

Faithful regularised classification over linguistic fuzzy rule activations

S. A. Davidsen ¹ and M. Padmavathamma ²

^{1,2}*Sri Venkateswara University, Dept of Computer Science, Tirupati, India.*

¹*Hamisoke Labs, R&D, Fredericia, Denmark.*

soren@hamisoke.com, profpadma@yahoo.com

Abstract

Achieving competitive accuracy while exposing exact prediction computations remains difficult for tabular classifiers. We introduce and evaluate FuzzyRuleSVM, a regularised classifier that maps each instance into linguistic fuzzy antecedent activations and trains a linear SVM in that bounded feature space. Unlike classical fuzzy SVMs, which fuzzify sample weights in the objective, FuzzyRuleSVM substitutes the feature space: every dimension is a named antecedent, and its coefficient supplies the signed class effect. The decision function therefore decomposes exactly into named rule-activation contributions. This is algebraic fidelity, not cognitive interpretability. On twenty-one benchmark and domain-specific datasets, including Steel Plates Faults as an industrial manufacturing case, FuzzyRuleSVM achieves competitive balanced accuracy while exposing exact rule-contribution decompositions, runtime profiles, and compactness/stability metrics. The contribution is faithful regularised classification over linguistic rule activations.

Keywords: Fuzzy rule classification, support vector machine, regularized classification, linguistic rule activation, algebraic explanation.

2020 MSC: 68T05, 68T37, 62H30.

1 Introduction

Accuracy and interpretability occupy opposite ends of the model-complexity axis in most machine learning surveys [4, 31]. Recent XAI work sharpens this distinction by separating interpretable representations from post-hoc explanation interfaces and by recommending ante-hoc models when the computation itself must be audited [29, 34]. Decision trees and rule lists are transparent but rarely match kernel or ensemble methods on accuracy; deep networks require post-hoc approximations such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) that explain a surrogate rather than the model itself. Rudin [30] argues that such methods are unreliable for high-stakes decisions, and Slack et al. [33] show that LIME and SHAP can be deceived to hide model bias from auditors. Fuzzy rule-based classifiers occupy a middle ground: they reason in linguistic terms (*IF glucose is high AND BMI is high THEN positive*) and remain active in diagnostic XAI research [2, 7, 32].

Support vector machines (SVMs) are among the most reliable non-ensemble baselines on structured data [9, 27], but their explanations are intrinsically opaque: a prediction decomposes into a sum over training-set support vectors, each of which is a raw input example rather than a human concept. The fuzzy SVM literature addresses a different problem: it assigns fuzzy membership scores to training samples to make the SVM more robust to outliers and class imbalance [5, 23, 36]. Interpretability is not the goal.

We take the opposite direction. Instead of adding fuzzy logic to the SVM training objective, we replace the feature space entirely. We first generate a finite bank of linguistic fuzzy rule antecedents from the training data, then compute, for every instance, the vector of rule firing strengths. A regularised linear SVM trained on this *rule activation matrix*

Corresponding Author: S. A. Davidsen

Received: July 2026; Revised: September 2026; Accepted: October 2026.

<https://doi.org/10.22111/ijfs.2026.56189.10009>, <https://ijfs.usb.ac.ir>, Print ISSN: 1735-0654

Table 1: Claim boundaries used in this revision. The paper evaluates algebraic and statistical properties of the representation; it does not claim a human-interpretability advantage without a user study.

Claim	Evidence reported	Boundary
Exact explanation fidelity	Full decomposition error $ f(x) - (b + \sum_k \beta_k \phi_k(x)) $ and top- k partial-display agreement	Fidelity is algebraic, not proof that users understand or trust the rules.
Competitive accuracy	Nested-CV comparisons against fuzzy, standard, glass-box, and GBDT baselines	The method is not predictively dominant over every baseline.
Auditable linguistic units	Named fuzzy antecedents, support-rule counts, average rule length, and industrial Steel Plates examples	Rules are predictive associations, not causal mechanisms or deployment validation.
Computational scope	Symbolic complexity, runtime, memory, and stage profiling	Length-2 enumeration grows quadratically and requires screening or adaptive discovery at high d .

inherits the max-margin geometry of standard SVMs while every feature in the margin computation is a named rule-activation contribution. The practical advantage is an exact audit trail: partitions and candidate rules are fitted inside each training fold, every retained column has a linguistic name, and the final optimizer remains a standard regularised linear SVM. The resulting decomposition has exact algebraic fidelity; whether users find the rule lists cognitively manageable remains an empirical human-computer interaction (HCI) question outside this study.

This architectural inversion (using fuzzy rules as features for an SVM rather than extracting rules from an SVM) is the core contribution. Prior work [11, 37] links SVM learning to fuzzy rule optimisation but does not build the decision function directly over an explicit linguistic rule activation space. The resulting explanation is therefore not post-hoc: it is the native computation of the classifier.

We are not aware of a prior evaluated classifier in the fuzzy-SVM or SVM-for-fuzzy-rules literature that combines these commitments: graded linguistic rule activations generated within the training fold, a regularised linear max-margin fit over those activations, and an exact additive decomposition over named antecedents. The contribution is the rule-space formulation and its evaluation, not a new optimiser. We make five concrete contributions:

1. **Data-adaptive rule activation map.** For each numeric feature we fit three quantile-anchored membership functions (low, medium, high) from training data and form conjunctive fuzzy antecedents up to a prescribed length, scoring and retaining the most discriminative.
2. **Sparsity-tuned classification in rule space.** A LinearSVC is trained on the rule activation matrix, with L1/L2 selection tuned by inner CV. L1 produces exact zeros; L2 fits are generally dense. The max-margin solver is an implementation choice, not a unique predictive advantage.
3. **Exact full-model additive decomposition.** Because the classifier is linear in rule space, every prediction decomposes exactly as $f(x) = \sum_k \beta_k \phi_k(x) + b$; each term is a named, auditable rule contribution, not an approximation.
4. **Comparative evaluation across twenty-one datasets.** We evaluate FuzzyRuleSVM against FURIA (primary fuzzy-rule baseline), EBM and RuleFit (modern interpretable tabular models), standard SVM, logistic regression, random-forest, histogram-gradient-boosting, XGBoost, and membership-space regularised baselines, using nested cross-validation with Wilcoxon signed-rank tests and Holm-Bonferroni correction.
5. **Algebraic interpretability analysis of compactness, stability, and efficiency.** We measure active support rules, average rule length, top- k sign agreement, exact decomposition error, cross-fold Jaccard stability, stage-level runtime, and memory use. These are properties that fidelity alone does not capture, and they reveal an association between dimensionality, explanation size, and compute.

We note that an L1-logistic regression baseline trained on the same atomic fuzzy membership features slightly exceeds FuzzyRuleSVM in mean balanced accuracy (0.833 vs. 0.830; $p_{\text{raw}} = 0.679$, $p_{\text{Holm}} = 1.000$). This null difference does not support a unique max-margin accuracy advantage. The strongest supported contribution is instead the exact named decomposition over conjunctive rule activations.

Our implementation of the classifier is available for use from <https://github.com/sorend/fysvm>, as well as pypi repositories. See the repository for details.

The paper is organised as follows. Section 2 places our approach in the context of existing fuzzy SVMs and fuzzy rule classifiers. Section 3 defines the model formally. Section 4 reports the comparative evaluation. Section 5 discusses implications and limitations. Section 6 concludes.

Table 2: Where fuzzy, rule, or interpretable objects enter comparable model families. FuzzyRuleSVM is positioned as a representation substitution: fuzzy rules are the native columns supplied to an otherwise standard regularised linear SVM.

Family	Object used	Where it enters	Difference from FuzzyRuleSVM
Fuzzy SVMs	Fuzzy sample memberships	Hinge-loss weights	Explanations still decompose over support vectors.
Fuzzy rule classifiers	Induced linguistic rule base	Native classifier/voting system	Training is greedy, association-rule based, or evolutionary rather than regularised max-margin over fixed activations.
SVM-compatible fuzzy systems	Rules selected or ranked by margin objectives	Rule construction or ranking	The direction runs from SVMs to rules, not from a fixed rule-activation map to an SVM.
RuleFit and rule ensembles	Crisp tree-derived rules	Sparse linear model features	Rule activations are binary and not linguistic fuzzy degrees.
EBM/GA2M	Additive feature or interaction shapes	Native additive terms	Explanations are shape functions, not named fuzzy antecedents.
FuzzyRuleSVM	Graded fuzzy antecedents	Native feature space for LinearSVC	Each margin term is a named rule activation with a signed coefficient.

