which our method is guaranteed to improve the runtime. Our approximation method also did not directly produce valid CEs; further work should be done to find a method of de-aggregating the CEs to produce valid CEs for the original QBAFs. Perhaps weighting the edges differently rather than assuming a equal distribution would help.

Acknowledgments

This research was partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934, ADIX) and by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Any views or opinions expressed herein are solely those of the authors.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: ACM SIGKDD, 2016, pp. 1135–1144.

[2] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.

[3] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, PLOS ONE 10 (2015) e0130140.

[4] P. Angelov, E. Soares, Towards explainable deep neural networks (xdnn), Neural Networks 130 (2020) 185–194.

[5] A. Rago, K. Cyras, J. Mumford, O. Cocarascu, Argumentation and Machine Learning, in: D. Gabbay, G. Kern-Isberner, G. R. Simari, M. Thimm (Eds.), Handbook of Formal Argumentation, Volume 3, 2024. arXiv:2410.23724.

[6] K. Čyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: A survey, in: Z.-H. Zhou (Ed.), IJCAI-21, 2021, pp. 4392–4399.

[7] H. Ayoobi, N. Potyka, F. Toni, SpArX: Sparse Argumentative Explanations for Neural Networks, in: ECAI, volume 372 of *Frontiers in Artificial Intelligence and Applications*, 2023, pp. 149–156.

[8] T. Mossakowski, F. Neuhaus, Modular semantics and characteristics for bipolar weighted argumentation graphs, CoRR (2018). arXiv:1807.06685.

[9] N. Potyka, Interpreting Neural Networks as Quantitative Argumentation Frameworks, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 6463–6470.

[10] P. Baroni, M. Romano, F. Toni, M. Aurisicchio, G. Bertanza, Automatic evaluation of design alternatives with quantitative argumentation, Argument & Computation 6 (2015) 24–49.

[11] A. Rago, F. Toni, M. Aurisicchio, P. Baroni, Discontinuity-free decision support with quantitative argumentation debates, in: 15th International Conference on the Principles of Knowledge Representation and Reasoning (KR), 2016.

[12] N. Potyka, Continuous dynamical systems for weighted bipolar argumentation, in: International Conference on Principles of Knowledge Representation and Reasoning (KR), 2018, pp. 148–157.

[13] L. Amgoud, J. Ben-Naim, Evaluation of arguments in weighted bipolar graphs, International Journal of Approximate Reasoning 99 (2018) 39–55.

[14] C. Fan, J. Liu, Y. Zhang, E. Wong, D. Wei, S. Liu, Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation, in: The Twelfth International Conference on Learning Representations, 2024.

[15] S. Han, J. Pool, J. Tran, W. J. Dally, Learning both weights and connections for efficient neural networks, in: Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15, MIT Press, Cambridge, MA, USA, 2015, p. 1135–1143.

[16] X. Yin, N. Potyka, F. Toni, Argument attribution explanations in quantitative bipolar argumentation frameworks, in: ECAI, volume 372, 2023, pp. 2898–2905.

[17] X. Yin, N. Potyka, F. Toni, Explaining arguments' strength: Unveiling the role of attacks and supports, in: K. Larson (Ed.), IJCAI-24, 2024, pp. 3622–3630.

[18] X. Yin, N. Potyka, A. Rago, T. Kampik, F. Toni, Contestability in quantitative argumentation, arXiv preprint arXiv:2507.11323 (2025).

[19] T. Kampik, N. Potyka, X. Yin, K. Čyras, F. Toni, Contribution functions for quantitative bipolar argumentation graphs: A principle-based analysis, International Journal of Approximate Reasoning 173 (2024) 109255.

[20] J. Delobelle, S. Villata, Interpretability of gradual semantics in abstract argumentation, in: G. Kern-Isberner, Z. Ognjanović (Eds.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Cham, 2019, pp. 27–38.

[21] L. S. Shapley, Notes on the N-Person Game II: The Value of an n-Person Game, Santa Monica, CA, 1951.

[22] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harvard Journal of Law and Technology 31 (2018) 841–887.

[23] R. A. Fisher, Iris, UCI Machine Learning Repository, 1936.

[24] National Institute of Diabetes and Digestive and Kidney Diseases, Diabetes Dataset, Kaggle, 1990.

