RESEARCH-ARTICLE

Online Activation Value-aware Clustering and Aggregation for Faithful Argumentative Explanations

UNGSIK KIM, Gyeongsang National University, Jinju, Kyongsangnam-do, South Korea

JIHO BAE, Gyeongsang National University, Jinju, Kyongsangnam-do, South Korea

SANG-MIN CHOI, Gyeongsang National University, Jinju, Kyongsangnam-do, South Korea

SUWON LEE, Gyeongsang National University, Jinju, Kyongsangnam-do, South Korea

Published: 10 November 2025

Citation in BibTeX format

CIKM '25: The 34th ACM International Conference on Information and Knowledge Management
November 10 - 14, 2025
Seoul, Republic of Korea

Conference Sponsors:
SIGWEB
SIGIR

Online Activation Value-aware Clustering and Aggregation for Faithful Argumentative Explanations

Ungsik Kim
Gyeongsang National University
Jinju-si, Gyeongsangnam-do, Republic of Korea
blpeng@gnu.ac.kr

Jiho Bae
Gyeongsang National University
Jinju-si, Gyeongsangnam-do, Republic of Korea
dream_cacao_jh@gnu.ac.kr

Sang-Min Choi
Gyeongsang National University
Jinju-si, Gyeongsangnam-do, Republic of Korea
jerassi@gnu.ac.kr

Suwon Lee*
Gyeongsang National University
Jinju-si, Gyeongsangnam-do, Republic of Korea
leesuwon@gnu.ac.kr

Abstract

Argumentative explainable artificial intelligence employs argumentation theory to explain the mechanisms of machine learning. Previous approaches for explaining deep learning models collectively compressed layers via clustering. However, this resulted in accumulated information loss across layers, thereby degrading the fidelity of explanations. We propose online activation value-aware clustering and aggregation, a compression algorithm that preserves the inference structure of the original neural network with greater fidelity. The proposed method sequentially compresses each layer, immediately recalculates activation values following compression, and rectifies inter-layer information loss using a singular-value-scaled ridge alignment approach. To evaluate the effectiveness of the proposed method, we introduce four novel quantitative metrics. Input–output fidelity and structural fidelity measure how accurately the compressed model preserves the original model predictions and internal activations. Input–output perturbation consistency and structural perturbation consistency assess the similarity of the changes induced by Gaussian-perturbed input data. Experiments on three benchmark datasets (Breast Cancer, California Housing, and HIGGS) demonstrate that our method achieves performance improvements ranging from 12.9% to 53.7% across the four metrics, demonstrating significantly higher explanation fidelity than existing approaches.

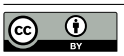
CCS Concepts

• **Computing methodologies** → **Semantic networks**; **Causal reasoning and diagnostics**.

Keywords

Explainable AI (XAI), Argumentative XAI, Model Compression, Singular Value, Aggregation Function, Online Activation Value

*Corresponding author.



CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761362

ACM Reference Format:
Ungsik Kim, Jiho Bae, Sang-Min Choi, and Suwon Lee. 2025. Online Activation Value-aware Clustering and Aggregation for Faithful Argumentative Explanations. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746252.3761362

1 Introduction

Explainable Artificial Intelligence (XAI) aims to make the decision-making processes of deep neural networks interpretable to humans [12, 14]. This is particularly important in high-stakes domains such as healthcare [25], finance [11], and stock prices [13], where model decisions must be understood and justified [7, 27]. To achieve this, various approaches, such as intrinsic methods [2, 17, 33] and post hoc methods [5, 20, 26, 29, 32], have been proposed. However, these approaches often fail to accurately capture the underlying mechanisms of the original models [3].

Argumentative XAI provides more intuitive and clear explanations than traditional XAI methods by mapping a model's internal reasoning process onto argumentation frameworks [9]. In particular, Potyka [23] theoretically demonstrates that Multi-Layer Perceptrons (MLPs) can be interpreted as Quantitative Bipolar Argumentation Frameworks (QBAFs), which quantitatively represent the relationships such as attack and support among arguments [6, 24].

Based on this theoretical foundation, SpArX [3] proposes a prototype model enabling argumentative explanations from tabular data using MLPs, while ProtoArgNet[4] introduced a model based on Sparse MLPs (SMLPs) to produce argumentative explanations from image data. However, prior studies have not adequately explored how to compress reasoning mechanisms while preserving them faithfully in argumentative explanations. Considering that high-quality explanations must simultaneously exhibit high fidelity to the original model and low cognitive complexity to the compressed model, a fundamental trade-off exists between compression ratio and fidelity [3]. Improving compression fidelity and efficiency has the potential to reduce cognitive complexity, making this research essential for advancing argumentative explanations towards practical applications.

We propose Online Activation Value-aware Clustering and Aggregation (OVCA), a novel compression method for building faithful argumentative explanations. Instead of simplifying all layers in one

shot (which do not consider cumulative error [18]), our method processes each layer sequentially and performs an online update of activations after each compression. The online activation values is not only used as input for the subsequent layer but also employed to solve a linear equation involving activation values of the original model, whose solution is then multiplied by the weight matrix. This procedure minimizes information loss and reduces cumulative error; additionally, singular values are incorporated as a form of stabilization to address cases where the linear equation does not yield a direct solution.

There is a lack of reliable metrics to accurately evaluate how faithfully these compression methods explain the original model. SpArX [3] introduces two metrics, input-output and structural unfaithfulness. If these two values of metrics are zero, the compressed model has the same mechanism as the original model. However, these two metrics are sensitive to external factors. For instance, increasing the model or batch size leads to corresponding increases in these metrics. Because of this increase, it is impossible to compare across models or datasets. Furthermore, these two metrics lack discrimination power when comparing across compression methods. We observed that differences in metrics value emerged starting from the sixth decimal place. It could be argued that existing methods sufficiently approximate the mechanism of the original model, resulting in metrics approaching zero. Nevertheless, this phenomenon can be observed when evaluations are conducted with small model or batch sizes, and the variation in metric performance is also minimal under these conditions.

To overcome the limitations of the current metrics, we propose four novel evaluation metrics, designed to address the limitations of existing input-output and structural unfaithfulness measures. Input-Output Fidelity (IOF) and Structural Fidelity (SF) quantitatively evaluate how closely the outputs or activation values of the compressed model align with those of the original model. Input-Output Perturbation Consistency (IOPC) and Structural Perturbation Consistency (SPC) assess fidelity more rigorously by quantitatively measuring how similarly, in both magnitude and direction, the outputs or activation values of the compressed model respond to input feature perturbations induced by Gaussian noise, compared to those of the original model.

IOF and SF can be used in global explanations and local explanations. IOPC and SPC can be used in global explanations, but not in local explanations. The reason is written in definition 12. Global explanations are methods that observe the mechanism through which a model generally makes predictions based on specific criteria and principles when given large amounts of data. In contrast, local explanations analyze the mechanism specifically to clarify why a particular prediction was made for an individual data instance. In other words, global explanations help in understanding the overall behavior of a model, while local explanations enhance detailed understanding and trust regarding individual cases. Our contributions are as follows:

- By proposing a novel compression technique, we minimize performance degradation caused by cumulative error and effectively preserve high fidelity even at higher compression ratios, thus contributing to stable and reliable argumentative explanations suitable for practical applications.
- We propose four quantitative metrics (IOF, SF, IOPC, and SPC) to systematically evaluate both global and local faithfulness of compressed models. These metrics provide a standardized framework for objectively validating and comparing the performance of various compression methodologies in future research.

