FLARE: A DIFFERENTIABLE AND PARSIMONIOUS FUZZY RULE LAYER FOR GROUNDING LEARNED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks learn representations that support accurate prediction, but the decision structure built on these representations is typically opaque. Attribution methods can identify input features associated with individual predictions, while rule-based models expose explicit decision logic; connecting these two forms of interpretability remains challenging. We introduce **FLARE**, a differentiable and parsimonious fuzzy rule layer for learned representations. FLARE replaces a conventional prediction head with a first-order Takagi–Sugeno system whose Gaussian antecedents define fuzzy regions of representation space and whose local linear consequents produce the prediction. An explicit regularisation objective encourages concentrated rule usage and sparse consequents. Because the complete predictor remains differentiable, its decisions can be traced through the representation model back to the input domain. Across tabular, image, and text experiments, FLARE achieves predictive performance close to conventional predictors while providing an explicit rule-structured decision layer. On breast ultrasound image classification and sentiment analysis, attribution of the fuzzy predictor consistently concentrates on independently defined lesion masks and sentiment annotations across multiple localisation measures and random seeds. FLARE therefore connects three levels that are usually studied separately: learned representations, explicit rule-based decisions, and the input evidence associated with those decisions.

1 INTRODUCTION

Modern neural networks (NNs) learn representations that support highly accurate predictions (LeCun et al., 2015), yet the decision structure operating on these representations is typically opaque and difficult to inspect (Guidotti et al., 2018). In contrast, post-hoc attribution methods can identify which input features contribute to an individual prediction (Sundararajan et al., 2017; Lundberg & Lee, 2017), but leave the underlying decision mechanism implicit. Rule-based models offer the complementary advantage of explicitly exposing decision logic (Takagi & Sugeno, 1985; Lakkaraju et al., 2016). When rules are defined over learned representations, however, two challenges remain: (*i*) the rule system must stay compact enough to inspect, and (*ii*) its representation-space conditions must connect back to meaningful evidence in the input domain. Post-hoc rule extraction does not naturally address both objectives jointly, because the representation and the resulting rule geometry are already fixed when the rules are constructed.

To address these limitations, we introduce here **FLARE** (Fuzzy Latent Attribution and Rule Extraction), a differentiable and parsimonious fuzzy rule layer for learned representations. Given a learned representation $z = h(x)$ of an input sample x, FLARE replaces a conventional prediction head with a first-order Takagi–Sugeno (TS) fuzzy system (Takagi & Sugeno, 1985), whose Gaussian antecedents define overlapping regions of the representation space and whose local linear consequents are combined through normalised rule firing. Because the system is differentiable, its antecedents and consequents can be optimised directly, while an explicit parsimony objective encourages concentrated rule usage and sparse local consequents. The resulting predictor also remains differentiable with respect to its input, allowing representation-space decisions to be attributed back through the representation model. FLARE therefore connects three levels that are usually treated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

separately: learned representations, explicit rule-structured decisions, and input features associated with those decisions.

We evaluate FLARE across tabular, vision, and language settings. In our experiments, we first train or fine-tune the representation model, then freeze it before fitting the fuzzy head, allowing us to study the rule-structured predictor without further adapting the representation. We examine whether the fuzzy layer preserves predictive performance, how its parsimony objective affects rule usage as the nominal rule count grows, and whether attribution of its predictions aligns with semantically meaningful input regions. For breast ultrasound, we compare image attributions with clinician-drawn lesion masks, while for the Stanford Sentiment Treebank binary sentiment classification task (SST-2), we compare token-level attributions with phrase annotations derived from the treebank. In both cases, the reference annotations are independent of the models being explained, allowing us to evaluate how well the resulting attributions align with meaningful input regions.

Our contribution is thus threefold. First, we formulate a first-order TS fuzzy system as a differentiable prediction layer over learned representations, using stable log-space rule firing and an explicit parsimony objective that promotes concentrated rule usage and sparse consequents. Second, we exploit the differentiability of the composed predictor to trace representation-space decisions back to the relevant continuous input domain, linking explicit rule structure to the evidence supporting a prediction. Third, we empirically characterise FLARE across tabular, vision, and language settings, studying the trade-off between rule parsimony and predictive performance and showing, on breast ultrasound image and sentiment analysis classification tasks, consistent agreement between FLARE attributions and independently defined lesion masks and sentiment annotations across multiple localisation measures and random seeds.

2 Related Work

Feature-attribution methods explain individual predictions through local surrogates (Ribeiro et al., 2016), Shapley-value allocation (Lundberg & Lee, 2017), or gradients (Simonyan et al., 2014; Sundararajan et al., 2017; Selvaraju et al., 2017; Chattopadhay et al., 2018), and remain a dominant form of post-hoc explanation (Guidotti et al., 2018; Adadi & Berrada, 2018; Barredo Arrieta et al., 2020). Such methods expose influential inputs while leaving the decision mechanism itself implicit, motivating models designed to be interpretable by construction (Rudin, 2019). At the same time, explanation quality lacks a universally agreed evaluation criterion (Doshi-Velez & Kim, 2017), motivating the complementary localisation and perturbation-based evaluations used in Section 3.4.

Extracting symbolic rules from trained networks has a long history (Andrews et al., 1995) and has re-emerged within neuro-symbolic Artificial Intelligence (AI) (d'Avila Garcez & Lamb, 2023). Fuzzy systems are particularly natural for learned representations because they replace hard predicates with graded memberships. First-order TS models (Takagi & Sugeno, 1985), rooted in fuzzy set theory (Zadeh, 1965), combine fuzzy antecedents with local linear consequents, and are commonly induced through fuzzy partitioning such as Fuzzy C-Means (Bezdek, 1981; Bezdek et al., 1984; Fuchs et al., 2020). These procedures are typically post-hoc: the representation is fixed before rule induction. FLARE instead expresses the TS system as a differentiable prediction layer, allowing us to optimise its antecedents, consequents, and rule usage directly (Appendix G relates this formulation to the post-hoc alternatives implemented in the framework).

A complementary line of literature seeks interpretable structure inside the representation itself, through disentangled factors (Higgins et al., 2017), human-defined concept directions (Kim et al., 2018), or explicit concept bottlenecks (Koh et al., 2020). FLARE is related in that prediction passes through an inspectable intermediate structure, but its fuzzy regions are discovered in representation space rather than supplied as semantic concepts. They therefore carry no intrinsic semantic interpretation, and their relation to the input domain must be established empirically. This distinction is important because unsupervised axis-aligned factors are not identifiable without appropriate inductive biases or supervision (Locatello et al., 2019). FLARE therefore does not assume that individual representation coordinates are disentangled.

Human-marked regions provide an external reference for evaluating explanations. In Natural Language Processing (NLP), researchers have used these rationales for both supervision and evaluation. For instance, Zaidan et al. (2007) introduced annotator rationales for sentiment classification, Lei

et al. (2016) incorporated extracted rationales as latent variables of the predictor, and ERASER (DeYoung et al., 2020) standardised evaluation of rationale agreement and faithfulness. Our text evaluation instead derives reference spans from constituent-level sentiment annotations in the Stanford Sentiment Treebank (Socher et al., 2013), which were collected independently of the models studied here. This provides a model-independent reference for testing whether attribution concentrates on regions associated with strong sentiment, without requiring explanation annotations collected specifically for FLARE.

A separate line of work has incorporated differentiable fuzzy inference into neural architectures, including fuzzy NNs whose intermediate units are interpreted as rules (de Campos Souza & Dragoni, 2024), fuzzy classifiers operating on deep convolutional features (Juang et al., 2024), and differentiable fuzzy rule-based representation learning (Zhang et al., 2026). More closely related to our setting, the generative-fuzzy pipeline of Grazioso et al. (2026) induces a post-hoc TS rule base on a frozen Variational AutoEncoder (VAE) representation and relates latent directions to input-space features through the decoder Jacobian. FLARE extends this line of work in two main directions: (*i*) it introduces a trainable TS predictor under an explicit parsimony objective, and (*ii*) it connects representation-space decisions to the input domain even when no decoder is available, by differentiating the composed predictor with respect to the input. FLARE therefore combines a trainable fuzzy decision layer, explicit regularisation of rule usage and consequent complexity, and input-domain attribution within a single differentiable framework.

3 METHOD

FLARE operates on a finite-dimensional numeric representation. Given an input x, a representation map h produces $z = h(x) \in \mathbb{R}^d$. This representation may arise from a deterministic bottleneck, a latent-variable model, or the measured features themselves, without requiring it to be probabilistic, generative, disentangled, or lower-dimensional than the input. In the trainable-head setting used in our experiments, a transformation fitted on the training representations maps z to standardised coordinates $\tilde{z}$, on which the fuzzy system is defined. FLARE then associates these coordinates with a first-order TS predictor and, when an attribution is required, differentiates a selected output of the resulting composed predictor with respect to the relevant continuous input domain. These components are connected through the shared representation z rather than forming a single fixed end-to-end architecture (see Figure 1). The representation model, preprocessing applied before h, representation-space transformation, and differentiated output vary by experiment and are summarised in Table 1. Throughout, d denotes the representation dimensionality, K the number of fuzzy rules, and C the output dimensionality. Note that for classification, C is the number of classes.