2 Related work

The closest prior work falls into four groups: fuzzy modifications of SVM training, fuzzy rule classifiers, sparse rule-based tabular models, and additive glass-box models. The distinction that matters for this paper is not whether a method contains fuzzy sets or rules, but where those objects enter the computation: as sample weights, induced rule bases, post-hoc explanation units, additive feature effects, or the native feature space of a max-margin classifier.

Fuzzy SVMs and SVM-compatible fuzzy models. Lin and Wang [23] introduced the first fuzzy SVM by assigning each training sample a scalar membership value that scales its contribution to the hinge loss. Samples near the decision boundary or with unclear class membership receive lower weights, improving robustness to noise and outliers. Subsequent work extended this weighting strategy to class-imbalance correction [5], kernel-induced clustering-based membership functions [36], and intuitionistic or shadowed-set representations. In these models, fuzzy logic changes how much each sample influences the margin; it does not change the representation on which the decision function is evaluated. The prediction still decomposes over support vectors rather than over linguistic concepts.

A second SVM–fuzzy thread uses margin objectives to learn, select, or rank fuzzy rules. Chen and Wang [11] connect fuzzy rule systems to reproducing-kernel Hilbert spaces and propose support-vector learning for fuzzy classifiers. Zhou and Gan [37] use an L2-SVM to select and rank fuzzy rules in high-dimensional space. ANFIS [21] and later fuzzy neural networks, including support-vector-based variants [24], combine fuzzy memberships with adaptive numeric layers. These approaches move from SVMs or adaptive networks toward rules. FuzzyRuleSVM takes the reverse direction: it fixes a finite bank of conjunctive linguistic antecedents first, then trains a regularised max-margin classifier over their activations.

Fuzzy rule classifiers. FURIA [20] induces unordered fuzzy rules with a RIPPER-like FOIL information-gain criterion, fuzzifies antecedent bounds, and applies rule stretching for uncovered instances. FARC-HD [2] mines fuzzy association rules with Apriori-style support/confidence thresholds, prunes them with weighted-relative-accuracy scoring, and tunes the surviving rules with a genetic algorithm (GA). IVTURS [32] extends FARC-HD by replacing type-1 fuzzy partitions with interval-valued type-2 sets, while Elkano et al. [14] extend FARC-HD to multiclass problems using n -dimensional overlap functions. Recent diagnostic-XAI reviews continue to treat interpretable fuzzy rule bases as useful domain-facing objects, especially in medicine [7, 17]. These systems are the most direct fuzzy-rule context because they also produce linguistically labelled rules. Their training logic, however, is greedy, association-rule-based, or evolutionary rather than a regularised max-margin objective over a fixed rule activation matrix.

Sparse rule ensembles and optimal rule lists. RuleFit [16] generates rules from gradient-boosted trees and learns a sparse linear model over those binary rule indicators; it is the closest non-fuzzy relative of FuzzyRuleSVM. Bayesian Rule Lists [22] and CORELS [3, 6] optimise compact *ordered* rule lists under probabilistic or formal optimality objectives. Optimal Sparse Decision Trees [19] minimise a global sparsity criterion over tree structures. All of these produce crisp, binary-valued rule activations: a rule either fires completely or not at all. FuzzyRuleSVM differs on two counts. Its rule activations are graded linguistic degrees in $[0, 1]$, so an active rule can contribute partially to the margin rather than firing all-or-nothing. Its training objective is a regularised max-margin criterion rather than a probabilistic or combinatorial search, which connects it directly to the classification geometry of the SVM.

Glass-box additive tabular models. GA2M [25] and its application to healthcare [8] fit additive shape functions over individual features and selected pairwise interactions; the Explainable Boosting Machine (EBM) [26] industrialises this approach via cyclic boosting with automatic interaction detection. Neural Additive Models [1] and NODE-GAM [10] replace the shape functions with neural networks, preserving additivity while gaining flexibility. These models explain predictions as sums of per-feature (or per-interaction) contributions visualised as learned shape curves. FuzzyRuleSVM instead explains predictions as sums of named fuzzy antecedent activations: “*glucose is high AND BMI is high*” fires at 0.81 with weight +0.90, contributing +0.73 to the margin. The explanation is a ranked list of verbalizable antecedent statements with learned signed effects rather than a collection of univariate shape plots, a format that may be natural in domains that already use linguistic rules; whether domain experts use it more accurately requires a separate user or expert study.

Research gap. The gap is therefore not the absence of fuzzy classifiers, rule classifiers, or max-margin fuzzy hybrids considered separately. It is the absence of a classifier combining graded linguistic rule activations, regularised max-margin training, and an exact additive decomposition over named antecedents. Classical FSVMs [5, 23, 36] fuzzify sample weights but leave explanations tied to support vectors. SVM-compatible fuzzy models [11, 24, 37] use margins to construct or rank rules, but the reduction runs from SVMs to rules rather than from rules to an SVM. FURIA, FARC-HD, and IVTURS induce linguistic rules but do not train a max-margin classifier in rule activation space. Sparse rule ensembles use crisp activations, while additive glass-box models sum feature effects rather than conjunctive fuzzy antecedents. FuzzyRuleSVM addresses this intersection without introducing a new optimiser; L1 regularisation provides rule selection when chosen.

3 The regularised max-margin fuzzy rule machine

Fig. 1 presents the complete six-stage processing pipeline. Starting from raw feature values, the model computes per-feature linguistic membership degrees, generates and scores conjunctive fuzzy antecedents, collects their firing strengths into a rule activation vector $\Phi(x)$, learns a regularised max-margin classifier over that vector, and decomposes each prediction as an exact signed sum of named rule-activation contributions. Rule generation and SVM training are independent stages: the SVM itself is an unchanged linear classifier with L1/L2 regularisation selected by inner cross-validation; only its input feature space is linguistically structured. The subsections below define each stage formally.

3.1 Fuzzy partitions and rule generation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$; $i \in \{1, \dots, n\}$ indexes samples and $j \in \{1, \dots, d\}$ indexes numeric features. For arbitrary binary labels, the implementation maps the second sorted class to +1. Let $\mathcal{A} = \{\text{low}, \text{medium}, \text{high}\}$. For each feature j , we fit three membership functions from the 5th, 50th, and 95th percentiles of the training fold, denoted $a_j \leq b_j \leq c_j$. For $z = x_j$, the default non-degenerate memberships are

$$\mu_{j,\text{low}}(z) = \begin{cases} 1, & z \leq a_j, \\ (b_j - z)/(b_j - a_j), & a_j < z < b_j, \\ 0, & z \geq b_j, \end{cases}$$

$$\mu_{j,\text{medium}}(z) = \min \left\{ \begin{cases} 0, & z \leq a_j, \\ (z - a_j)/(b_j - a_j), & a_j < z < b_j, \\ 1, & z \geq b_j, \end{cases}, \begin{cases} 1, & z \leq b_j, \\ (c_j - z)/(c_j - b_j), & b_j < z < c_j, \\ 0, & z \geq c_j, \end{cases} \right\},$$