2 Related Work

2.1 Post-hoc Explanation Methods

Post-hoc XAI aims to interpret models, already trained, by analyzing their inputs, outputs, and internal behavior without modifying the model itself. Popular methods include feature attribution techniques such as LIME [26] and SHAP [20], which highlight influential input features by perturbing input samples or approximating local gradients. Although these methods are widely adopted for their ease of use, they often do not capture the full reasoning process of the model and can be brittle under adversarial or out-of-distribution perturbations [34]. Other post hoc approaches involve surrogate modeling, such as training simpler models (e.g., decision trees) to mimic the predictions of complex models. However, these surrogates rarely reflect the internal structure and can lead to misleading conclusions [19]. Our method can be viewed as a structured surrogate that retains internal semantics and interpretability through principled layerwise aggregation.

2.2 Argumentative XAI

Argumentative explanations aim to transform neural networks into QBAFs [9, 23]. Here, neurons correspond to arguments, and edges between neurons represent relations such as attack or support among arguments. SpArx [3] proposes explaining in tabular data to transform MLPs to QBAFs. ProtoArgNet proposes explaining in image data to transform SMLPs to QBAFs. Recent works discuss how QBAFs explain specific domains in deep learning. On the other side, our work focuses on improving and discussing the compression process, which represents the common underlying logic shared by these recent works.

2.3 Model Compression

Classical compression methods include pruning [15, 18], quantization [1], and knowledge distillation (KD) [16, 21, 21, 36]. These methods aim to reduce model size and inference cost. However, these methods generally ignore mechanism preservation [3]. Classical compression methods prioritize prediction accuracy. We need compression methods that prioritize prediction similarity. For an extreme example, if the accuracy of the original model is 70%, the classical compression methods consider compression successful if the accuracy improves beyond 70%. In contrast, our desired compression method does not aim to surpass this 70% accuracy but rather aims to closely match the original prediction behavior.

2.4 Activation Value-aware SVD

Some works compress the model with Activation Value-aware Singular Value Decomposition (ASVD) [28, 35]. These works demonstrate that the compression with activation distribution is possible and effective. Singular Value Decomposition (SVD) reduces the

number of parameters and computational complexity by approximating a high-rank matrix with a low-rank representation [10], such as $W \approx \tilde{U}\,\Sigma\,\tilde{V}^{\top}$. In our approach, we use only the singular values without computing the matrices U and V. Previous aggregation methods sum all weights directly. In contrast, our aggregation method derives a projection matrix that maps original weights onto compressed weights.

2.5 Fidelity Metrics in XAI

Faithfulness metrics are used to assess how well an explanation reflects the original model. While many XAI works rely on deletion-based metrics such as MoRF or ROAR, these can produce misleading results under distribution shift or input corruption [38]. Moreover, such methods assume monotonicity of importance scores, which may not hold in nonlinear models [34]. Alternatives such as sensitivity analysis and feature perturbation have been proposed to mitigate these issues. Inspired by recent advances in robust fidelity estimation [37], our metrics IOPC and SPC emphasize perturbation-aware fidelity by measuring response similarity under input noise. These metrics provide more reliable estimates of model–explanation alignment, especially for evaluating compressed models across local and global behaviors.

3 Preliminaries

Definition 1, 2, 4, 5, 6, 7 are from SpArX [3].

Definition 1 (Multi-Layer Perceptron (MLP)). *An MLP* $\mathcal{M}$ is a tuple (V, E, B, W, φ). (V, E) is a directed graph. $V = \biguplus_{l=0}^{d+1} V_l$ consists of (ordered) layers of neurons; for $0 \le l \le d+1$, $V_l = \{v_{l,1}, \ldots, v_{l,|V_l|}\}$ is the set of neurons in layer l. We call V_0 the input layer, V_{d+1} the output layer, and V_l (for $1 \le l \le d$) the l-th hidden layer; d is the depth of the MLP. $E \subseteq \bigcup_{l=0}^{d}(V_l \times V_{l+1})$ is a set of edges between adjacent layers; if $E = \bigcup_{l=0}^{d}(V_l \times V_{l+1})$, then the MLP is called fully connected. $B = \{b^1, \ldots, b^{d+1}\}$ is a set of bias vectors, where for $1 \le l \le d+1$, $b^l \in \mathbb{R}^{|V_l|}$. $W = \{W^0, \ldots, W^d\}$ is a set of weight matrices, where for $1 \le l \le d+1$, $W^l \in \mathbb{R}^{|V_{l+1}| \times |V_l|}$, such that $W^l_{j,i} = 0$ when $(v_{l+1,j}, v_{l,i}) \notin E$. $\varphi : \mathbb{R} \to \mathbb{R}$ is an activation function.

Definition 2 (Quantitative Argumentation Framework (QBAF)). A QBAF with domain $\mathcal{D} \subseteq \mathbb{R}$ is a tuple (A, E, B, w) that consists of

- $\mathcal{A} = \{a_1, \ldots, a_{|A|}\}$ is a set of arguments;
- $E \subseteq \mathcal{A} \times \mathcal{A}$ is a set of directed edges;
- $B : \mathcal{A} \to \mathcal{D}$ is a base score;
- $w : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ is a weight function.

Edges with negative/positive weights are called attack and support edges, denoted by Att/Sup, respectively.

To interpret the arguments in an edge weighted QBAF, we considered a modular semantics based on the relationship between QBAFs and MLPs noted earlier to interpret the arguments in an edge weighted QBAF [23].

Definition 3 (Clustering of MLP).

- $C = \{C_1, C_2, \ldots, C_d\}$
- $C_l = \{c_{l,1}, c_{l,2}, \ldots, c_{l,|C_l|}\}$
- $c_{l,i} = \{v^{\mathcal{M}}_{l,j} | 1 \le i \le |V^{\mathcal{M}}_l|, label(v^{\mathcal{M}}_{l,j}) = i\}$

Given MLP $\mathcal{M}$ and compression ratio γ, C is the set of clustering results of all hidden layers. C_l is the clustering result of lth hidden layer, and $|C_l| = max(1, \lfloor \gamma |V^{\mathcal{M}}_l| \rfloor) = |V^{\mu}_l|$. The label function assigns each node to the corresponding cluster label. For example, if we cluster $V^{\mathcal{M}}_1 = \{1, 2, 5, 6, 8, 9, 10, 13, 14, 15\}$ with $\gamma = 0.4$ and clustering results are $C_1 = \{\{1, 2\}, \{5, 6\}, \{8, 9, 10\}, \{13, 14, 15\}\}$, then $label(v^{\mathcal{M}}_{1,1}) = 1, label(v^{\mathcal{M}}_{1,2}) = 1$. Each $c_{l,i}$ is a non-empty set.