3.1 REPRESENTATION AND STANDARDISATION

Given $z = h(x) \in \mathbb{R}^d$, FLARE places no further restriction on the representation map h. In every reported experiment using the trainable fuzzy head, z is standardised component-wise using statistics estimated from the training split:

$$\tilde{z} = \frac{z - \mu_{\text{train}}}{\sigma_{\text{train}}}, \tag{1}$$

where $\mu_{\text{train}}, \sigma_{\text{train}} \in \mathbb{R}^d$ are the training mean and standard deviation and are held fixed thereafter. The TS system operates on $\tilde{z}$, so its antecedent centres, widths, and local consequents are defined in these standardised coordinates. For attribution, the same transformation remains part of the composed predictor and is therefore included in the differentiated computation.

How the representation is related back to the input depends on the model structure available below z. If a map from representation space to the input domain is available, such as the decoder $g_\theta : \mathbb{R}^d \to \mathcal{X}$ of a VAE (Kingma & Welling, 2014) or the generator of a GAN-based model (Goodfellow et al., 2014), variations in z can be propagated through that map. For a VAE, we take z to be the posterior mean rather than a stochastic sample. When no such map is available, as in the vision and language case studies, we instead differentiate a selected output of the composed predictor with respect to the relevant continuous input. The choice of representation model, pooling or projection operation, and dimensionality d is specific to each experiment and is reported in Section 4 and Appendix B.

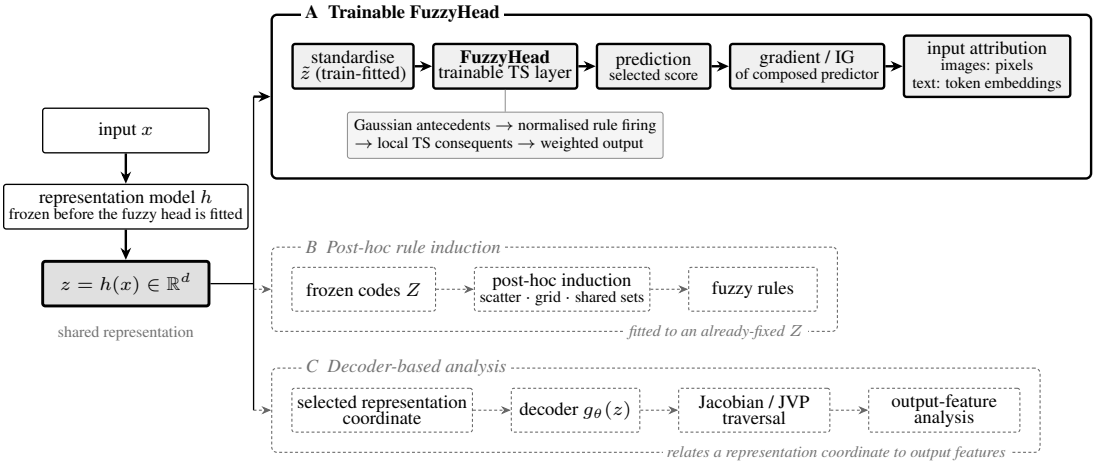

Figure 1: Overview of FLARE around the shared representation $z = h(x)$. **A** shows the trainable-head route: training representations determine the standardisation in Equation 1, the TS head of Equations 2–5 operates on $\tilde{z}$, and a selected score of the composed predictor can be differentiated with respect to the continuous input. **B** shows post-hoc rule induction on an already-fixed representation, while **C** shows how a representation coordinate can be related to output features through a generative decoder.

3.2 The trainable Takagi–Sugeno head and its parsimony objective

Given the standardised representation $\tilde{z} \in \mathbb{R}^d$, FLARE defines a first-order Takagi–Sugeno system (Takagi & Sugeno, 1985) with K rules. Each rule R_k has the form

$$R_k : \quad \text{IF } \tilde{z}_1 \text{ is } A_1^k \wedge \cdots \wedge \tilde{z}_d \text{ is } A_d^k \quad \text{THEN } \mathbf{f}_k(\tilde{z}) = \mathbf{A}_k \tilde{z} + \mathbf{b}_k, \tag{2}$$

where A_i^k denotes the fuzzy set associated with coordinate i in rule k, $\mathbf{A}_k \in \mathbb{R}^{C \times d}$ is the consequent matrix, and $\mathbf{b}_k \in \mathbb{R}^C$ is the corresponding bias. All d representation coordinates appear in every rule antecedent, as FLARE does not currently include a don't-care mechanism for removing individual antecedent conditions during training.

Each rule uses Gaussian membership functions, combined through the product t-norm. To avoid forming the product explicitly, FLARE computes the rule firing strength in log space:

$$\log \beta_k(\tilde{z}) = -\frac{1}{2} \sum_{i=1}^{d} \left(\frac{\tilde{z}_i - c_{k,i}}{\sigma_{k,i} + \varepsilon} \right)^2, \qquad \sigma_{k,i} = \exp(\ell_{k,i}), \qquad \varepsilon = 10^{-9}, \tag{3}$$

where $c_{k,i}$ and $\sigma_{k,i} > 0$ are the centre and width of the Gaussian membership associated with coordinate i of rule k, and $\ell_{k,i}$ is its trainable log-width. The normalised firing strength is then

$$w_k(\tilde{z}) = \frac{\exp(\log \beta_k(\tilde{z}))}{\sum_{j=1}^{K} \exp(\log \beta_j(\tilde{z}))} = \mathrm{softmax}_k(\log \boldsymbol{\beta}(\tilde{z})) \,. \tag{4}$$

This is equivalent to $w_k = \beta_k / \sum_j \beta_j$, but avoids directly multiplying many small membership values, which can underflow in finite precision as d increases. The quantities w_k are normalised rule-firing strengths rather than probabilities of latent events.

The head combines the local consequents using these firing strengths:

$$\mathbf{f}(\tilde{z}) = \sum_{k=1}^{K} w_k(\tilde{z}) \, (\mathbf{A}_k \tilde{z} + \mathbf{b}_k) \in \mathbb{R}^C. \tag{5}$$

For classification, $\mathbf{f}(\tilde{z})$ is interpreted as a vector of logits and passed directly to the cross-entropy loss, and no softmax is applied inside the FuzzyHead. For regression, the same weighted output is used directly as the prediction. Equations 2–5 are differentiable with respect to the antecedent

centres, log-widths, consequent matrices, and biases, although individual experiments may keep a subset of these parameters fixed.

Predictive accuracy alone does not constrain the complexity of the TS head. We therefore regularise two complementary aspects of the rule base: the complexity of the local consequents and the distribution of predictions across the available rules. The former is controlled by an ℓ_1 penalty on the consequent slopes. For the latter, given a batch of n representations, we define the mean usage of rule k as

$$\bar{w}_k = \frac{1}{n} \sum_{m=1}^{n} w_k\left(\tilde{z}^{(m)}\right).$$

Because the firing strengths are normalised across rules, $\sum_{k=1}^{K} \bar{w}_k = 1$. For numerical stability, we set $p_k = \max(\bar{w}_k, \varepsilon)$, with $\varepsilon = 10^{-9}$, and compute the usage entropy as

$$H(\bar{\mathbf{w}}) = -\sum_{k=1}^{K} p_k \log p_k.$$

The complete training objective is thus

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\ell_1} \frac{1}{KCd} \sum_{k=1}^{K} \sum_{o=1}^{C} \sum_{i=1}^{d} |[\mathbf{A}_k]_{o,i}| + \lambda_H \frac{H(\bar{\mathbf{w}})}{\log K}, \tag{6}$$

with $\lambda_{\ell_1} \geq 0$ and $\lambda_H \geq 0$. The ℓ_1 term is the mean absolute value of the KCd consequent slopes and does not penalise the biases $\mathbf{b}_k$, encouraging simpler local linear models. The entropy term instead acts on aggregate rule usage, with division by $\log K$ scaling it relative to the maximum entropy of a K-rule distribution. Because this term is added with a non-negative coefficient to an objective that is minimised, increasing λ_H favours concentrated rather than uniform rule usage.

We quantify this concentration through the following *effective rule count*: $K_{\text{eff}} = \exp(H(\bar{\mathbf{w}}))$, which equals K under uniform rule usage and approaches 1 as usage concentrates on a single rule, up to the numerical clipping used in the entropy calculation. We additionally report the *active rule count* as $K_{\text{active}} = |\{k : \bar{w}_k > 0.01\}|$. Both statistics depend on the samples used to estimate $\bar{\mathbf{w}}$; throughout the manuscript, we compute them from the training representations.