[25] W. N. Street, W. H. Wolberg, O. L. Mangasarian, Breast Cancer Wisconsin (Diagnostic), UCI Machine Learning Repository, 1993.

[26] ProPublica, Compas recidivism risk score data and analysis, GitHub, 2016.

[27] C. Spearman, The proof and measurement of association between two things. (1961).

[28] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.

Appendix

A. Runtime graphs

Here we give the full results for the runtime of our modified CE method compared to the original CE method. In the plots, the x-axis is the sparsification percentage, and the y-axis is the runtime (in seconds). The red line represents the average runtime (over the test dataset) of the original CE algorithm, and the black line is the runtime using our approximation method with various levels of sparsification.

- In Figure 8, we see the results for the Iris dataset.
- In Figure 9, we see the results for the Diabetes dataset.
- In Figure 10 we see the results for the Cancer dataset.
- In Figure 11, we see the results for the COMPAS dataset.

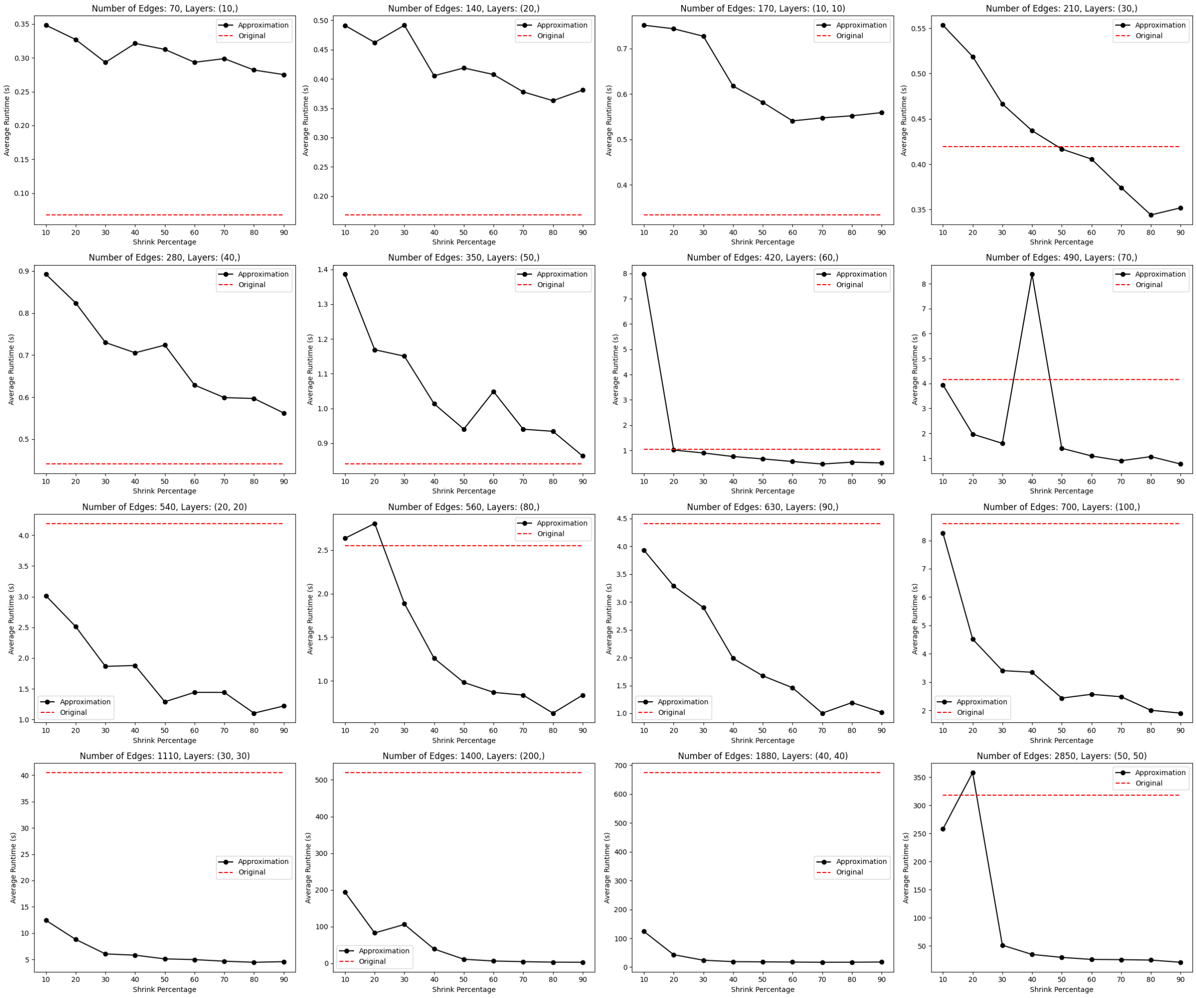


Figure 8: Runtime (in s) of the original CE method and our approximation method for the Iris dataset with various MLP sizes.