Definition 4 (Parameters of Clustered MLP). Given an MLP $\mathcal{M}$, let $(V^{\mathcal{M}}, E^{\mathcal{M}})$ be the graphical structure of the corresponding classical MLP μ. Then for the cluster and edge aggregation functions Agg^b and Agg^e, respectively, μ is

$$(V^{\mu}, E^{\mu}, B^{\mu}, W^{\mu}, \varphi)$$

with parameters B^{μ}, W^{μ} as follows:

- For every cluster-neuron $v_C \in V^{\mu}$, the bias in B^{μ} of v_C is $\mathrm{Agg}^b(C)$;
- For every edge $(v_{C1}, v_{C2}) \in E^{\mu}$, the weight in W^{μ} of the edge is $\mathrm{Agg}^e((C1, C2))$.

Definition 5 (Graphical Structure of Clustered MLP). See Figure 1; Given an MLP $\mathcal{M}$ and a clustering $C = \uplus_{l=1}^{d} C_l$ of $\mathcal{M}$, the graphical structure of the corresponding clustered MLP μ is a directed graph (V^{μ}, E^{μ}) with:

$$V^{\mu} = \bigoplus_{l=0}^{d+1} V^{\mu}_l$$

Comprising (ordered) layers of cluster-neurons such that:

- The input layer V^{μ}_0 consists of a singleton cluster-neuron $v_{\{0,i\}}$ for every input neuron $v_i \in V_0$.
- The l-th hidden layer of μ (for $0 < l < d+1$) consists of a cluster-neurons c_l.
- The output layer V^{μ}_{d+1} consists of a singleton cluster-neuron $v_{\{d+1,j\}}$ for every output neuron $v_{d+1,j} \in V_{d+1}$.
- $E^{\mu} = \bigcup_{l=0}^{d}(V^{\mu}_l \times V^{\mu}_{l+1})$.

Definition 6 (Structural Unfaithfulness). The local structural unfaithfulness of μ to $\mathcal{M}$ with respect to input x and dataset Δ is:

$$\mathcal{L}^{\mathcal{M}}_s = \sum_{x' \in \Delta} \pi_{x',x} \sum_{l=1}^{d+1} \sum_{j=1}^{|C_l|} \sum_{v_{l,i} \in c_{l,j}} (O^{\mathcal{M}}_{l,i}(x') - O^{\mu}_{l,i}(x'))^2.$$

The global structural unfaithfulness is defined analogously by removing the similarity terms $\pi_{x',x}$.

Definition 7 (Cognitive Complexity).

$$\Omega(\mu) = \prod_{0 < l \le d+1} |V^{\mu}_l|.$$

A larger $\Omega(\mu)$ suggests a more complex (hence less interpretable) model. Therefore, balancing fidelity (IOF, SF, IOPC, SPC) against cognitive complexity $\Omega(\mu)$ is thus crucial for creating explanations that are both accurate and comprehensible.

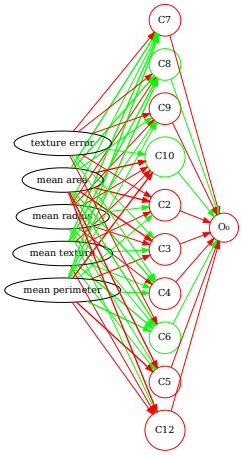


Figure 1: The positive and negative weight relationships between neuron clusters (C2–C12) and the output node (O_0) as an argumentation graph, after compressing a MLP trained on the Breast Cancer dataset by 80% with the proposed OVCA algorithm.

4 Proposed Method

Figure 2 is an overview of our method. We consider activation value from both offline and online perspectives. In a scenario where layer 1 is compressed and layer 2 needs to be compressed subsequently, clustering utilizes activation values obtained from layer 2. However, the activation values obtained from the original model differ from those obtained from a model where layer 1 is compressed. We call the activation value obtained from the original model as offline activation value and the recalculated activation value as online activation values. A key difference between offline and online activation values is the consideration of clustering information from the previous layer. We perform clustering and aggregation with online activation values, thus, we minimize the information loss and accumulated error.

Algorithm 1 is a four-step pseudo code for local explanation. The reason for sampling Δ' is that it is difficult for the results of clustering and aggregation to be stable with only one piece of data [3]. The PERTURB function generates data by adding some noise to the data x. Let pi be a weight variable that is based on the distance between x and x'. We use this pi to cluster. Step 2 is the clustering based on the weights pi and the activation values $A^{\mathcal{M}_l}$. Steps 3 and 4 are the aggregation process based on the clustering results. We used k-means as our clustering algorithm. Steps 3 and 4 are the aggregation process based on the clustering results. In particular, the MERGENODES function in step 3 is equivalent to (i), (ii), and step 4 is (iii) in definition 9. Global explanation first uses the dataset Δ instead of the data x_0. Therefore, it doesn't calculate π and doesn't utilize pi during clustering and aggregation. Other than that, the behavior is the same.

Definition 8 (Global Aggregation Functions)**.**

(i) **Bias aggregation**

$$\mathrm{Agg}_b(C_l) = \{b^{\mu}_{l,i} | b^{\mu}_{l,i} = \frac{1}{|c_{l,i}|} \sum_{v^{\mathcal{M}}_{l,j} \in c_{l,i}} b^{\mathcal{M}}_{l,j}, c_{l,i} \in C_l, \}$$

(ii) **Incoming-weight aggregation**

$$\mathrm{Agg}^{e}_{\mathrm{in}}((V^{\mu}_{l-1}, C_l)) = \{e^{\mu}_{l-1,i,j} \mid e^{\mu}_{l-1,i,j} = \frac{1}{|c_{l,j}|} \sum_{v_{l,k} \in c_{l,j}} W^{l-1}_{k,i},$$

$$v^{\mu}_{l-1,i} \in V^{\mu}_{l-1}, c_{l,j} \in C_l\}$$

(iii) **Outgoing-weight aggregation**

$$\mathrm{Agg}^{\mathrm{e}}_{\mathrm{out}}((C_l, V^{\mathcal{M}}_{l+1})) = \{e^{\mu}_{l,i,j} \mid e^{\mu}_{l,i,j} = W^{l}_{(j,:)} \cdot \theta_{(:,i)},$$

$$c_{l,i} \in C_l,\ v^{\mathcal{M}}_{l+1,j} \in V^{\mathcal{M}}_{l+1}\}$$

The aggregation function for edges operates in two stages. First performing Agg^{e}_{in} function, and then Agg^{e}_{out} function. After processing Agg^{e}_{in}, weights between V^{l-1} and V^{l} are compressed, enabling the recalculation of activation values. $A^{\mu}_{l} \in \mathbb{R}^{|C_l|}$ denote the recalculated online activation values. $A^{\mathcal{M}}_{l} \in \mathbb{R}^{|V^{\mathcal{M}}_{l}|}$ denote the original model's activation values. The main idea is to solve a linear system between A^{μ}_{l} and $A^{\mathcal{M}}_{l}$ to preserve the original's information in a compressed state. $\theta = \left(A^{\mu\top}_{l} A^{\mu}_{l}\right)^{-1} A^{\mu\top}_{l} A^{\mathcal{M}}_{l}$ denotes the solution to the linear system.

However, if A^{μ}_{l} is rank-deficient, the linear system may have no solution (ill-posed) or may not have a unique solution. To maintain the full-rank condition of the online activation values, we add a regularization matrix. The regularization matrix is λI. λ is the product of a hyperparameter and the maximum singular value. Additionally, this method prevents the compressed weights or model from becoming ill-conditioned which amplify changes in the output in response to small changes in the input.