We use two training regimes, both with the representation fixed before fitting the fuzzy head. In the *prototype* regime used in the neural experiments, we first train the representation model with a conventional linear prediction head and select the training epoch based on validation performance. The representation is then frozen, the training codes are standardised, and one rule centre is placed at the mean standardised representation of each class. Antecedent centres and widths remain fixed, while only the consequent parameters $\mathbf{A}_k$ and $\mathbf{b}_k$ are optimised. Consequently, the firing weights are fixed functions of $\tilde{z}$, so the rule-usage entropy is constant with respect to the trainable parameters and does not affect optimisation in this regime.

The *learned-antecedent* regime is used for the rule-count sweeps in Appendix A. Here h is the identity composed with a feature standardiser fitted on the training split. Rule centres are initialised by K-Means on the standardised training data, after which the centres, log-widths, consequent matrices, and biases are jointly optimised under Equation 6. In this regime, the entropy term can reshape the antecedent geometry and thereby alter the rule-usage distribution, making it possible to evaluate the effect of the parsimony objective. For illustration, Appendix A also reports a linguistic rendering of one learned Iris rule, obtained by assigning descriptive labels to its learned antecedent parameters.

3.3 Input attribution

The coordinates of a learned representation generally have no direct interpretation in the original input domain. In the decoder-free setting used in the two main case studies, FLARE therefore attributes a selected output of the complete fuzzy predictor to the continuous input on which differentiation is performed. Writing $S(z) = (z - \mu_{\text{train}})/\sigma_{\text{train}}$ for the standardisation in Equation 1, the composed predictor is $F(x) = \mathbf{f}(S(h(x)))$. For a selected scalar output $F_c(x)$, we compute Integrated Gradients (Sundararajan et al., 2017) relative to a baseline x_0 as follows:

$$\text{IG}_i(x; x_0) = (x_i - x_{0,i}) \int_0^1 \frac{\partial F_c(x_0 + \alpha(x - x_0))}{\partial x_i} \, \mathrm{d}\alpha. \tag{7}$$

The integral is approximated numerically using the step counts reported in Appendix B. Importantly, attribution is computed for the selected output of the fuzzy predictor as a whole, rather than for an individual rule, firing strength, or consequent.

For images, x in Equation 7 denotes the continuous image tensor supplied to the representation model. For text, token identifiers are discrete, so attribution is computed with respect to the continuous token-embedding tensor instead; Equation 7 therefore applies in embedding space. The attribution vector associated with each token is reduced to a scalar score using its ℓ_2 norm over the embedding dimension.

FLARE also supports two additional analysis routes beyond the main attribution setting. When a decoder g_θ is available, a representation coordinate can be related to output features through the decoder Jacobian and finite traversal of that coordinate, as described in Appendix F. This analysis characterises the representation coordinate rather than an individual fuzzy rule. Separately, FLARE supports post-hoc rule induction on an already-fixed representation matrix (Appendix G). These routes are distinct from the trainable FuzzyHead: the product t-norm and softmax-normalised firing in Equations 3–4 specifically define the trainable TS head.

3.4 Evaluating an attribution

We evaluate attributions along two complementary axes. *Agreement with a reference annotation* measures whether attribution concentrates within an independently defined region. Given the attribution scores a_i and an annotated region M, we report the fraction of absolute attribution mass inside M, whether the position with the largest $|a_i|$ lies in M (the pointing game), and the Intersection over Union (IoU) and Dice coefficient between M and an equally sized top-attribution region obtained by ranking positions by $|a_i|$. These metrics measure localisation relative to the reference annotation, but do not, by themselves, establish faithfulness to the predictor.

Faithfulness is evaluated by perturbing the input according to the attribution ranking. For deletion, positions are progressively replaced by a baseline, and the area under the resulting target-score curve is recorded. For insertion, positions are progressively restored from the baseline. A more informative ranking should therefore yield a lower deletion Area Under the Curve (AUC) and a higher insertion AUC than a random ordering.

Perturbation scores depend on three choices: the baseline, the unit removed or restored, and the ranking convention. In the signed-attribution experiments, positions are ranked by decreasing attribution to prioritise features that support the selected target score, whereas annotation agreement uses attribution magnitude regardless of the sign. For text, each token score is an ℓ_2 norm and is therefore non-negative, so the distinction between signed and magnitude ranking disappears. In the image experiments, the reported perturbation conventions also differ in the unit perturbed. In particular, differences between them therefore reflect the convention as a whole rather than the ranking criterion alone. Section 4 and Table 1 specify the conventions used in each experiment.

4 Experimental Setup

The two main case studies use ten random seeds and a validation split for model selection, with hyperparameters fixed before the test evaluation. Comparisons across seeds are paired, and we report two-sided Wilcoxon signed-rank tests alongside paired t-tests. For $n = 10$, the smallest attainable two-sided Wilcoxon p-value is approximately 0.0020. Table 1 summarises the representation, fuzzy-head regime, attribution target, perturbation domain, and evaluation population for each experiment. Full dataset, optimisation, and reproducibility details are provided in Appendix B.

For the image experiments, $z \in \mathbb{R}^{16}$ is obtained by linearly projecting the penultimate representation of an ImageNet-pre-trained ResNet-18 (He et al., 2016), with its final residual block fine-tuned. For SST-2, $z \in \mathbb{R}^{16}$ is a linear projection of the masked mean of ModernBERT-base contextual token representations over non-padding positions (Warner et al., 2024). We use a compact $d = 16$ representation in both case studies to keep the fuzzy rule geometry tractable. The SST-2 model fine-tunes all 22 ModernBERT encoder layers for two epochs with AdamW, a learning rate of 2×10^{-5}, and batch size 32. The complete optimisation details are provided in Appendix H.

Table 1: Summary of experiments. Representations are frozen before FuzzyHead fitting; Integrated Gradients uses a zero baseline. Full protocols are given in Appendix B.

Experiment	Representation z	Regime	Target	Domain (unit)	Evaluated on
BUSI (§4.1)	ResNet-18, d=16	prototype	fuzzy	pixels (C-P, PX)	correct lesion images
SST-2 (§4.2)	ModernBERT, d=16	prototype	fuzzy	embeddings (TOK)	correct annotated sentences
Pneumonia (D)	ResNet-18, d=16	prototype	fuzzy	pixels (C-P, PX)	correct images
PBMC (E)	Transformer CLS, d=64	prototype	*base*	gene values	fuzzy-correct cells
Iris, Digits (A)	standardised features	learned	—	—	test (macro-F1); train (K_{eff})

The four neural experiments use the prototype regime of Section 3.2, with one rule per class. Rule centres are placed at the class means of the standardised training representations, while centres and widths are fixed, and only the local consequents are optimised. Predictive performance is compared with the representation model's original linear head. This comparison concerns the complete predictors rather than fuzzy and linear decision geometries under identical training conditions: the linear head is trained jointly with the backbone and receives the post-ReLU bottleneck, whereas the TS head is fitted after freezing the backbone and receives the standardised pre-ReLU representation of Equation 1. Appendix A separately considers the learned-antecedent setting in which centres and widths are also optimised.

For BUSI (Al-Dhabyani et al., 2020), we evaluate attribution against clinician-drawn lesion masks. Lesion-free images are excluded from localisation metrics, leaving 35.4 ± 2.1 evaluable test images per seed. Because the dataset provides no patient or lesion identifier, we define the stratified splits at the image level. For SST-2 (Socher et al., 2013), reference spans are derived from constituent-level sentiment annotations: a token is marked when it belongs to a phrase of at most two words with sentiment below 0.30 or above 0.70. We fixed these thresholds on 400 validation sentences before training the model. Across the 150 evaluated test sentences per seed, the resulting annotated fraction is 0.1840 ± 0.0096 and serves as the chance level for attribution mass in Table 2. We exclude sentences without marked tokens from localisation evaluation. Additional experiments in the appendix broaden the evaluation beyond the two main case studies: PneumoniaMNIST provides a second image-classification setting (Appendix D), PBMC extends the neural evaluation to single-cell gene-expression classification (Appendix E), and Iris and Digits are used to study learned antecedents and parsimony across different rule counts (Appendix A).