We can see from these plots that in general when the MLP is small (a low number of edges), the original CE method is faster than our approximation method. However, as the MLP size increases, our method can outperform the original CE method.

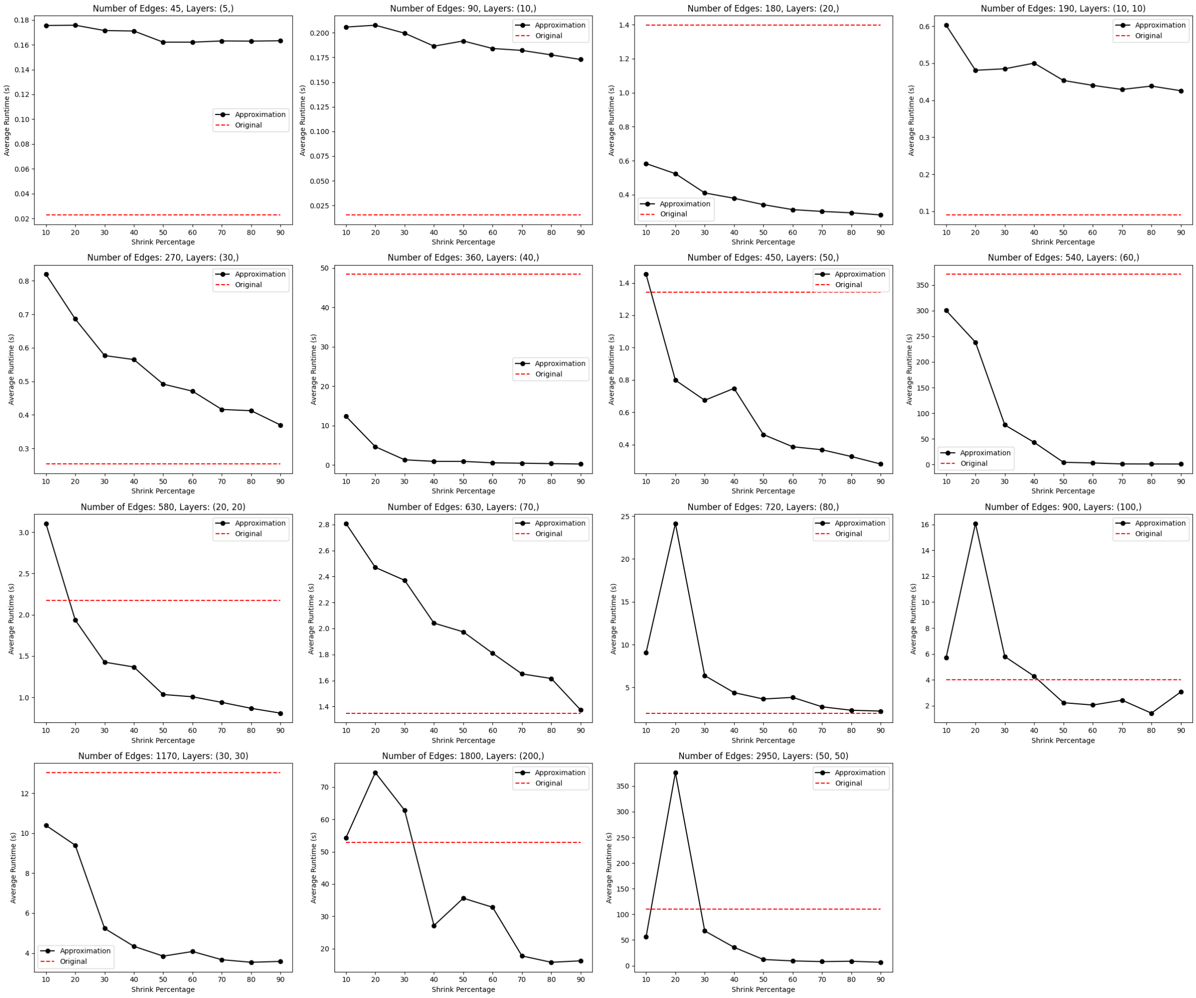


Figure 9: Runtime (in s) of the original CE method and our approximation method for the Diabetes dataset with various MLP sizes.