Theorem 1. *The linear system* $(A^{\mu\top}A^{\mu} + \lambda I)\theta = A^{\mu\top}_{l} A^{\mathcal{M}}_{l}$ *has a unique solution* θ.

Proof. Suppose $\lambda > 0$. Let A^{μ} be the compressed activation value matrix and $A^{\mathcal{M}}$ be the original activation value matrix. We show that the solution of the linear system always exists and is unique.

For any vector x, the following holds:

$$x^{\top}(A^{\mu\top}A^{\mu})x = (A^{\mu}x)^{\top}(A^{\mu}x) = \|A^{\mu}x\|^2_2 \geq 0.$$

Thus, the matrix $A^{\mu\top}A^{\mu}$ is positive semi-definite, and all eigen values σ^2_i of $A^{\mu\top}A^{\mu}$ satisfy:

$$\sigma^2_i \geq 0.$$

Since $\lambda > 0$, the eigenvalues of the matrix λI are all equal to λ, which are strictly positive. Therefore, the eigen values of the matrix $A^{\mu\top}A^{\mu} + \lambda I$ are:

$$\sigma^2_i + \lambda > 0.$$

Hence, the matrix $A^{\mu\top}A^{\mu} + \lambda I$ is positive definite, implying that for any $x \neq 0$,

$$x^{\top}(A^{\mu\top}A^{\mu} + \lambda I)x > 0.$$

This means that the matrix $(A^{\mu\top}A^{\mu} + \lambda I)$ is full rank and invertible (i.e., $(A^{\mu\top}A^{\mu} + \lambda I)^{-1}$ always exists).

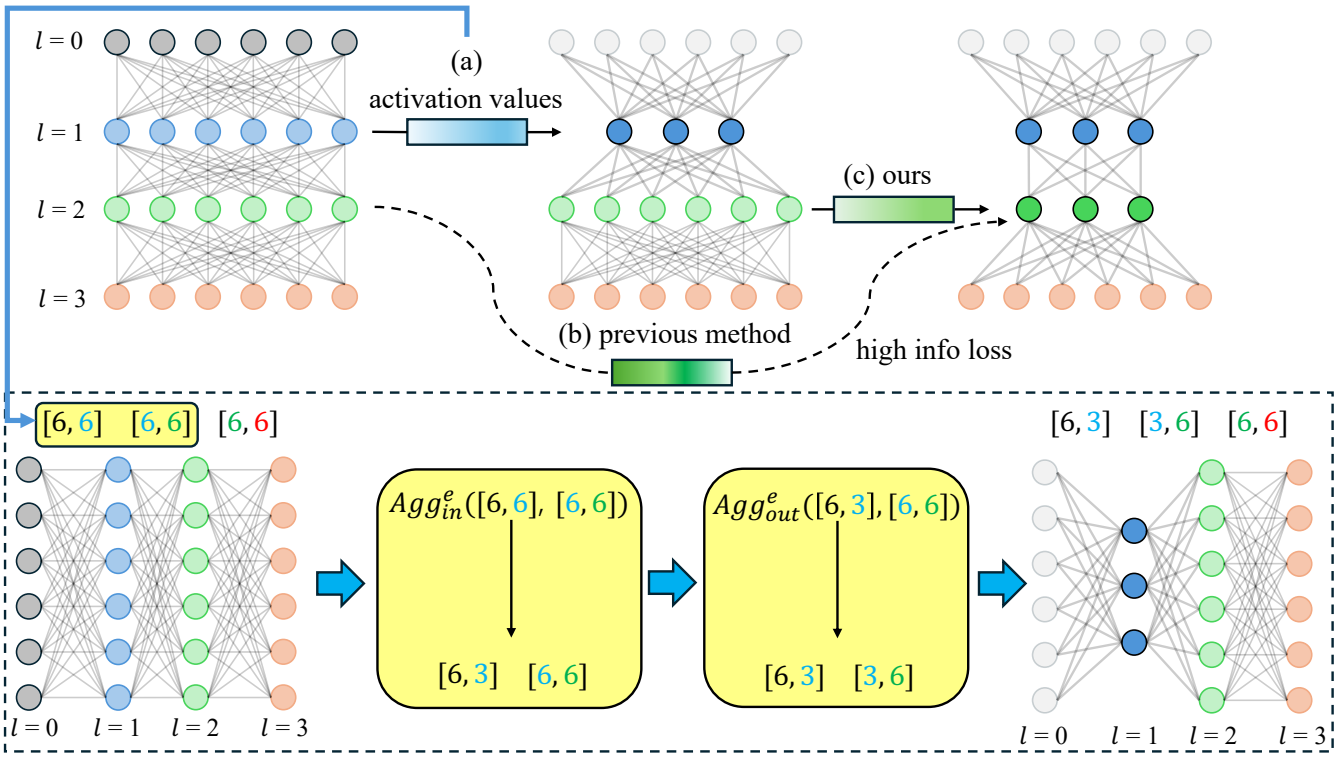


Figure 2: An overview of the proposed method; (a) After determining how to combine nodes by clustering, our method extracts activation values through the forward of MLP, and compresses layer $l = 1$ by 50% through Agg^e_{in} and Agg^e_{out}.(See Definition 8); (b) Prior method use activation values from MLP models for compression, which breaks the connectivity between layers and causes high information loss. (The previous method does not compress with method (a)); (c) The proposed method extracts and clusters the activation values of the compressed model from $l = 2$ to $l = 1$ instead of the original model at $l = 2$.

Therefore, the linear equation

$$(A^{\mu\top}A^{\mu} + \lambda I)x = A^{\mu\top}A^{\mathcal{M}}$$

always has a unique solution. □

If all activation values in a layer are negative, then λ can be less than or equal to zero. However, this scenario is uncommon and realistically does not occur. Nevertheless, to mitigate these exception cases, it is possible to create a ridge regression matrix by squaring the singular values or to create a lasso regression matrix by taking the absolute values. However, we did not experiment with ridge or lasso methods, since ReLU activation does not produce negative values, rendering these regularization approaches meaningless. Despite this approach, it is impossible to compute if the activation values of a layer are all zero. However, this is an unusual result that must not occur during the training process, regardless of whether the model is compressed or not, so it is not considered a typical case and is not considered in this study.

Definition 9 (Local Aggregation Functions). Fix an input x and its perturbed neighborhood $\Delta' = \{x'_1, \ldots, x'_m\}$ with weights $\pi_{x'} = \exp(-\|x' - x\|^2/\sigma^2)$.

(i) **Bias aggregation**

$$\mathrm{Agg}_b(C_l) = \{b^{\mu}_{l,i} \mid b^{\mu}_{l,i} = \frac{1}{|c_{l,i}|} \sum_{v^{\mathcal{M}}_{l,j} \in c_{l,i}} b^{\mathcal{M}}_{l,j}, c_{l,i} \in C_l\}$$

(ii) **Incoming-weight aggregation**

$$\mathrm{Agg}_{\mathrm{in}}((V^{\mu}_{l-1}, C_l)) = \{e^{\mu}_{i,j} \mid e^{\mu}_{i,j} = \frac{\pi_{x'}}{|c_{l,j}|} \sum_{v_{l,k} \in c_{l,j}} W^{l-1}_{k,i},$$
$$v^{\mu}_{l-1,i} \in V^{\mu}_{l-1}, c_{l,j} \in C_l\}$$

(iii) **Outgoing-weight aggregation**

$$\mathrm{Agg}^{\mathrm{e}}_{\mathrm{out}}((C_l, V^{\mathcal{M}}_{l+1})) = \{e^{\mu}_{l,i,j} \mid e^{\mu}_{l,i,j} = W^{l}_{(j,:)} \cdot \theta_{(:,i)},$$
$$c_{l,i} \in C_l,\ v^{\mathcal{M}}_{l+1,j} \in V^{\mathcal{M}}_{l+1}\}$$

5 Evaluation Metric

We measure how the compressed model μ closely reproduces the original model $\mathcal{M}$ using four metrics. All four metrics use the following function: $score(x) = \frac{1}{1+x}$, where higher values indicate greater alignment with the original model. $\varepsilon > 0$ prevents division by zero.