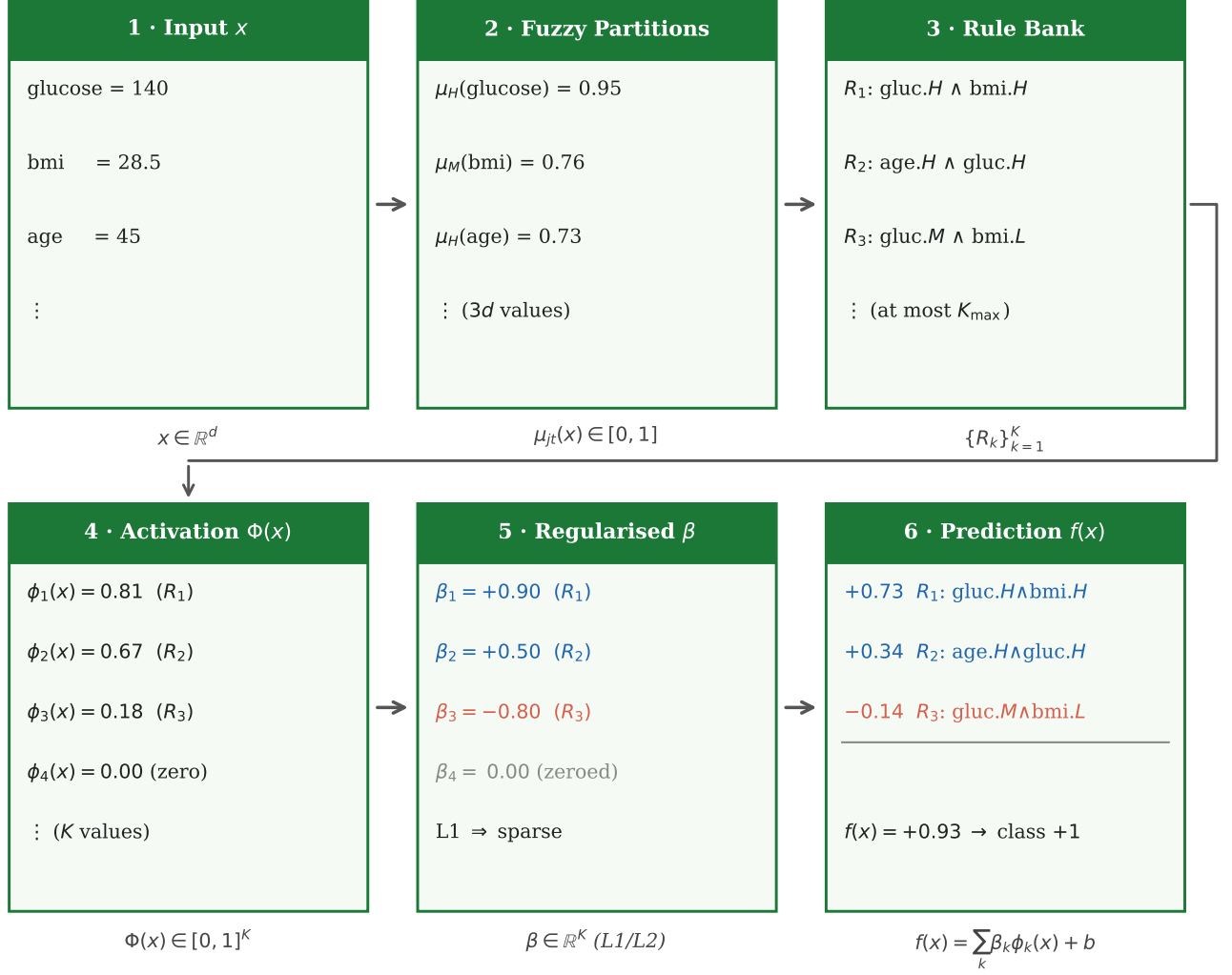


Figure 1: Architecture of FuzzyRuleSVM as a six-stage pipeline. (1) Raw input values are (2) mapped to per-feature linguistic membership degrees via quantile-anchored low/medium/high fuzzy partitions (H = high, M = medium, L = low). (3) Conjunctive fuzzy antecedents are generated, scored, and at most $K_{\max}$ retained; the realised count is K . (4) Each firing strength $\phi_k(x)$ aggregates its memberships, giving $\Phi(x) \in [0, 1]^K$. (5) A regularised max-margin classifier assigns weights β_k ; L1/L2 selection is tuned by inner cross-validation (CV). (6) The prediction decomposes exactly as $f(x) = \sum_k \beta_k \phi_k(x) + b$. Values are illustrative and consistent with Section 3.6.

and

$$\mu_{j,\text{high}}(z) = \begin{cases} 0, & z \leq b_j, \\ (z - b_j)/(c_j - b_j), & b_j < z < c_j, \\ 1, & z \geq c_j. \end{cases}$$

Thus low and high are shoulder functions: values below the lower anchor remain fully low, and values above the upper anchor, remain fully high. Values outside the training range are not extrapolated beyond $[0, 1]$; all memberships are clipped to that interval.

Duplicate quantiles produce degenerate shoulders rather than undefined divisions. If an increasing segment has identical endpoints, its value is the step function $\mathbb{I}\{z \geq \text{endpoint}\}$; if a decreasing segment has identical endpoints, its value is $\mathbb{I}\{z \leq \text{endpoint}\}$. The medium membership remains the minimum of the corresponding increasing and decreasing components, so a constant feature with $a_j = b_j = c_j$ assigns medium membership one only at the shared value and zero elsewhere. The estimator operates on numeric arrays; missing values are imputed before fitting in the evaluation pipeline, and nonnumeric features must be encoded numerically before fitting or are rejected by the numeric input validation. The instance x is represented by the membership triplet

$$\mu_{j,\text{low}}(x_j), \quad \mu_{j,\text{medium}}(x_j), \quad \mu_{j,\text{high}}(x_j).$$

We generate conjunctive candidates by crossing feature-term pairs. Let ℓ_k denote the number of antecedents in rule R_k , with $1 \leq \ell_k \leq L$, distinct ordered indices $j_1 < \dots < j_{\ell_k}$, and terms $A_{k\ell} \in \mathcal{A}$. A candidate rule is

$$R_k : \text{IF } x_{j_1} \text{ is } A_{k1} \text{ AND } \dots \text{ AND } x_{j_{\ell_k}} \text{ is } A_{k\ell_k}.$$

The rule does not store a fixed consequent. Its learned “THEN” part is the signed margin contribution $\beta_k \phi_k(x)$: positive coefficients support the positive one-vs-rest class, negative coefficients oppose it, and the magnitude depends on both β_k and the instance-specific activation. During candidate selection, each antecedent is scored on the training fold by

$$|\bar{\phi}_k^+ - \bar{\phi}_k^-| \bar{\phi}_k / \ell_k,$$

where $\bar{\phi}_k^+$ and $\bar{\phi}_k^-$ are the class-conditional mean firing strengths and $\bar{\phi}_k = n^{-1} \sum_i \phi_k(x_i)$ is fuzzy coverage. Rules below $\rho_{\min}$ are discarded and at most $K_{\max}$ candidates are retained; write $K \leq K_{\max}$ for the realised bank size. If no candidate passes coverage, the implementation falls back to the ranked single-feature candidates. The reported harness uses $\rho_{\min} = 0.01$.

3.2 Rule activation matrix and decision function

The firing strength of rule R_k on instance x is

$$\phi_k(x) = T(\mu_{j_1, A_{k1}}(x_{j_1}), \dots, \mu_{j_{\ell_k}, A_{k\ell_k}}(x_{j_{\ell_k}})),$$

where T is an aggregation operator. The default is the minimum t-norm; the product t-norm and a smooth minimum are implemented alternatives. The smooth minimum is not a t-norm because it does not satisfy $T_\eta(v, 1) = v$; it uses

$$T_\eta(v_1, \dots, v_{\ell_k}) = -\eta \log \left(\frac{1}{\ell_k} \sum_{m=1}^{\ell_k} e^{-v_m/\eta} \right),$$

with fixed temperature $\eta = 0.1$. Stacking all K rules gives the rule activation vector

$$\Phi(x) = [\phi_1(x), \phi_2(x), \dots, \phi_K(x)].$$

All implemented aggregations are clipped to $[0, 1]$, so $\Phi(x) \in [0, 1]^K$. For a rule length limit L , the unfiltered candidate count is $\sum_{\ell=1}^L 3^\ell \binom{d}{\ell}$; for $L = 2$ this is $3d + 9 \binom{d}{2}$. A linear max-margin classifier trained on $\Phi(\cdot)$ defines

$$f(x) = \sum_{k=1}^K \beta_k \phi_k(x) + b.$$

The binary prediction is $\hat{y} = +1$ when $f(x) \geq 0$ and -1 otherwise.

3.3 Class-weighted linearSVC training and sparsity

Let $q_k = 1 + \tau(\ell_k - 1)$ be the rule-length factor. In the reported activation coordinates, **LinearSVC** solves the class-weighted squared-hinge problem

$$\begin{aligned} \min_{\beta, b} \quad & \Omega_p(q \odot \beta, b) \\ & + C \sum_{i=1}^n w_i [\max(0, 1 - y_i(\beta^\top \Phi(x_i) + b))]^2, \\ \Omega_1(q \odot \beta, b) = & \sum_k q_k |\beta_k| + |b|, \\ \Omega_2(q \odot \beta, b) = & \frac{1}{2} \left(\sum_k q_k^2 \beta_k^2 + b^2 \right). \end{aligned} \tag{1}$$

where $p \in \{1, 2\}$ is selected by inner CV and $w_i = n/(2n_{y_i})$, with n_{y_i} the size of sample i 's binary class. This is the **LinearSVC** convention for `class_weight='balanced'`, `loss='squared_hinge'`, `dual=False`, and default unit intercept scaling. Under L1, many β_k are exactly zero. Under L2, coefficients are usually dense, so active-rule counts use $|\beta_k| > 10^{-10}$ and denote non-negligible rather than structurally sparse coefficients.

The implementation realises q_k , by dividing activation column k by q_k , before fitting and transforming the learned coefficient back to the original activation scale. The reported experiments use $\tau = 0.35$. This weighted regularisation makes the longer rules more expensive, while leaving each reported $\beta_k \phi_k(x)$ on the original activation scale.