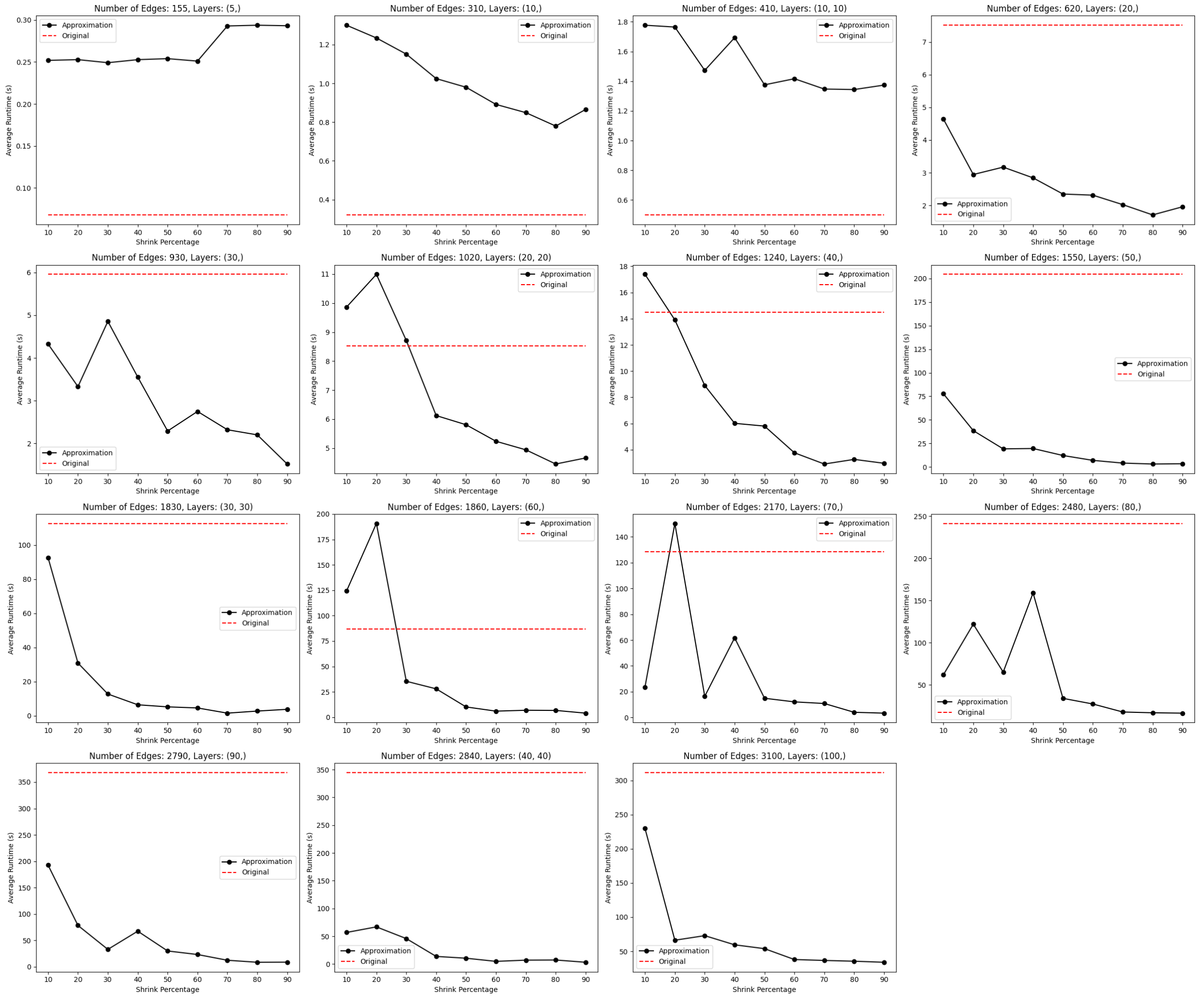


Figure 10: Runtime (in s) of the original CE method and our approximation method for the Cancer dataset with various MLP sizes.