Algorithm 1 OVCA for local explanation

Require: trained MLP $\mathcal{M}$ with layers V, preserve ratio $\gamma \in (0, 1]$, neighborhood size N, noise scale σ_{noise}, kernel width σ, and ridge λ
Ensure: compressed model μ
$\mu \leftarrow \mathcal{M}$ ▷ initial copy
$x_0 \leftarrow$ target input
(1) Local sample generation
$\Delta' \leftarrow \textsc{Perturb}(x_0, N, \sigma_{\text{noise}})$
append x_0 to Δ'
$\pi \leftarrow \exp(-\|\Delta' - x_0\|^2/\sigma^2)$
for $l = 1$ **to** d **do** ▷ iterate hidden layers
(2) π-weighted clustering
$A_l^{\mathcal{M}} \leftarrow O_l^{\mathcal{M}}(\Delta')$
$|C_l| \leftarrow \max(1, \lfloor \gamma\, |V_l| \rfloor)$
$C_l \leftarrow \textsc{Clustering}(A_l, |C_l|, sample\_weight = \pi)$
(3) Merge neurons
$(W_l^{\mu}, b_l^{\mu}) \leftarrow \textsc{MergeNodes}(W^l, b^l, C_l, \pi)$
Replace layer l of μ with (W_l^{μ}, b_l^{μ})
(4) Ridge Alignment
if $l < d$ **then**
$A_l^{\mu} \leftarrow O_l^{\mu}(\Delta')$
$\theta = \left(A_l^{\mu\top} A_l^{\mu} + \lambda I\right)^{-1} A_l^{\mu\top} A_l^{\mathcal{M}}$
$W^{l+1} \leftarrow \theta W^{l+1}$ ▷ update layer l+1 in μ
end if
end for
return μ

Definition 10 (Input-Output Fidelity; IOF)**.**

$$IOF = Score\left(\frac{(O_{d+1}^{\mathcal{M}}(\Delta) - O_{d+1}^{\mu}(\Delta))^2}{(O_{d+1}^{\mathcal{M}}(\Delta))^2 + \varepsilon}\right)$$

If $IOF = 1$, the compressed model perfectly reproduces the output probabilities of the original model. This metric is similar to input-output unfaithfulness [3].

Definition 11 (Structural Fidelity; SF)**.**

$$SF = Score\left(\frac{\sum_{(l,i,j)\in\mathcal{S}} \left(O_{l,j}^{\mathcal{M}}(\Delta) - O_{l,i}^{\mu}(\Delta)\right)^2}{\sum_{(l,i,j)\in\mathcal{S}} (O_{l,j}^{\mathcal{M}}(\Delta))^2 + \varepsilon}\right).$$

If SF = 1, the compressed model perfectly reproduces the hidden activation values of the hidden layers in the original model. $\mathcal{S} = \left\{(l, i, j) \mid C_l \in C,\ c_{l,i} \in C_l,\ v_{l,j}^{\mathcal{M}} \in c_{l,i}\right\}$

Definition 12 (Input-Output Perturbation Consistency; IOPC)**.** Given a base input x and a perturbed input $x' = x + \delta$, let

$$IOPC = Score\left(\frac{(\delta^{\mathcal{M}} - \delta^{\mu})^2}{(\delta^{\mathcal{M}})^2 + \varepsilon}\right).$$

If IOPC = 1, the compressed model perfectly tracks the changes in output probabilities of the original model induced by input perturbations [22, 34]. $\delta^{\mathcal{M}} := O_{d+1}^{\mathcal{M}}(\Delta) - O_{d+1}^{\mathcal{M}}(\Delta')$ be the change of ouptut of the original model. $\delta^{\mu} := O_{d+1}^{\mu}(\Delta) - O_{d+1}^{\mu}(\Delta')$ be the change of output of the compressed model. Achieving a high score on the IOPC metric requires accurately matching both the direction and magnitude of output changes in the original model. If both the original and compressed models initially output 0.8, and the perturbed input causes the output of the original model to increase to 0.9 while the output of the compressed model decreases to 0.7, the IOPC score becomes significantly low. This example illustrates a case where the compressed model fails to accurately track the mechanism of the original model. However, IOPC is not appropriate for assessing local explanations. Requiring consistency under perturbation is less meaningful since the local explanation needs only to be faithful to the original model's behavior on that specific input.

Definition 13 (Structural Perturbation Consistency; SPC)**.**

$$SPC = Score\left(\frac{\sum_{(l,i,j)\in\mathcal{S}} \left(\delta_{l,j}^{\mathcal{M}} - \delta_{l,i}^{\mu}\right)^2}{\sum_{(l,i,j)\in\mathcal{S}} \left(\delta_{l,j}^{\mathcal{M}}\right)^2 + \varepsilon}\right).$$

If SPC = 1, the compressed model perfectly tracks the changes in hidden activation values of hidden layers in the original model induced by input perturbations. $\mathcal{S}$ is identical to that in definition 11. $\delta_{l,j}^{\mathcal{M}}$ and $\delta_{l,i}^{\mu}$ is identical to that in definition 12.

We use perturbation-based metrics for IOPC and SPC, not deletion-based metrics. Widely used deletion-based metrics such as MoRF, ROAR, and $Fid_{+/-}$ may yield misleading results due to their tendency to produce out-of-distribution (OOD) inputs when masking important features. As Zheng et al. [38] point out, this OOD behavior can lead the original model $\mathcal{M}$ to behave erratically, thus corrupting fidelity measurements.

6 Experiments

We used three complementary datasets: Breast Cancer [31], HIGGS [30], and California Housing.

This dataset selection reflects a deliberate balance across domain diversity, data scale, class structure, and task type. The Breast Cancer dataset is a small-scale binary classification task (569 samples) that can be effectively modeled using a MLP. As such, it serves as a suitable benchmark for evaluating whether argument-based explanations can accurately reflect a well-performing MLP. California Housing is a medium-difficulty regression task, allowing us to assess whether the proposed approach generalizes beyond classification. In contrast, the HIGGS dataset, a large-scale multi-class classification task with 581,000 samples, is known to be challenging for MLPs. We believe that explanation methods should remain informative even when model performance is suboptimal, particularly by offering insight into prediction failures. HIGGS thus provides a valuable test case for such scenarios.

6.1 Global Explanation

Table 1 is global explanation experiments. In terms of IOF, OVCA consistently achieved higher scores across all dataset-model combinations. Notably, in the shallow Breast Cancer-S model (L1 H64), IOF improved from 0.6486 to 0.9966—a remarkable 53.7 percentage point increase. This result indicates that the linear system method is more effective than the sum method used in prior studies.