In Table 1, *prototype* denotes the fixed-prototype regime described above, whereas *learned* denotes the K-Means-initialised regime in which centres, widths, and consequents are all trainable. The attribution target is either the selected FuzzyHead logit (*fuzzy*) or, for PBMC, the base Transformer's classifier (*base*). For image perturbations, C-P denotes a single channel-position coordinate ranked by signed attribution, while PX denotes a whole pixel ranked by attribution magnitude; TOK denotes a complete token embedding. Attribution and perturbation step counts are reported in Appendix B. Across the four neural datasets, the mean accuracy of the TS predictor differs from that of the corresponding conventional linear predictor by less than one percentage point. On BUSI, FLARE reaches 0.9272 ± 0.0199 test accuracy compared with 0.9313 ± 0.0175 for the linear predictor, a paired difference of -0.0041 (Wilcoxon $p = 0.52$; paired t-test $p = 0.46$). On SST-2, the corresponding accuracies are 0.9469 ± 0.0033 and 0.9455 ± 0.0045, respectively, giving a paired difference of $+0.0014$ (Wilcoxon $p = 0.46$; paired t-test $p = 0.30$). The appendix experiments show similarly small differences: $+0.0087$ on PneumoniaMNIST ($p = 0.094$; Appendix D) and $+0.0017$ on PBMC ($p = 0.43$; Appendix E). None of these paired comparisons provides evidence of a difference in predictive accuracy at conventional significance levels. These results should not be interpreted as an equivalence test: with ten seeds, failure to reject a difference does not establish equal performance, and, as discussed in Section 4, the comparison concerns the complete conventional and rule-structured predictors rather than isolating their decision heads under identical training conditions.

4.1 Breast ultrasound: attribution against clinician-drawn masks

Across all four localisation measures, FLARE attribution shows greater agreement with the clinician-drawn lesion masks than the random baseline (Table 2). Each comparison attains the minimum two-sided Wilcoxon p-value possible with ten paired seeds, $p = 0.0020$, with paired t-test

Table 2: Agreement between FLARE attribution and independently defined reference annotations. Values are mean ± standard deviation over ten seeds; *ratio* denotes the FLARE value divided by the random-ordering baseline, and p the two-sided paired Wilcoxon signed-rank p-value. Positions are ranked by attribution magnitude; IoU and Dice compare the annotation with an equally sized top-attribution region.

Measure	FLARE attribution	Random ordering	Ratio	Wilcoxon p
BUSI: clinician-drawn lesion masks				
Pointing game	0.5143 ± 0.0534	0.0899 ± 0.0451	$5.72\times$	0.0020
Mass in mask	0.1646 ± 0.0158	0.1015 ± 0.0091	$1.62\times$	0.0020
IoU	0.1589 ± 0.0197	0.0564 ± 0.0056	$2.82\times$	0.0020
Dice	0.2573 ± 0.0276	0.1015 ± 0.0089	$2.54\times$	0.0020
SST-2: Stanford Sentiment Treebank phrase annotations				
Pointing game	0.4293 ± 0.0316	0.1733 ± 0.0311	$2.48\times$	0.0020
Mass in spans	0.2334 ± 0.0122	0.1837 ± 0.0082	$1.27\times$	0.0020
IoU	0.2734 ± 0.0217	0.1190 ± 0.0080	$2.30\times$	0.0020
Dice	0.3623 ± 0.0249	0.1788 ± 0.0086	$2.03\times$	0.0020

$p < 10^{-4}$. The largest relative difference is observed for the pointing game: the highest-ranked pixel lies inside the lesion in 0.5143 ± 0.0534 of evaluated images, compared with 0.0899 ± 0.0451 under random ordering. Attribution mass, IoU, and Dice likewise exceed their random baselines by factors of 1.62, 2.82, and 2.54, respectively.

Under the signed channel-position convention, attribution-guided deletion produces a lower AUC than random ordering (0.6929 ± 0.1491 versus 0.8668 ± 0.0462, Wilcoxon $p = 0.0020$), indicating that removing highly ranked positive-attribution coordinates reduces the selected score more rapidly. Under the magnitude-ranked whole-pixel convention, the corresponding difference is smaller and not statistically significant (0.7930 ± 0.0896 versus 0.8245 ± 0.0709, $p = 0.13$), whereas insertion distinguishes attribution from random under both conventions. Because the two deletion protocols differ in both ranking criterion and perturbation unit, this contrast cannot be attributed to either factor in isolation; instead, it shows that perturbation-based faithfulness depends on the evaluation convention. Qualitative examples and a comparison with Grad-CAM++ Chattopadhay et al. (2018) on the same fuzzy predictors and evaluation samples are provided in Appendix C. The comparison further separates agreement with the clinician masks from predictor-specific faithfulness.

4.2 Sentiment: attribution against treebank phrase annotations

The SST-2 case study evaluates the same attribution framework in a different modality and architecture. Its reference annotations come from constituent-level sentiment labels in the Stanford Sentiment Treebank, collected independently of FLARE and the models evaluated here. They therefore provide a model-independent reference for assessing whether attribution concentrates on short phrases associated with strong sentiment.

FLARE attribution exceeds the random-ordering baseline on all four localisation measures (Table 2), with each comparison attaining the minimum two-sided Wilcoxon p-value possible with ten paired seeds, $p = 0.0020$, and paired t-test $p < 10^{-4}$. The highest-ranked token lies inside an annotated span in 0.4293 ± 0.0316 of evaluated sentences, compared with 0.1733 ± 0.0311 under random ordering. Attribution mass within the spans is 0.2334 ± 0.0122, compared with 0.1837 ± 0.0082 for random ordering and an observed annotated-token fraction of 0.1840 ± 0.0096. IoU and Dice show the same pattern, exceeding their random baselines by factors of 2.30 and 2.03, respectively.

Attribution-guided deletion produces a lower target-score AUC than random ordering (0.5883 ± 0.0211 versus 0.6991 ± 0.0133), while attribution-guided insertion produces a higher AUC (0.8176 ± 0.0131 versus 0.7236 ± 0.0142); both comparisons attain $p = 0.0020$. Because each token attribution is reduced to an ℓ_2 norm over the embedding dimensions, the resulting scores are non-negative, so signed and magnitude rankings coincide. Unlike the magnitude-ranked whole-pixel deletion result on BUSI, both perturbation measures distinguish attribution from random ordering in this text setting. The present experiments do not isolate which modality- or protocol-specific factors account for this difference.

5 DISCUSSION

Across breast ultrasound image classification and sentiment analysis, attribution of the composed fuzzy predictor is more concentrated on independently defined reference annotations than a random ordering under every localisation measure considered. BUSI uses clinician-drawn lesion masks, whereas SST-2 uses constituent-level sentiment annotations collected independently of the evaluated models. Absolute localisation scores are not directly comparable across datasets, but the consistent effect direction supports FLARE's central premise: an explicit rule-structured predictor in representation space can remain differentiably connected to meaningful evidence in the input domain. The BUSI comparison with Grad-CAM++ in Appendix C further shows that localisation should not be equated with predictor-specific faithfulness: Grad-CAM++ attains higher IoU and Dice while remaining nearly invariant to the selected output and to randomisation of the fuzzy consequents. More generally, attribution methods can inherit strong spatial structure from the representation model itself, so agreement with an external annotation should be complemented by sensitivity tests to the predicted output and model parameters. Broader generalisation will require controlled evaluation across additional modalities and architectures.

The fuzzy head provides explicit local decision rules while maintaining predictive performance close to the corresponding conventional predictors in the experiments studied here. When the nominal number of rules exceeds the number of classes, the entropy regulariser reduces effective rule usage across the Iris and Digits sweeps (Appendix A). This effect is limited when the unregularised model is already compact. For instance, at Iris $K = 3$, the penalty produces a negligible increase in K_{eff}, while at Digits $K = 30$ and $K = 40$ stronger concentration costs about 1.1 macro-F1 points. Parsimony is therefore a tunable complexity–performance trade-off rather than a cost-free reduction in rule usage. The rules are explicit in representation coordinates, but those coordinates are not intrinsically semantic; Appendix A provides an illustrative linguistic rendering of one learned Iris rule. Interpretability therefore arises from combining explicit decision structure with empirical analysis of how that structure relates to the input, rather than assuming semantic meaning for individual representation coordinates.

Several aspects constrain the scope of these conclusions. Predictive comparisons are not equivalence tests and involve complete predictors trained under different head-fitting regimes. The main case studies use fixed antecedents, whereas learned antecedents are evaluated separately in the tabular sweeps. The SST-2 spans depend on a pre-specified thresholding rule, and BUSI uses image-level splits because patient identifiers are unavailable; its snapshot also contains one duplicated ultrasound image under two diagnostic entries, but sensitivity analyses show no material effect on the reported conclusions. Perturbation-based faithfulness is protocol-dependent, and the two BUSI deletion conventions change both ranking criterion and perturbation granularity. Finally, the decoder-based and post-hoc routes are not evaluated in the main case studies, while the PBMC attribution analysis explains the base Transformer rather than the fuzzy predictor. The attribution claims therefore remain specific to the evaluated fuzzy-predictor settings.