The implementation also supports two optional high-dimensional discovery paths, neither used in the primary fixed-budget result. The *screened* path applies analysis-of-variance (ANOVA) F or mutual-information screening within a training fold before enumerating length-2 rules. The *column-generation-inspired* path starts from covered length-1 rule, solves an auxiliary L1-hinge linear program (LP), and proposes length-2 rules with a large proxy reduced gradient $|r_k| = |\sum_i \alpha_i y_i \phi_k(x_i)|$. Its product-activation shortlist costs $O(d^2 n)$ per iteration; candidates are then rescored with the configured aggregation. The final estimator remains the squared-hinge **LinearSVC** in (1), so pricing is a discovery heuristic rather than an optimality certificate. Section 4.10 evaluates only the primary atomic-rule budget limitation; adaptive discovery remains future work.

3.4 Computational complexity

Let $M_L = \sum_{\ell=1}^L 3^\ell \binom{d}{\ell}$ denote the exhaustive candidate pool, $\bar{\ell}$ the average retained rule length, and $K \leq K_{\max}$ the retained activation columns. Table 3 separates the one-time fit stages from prediction. The analysis uses conservative bounds: training-fold quantiles are reported as $O(nd \log n)$ even though selection-based implementations can do better. After candidate generation, the final optimisation is delegated to liblinear through **LinearSVC**; its wall-clock cost depends on solver iterations, but the data matrix supplied to it has shape $n \times K$.

Table 3: Symbolic computational costs for the default enumeration path. The optional pricing path is included only to delimit the implemented high-dimensional extension; it is not a final optimality certificate.

Stage	Time	Memory / note
Partition fitting	$O(nd \log n)$	Three quantile anchors per feature.
Membership tensor	$O(nd)$	$O(3nd)$ floating-point memberships.
Exhaustive candidate count	$M_L = \sum_{\ell=1}^L 3^\ell \binom{d}{\ell}$	For $L = 2$, $3d + 9 \binom{d}{2}$.
Candidate scoring	$O(nM_L \bar{\ell})$	Can be streamed before truncation; stores candidate metadata.
Retained activation matrix	$O(nK \bar{\ell})$	$O(nK)$ dense activation matrix.
LinearSVC fit	solver-dependent over $n \times K$	Standard liblinear optimisation; no new SVM solver.
Prediction	$O(n_{\text{test}} K \bar{\ell}) + O(n_{\text{test}} K)$	Compute rule activations and margins.
Optional pricing iteration	approximately $O(d^2 n)$ LP/shortlist costs	plus Heuristic column discovery for high-dimensional settings.

3.5 Multiclass extension

The base estimator is binary. For a G -class problem with labels $\mathcal{Y} = \{1, \dots, G\}$, we use one-vs-rest decomposition. For each class c , a binary FuzzyRuleSVM is fitted with labels

$$y_i^{(c)} = \begin{cases} +1, & y_i = c, \\ -1, & y_i \neq c, \end{cases}$$

yielding K_c class-specific rules $R_k^{(c)}$, activations $\phi_k^{(c)}(x)$, coefficients $\beta_k^{(c)}$, and margin

$$f_c(x) = \sum_{k=1}^{K_c} \beta_k^{(c)} \phi_k^{(c)}(x) + b_c.$$

Prediction uses $\hat{y} = \arg \max_{c \in \mathcal{Y}} f_c(x)$. The additive decomposition is exact separately for each class margin; the implementation can report the predicted-class margin and its gap to the largest competing margin. All multiclass experiments use `scikit-learn`'s `OneVsRestClassifier` wrapper.

3.6 Exact additive prediction decomposition

Because every activation $\phi_k(x)$ is named, the prediction decomposes additively without any further approximation:

$$f(x) = \beta_1 \phi_1(x) + \beta_2 \phi_2(x) + \dots + \beta_K \phi_K(x) + b.$$

A prediction on a test instance from the Pima Diabetes dataset, for example, reads:

“IF glucose is high AND BMI is high” fires at 0.81, weight +0.90, contribution +0.73.
“IF age is high AND glucose is high” fires at 0.67, weight +0.50, contribution +0.34.
“IF glucose is medium AND BMI is low” fires at 0.18, weight −0.80, contribution −0.14.
Net margin: +0.93; class: high risk.

This is not a post-hoc surrogate; it is the model’s own computation. Algebraic fidelity is exact by construction: the full nonzero decomposition $f(x) = \sum_k \beta_k \phi_k(x) + b$ reproduces the decision value without approximation. Top- k displays (such as the waterfall in Fig. 2) are partial summaries of this full decomposition; exactness holds only when all nonzero rule contributions are included.

4 Experiments

4.1 Datasets and evaluation protocol

We evaluate on twenty-one benchmark and domain-specific datasets drawn from the UCI ML Repository [13] and `scikit-learn`’s built-in collection [27]. The datasets are predominantly medical: breast cancer (three variants), heart disease (four datasets), Parkinson’s disease (two variants), diabetes, diabetic retinopathy, liver disease, dermatology, SPECT/SPECTF cardiac imaging, and mammographic mass. Steel Plates Faults adds an industrial manufacturing-defect case. General-domain datasets (Iris, Wine, Digits) round out the benchmark. Sample sizes range from 106 (Breast Tissue) to 1941 (Steel Plates Faults) and feature dimensionality from 3 (Haberman’s Survival) to 754 (Parkinson’s Disease Classification). Tasks include fifteen binary and six multiclass problems.

We apply five-fold stratified outer CV with shuffling and root seed 0. Within each outer training fold, three-fold inner CV selects hyperparameters by mean balanced accuracy, using macro-F1 as a tie-breaker in the main fuzzy, standard, and modern comparisons; the membership and budget analyses use balanced accuracy with deterministic grid-order ties. Missing values are imputed with the training-fold median, and all fuzzy partitions are fitted only on training data. A deterministic registry lookup keyed by the canonical dataset slug fixes dataset seeds. For G classes, balanced accuracy is $G^{-1} \sum_g \text{TP}_g / (\text{TP}_g + \text{FN}_g)$ and macro-F1 is $G^{-1} \sum_g \text{F1}_g$, with undefined class F1 set to zero.

4.2 Statistical analysis

The paired unit for inferential tests is the per-dataset outer-CV mean, not individual folds. For two-model comparisons, we use Wilcoxon signed-rank tests [35] on the vector of per-dataset differences. For families with more than two models, we report Friedman omnibus tests [12, 15, 28]. Holm–Bonferroni correction [18] is applied separately within each comparison family and metric; both p_{raw} and p_{Holm} are reported when pairwise tests are used. Bootstrap intervals in figures are descriptive summaries of fold variability and are not used as the primary inferential test.