Table 1: Global explanation results at 80% compression on the Breast Cancer, HIGGS, and California Housing datasets for model sizes S, M, and L (L: layers; H: hidden nodes). Δ(%): relative improvement ((Ours−Original)/Original×100). Gaussian noise (p=0.05) applied for IOPC and SPC.

Dataset / Model	Method	IOF↑	ΔIOF(%)	SF↑	ΔSF(%)	IOPC(p=0.05)↑	ΔIOPC(%)	SPC(p=0.05)↑	ΔSPC(%)
Breast Cancer									
S (L1 H64)	Original	0.6486	–	**0.9947**	–	0.6775	–	0.5801	–
	Ours	**0.9966**	53.7↑	0.9944	0.0	**0.8160**	20.4↑	**0.6196**	6.8↑
M (L3 H128)	Original	0.9280	–	0.9956	–	0.7498	–	0.6661	–
	Ours	**0.9997**	7.7↑	**0.9965**	0.1↑	**0.9420**	25.6↑	**0.7003**	5.1↑
L (L5 H256)	Original	0.9884	–	0.9978	–	0.9235	–	0.7979	–
	Ours	**0.9998**	1.2↑	**0.9979**	0.0	**0.9491**	2.8↑	**0.8005**	0.3↑
HIGGS									
S (L3 H128)	Original	0.8676	–	0.6615	–	**0.5356**	–	0.5387	–
	Ours	**0.8930**	2.9↑	**0.6745**	2.0↑	0.5263	1.7↓	**0.5403**	0.3↑
M (L5 H256)	Original	0.8928	–	0.7234	–	0.5461	–	0.5648	–
	Ours	**0.9429**	5.6↑	**0.7897**	9.2↑	**0.6070**	11.1↑	**0.6032**	6.8↑
L (L7 H512)	Original	0.9417	–	0.7726	–	0.6162	–	0.6153	–
	Ours	**0.9717**	3.2↑	**0.8719**	12.9↑	**0.6947**	12.7↑	**0.6908**	12.3↑
California									
S (L2 H64)	Original	0.8446	–	0.7973	–	0.5408	–	0.5561	–
	Ours	**0.9460**	12.0↑	**0.8240**	3.3↑	**0.5869**	8.5↑	**0.5722**	2.9↑
M (L3 H128)	Original	0.8365	–	0.7873	–	0.5171	–	0.5295	–
	Ours	**0.9658**	15.5↑	**0.8647**	9.8↑	**0.6189**	19.7↑	**0.5954**	12.5↑
L (L7 H256)	Original	0.9427	–	0.8555	–	0.6514	–	0.6344	–
	Ours	**0.9924**	5.3↑	**0.9072**	6.0↑	**0.9005**	38.2↑	**0.8469**	33.5↑

In terms of SF, the benefits of OVCA were more pronounced for deeper networks. For instance, in HIGGS (L7 H512), SF increased from 0.7726 to 0.8719 (+12.9%). Similarly, in California Housing (L3 H128), SF rose from 0.7873 to 0.8647 (+9.8%). These results suggest that OVCA effectively mitigates error propagation between layers, leading to better preservation of internal activation patterns. In the breast cancer dataset, there is no difference between the previous method and the proposed method. This may explain why structural unfaithfulness-based SF does not effectively differentiate between the methods.

In terms of SPC (p = 0.05), we evaluated them by injecting Gaussian noise ($p = 0.05$). OVCA outperformed the original method across datasets. The improvement was smaller for the Breast Cancer dataset compared to others. Nevertheless, OVCA structurally mitigated information loss clearly. This indicates that SPC is more discriminative compared to the previous SF results. However, as SPC and SF measure inherently different aspects, assessing with both metrics is important.

In terms of IOPC (p = 0.05), OVCA generally achieved higher performance across most cases, exhibiting substantial improvements. However, in the HIGGS-S, the performance is decreased. This result highlights that performance improvements were inconsistent across datasets, indicating potential limitations of the proposed perturbation metrics. Nevertheless, these findings suggest that OVCA better maintains alignment with the original model behavior compared to the original method.

Figure 3 visualizes the layer-wise structure unfaithfulness for HIGGS and California Housing experiments. While the prior method leads to linearly increasing error with depth, OVCA keeps per-layer errors suppressed below 0.01, resulting in a much flatter accumulation curve. This supports the theoretical claim that repeating the compress-align-transfer procedure at each layer enables immediate error correction.

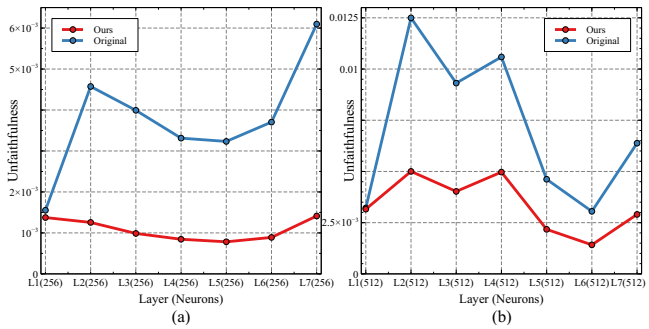


Figure 3: Comparison of the layer-wise structural unfaithfulness between the proposed method and the orginal method. (a) L7 H256 model in California Housing dataset. (b) L7 H512 model in HIGGS dataset.

Figure 4 shows changes in performance per compression ratio. We set the compression ratio from 1% to 90%. The reason for this

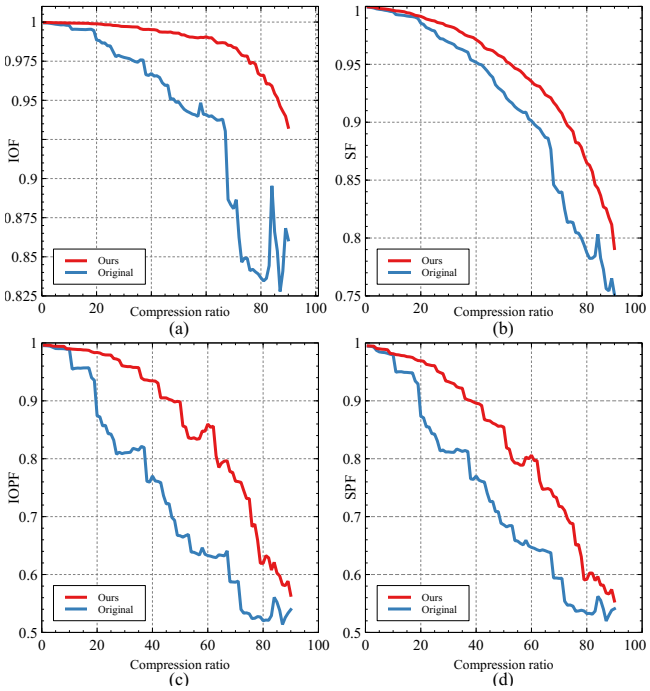


Figure 4: Experimental results for the L3 H128 model on the California Housing dataset across compression rates from 1% to 90% for each metric. (a) IOF, (b) SF, (c) IOPC with a Gaussian noise perturbation of 0.05, and (d) SPC with the same perturbation.

is a vibration of performance values, and this makes interpretation difficult. The vibration of performance values occurs when the compressed model does not have a mechanism like the original model. The vibration of the prior method begins from about 70% compression ratio. This means that the prior method fails to preserve information of the original model from compression ratios exceeding about 70%. Indeed, from compression ratios exceeding 90%, the vibration of performance values occurs in the proposed method, resulting in cases where the performance value of the prior method is higher than the proposed method. Apart from these specific interpretations, in terms of a general perspective, the proposed method preserves performance as the compression ratio increases.