6 CONCLUSION

FLARE introduces a differentiable first-order TS prediction layer over learned representations, together with an explicit parsimony objective and a mechanism for relating representation-space decisions to evidence in the input domain. Across tabular, vision, and language experiments, the rule-structured predictor remains close in predictive performance to conventional predictors while exposing explicit local fuzzy rules. On breast ultrasound image classification and sentiment analysis, attribution of the composed predictor consistently concentrates on independently defined lesion masks and sentiment annotations. The experiments also identify two practical boundaries: stronger parsimony can trade predictive performance for a smaller effective rule set, while attribution quality depends on how it is evaluated, and localisation against reference annotations does not necessarily imply predictor-specific faithfulness. FLARE therefore provides a general approach for introducing inspectable rule structure into learned representations while preserving differentiability from prediction back to the input.

AI USE STATEMENT

Generative AI tools were used to inspect and summarise the source repository, perform a final consistency check between the implementation and manuscript, assist with bibliographic search and metadata retrieval, and edit the final version of the manuscript. AI tools were not used to design the reported experimental evaluation or to produce or modify its numerical results. The authors reviewed all AI-assisted output, checking implementation claims against the source code, numerical claims against the experiment artefacts, and references against publisher or proceedings records. The authors take responsibility for the final content.

ETHICS STATEMENT

All experiments use existing publicly released datasets obtained through their standard distributions. The two medical imaging datasets are distributed under CC BY 4.0. No new human-subject data or annotations were collected, and no attempt was made to identify individuals. The medical imaging experiments provide a methodological evaluation of attribution and rule structure and do not constitute clinical validation or support diagnostic use.

REPRODUCIBILITY STATEMENT

The implementation is provided as anonymised supplementary material. Section 3 specifies the model and objective, while Section 4, Table 1, and Appendix B document the experimental protocols, including datasets, splits, optimisation, head fitting, attribution settings, and perturbation step counts. We include per-seed results underlying the reported aggregates as tracked artefacts, and Appendix H records software versions, execution provenance, and known reproducibility limitations. Re-execution in the original environment reproduced the stored per-seed results. We do not claim bitwise reproducibility across different hardware, drivers, or library versions.

REFERENCES

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. doi: 10.1109/ACCESS.2018.2870052.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. doi: 10.1016/j.dib.2019.104863.

E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998.

Robert Andrews, Joachim Diederich, and Alan B. Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6):373–389, 1995. doi: 10.1016/0950-7051(96)81920-4.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. doi: 10.1016/j.inffus.2019.12.012.

James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981. doi: 10.1007/978-1-4757-0450-1.

James C. Bezdek, Robert Ehrlich, and William Full. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203, 1984. doi: 10.1016/0098-3004(84)90020-7.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, 2018. doi: 10.1109/WACV.2018.00097.

Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic AI: the 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, 2023. doi: 10.1007/s10462-023-10448-w.

Paulo Vitor de Campos Souza and Mauro Dragoni. IFNN: Enhanced interpretability and optimization in FNN via Adam algorithm. *Information Sciences*, 678:121002, 2024. doi: 10.1016/j.ins.2024.121002.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4443–4458, 2020. doi: 10.18653/v1/2020.acl-main.408.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. doi: 10.48550/arXiv.1702.08608.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

Caro Fuchs, Simone Spolaor, Marco S. Nobile, and Uzay Kaymak. pyFUME: A Python package for fuzzy model estimation. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8, 2020. doi: 10.1109/FUZZ48607.2020.9177565.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NeurIPS)*, pp. 2672–2680, 2014.

Matteo Grazioso, Elisa Degara, Marco S. Nobile, and Andrea Tangherloni. A generative-fuzzy approach for transparent rule induction from latent embeddings. In *International Joint Conference on Neural Networks (IJCNN)*, 2026. to appear.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51 (5):93:1–93:42, 2018. doi: 10.1145/3236009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations (ICLR)*, 2017.

Chia-Feng Juang, Yun-Wei Cheng, and Yeh-Ming Lin. Visually interpretable fuzzy neural classification network with deep convolutional feature maps. *IEEE Transactions on Fuzzy Systems*, 32 (3):1063–1077, 2024. doi: 10.1109/TFUZZ.2023.3318086.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2668–2677, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6114.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 5338–5348, 2020.

Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1675–1684. Association for Computing Machinery, 2016. doi: 10.1145/2939672.2939874.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 107–117, 2016. doi: 10.18653/v1/D16-1011.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 4114–4124, 2019.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 4765–4774, 2017.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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 (KDD)*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations (ICLR), Workshop Track*, 2014. arXiv:1312.6034.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642, 2013. URL https://aclanthology.org/D13-1170/.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 3319–3328, 2017.

Tomohiro Takagi and Michio Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15(1):116–132, 1985. doi: 10.1109/TSMC.1985.6313399.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 5998–6008, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018. doi: 10.18653/v1/W18-5446.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024. doi: 10.48550/arXiv.2412.13663.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023. doi: 10.1038/s41597-022-01721-8.

Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965. doi: 10.1016/S0019-9958(65)90241-X.

Omar Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 260–267, 2007. URL https://aclanthology.org/N07-1033/.

Wei Zhang, Zhaohong Deng, Guanjin Wang, and Kup-Sze Choi. Fuzzy rule-based differentiable representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 37(2):603–616, 2026. doi: 10.1109/TNNLS.2025.3609722.

A RULE-COUNT SWEEPS ON IRIS AND DIGITS

These experiments examine how the parsimony objective of Equation 6 affects rule usage as the nominal rule count K increases. In contrast to the main neural experiments, the antecedent centres and widths are trainable rather than fixed at class prototypes. No learned representation is used, and h is the identity composed with a feature standardiser fitted on the training split, so the fuzzy rule geometry is defined directly over the standardised measured features. The sweeps therefore isolate the setting in which the entropy term can modify the antecedent geometry and, consequently, the distribution of rule usage.

A.1 RULE-COUNT AND PARSIMONY SWEEPS

We evaluate Iris (Fisher, 1936) and Digits (Alpaydin & Kaynak, 1998), both obtained through scikit-learn (Pedregosa et al., 2011), using a $60/20/20$ train/validation/test split. Antecedent centres are initialised with K-Means (`n_init`=20), and centres, log-widths, and consequents are jointly optimised for 500 epochs with learning rate 10^{-2} and no epoch selection. We select the regularisation weights once per dataset on the validation split and hold them fixed across all values of K. Iris uses $\lambda_{\ell_1} = 0.1$ and $\lambda_H = 0.05$, whereas Digits uses $\lambda_{\ell_1} = 0.001$ and $\lambda_H = 0.05$. We report $n = 10$ seeds for Iris and $n = 15$ for Digits. In addition to K_{eff} and the active rule count defined in Section 3.2, we measure *consequent sparsity* as the fraction of entries of the consequent matrices with magnitude below 10^{-2}; unlike the usage statistics, this quantity depends only on the learned parameters.

As an illustrative example, the highest-usage rule for Iris with $K = 3$ and seed 0 is rendered as

> **R3:** IF sepal length is High AND sepal width is High AND petal length is Medium AND petal width is Low THEN versicolor.

The linguistic antecedent terms and class label are descriptive summaries of the learned parameters rather than the executable Gaussian memberships and affine consequent.

Table 3: Rule-count sweep comparing the unregularised and regularised models. Values are mean $\pm$ standard deviation over $n = 10$ seeds for Iris and $n = 15$ for Digits. Macro-F1 is evaluated on the test split, whereas K_{eff} is computed from rule usage on the training representations. Consequent sparsity is shown separately in Figure 3.

		K_{eff}		test macro-F1	
Dataset	K	baseline	regularised	baseline	regularised
Iris	3	2.999 ± 0.002	2.999 ± 0.001	0.946 ± 0.033	0.943 ± 0.032
	6	3.925 ± 0.363	3.309 ± 0.361	0.940 ± 0.035	0.950 ± 0.033
	9	4.364 ± 0.463	3.452 ± 0.464	0.950 ± 0.033	0.953 ± 0.032
	12	4.823 ± 0.540	3.460 ± 0.479	0.943 ± 0.032	0.933 ± 0.035
Digits	10	7.901 ± 0.367	7.149 ± 0.396	0.950 ± 0.009	0.947 ± 0.010
	20	14.157 ± 0.473	12.273 ± 0.574	0.950 ± 0.010	0.950 ± 0.014
	30	20.941 ± 1.124	13.948 ± 0.579	0.957 ± 0.010	0.946 ± 0.012
	40	28.513 ± 1.033	14.724 ± 0.633	0.958 ± 0.012	0.947 ± 0.011

For both datasets, regularisation reduces the effective rule count whenever the nominal number of rules exceeds the number of classes, and the reduction becomes larger as K increases. On Digits at $K = 40$, for example, the baseline uses 28.5 effective rules on average, compared with 14.7 under regularisation. The boundary case occurs for Iris at $K = 3$, where the unregularised model already uses all three available rules essentially. Here the regularised model has a marginally larger K_{eff}, with a paired difference of only $+0.0006$. The direction is consistent across the ten seeds, yielding the minimum Wilcoxon p-value, although the magnitude is negligible and the paired t-test is not significant ($p = 0.082$).