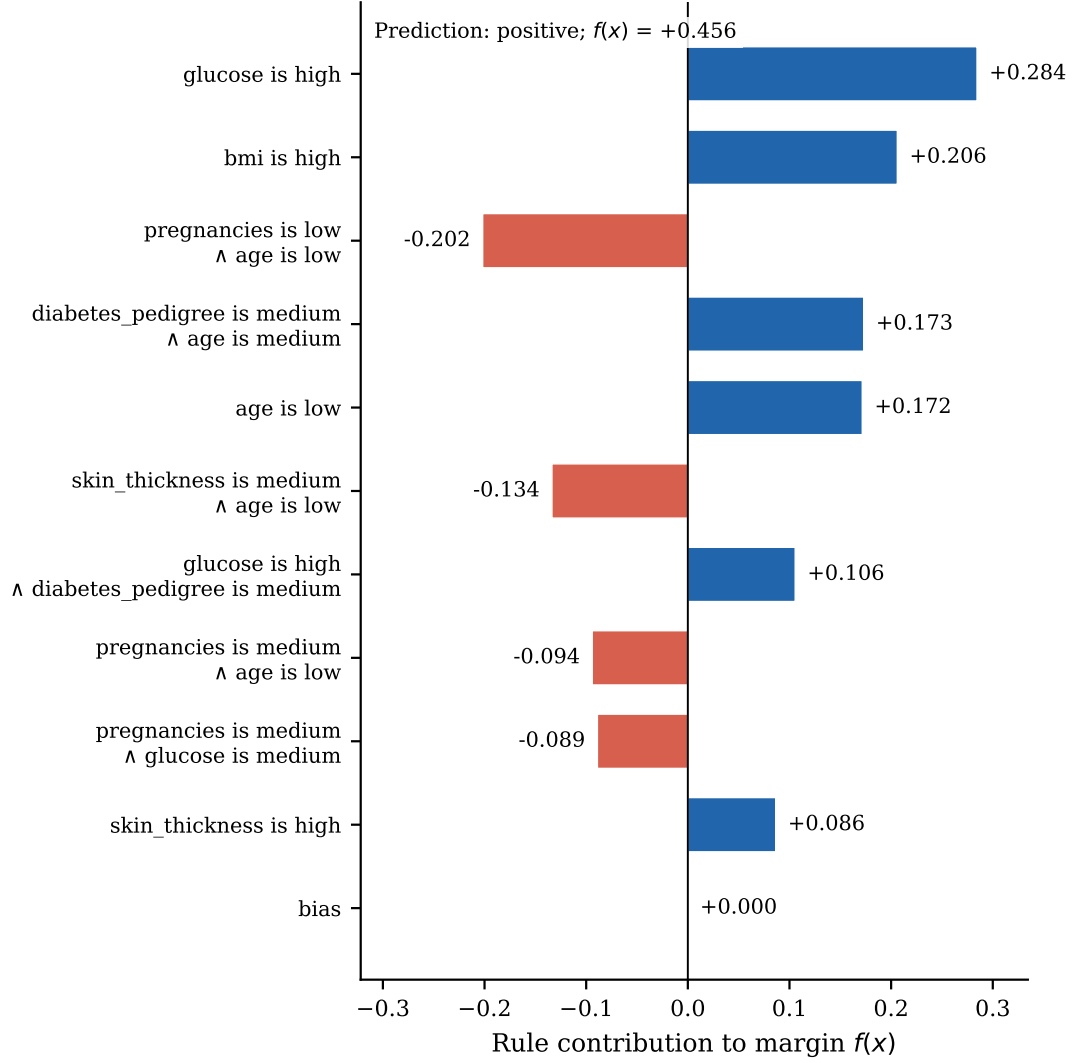


Figure 2: Rule contribution waterfall for a median-complexity Pima Diabetes test instance. Each displayed bar is $\beta_k \phi_k(x)$; blue bars support the positive class and red bars oppose it. The bars are a top- k partial summary; all nonzero contributions plus the bias, not only the displayed bars, sum to the net margin exactly.

Table 4: Benchmark dataset composition. n and d are prepared sample and feature counts; G is the number of classes. Inputs are numeric or encoded, with median imputation fitted inside each outer-training fold. Original SPECT files are pooled; Cleveland targets > 0 and nonnormal Arrhythmia labels are binarized; Steel one-hot fault columns become class labels.

Dataset	n	d	G	Source
Breast Cancer Diagnostic	569	30	2	sklearn
Breast Cancer Original	699	9	2	UCI
Mammographic Mass	961	5	2	UCI
Breast Tissue	106	9	6	UCI
Heart Disease Cleveland	303	13	2	UCI
Statlog Heart	270	13	2	UCI
SPECT Heart	267	22	2	UCI
SPECTF Heart	267	44	2	UCI
Pima Indians Diabetes	768	8	2	OpenML/UCI
Diabetic Retinopathy Debrecen	1151	19	2	UCI
Parkinsons	195	22	2	UCI
Parkinson’s Disease Classification	756	754	2	UCI
Indian Liver Patient Dataset	583	10	2	UCI
Dermatology	366	34	6	UCI
Haberman’s Survival	306	3	2	UCI
Vertebral Column 2-Class	310	6	2	UCI
Arrhythmia Binary	452	279	2	UCI
Iris	150	4	3	sklearn
Wine	178	13	3	sklearn
Digits	1797	64	10	sklearn
Steel Plates Faults	1941	27	7	UCI

4.3 Baselines

Fuzzy rule classifiers. We use FURIA as the primary fuzzy-rule baseline because it is a direct linguistic fuzzy-rule classifier and its greedy induction is not governed by the genetic-algorithm budgets that complicate FARC-HD and IVTURS reproduction. FuzzyRuleSVM and FURIA use the same outer folds, training-fold median imputation, fold-local fuzzy partitions where relevant, and inner selection.

FURIA [20] is implemented via the Python port in `asieriko/ex_fuzzy_farchd`, with `min_no` $\in \{1, 2, 3\}$. A sanity check against published FURIA results found close values on Iris, Wine, Wisconsin Breast Cancer, and Heart Statlog, but a lower Pima result (0.712 balanced accuracy versus 0.747 published accuracy). Because the original reports plain accuracy under repeated random splits whereas ours uses balanced accuracy under nested CV, this is a practical comparison to the Python port, not proof of equivalence to Java/Weka.

FARC-HD [2] and IVTURS [32] remain important fuzzy-rule systems and are discussed in Related Work and Limitations. We do not use reduced-budget FARC-HD or IVTURS runs as main empirical evidence, because their canonical KEEL settings use 20 000 GA fitness evaluations with population 50 and the available Python approximation does not reproduce the complete reference IVTURS of a reasoning procedure. A fair full-budget reproduction, is therefore outside the present revision rather than a hidden negative result.

Standard classifiers. We report a standardised linear SVM, radial-basis-function (RBF) SVM, L2/L1/elastic-net logistic regression, random forest, histogram gradient boosting, and XGBoost. The RBF grid spans C and γ ; tree ensembles are evaluated without scaling. These baselines measure whether the exact rule-space decomposition comes at a predictive cost relative to mainstream and modern GBDT classifiers.

FuzzyRuleSVM configuration. All primary grids tune $C \in \{0.3, 1.0, 3.0\}$ and $p \in \{1, 2\}$, use $L = 2$ for $d \leq 32$ and $L = 1$ otherwise, and set $K_{\max} = K_0 = \min(256, \max(24, 3d))$. This fixed policy supplies the fuzzy, standard, and membership comparisons. The separately reported modern-baseline family tunes $K_{\max} \in \{K_0, 512\}$ for $d \leq 100$; above 100 features it screens to 16 or 32 features by ANOVA, permits length-2 rules, and tunes $K_{\max} \in \{64, 128, 256, 512, 1024, 2048\}$. All configurations use $\rho_{\min} = 0.01$, minimum t-norm, 5/50/95 anchors, and $\tau = 0.35$. Family-specific FuzzyRuleSVM means are therefore reported separately rather than pooled.

Table 5: Main predictive results. BA = balanced accuracy; FRS = FuzzyRuleSVM. Family-specific FRS reference BA is 0.834 for FURIA (19 matched datasets), 0.830 for standard and membership comparisons, and 0.829 for modern comparisons; positive Δ favours that reference. The final column reports $p_{\text{raw}}/p_{\text{Holm}}$.

Comparison family	Model	BA	Macro F1	Δ vs. FRS	W/T/L; $p_{\text{raw}}/p_{\text{Holm}}$
This work	FuzzyRuleSVM	0.830	0.817	—	—
Fuzzy rules	FURIA [20]	0.782	0.782	+0.052	19/0/0; < .001/ < .001
Standard	RBF SVM	0.836	0.822	−0.005	9/1/10; .376/1.000
Standard	Linear SVM	0.826	0.807	+0.004	10/0/10; .571/1.000
Standard	Logistic L2	0.826	0.809	+0.004	9/0/11; .837/1.000
Modern interpretable	EBM [26]	0.838	0.825	−0.009	10/0/10; .784/.784
Modern interpretable	RuleFit [16]	0.796	0.794	+0.033	17/0/3; .002/.005
Membership ablation	MembershipSVM	0.829	0.815	+0.001	13/1/6; .107/.322
Membership ablation	MembershipLogisticL1	0.833	0.817	−0.002	9/2/9; .679/1.000

Table 6: Generated standard and GBDT baseline summary over the available recommended-comparison artifacts.

Model	Balanced accuracy	Macro F1	Fit time (s)
RBF SVM	0.833	0.816	0.014
FuzzyRuleSVM	0.826	0.810	0.458
Elastic-Net Logistic Regression	0.825	0.803	0.131
Logistic Regression L2	0.822	0.802	0.029
Logistic Regression L1	0.820	0.797	0.152
Linear SVM	0.815	0.796	0.017
Random Forest	0.810	0.809	1.032
XGBoost	0.804	0.806	0.235
Histogram Gradient Boosting	0.797	0.800	0.696

4.4 Main predictive results

Table 5 consolidates the retained benchmark comparisons from the prior full artifact set, while Table 6 is regenerated from the current comparison artifacts after adding Steel Plates Faults and XGBoost. The cleanest fuzzy-rule comparison is FURIA, because reduced-budget FARC-HD and IVTURS results are omitted from main claims unless canonical settings are reproduced.