Table 2 presents ablation studies that assess the contribution of each OVCA component. IOF and SF perform better without λ, but the difference is not significant. IOPC and SPC are much better with λ, with Adaptive λ using singular value performing best. Interpreting this in terms of bias and variance, we can see that the addition of regularization increased the bias, resulting in a small drop in IOF and SF, but decreased the variance, resulting in a large increase in performance on perturbation-based metrics. Furthermore, applying strong or weak regularization yields limited improvement, while adjusting the regularization strength per layer based on the singular values of the activation outputs yields significantly better performance. The only difference between using the sum method and using the linear system is recalculating the activation. In this case, the prior method performed slightly better.

Table 2: Ablation study of L7 H256 model on the California Housing dataset for global explanations (mean ± standard deviation across five random seeds).

Variant	IOF	SF
Full	0.9925 ± 0.0003	0.9094 ± 0.0022
w/o Adaptive λ	0.9927 ± 0.0002	**0.9095** ± 0.0022
w/o λ	**0.9928** ± 0.0002	**0.9095** ± 0.0022
w/o Alignment	0.9396 ± 0.0136	0.8551 ± 0.0065
Original	0.9437 ± 0.0069	0.8587 ± 0.0065

Variant	IOPC	SPC
Full	**0.9088** ± 0.0149	**0.8564** ± 0.0138
w/o Adaptive λ	0.8881 ± 0.0065	0.8557 ± 0.0133
w/o λ	0.8357 ± 0.0086	0.8545 ± 0.0129
w/o Alignment	0.6394 ± 0.0383	0.6249 ± 0.0326
Original	0.6521 ± 0.0175	0.6333 ± 0.0237

6.2 Local Explanation

Table 3 is the performance evaluation of our proposed method for local explanations. In previous work, data x was sampled to create Δ' and compressed using distance-based weighting. We compare our proposed method with prior methods in two cases, one with no weighting and the other with no sampling at all. When weighted, our proposed method consistently performs better. Even without weighting, the performance numbers are the same with and without weighting. This is due to rounding to the nearest 5 decimal places for the change in performance with and without weighting. Although the difference is too minimal, weighting performed slightly better. Surprisingly, the prior method outperformed the proposed one when no sampling is done.

6.3 Time Usage

Table 4 measures the execution time of the existing and proposed methods. In global explanations, the time was measured using the dataset Δ used to train the model, and for local explanations, the time was measured by sampling data x and creating 100 dataset Δ'. Compared to the original method, the runtime of proposed method increased by up to 19% for global explanations. Local explanation increased by up to 8%. Since the clustering operation is the bottleneck in both methods, the time complexity of both methods is the same in big-O notation. However, the recalculation of the activation value adds $O(Nl|V_l^\mu|)$ to the execution time.

7 Conclusion

In this paper, we proposed a novel compression method (OVCA) using online activation values. To address shortcomings of prior evaluation metrics (input-output unfaithfulness, structural unfaithfulness), we introduced 2 metrics (IOF, SF) robust across different model sizes and batch sizes, and 2 metrics (IOPC, SPC) enabling more rigorous assessments.

Table 3: Local explanation results at 80% compression on Breast Cancer, HIGGS, and California Housing datasets for model sizes S, M, and L (L: layers; H: hidden nodes). "w/o W": clustering without distance-based weighting; "w/o S": clustering and aggregation on a single input only. Δ(%): ((Ours–Original)/Original×100).

Dataset / Model	Method	IO	ΔIO (%)	w/o W	ΔIO (%)	w/o S	ΔIO (%)	SF	ΔSF (%)	w/o W	ΔST (%)	w/o S	ΔSF (%)
Breast Cancer													
S (L1 H64)	Original	0.9997	–	0.9997	–	**0.9998**	–	0.9963	–	0.9963	–	**0.9965**	–
	Ours	**0.9999**	0.02↑	**0.9999**	0.02↑	0.9997	0.00	0.9963	0.00	0.9963	0.00	0.9962	0.02↓
M (L3 H128)	Original	1.0000	–	1.0000	–	0.9999	–	0.9965	–	0.9965	–	**0.9962**	–
	Ours	1.0000	0.00	1.0000	0.00	0.9999	0.00	**0.9967**	0.02↑	**0.9967**	0.02↑	0.9961	0.01↓
L (L5 H256)	Original	1.0000	–	1.0000	–	**0.9999**	–	0.9969	–	0.9969	–	**0.9964**	–
	Ours	1.0000	0.00	1.0000	0.00	0.9998	0.01↓	**0.9972**	0.02↑	**0.9972**	0.02↑	0.9960	0.03↓
HIGGS													
S (L3 H128)	Original	0.9944	–	0.9944	–	**0.9856**	–	0.9772	–	0.9772	–	**0.9678**	–
	Ours	**0.9987**	0.42↑	**0.9987**	0.42↑	0.9800	0.57↓	**0.9874**	1.03↑	**0.9874**	1.03↑	0.9533	1.49↓
M (L5 H256)	Original	0.9966	–	0.9966	–	**0.9868**	–	0.9841	–	0.9841	–	**0.9628**	–
	Ours	**1.0000**	0.33↑	**1.0000**	0.33↑	0.9839	0.29↓	**0.9934**	0.94↑	**0.9934**	0.94↑	0.9485	1.48↓
L (L7 H512)	Original	0.9973	–	0.9973	–	**0.9851**	–	0.9864	–	0.9864	–	**0.9566**	–
	Ours	**1.0000**	0.26↑	**1.0000**	0.26↑	0.9653	2.01↓	**0.9953**	0.90↑	**0.9953**	0.90↑	0.9157	4.27↓
California													
S (L2 H64)	Original	0.9570	–	0.9570	–	**0.9501**	–	0.9787	–	0.9787	–	**0.9663**	–
	Ours	**0.9915**	3.60↑	**0.9915**	3.60↑	0.9336	1.74↓	**0.9807**	0.21↑	**0.9807**	0.21↑	0.9621	0.43↓
M (L3 H128)	Original	0.9970	–	0.9970	–	**0.9871**	–	0.9851	–	0.9851	–	**0.9700**	–
	Ours	**0.9996**	0.26↑	**0.9996**	0.26↑	0.9858	0.12↓	**0.9899**	0.48↑	**0.9899**	0.48↑	0.9645	0.57↓
L (L7 H256)	Original	0.9993	–	0.9994	–	**0.9922**	–	0.9947	–	0.9950	–	**0.9810**	–
	Ours	**1.0000**	0.06↑	**1.0000**	0.06↑	0.9899	0.23↓	**0.9964**	0.16↑	**0.9964**	0.14↑	0.9724	0.87↓

Table 4: Comparison of execution times between the original and proposed method (mean ± std, unit: ms). Experimental setup includes fixed threads/cores, warm-up runs, and randomized execution order (15 iterations).