The reduction in effective rule usage is accompanied by little change in predictive performance at moderate rule counts, but the trade-off becomes visible for larger models. On Digits, the regularised

Table 4: Paired seed-level comparisons corresponding to Table 3, reported as regularised minus baseline. Confidence intervals are 95% percentile bootstrap intervals over 10,000 resamples of the paired mean difference. Wilcoxon values are two-sided signed-rank p-values; the minimum attainable values are 0.0020 for $n = 10$ and approximately 0.0001 for $n = 15$.

Dataset	K	ΔK_{eff}	95% CI	Wilcoxon	t-test
Iris	3	$+0.0006$	$[+0.0001, +0.0012]$	0.0020	0.0816
	6	-0.616	$[-0.865, -0.398]$	0.0020	0.0009
	9	-0.911	$[-1.212, -0.647]$	0.0020	0.0002
	12	-1.363	$[-1.667, -1.100]$	0.0020	< 0.0001
Digits	10	-0.753	$[-0.851, -0.654]$	0.0001	< 0.0001
	20	-1.884	$[-2.150, -1.634]$	0.0001	< 0.0001
	30	-6.993	$[-7.501, -6.462]$	0.0001	< 0.0001
	40	-13.789	$[-14.336, -13.239]$	0.0001	< 0.0001
Test macro-F1, same pairing					
Iris	all	-0.003 to $+0.010$	spans 0	0.66 to 1.00	0.34 to 0.79
Digits	10, 20	-0.001 to $+0.001$	spans 0	0.85	0.83
Digits	30	-0.0111	$[-0.0155, -0.0072]$	0.0001	0.0002
Digits	40	-0.0106	$[-0.0157, -0.0058]$	0.0015	0.0013

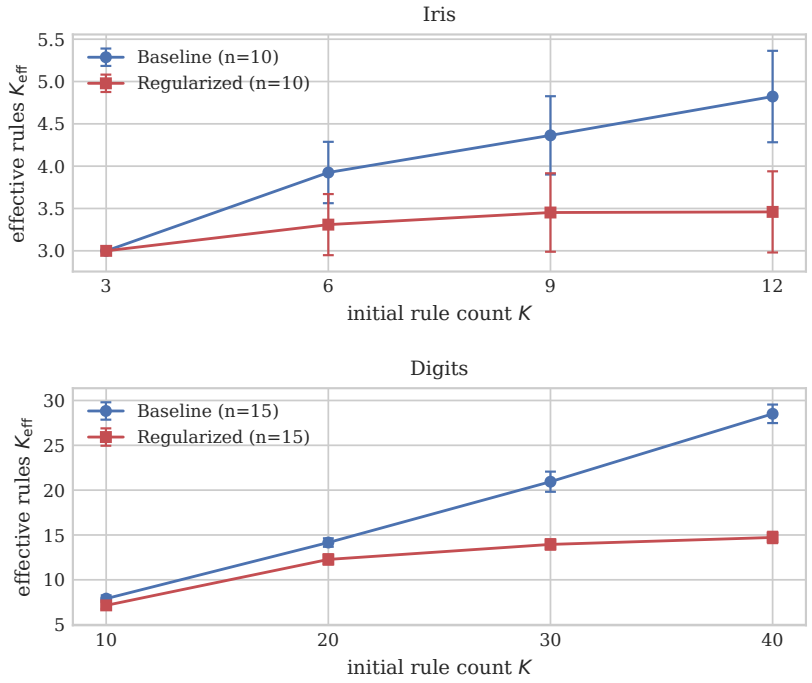

Figure 2: Effective rule count as a function of nominal rule count for the baseline and regularised models. K_{eff} is computed from rule usage on the training representations. Error bars show one standard deviation over $n = 10$ seeds for Iris and $n = 15$ for Digits. The vertical axes use dataset-specific ranges; comparisons are intended within each dataset rather than between absolute effective rule counts across datasets.

and baseline models have similar macro-F1 at $K = 10$ and $K = 20$, whereas at $K = 30$ and $K = 40$ regularisation reduces macro-F1 by approximately 1.1 percentage points, with both paired tests indicating a difference. These results show that the parsimony objective can substantially concentrate rule usage when excess rules are available, but increasingly strong concentration can incur a predictive cost.

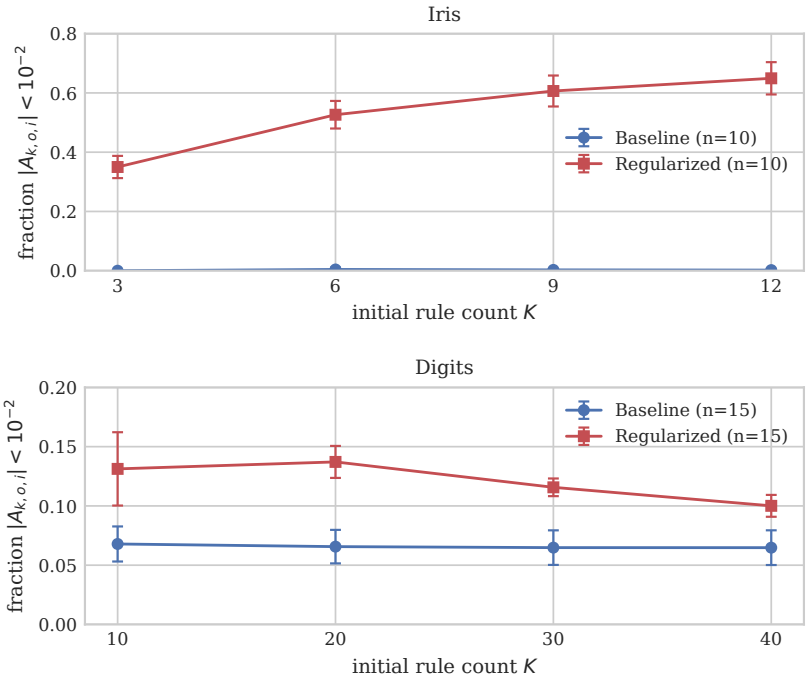

Figure 3: Fraction of consequent coefficients with magnitude below 10^{-2} for the same runs as Figure 2. This statistic reflects the sparsifying effect of the ℓ_1 term in Equation 6. Error bars show one standard deviation across seeds. The vertical axes use dataset-specific ranges, both starting at zero; comparisons are intended within each dataset rather than between absolute sparsity levels across datasets.

B Full experimental protocol

This appendix provides the experimental details for the four neural experiments. Section 4 states the common protocol, and Table 1 summarises the per-experiment attribution conventions.

SST-2 (Socher et al., 2013; Wang et al., 2018) is used via its GLUE release. We split the training set $90/10$ into training and validation ($60{,}614$ and $6{,}735$ sentences), redraw the split for each seed, and use the official validation split of 872 sentences as the test set because the GLUE test labels are not public. The encoder is ModernBERT-base (Warner et al., 2024), with 149M parameters and a maximum sequence length of 64 tokens. Sentence representations are obtained by masked mean pooling over contextual token representations at non-padding positions, followed by a linear projection from 768 features to $d = 16$. We fine-tune all 22 encoder layers with AdamW at a learning rate of 2×10^{-5} and a batch size of 32 for two epochs over ten seeds, restoring the epoch with the best validation macro-F1. Localisation is computed over non-special tokens only, whereas deletion and insertion operate over all T positions, including special tokens.

BUSI (Al-Dhabyani et al., 2020) contains 780 ultrasound images (437 benign, 210 malignant, 133 normal) with clinician-drawn lesion masks and is distributed under CC BY 4.0. We resize images to 128×128, and the task is binary: normal versus abnormal. We stratify the split and redraw it for each seed into 468 training, 117 validation, and 195 test images. The reported runs use the dataset snapshot identified by the manifest provided with the supplementary material, in which one pre-merged lesion mask is stored per image. Because the materialised filenames differ from those of other BUSI distributions and the split procedure operates on the resulting ordered file list, reproducing the exact reported splits requires this snapshot rather than a newly downloaded copy.

We use PneumoniaMNIST (Yang et al., 2023) with its official train/validation/test splits ($4{,}708$, 524, and 624 images) and upsample from 28×28 to 128×128. For both datasets, we replicate grayscale images across three channels and normalise them with ImageNet statistics. The backbone is an

ImageNet-pre-trained ResNet-18 (He et al., 2016), with its final residual block fine-tuned and batch-normalisation statistics frozen. We project the 512 penultimate features linearly to $d = 16$. Training runs for 20 epochs at learning rate 10^{-4}, with batch size 32 for BUSI and 64 for PneumoniaMNIST, selecting the epoch with the best validation macro-F1.