Against standard classifiers, FuzzyRuleSVM is competitive rather than dominant. RBF SVM has the highest family mean (0.836 vs. 0.830), but no pairwise comparison survives Holm correction. In the adaptive modern family, balanced EBM has the highest mean (0.838 vs. 0.829), with a 10/0/10 split and $p_{\text{Holm}} = 0.784$. FuzzyRuleSVM outperforms RuleFit (0.829 vs. 0.796; $p_{\text{raw}} = 0.002$, $p_{\text{Holm}} = 0.005$). The membership ablation in Section 4.7 further narrows the claim.

Fig. 3 preserves the dataset-level uncertainty behind the two most important accuracy comparisons. Panel (a) shows that the aggregate FURIA win is broad but not equally reliable on every dataset: five small positive deltas have bootstrap intervals crossing zero. Panel (b) shows the narrower claim against RBF SVM: most intervals cross zero, and four datasets favour RBF SVM clearly. This pattern supports the paper’s central framing: FuzzyRuleSVM offers a faithful linguistic rule-space decomposition with competitive accuracy, not universal predictive dominance.

4.5 Runtime and memory

The implementation records fit and prediction wall-clock time and Python allocation peak memory for every compared model. FuzzyRuleSVM additionally exposes a `fit_profile_` dictionary with stage-level timings for feature screening, partition fitting, membership construction, rule generation, activation-matrix construction, LinearSVC fitting, and finalisation. Table 7 and Table 8 are regenerated from the current run artifacts. The values should be read as empirical implementation measurements under the reported environment, not as machine-independent constants; the symbolic costs in Table 3 define the scaling boundary.

4.6 Algebraic fidelity, compactness, and stability

The interpretability of FuzzyRuleSVM operates along three dimensions that must not be conflated. *Algebraic fidelity*: the additive decomposition $f(x) = \sum_k \beta_k \phi_k(x) + b$ is exact by construction (it is the model’s own computation, not an approximation) and is therefore not vulnerable to the surrogate-fidelity failures that affect post-hoc explainers such

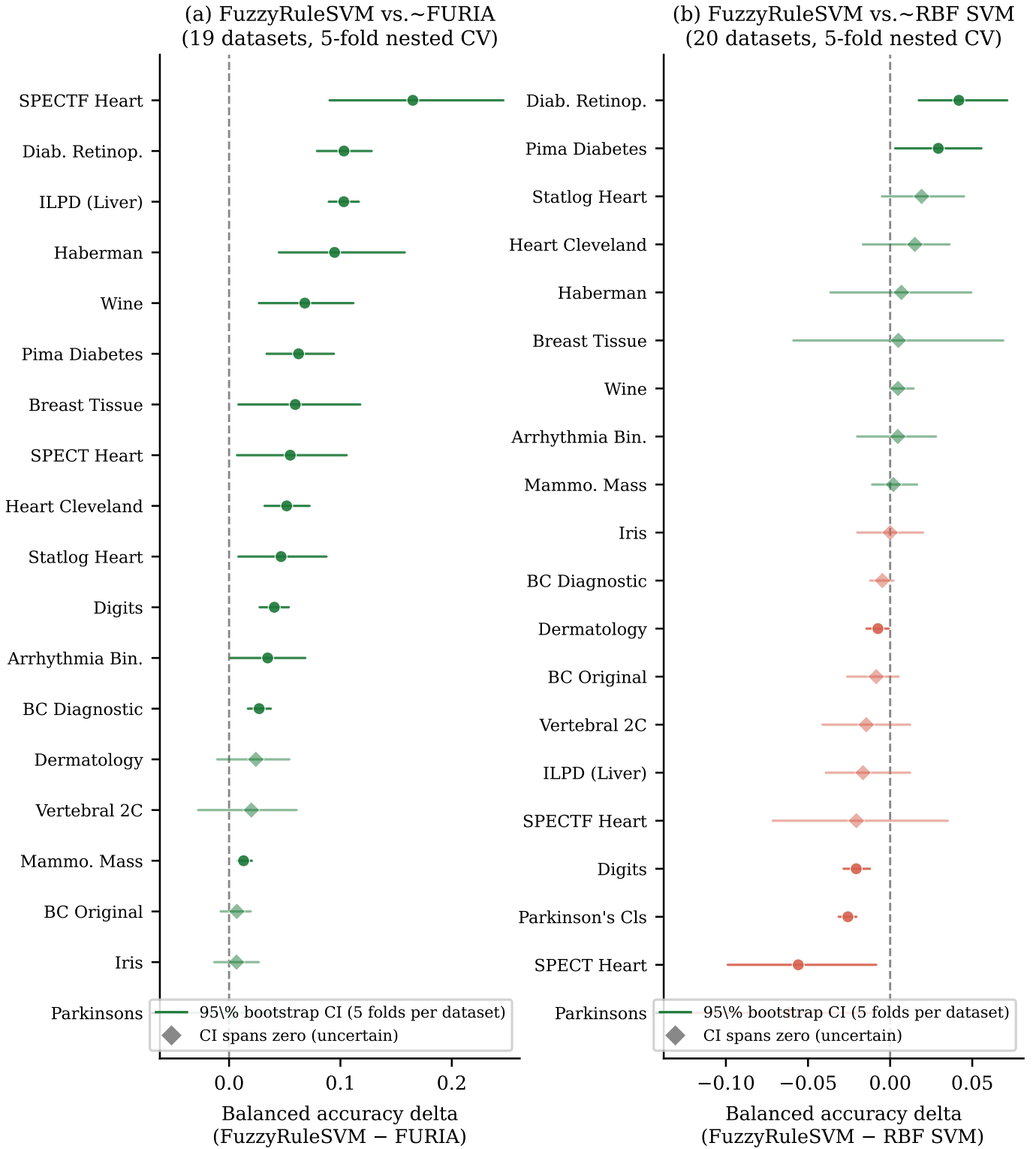


Figure 3: Per-dataset balanced-accuracy deltas with 95% bootstrap confidence intervals (2000 resamplings of the 5 per-fold values). Hollow diamonds mark datasets where the CI crosses zero. Bootstrap CIs from 5 folds quantify within-dataset fold variability, not asymptotic guarantees; regenerated comparison artifacts include the added Steel Plates Faults case.

Table 7: Generated runtime and Python allocation peak-memory summary. Values are means over available dataset-level summaries.

Model	Fit (s)	Predict (s)	Fit MB	Predict MB
RBF SVM	0.014	0.003	1.291	0.164
Linear SVM	0.017	0.001	1.290	0.164
Logistic Regression L2	0.029	0.001	1.293	0.164
Elastic-Net Logistic Regression	0.131	0.001	1.291	0.164
Logistic Regression L1	0.152	0.001	1.291	0.164
XGBoost	0.235	0.002	1.300	0.114
FuzzyRuleSVM	0.458	0.012	2.778	0.573
Histogram Gradient Boosting	0.696	0.006	3.490	0.117
Random Forest	1.032	0.016	1.428	0.130
EBM	11.887	0.004	5.432	0.242
RuleFit	13.598	0.443	4.721	0.451

Table 8: Generated FuzzyRuleSVM fit-stage breakdown on representative datasets. Times are fold means in seconds.

Dataset	d	Partition	Membership	Rules	Activation	SVM fit
haberman_survival	3	2.2e-04	3.5e-04	0.003	3.9e-04	0.003
heart_cleveland	13	3.0e-04	0.001	0.061	0.001	0.004
parkinsons_disease_classification	754	0.011	0.075	0.177	0.004	0.165

as LIME and SHAP [33]. *Compactness*: the number of rules required to cover a given fraction of the decision mass is an empirical property that varies across datasets. Displaying the top- k rules yields a partial summary of the full decomposition; algebraic exactness holds only when all nonzero contributions are included. *Stability*: the consistency of the support-rule set across training folds, measured by mean pairwise Jaccard overlap. As Rudin et al. [31] argue, algebraic exactness does not imply stability, causal validity, or human comprehensibility; these require separate empirical assessment.

For test instance i , let $c_{ik} = \beta_k \phi_k(x_i)$ and $m_i = \sum_k |c_{ik}|$, excluding the bias. Full-decomposition fidelity is measured as $\max_i |f(x_i) - (b + \sum_k c_{ik})|$ and the corresponding mean absolute error. If $q_{i,(1)} \geq \dots \geq q_{i,(K)}$ sorts $|c_{ik}|$, then the top-5 share is $\sum_{s=1}^{\min(5,K)} q_{i,(s)} / m_i$, and rules-for-90% is the smallest r with $\sum_{s=1}^r q_{i,(s)} \geq 0.9m_i$; both are zero when $m_i = 0$. Top- k sign agreement compares $\text{sign}(b + \sum_{s=1}^k c_{i,(s)})$ with the full model’s sign and therefore quantifies partial-display faithfulness rather than exact explanation. Values are averaged over test instances, one-vs-rest estimators, folds, then datasets. Active counts use $|\beta_k| > 10^{-10}$. Jaccard compares class-prefixed linguistic rule strings with $|\beta_k| \geq 10^{-8}$ across fold pairs and assigns one when both sets are empty; it does not compare fitted membership anchors.