Global Explanations			
	Original (ms)	OVCA (ms)	relative
cancer	4.0 ± 0.2	4.3 ± 0.2	1.08
housing	476.1 ± 10.6	566.8 ± 28.2	1.19
HIGGS	19 390.8 ± 116.1	20 325.5 ± 93.7	1.05

Local Explanations			
	Original (ms)	OVCA (ms)	relative
cancer	3.55 ± 0.4	3.57 ± 0.1	1.00
housing	18.6 ± 1.2	19.2 ± 0.5	1.03
HIGGS	18.0 ± 0.5	19.5 ± 0.5	1.08

Overall, OVCA outperformed the original method across all four proposed metrics. In global explanation, performance improvements of up to 53.7% in IOF, 12.9% in SF, 38.2% in IOPC and 33.5% in SPC were achieved, and Local explanation saw improvements of up to 3.6% in IOF and 1.03% in SF, clearly demonstrating that our proposed method enhances the fidelity of compressed models compared to existing methods. This suggests that it is possible to generate argumentative explanations that are faithful to existing models with lower cognitive complexity. Moreover, the processing time remained below 20%, suggesting that the proposed approach can reliably provide argumentative XAI in the target application domain.

A limitation of this study is that the effectiveness of the proposed method has been only validated exclusively with tabular data. Recent work, such as ProtoArgNet[4], based on the ProtoPNet[8], demonstrates that argumentative Explanation is feasible for image classification tasks as well. Therefore, future research should validate our proposed method across diverse tasks, including image classification tasks like ProtoArgNet. Also, in the local explanations experiment, the performance of compression without sampling, i.e., with only a single input, was rather poor. The reason is that the size of the matrix is too small to preserve the information of the original model with a linear system. Solving this problem to improve performance without local sampling is key work to speed up explanation generation in the future. Additionally, since IOPC and SPC are metrics not applicable for local explanation, there remains a shortage of suitable metrics for evaluating compression methods in local explanation. Future research should strive to introduce more metrics tailored specifically for argumentative XAI. We hope that this work lays foundational groundwork for these efforts and broadly contributes to the advancement and growth of the argumentative XAI ecosystem.

8 Generative AI Disclosure

We did not use any generative AI tools beyond standard spelling, grammar checking functions that are exempt from disclosure under the ACM Authorship Policy.

References

[1] Yamato Arai and Yuma Ichikawa. 2025. Quantization Error Propagation: Revisiting Layer-Wise Post-Training Quantization. *arXiv preprint arXiv:2504.09629* (2025).

[2] Sercan Ö Arik and Tomas Pfister. 2021. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 6679–6687.

[3] Hamed Ayoobi, Nico Potyka, and Francesca Toni. 2023. Sparx: Sparse argumentative explanations for neural networks. In *ECAI 2023*. IOS Press, 149–156.

[4] Hamed Ayoobi, Nico Potyka, and Francesca Toni. 2025. ProtoArgNet: Interpretable Image Classification with Super-Prototypes and Argumentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 1791–1799.

[5] Oren Barkan, Yonatan Toib, Yehonatan Elisha, and Noam Koenigstein. 2024. A Learning-based Approach for Explaining Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 98–108.

[6] Pietro Baroni, Marco Romano, Francesca Toni, Marco Aurisicchio, and Giorgio Bertanza. 2015. Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation* 6, 1 (2015), 24–49.

[7] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.

[8] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems* 32 (2019).

[9] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. 2021. Argumentative XAI: a survey. *arXiv preprint arXiv:2105.11266* (2021).

[10] James W Demmel. 1997. *Applied numerical linear algebra*. SIAM.

[11] Jean Dessain, Nora Bentaleb, and Fabien Vinas. 2023. Cost of explainability in ai: An example with credit scoring models. In *World Conference on Explainable Artificial Intelligence*. Springer, 498–516.

[12] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[13] Kelvin Du, Rui Mao, Frank Xing, and Erik Cambria. 2024. Explainable stock price movement prediction using contrastive learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 529–537.

[14] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[15] Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems* 28 (2015).

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[17] Serdar Kadioglu, Elton Yechao Zhu, Gili Rosenberg, John Kyle Brubaker, Martin JA Schuetz, Grant Salton, Zhihuai Zhu, and Helmut G Katzgraber. 2025. BoolXAI: Explainable AI Using Expressive Boolean Formulas. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 28900–28906.

[18] Gui Ling, Ziyang Wang, and Qingwen Liu. 2024. SlimGPT: Layer-wise Structured Pruning for Large Language Models. *Advances in Neural Information Processing Systems* 37 (2024), 107112–107137.

[19] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[20] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[21] Amin Parchami-Araghi, Moritz Böhle, Sukrut Rao, and Bernt Schiele. 2024. Good teachers explain: Explanation-enhanced knowledge distillation. In *European Conference on Computer Vision*. Springer, 293–310.

[22] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421* (2018).

[23] Nico Potyka. 2021. Interpreting neural networks as quantitative argumentation frameworks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6463–6470.

[24] Antonio Rago, Francesca Toni, Marco Aurisicchio, Pietro Baroni, et al. 2016. Discontinuity-Free Decision Support with Quantitative Argumentation Debates. *KR* 16 (2016), 63–73.

[25] Md Mahmudur Rahman and Sanjay Purushotham. 2022. Fair and interpretable models for survival analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1452–1462.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[27] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[28] Charbel Sakr and Brucek Khailany. 2024. Espace: Dimensionality reduction of activations for model compression. *arXiv preprint arXiv:2410.05437* (2024).

[29] Alisa Smirnova, Jie Yang, and Philippe Cudre-Mauroux. 2024. XCrowd: Combining Explainability and Crowdsourcing to Diagnose Models in Relation Extraction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2097–2107.

[30] Daniel Whiteson. 2014. HIGGS. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5V312.

[31] William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. 1993. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DW2B.

[32] Akihiro Yamaguchi, Ken Ueno, Ryusei Shingaki, and Hisashi Kashima. 2024. Learning Counterfactual Explanations with Intervals for Time-series Classification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4158–4162.

[33] Yang Yang, Wendi Ren, and Shuang Li. 2024. Hyperlogic: Enhancing diversity and accuracy in rule learning with hypernets. *Advances in Neural Information Processing Systems* 37 (2024), 3564–3587.

[34] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. 2019. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems* 32 (2019).

[35] Zhihang Yuan, Yuzhang Shang, Yue Song, Qiang Wu, Yan Yan, and Guangyu Sun. 2023. Asvd: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821* (2023).

[36] Luca Zampierin, Ghouthi Boukli Hacene, Bac Nguyen, and Mirco Ravanelli. 2024. Skill: Similarity-aware knowledge distillation for speech self-supervised learning. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 675–679.

[37] Xu Zheng, Farhad Shirani, Zhuomin Chen, Chaohao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. 2024. F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI. *arXiv preprint arXiv:2410.02970* (2024).

[38] Xu Zheng, Farhad Shirani, Tianchun Wang, Wei Cheng, Zhuomin Chen, Haifeng Chen, Hua Wei, and Dongsheng Luo. 2023. Towards robust fidelity for evaluating explainability of graph neural networks. *arXiv preprint arXiv:2310.01820* (2023).