In the four neural experiments, the fuzzy head follows the prototype regime of Section 3.2, with $K = 2$ rules, one centre at each class mean in the standardised representation space, fixed antecedent centres and widths with $\sigma_{k,i} = 2$, and only the consequent parameters $\mathbf{A}_k$ and $\mathbf{b}_k$ optimised. The head is trained with Adam at learning rate 0.05 for 500 full-batch steps; Appendix E uses different values for both the antecedent width and the learning rate in the single-cell experiment. The baseline in each predictive comparison is the representation model's original linear head, trained jointly with the backbone and operating on the post-ReLU bottleneck.

For attribution, the two image datasets use 12 Integrated Gradients steps and 100 deletion/insertion steps; SST-2 uses 16 Integrated Gradients steps and $\min(40, T)$ perturbation steps for a sentence of length T. The baseline is zero in every case, corresponding to the channel-mean image after ImageNet normalisation and to the zero embedding vector for text. Attribution is evaluated on 150 correctly classified test sentences per seed for SST-2 and 40 correctly classified images per seed for each image dataset; BUSI localisation further excludes lesion-free images, leaving 35.4 ± 2.1 evaluable images per seed. Image perturbations use both the $3 \times 128 \times 128$ channel-position array ranked by signed attribution and the 128×128 pixel array ranked by attribution magnitude, with all three channels perturbed jointly in the latter case. Text perturbations replace one complete token-embedding vector at a time.

C Additional BUSI attribution analyses

Figure 4 complements the aggregate BUSI results with qualitative examples selected independently of their visual appearance. The examples were fixed before the Grad-CAM++ analysis and correspond to the lower quartile, median, and upper quartile of per-image Integrated Gradients IoU within the selected seed.

To assess whether the BUSI findings depend specifically on Integrated Gradients, we repeated the localisation and faithfulness analyses with Grad-CAM++ (Chattopadhay et al., 2018) while keeping the trained predictor, selected output, test splits, lesion masks, and evaluation population fixed. Grad-CAM++ uses the same selected output from the full fuzzy predictor and the final spatial ResNet-18 block, `layer4`, whose 4×4 feature map is bilinearly upsampled to the 128×128 input grid. Across all ten seeds, we evaluate both methods on exactly the same 35.4 ± 2.1 correctly classified lesion-containing images per seed.

Table 5: Integrated Gradients (IG) and Grad-CAM++ on the same BUSI fuzzy predictors and evaluation samples. Values are the mean $\pm$ standard deviation over ten seeds. Δ denotes Grad-CAM++ minus IG and p the two-sided paired Wilcoxon signed-rank p-value. Higher values are preferable except for deletion AUC. Faithfulness uses the magnitude-ranked whole-pixel perturbation protocol.

Measure	IG	Grad-CAM++	Random	Δ	p
Pointing ↑	0.5143 ± 0.0534	0.3938 ± 0.0908	0.0899 ± 0.0451	-0.1205	0.0156
Mass ↑	0.1646 ± 0.0158	0.1343 ± 0.0119	0.1015 ± 0.0091	-0.0303	0.0020
IoU ↑	0.1589 ± 0.0197	0.2581 ± 0.0433	0.0564 ± 0.0056	$+0.0992$	0.0020
Dice ↑	0.2573 ± 0.0276	0.3583 ± 0.0518	0.1015 ± 0.0089	$+0.1009$	0.0020
Deletion AUC ↓	0.7930 ± 0.0896	0.8934 ± 0.0539	0.8245 ± 0.0709	$+0.1004$	0.0039
Insertion AUC ↑	0.9090 ± 0.0369	0.9089 ± 0.0501	0.8280 ± 0.0629	-0.0001	1.0000

The two attribution methods produce substantially different spatial maps (mean Spearman $\rho = 0.049 \pm 0.064$) and separate the localisation measures in opposite directions. Grad-CAM++ has higher equal-area IoU and Dice, whereas Integrated Gradients has higher pointing-game accuracy and attribution mass inside the lesion. Under the common whole-pixel perturbation protocol, Grad-CAM++ also has a higher deletion AUC than Integrated Gradients ($p = 0.0039$), indicating weaker deletion faithfulness, and is worse than the random ordering on deletion ($p = 0.0195$). Insertion AUC is essentially identical for both methods, and both exceed the random baseline.

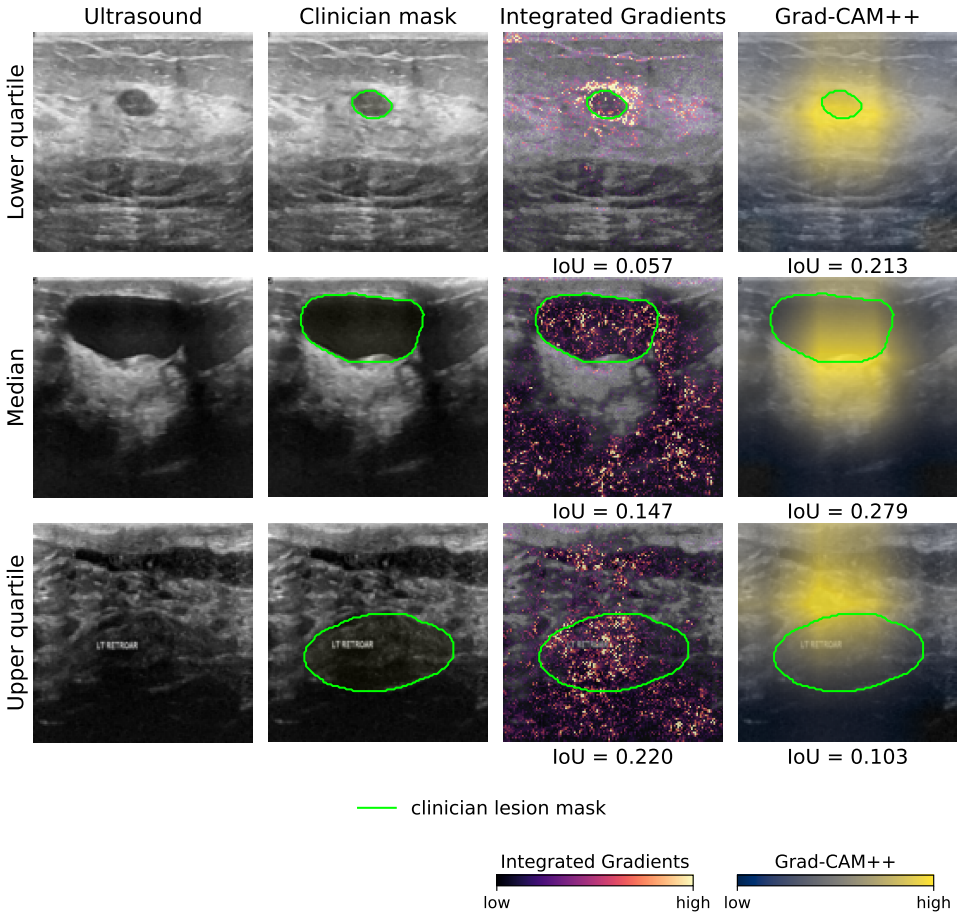

Figure 4: Qualitative BUSI examples for Integrated Gradients and Grad-CAM++ applied to the same selected FuzzyHead output. Rows show pre-declared examples at the lower quartile, median, and upper quartile of per-image Integrated Gradients IoU; we fixed their selection before adding Grad-CAM++. Green contours denote clinician-provided lesion masks. IoU compares each lesion mask with an equally sized top-attribution region derived from the corresponding attribution map. Attribution maps are independently normalised for visualisation, so colour intensity is not quantitatively comparable between methods or across rows; all reported metrics use the original attribution values. Lesion masks are used only for evaluation and display and never enter either attribution method.

The higher Grad-CAM++ IoU and Dice should not, however, be interpreted as stronger predictor-specific attribution. Recomputing the maps for the opposite output class or after randomising the fuzzy consequent parameters yields correlations of 0.987 ± 0.006 and 0.993 ± 0.004, respectively, with changes below 0.009 on every localisation measure. Grad-CAM++ is also almost identical to the gradient-free channel mean of the same `layer4` activations (Spearman $\rho = 0.997 \pm 0.002$), which produces nearly the same localisation scores. These controls indicate that, in this architecture, Grad-CAM++ maps are dominated by spatial backbone activations rather than by the specific fuzzy output being attributed.

This behaviour is consistent with the architecture. At `layer4`, the selected score is read downstream through global average pooling, and the gradient is constant across the 4×4 spatial grid within each channel. Grad-CAM++ therefore combines non-negative feature maps from only 16 native spatial positions before $32\times$ upsampling. Consequently, a smooth spatial map can overlap a contiguous lesion well without being strongly specific to the selected output. The comparison therefore illustrates why agreement with an external reference annotation and perturbation-based faithfulness should be assessed separately rather than collapsed into a single notion of explanation quality.