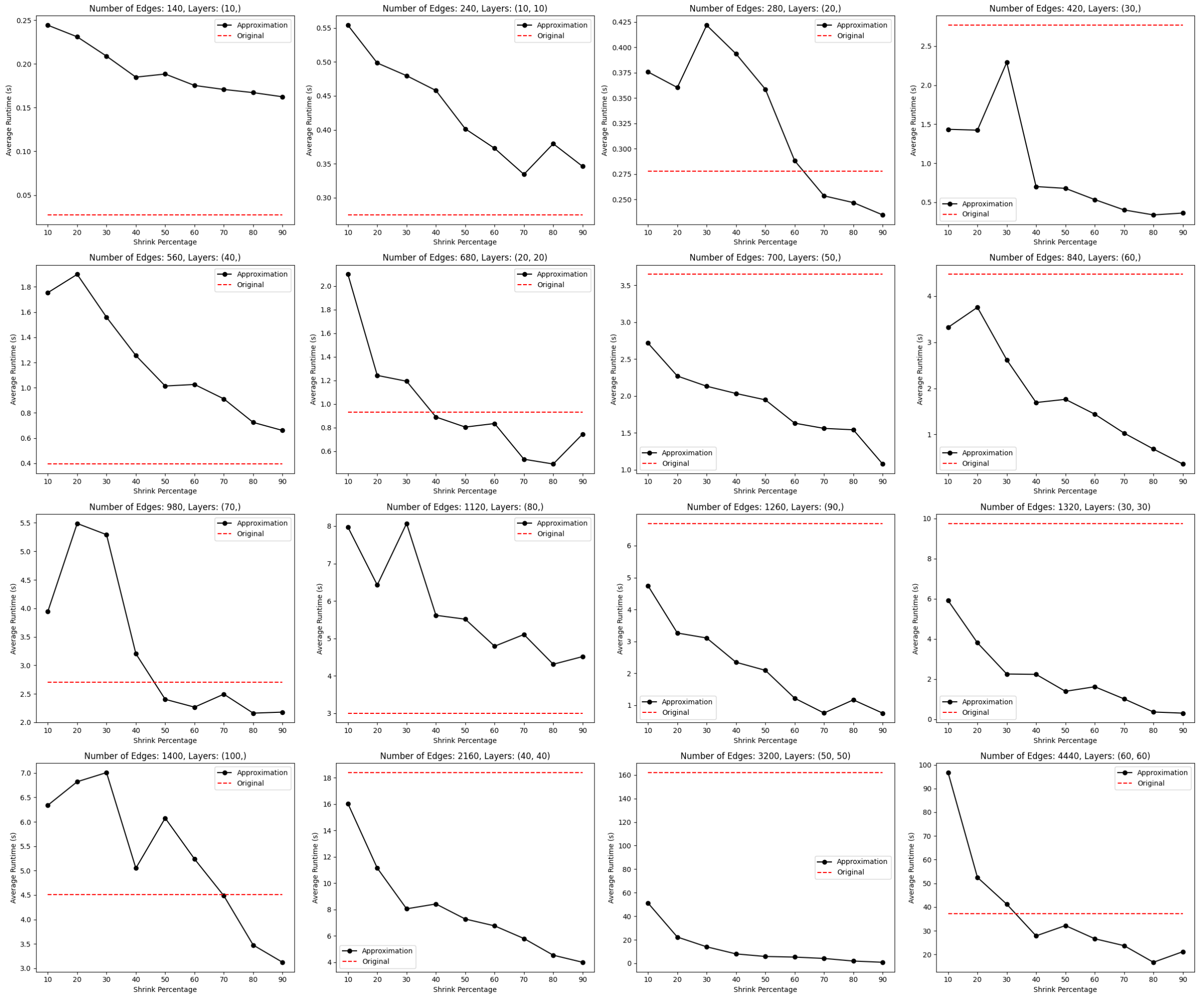


Figure 11: Runtime (in s) of the original CE method and our approximation method for the COMPAS dataset with various MLP sizes.