Table 9 reports the expanded generated metrics, including candidate and retained rule counts, average support-rule length, exact decomposition error, top- k partial-display agreement, residual contribution mass, and support-rule Jaccard. These quantities delimit what the explanation interface faithfully shows: the complete nonzero decomposition is exact up to numerical roundoff, whereas a short top- k display is a lossy summary whose sign agreement and residual mass must be reported.

Fig. 4 visualises compactness and stability across the available dataset artifacts. Panel (a) shows the fold-level distribution of rules needed for 90% decision mass: ten datasets have median values below ten rules, thirteen fall below fifteen, and three exceed thirty. The largest median occurs on Parkinson’s Classification (45.4), while SPECT Heart and Dermatology also exceed thirty. Wide IQR boxes for SPECT Heart, Breast Tissue, and Parkinsons reflect genuine variation across CV folds; the support-rule composition changes with the training partition, confirming that rule stability is a separate property from algebraic fidelity.

Panel (b) shows that compactness and stability are related but not identical: compact datasets such as Iris and Pima Diabetes have high Jaccard stability, but Parkinsons remains compact while unstable. Every dataset with median rules above 30 has Jaccard stability below 0.38. Panel (c) shows an association between feature dimensionality d and explanation size: low-dimensional datasets usually remain below ten rules for 90% coverage, while very high-dimensional datasets require substantially larger rule sets.

Table 13 illustrates global model inspection on Pima Diabetes, where the median explanation needs seven rules. All five negative rules fire more strongly on non-diabetic test instances. Two positive rules, *glucose is high* and *bmi is medium AND age is high*, fire more strongly on the diabetic class. Three positive rules marked $\dagger$ instead have larger mean activation in the negative class. Correlated predictors can produce such coefficient–activation sign disagreements in a regularised linear model, so individual terms must be interpreted within the complete signed sum. These are predictive associations in this dataset, not causal medical findings.

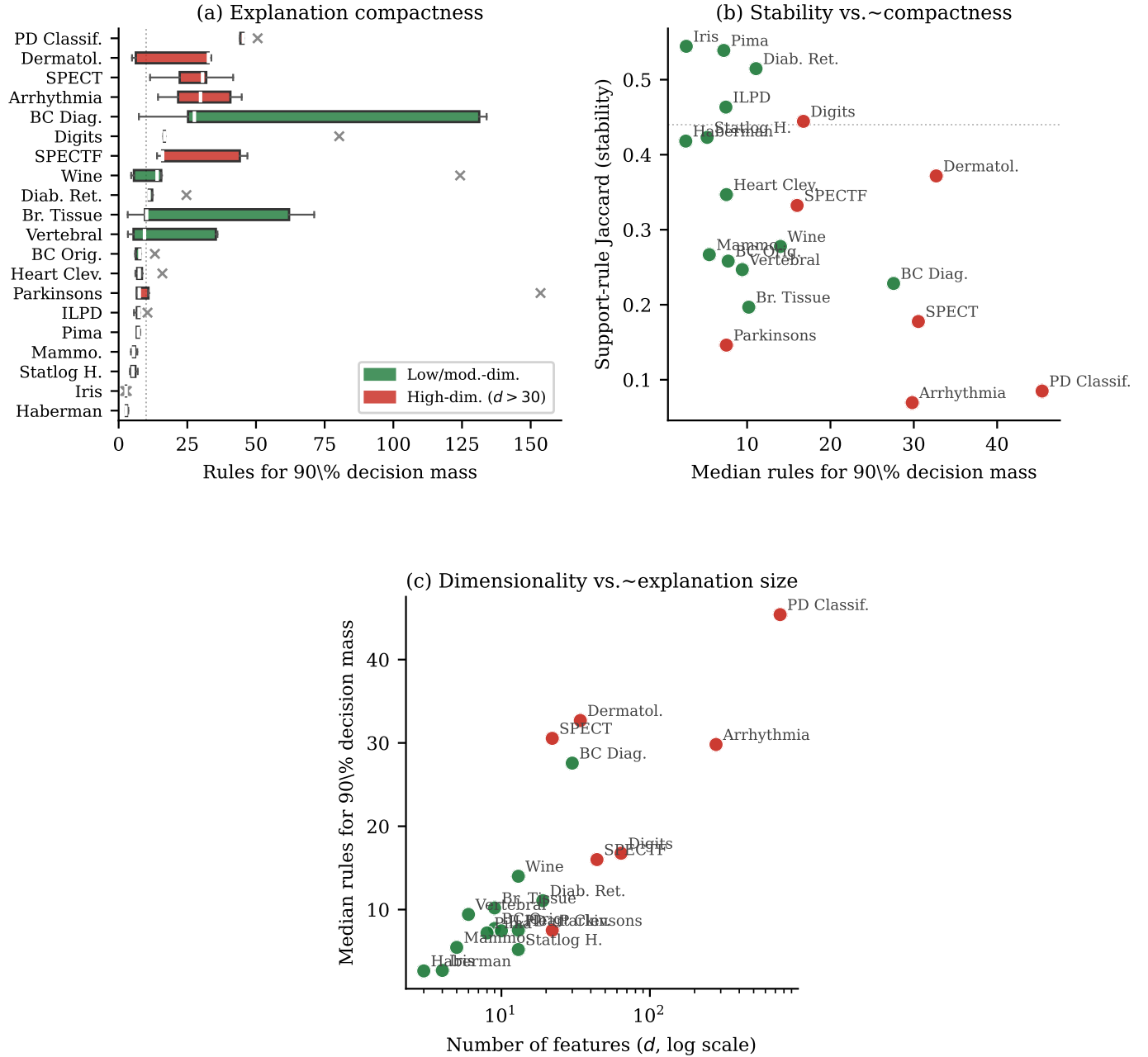


Figure 4: Interpretability analysis across the available dataset artifacts. (a) Fold-level distributions of rules needed for 90% absolute contribution mass (each box spans 5 folds; red = $d > 30$). (b) Support-rule Jaccard stability vs. median rules for 90%. (c) Feature dimensionality d is associated with larger explanations.

Table 9: Generated algebraic fidelity, compactness, and stability metrics for FuzzyRuleSVM.

Metric	Mean	Median	IQR
Candidate rules	961.667	430.000	[151.800–1538.000]
Retained rules	77.095	39.000	[27.000–90.000]
Active support rules	45.639	21.800	[17.067–68.600]
Average support-rule length	1.269	1.262	[1.112–1.371]
Rules for 90% contribution	16.868	11.476	[5.783–22.529]
Top-5 contribution share	0.657	0.694	[0.459–0.874]
Top-10 residual mass	0.196	0.157	[0.019–0.335]
Top-10 sign agreement	0.916	0.970	[0.848–0.993]
Full-decomposition max error	8.3e-16	6.1e-16	[3.6e-16–1.1e-15]
Support-rule Jaccard	0.434	0.445	[0.356–0.501]

Table 10: Generated design-choice sensitivity summary. Delta is variant minus FRS-Default; Holm correction is over the ablation family.

Variant	BA	Δ	W/T/L	p_{Holm}
MembershipSVM	0.833	+0.002	10/1/9	1.000
FRS-SoftminTnorm	0.831	+0.000	9/5/7	1.000
FRS-ProductTnorm	0.830	+0.000	8/5/8	1.000
FRS-Default	0.830	—	—	—
FRS-NoPenalty	0.830	-0.001	8/1/11	1.000
FRS-1TermOnly	0.828	-0.002	10/0/10	1.000
LinearSVC-Raw	0.825	-0.006	11/0/9	1.000
FRS-WideQuantiles	0.825	-0.006	5/1/14	1.000
FRS-L2Forced	0.824	-0.007	7/0/13	0.778

In deployment, rule-level displays should flag such cases explicitly. A rule should be marked for caution when the sign of β_k disagrees with class-conditional activation summaries, for example when $\beta_k > 0$ but $\bar{\phi}_+ < \bar{\phi}_-$. Users should then inspect the full contribution list and the competing rules before interpreting that rule as class evidence. The correct unit of audit is the complete signed sum, not an isolated coefficient.

4.7 Ablation: Conjunctive rules vs. raw membership features

To isolate conjunctive rule generation and the max-margin objective, we compare FuzzyRuleSVM against two membership-space baselines on the same outer folds, preprocessing, fixed rule policy, and balanced-accuracy inner selection. MembershipSVM applies the same partition to $3d$ atomic membership features and trains LinearSVC without conjunctions. MembershipLogisticL1 uses those features with L1-logistic regression.