D PneumoniaMNIST analysis

PneumoniaMNIST is included as an additional image-classification experiment following the protocol of Section 4. Unlike BUSI, it provides no lesion-level annotation, so we evaluate predictive performance and perturbation-based faithfulness but do not report localisation metrics.

Across ten seeds, the TS head reaches 0.8891 ± 0.0081 test accuracy, compared with 0.8804 ± 0.0126 for the conventional linear predictor on the same representation backbone. The paired difference is $+0.0087$ (Wilcoxon $p = 0.094$; paired t-test $p = 0.068$). The validation-selected epoch ranges from 3 to 18 across seeds.

Attribution-guided insertion yields a higher AUC than random ordering under both perturbation conventions: 0.7437 ± 0.0615 versus 0.7039 ± 0.0578 for signed ranking over channel-position coordinates ($p = 0.0020$), and 0.7117 ± 0.0531 versus 0.6961 ± 0.0575 for magnitude ranking over whole pixels ($p = 0.0020$). Deletion is more sensitive to the convention. Under signed channel-position ranking, attribution produces a lower AUC than random ordering (0.6951 versus 0.7040, $p = 0.037$), whereas under magnitude-ranked whole-pixel perturbation it produces a slightly higher AUC (0.6979 versus 0.6956, $p = 0.014$). Because the two protocols differ in both ranking criterion and perturbation unit, this contrast reflects sensitivity to the evaluation convention as a whole rather than either factor in isolation, consistent with the considerations in Section 3.4.

E Gene-token Transformer on single-cell data

This experiment extends the neural evaluation of the FuzzyHead to single-cell gene-expression data using a Transformer (Vaswani et al., 2017) over gene tokens from FCGR3A+ monocytes in a peripheral blood mononuclear cell (PBMC) dataset. It is included because a ten-gene interferon-response signature is available as a reference for gene-level attribution (Grazioso et al., 2026).

Unlike the two main case studies, the attribution target here is the *base* gene-token Transformer rather than the fuzzy predictor. Integrated Gradients and the deletion and insertion analyses operate on the Transformer's own linear classifier over its 64-dimensional pooled CLS representation, and the attention weights used for comparison are taken from the same model. The TS head is fitted separately on the standardised CLS representation and determines which held-out cells enter the attribution analysis through its classification correctness. It therefore defines the evaluation population but is not itself the model being explained. The comparison below is consequently between Integrated Gradients and attention as explanations of the base Transformer, rather than explanations of the fuzzy predictor itself.

The experimental protocol differs from the main case studies in several aspects. We normalised and log-transformed counts before cell-type selection. Each seed uses a $75/25$ train/test split without a validation split or model selection, and the Transformer is trained for a fixed 60 full-batch epochs. The 256 genes with highest variance are selected using the training split only, and gene standardis-

ation is likewise fitted on training cells only. The fuzzy head uses prototype centres, width 2.5, and consequent learning rate 0.02. Integrated Gradients uses 20 steps, and deletion and insertion use 100 perturbation steps.

Across ten seeds, the Transformer's linear classifier reaches 0.9804 ± 0.0082 test accuracy, while the TS head fitted on the same representation reaches 0.9821 ± 0.0059, a paired difference of $+0.0017$ (Wilcoxon $p = 0.43$). Integrated Gradients recovers 5.31 ± 1.38 of the ten reference genes, compared with 3.63 ± 1.20 for attention, corresponding to a paired difference of 1.68 genes (Wilcoxon $p = 0.0098$; paired t-test $p = 0.0047$). Both rankings outperform random ordering under deletion and insertion, and Integrated Gradients outperforms attention on both measures ($p = 0.0020$). Their top-50 gene sets overlap by 19.8 ± 3.5 genes on average, indicating partial but far from complete agreement between the two explanation methods.

The reference signature consists of ten interferon-response genes identified in a prior work on the same dataset and cell type (Grazioso et al., 2026), where it was also compared with differential-expression results and known interferon-stimulated-gene biology. Because this reference was not obtained as an independent third-party annotation and is related to the same methodological line of work, we treat the result as complementary evidence rather than as a third counterpart to the clinician masks and treebank annotations used in the main case studies. No analogous attention comparison is available for SST-2 because its representation is obtained by mean pooling rather than from a dedicated pooling token.

F Decoder-based analysis of representation coordinates

When a representation model provides a decoder $g_\theta : \mathbb{R}^d \to \mathcal{X}$, FLARE can relate individual representation coordinates to features of the reconstructed output. This route is not used in the reported experiments, whose representation models do not provide a decoder. Still, we include it here to specify the decoder-based analysis referred to in Section 3.3.

For a representation z, the local effect of coordinate i is characterised through the decoder Jacobian,

$$J_{g_\theta}(z) = \nabla_z g_\theta(z) = \left[\frac{\partial \hat{x}_l}{\partial z_i}\right]_{l=1,\dots,D;\, i=1,\dots,d}, \qquad \hat{x} = g_\theta(z). \tag{8}$$

The column associated with coordinate i is obtained using a forward-mode Jacobian–vector product, avoiding construction of the full $D \times d$ Jacobian. We evaluate the derivative at reference representations, including class-wise and overall mean codes, and average across these points. Because this derivative describes only a local change, it is complemented by a finite traversal of the same coordinate. Starting from a fixed reference representation, coordinate i is varied across its observed range while the remaining coordinates are held fixed, and the resulting change in the decoded output is recorded. Agreement between the Jacobian-derived direction and the finite traversal is summarised by their signed Pearson correlation, providing a consistency measure between local sensitivity and the effect of a larger displacement along the same coordinate.

This analysis characterises a representation coordinate rather than an individual fuzzy rule: no rule index enters the Jacobian, traversal, or consistency calculation. It therefore provides an additional route for interpreting decoder-bearing representations, distinct from the input attribution used in the main experiments.

G Post-hoc rule-induction alternatives

FLARE also provides post-hoc methods for inducing fuzzy rules from an already-fixed representation matrix Z. We do not use these methods in the reported experiments, and they differ from the trainable TS head in Section 3.2. *Scatter* partitioning clusters Z with fuzzy C-Means (Bezdek, 1981; Bezdek et al., 1984) and induces a TS rule base through pyFUME (Fuchs et al., 2020), using the minimum t-norm for antecedent conjunction and weighted least squares for the consequents. *Grid* and *shared-set* partitioning instead define fuzzy sets coordinate-wise, using either a separate set family for each coordinate or a common family shared across coordinates, with zero-order class votes as consequents.

These methods define different fuzzy systems rather than alternative optimisation procedures for Equations 2–5. In particular, the product t-norm of Equation 3 and softmax-normalised firing of Equation 4 apply specifically to the trainable fuzzy head.

H REPRODUCIBILITY

The supplementary material includes the implementation and tracked per-seed artefacts underlying the reported aggregate results. Means, standard deviations, paired Wilcoxon and t-test p-values, and bootstrap intervals are computed from these per-seed values using the procedures described in the corresponding sections and captions. The Iris and Digits paired analyses are additionally recorded in `results/seed_paired_significance.txt`; the original neural experiments derive their paired tests from the tracked per-seed CSVs. The additional BUSI Grad-CAM++ comparison and duplicate-sensitivity analysis are accompanied by dedicated scripts and per-seed artefacts in the supplementary material.

The result CSVs do not store the complete resolved command-line configuration. For SST-2, the executed configuration fine-tunes all 22 encoder layers at learning rate 2×10^{-5} with batch size 32 for two epochs over ten seeds, as reported in Section 4.

Exact reproduction of the BUSI splits requires the dataset snapshot identified by the supplied manifest because its materialised filenames differ from those of other BUSI distributions and the split procedure operates on the ordered file list. This snapshot also contains one ultrasound image duplicated under benign and malignant entries; both map to the abnormal class in our binary task. The two copies cross the train/test boundary in five of the ten reported splits. Removing the test-side duplicate changes the ten-seed mean FuzzyHead accuracy from 0.9272 to 0.9270 and leaves every reported Wilcoxon conclusion unchanged. A grouped-split refit likewise has no material effect on the reported conclusions, so we retain the original runs.

The environment uses `torch` 2.7.0, `numpy` 2.4.4, `scipy` 1.17.1, `pyFUME` 0.2.25, and `simpful` 2.12.0. The latter two packages are required only by the post-hoc methods of Appendix G and therefore do not affect the numerical results reported in this paper. For the image experiments, cuDNN is disabled because the `torch` 2.7.0 CUDA 13.0 environment used for the reported runs produces a sub-library version mismatch during convolution. Re-execution in the same environment reproduced the stored per-seed results exactly; we do not claim bitwise reproducibility across different hardware, drivers, or library versions.