Semantic expressiveness. Conjunctive rule generation primarily improves semantic auditability. MembershipSVM returns atomic activations such as “*glucose high*” and “*BMI medium*” separately; FuzzyRuleSVM represents their interaction as “*glucose is high AND BMI is medium*” with a signed margin contribution. This distinction does not establish reduced user effort or improved trust calibration.

Accuracy. The membership rows in Table 5 report the predictive results. FuzzyRuleSVM leads MembershipSVM by +0.001 mean balanced accuracy (13/1/6 wins/ties/losses, $p_{\text{raw}} = 0.107$, $p_{\text{Holm}} = 0.322$). MembershipLogisticL1 reaches 0.833 versus 0.830 for FuzzyRuleSVM (9/2/9, $p_{\text{raw}} = 0.679$, $p_{\text{Holm}} = 1.000$). Neither difference is significant, so the experiment does not support a unique predictive benefit for the max-margin rule-space formulation.

Association with dimensionality. The largest positive deltas against MembershipSVM occur on Heart Disease Cleveland ($d=13$, +0.030), Parkinsons ($d=22$, +0.024), and SPECT Heart ($d=22$, +0.017). Under the fixed policy, datasets with $d > 32$ use only atomic rules; at $d = 754$, FuzzyRuleSVM retains at most 256 of 2262 memberships and is therefore a selected subset of MembershipSVM. These differences are consistent with a representation-coverage limitation, not a causal interaction mechanism. The supported benefit is semantic and algebraic: conjunctive antecedents name interacting features whose signed effects contribute exactly to the margin, without establishing cognitive simplicity.

Table 11: Generated FuzzyRuleSVM selected-hyperparameter frequencies across outer folds.

Parameter	Most frequent value	Fold count
C	0.3	73
max_rule_length	2	80
max_rules	24	25
penalty	l1	67

Table 12: Generated Steel Plates Faults industrial case-study result.

Model	Balanced accuracy	Macro F1	Fit time (s)
RBF SVM	0.809	0.763	0.046
Random Forest	0.800	0.794	1.374
Histogram Gradient Boosting	0.794	0.813	2.919
XGBoost	0.789	0.807	0.542
Logistic Regression L2	0.767	0.664	0.136
FuzzyRuleSVM	0.767	0.709	6.030
Logistic Regression L1	0.763	0.659	0.381
Elastic-Net Logistic Regression	0.763	0.659	0.338
Linear SVM	0.760	0.673	0.056

4.8 Design-choice sensitivity

The broader ablation suite varies one design choice at a time: length penalty, forced L2, product and smooth-min aggregation, 25/50/75 rather than 5/50/95 quantile anchors, one-term-only rules, membership-only SVM, and raw-feature LinearSVC. Table 10 reports the regenerated balanced-accuracy deltas and Holm-corrected tests. The selected-parameter frequencies in Table 11 summarise how often outer folds choose each FuzzyRuleSVM setting. The main interpretation is deliberately modest: the default design is robust within the tested grid, but the high-dimensional budget remains a real boundary condition rather than a tunable nuisance.

4.9 Domain-specific case study

Steel Plates Faults adds an industrial manufacturing example to the otherwise medical-heavy benchmark. The task classifies seven surface-defect types from 27 numeric geometric and image-derived features. This setting admits concrete length-2 antecedents about plate thickness, defect perimeter, position, orientation, and luminosity. The predictive result is useful but not favourable to FuzzyRuleSVM: in nested CV, its balanced accuracy is 0.767 and macro-F1 is 0.709, compared with 0.809 and 0.763 for RBF SVM (Table 12). The case therefore demonstrates the model’s audit representation, not industrial predictive superiority.

For prediction-level inspection, we fitted the modal Steel Plates configuration ($C = 1$, L1 penalty, $L = 2$, $K_{\max} = 81$) on a fixed stratified 80/20 split (seed 42). From the 389 held-out plates, we selected one correct prediction with a large winner–runner-up gap and a strong two-antecedent contribution. This deliberately illustrative case is not an estimate of typical compactness or reliability; those properties are quantified over all outer folds in Table 9.

For plate 325, the model predicts Z Scratch over Bumps by a score gap of 1.80. Its Z-Scratch score decomposes exactly as $1.31 = -0.82 + 8.28 - 6.16$ (bias plus all supporting and opposing contributions, rounded). *A400 is low AND plate thickness is medium* fires at 0.99 with weight 2.66 and contributes +2.64; *plate thickness is medium* fires equally strongly but contributes −2.50. The conjunction therefore changes the signed effect represented by the atomic thickness term. This predictive association does not identify a physical cause or establish deployment performance.

4.10 High-dimensional limitation

FuzzyRuleSVM is intended primarily for low- to moderate-dimensional tabular data where linguistic partitions are meaningful and pairwise enumeration remains inspectable. The length-2 pool $3d + 9\binom{d}{2}$ has 276 candidates for Pima Diabetes ($d = 8$) but grows quadratically. The fixed policy therefore uses $L = 2$ for $d \leq 32$ and $L = 1$ otherwise. High-dimensional fits consequently become regularised selectors over atomic memberships rather than conjunctive IF-AND banks.

Fig. 5 shows this limitation directly. In a forced $L = 1$ budget sweep, Pima and SPECTF Heart exhaust their atomic candidate pools before the budget binds. At $K_{\max} = 256$, Arrhythmia ($d = 279$) retains 31% and Parkinson’s Disease Classification ($d = 754$) 11%. Increasing the budget helps Parkinson’s but hurts Arrhythmia, so no single larger default resolves this regime.

Table 13: Top 10 support rules by $|\beta_k|$ for the Pima Diabetes model (80/20 stratified split, random state 42). $\bar{\phi}_+$, and $\bar{\phi}_-$ are mean test activations for diabetes and non-diabetes; ℓ_k is antecedent length and DPF is diabetes pedigree function. Rules marked † activate more in the opposite class and require interpretation within the full signed sum.

Rule	β_k	ℓ_k	ϕ_+	ϕ_-
<i>Supports positive class (diabetes)</i>				
IF age is low †	+0.686	1	0.111	0.422
IF glucose is high	+0.601	1	0.388	0.142
IF insulin is medium AND age is low †	+0.565	2	0.023	0.160
IF bmi is medium AND age is high	+0.503	2	0.236	0.100
IF blood pressure is medium AND skin thickness is medium †	+0.378	2	0.226	0.285
<i>Supports negative class (no diabetes)</i>				
IF skin thickness is medium AND age is low	-0.936	2	0.033	0.240
IF glucose is low	-0.903	1	0.075	0.376
IF pregnancies is low AND age is low	-0.807	2	0.039	0.255
IF blood pressure is medium AND DPF is low	-0.471	2	0.163	0.238
IF bmi is low	-0.376	1	0.104	0.347

The high-dimensional result is therefore a boundary condition, not a second central contribution of this paper. Adaptive pricing or screened pairwise discovery may extend the approach, but those mechanisms require their own evaluation of fit cost, candidate-selection bias, and deployable inner-CV model choice.

5 Discussion

The FURIA comparison is the cleanest fuzzy-rule result in this study. FURIA uses greedy induction and is unaffected by the GA budget restriction. FuzzyRuleSVM outperforms it on every matched dataset ($p_{\text{raw}} < 0.001$, $p_{\text{Holm}} < 0.001$). Architecturally, FURIA uses fuzzy rule votes, whereas FuzzyRuleSVM assigns signed linear weights to fixed activations. The membership-logistic ablation precludes a stronger causal claim about max-margin learning: atomic membership classification matches FuzzyRuleSVM’s accuracy. The remaining distinction is representational: conjunctive versus atomic additive contributions.

Against modern interpretable baselines, the evidence is mixed. The Friedman test across FuzzyRuleSVM, EBM, and RuleFit is significant ($\chi^2 = 15.98$, $p < 0.001$ for balanced accuracy; $\chi^2 = 12.90$, $p = 0.002$ for macro-F1). FuzzyRuleSVM outperforms RuleFit on both metrics (balanced accuracy: $p_{\text{raw}} = 0.002$, $p_{\text{Holm}} = 0.005$; macro-F1: $p_{\text{raw}} = 0.012$, $p_{\text{Holm}} = 0.024$). Balanced EBM has the highest mean balanced accuracy (0.838 vs. 0.829), but its paired comparison is not significant ($p_{\text{raw}} = p_{\text{Holm}} = 0.784$). Its largest deltas occur on Arrhythmia, Parkinson’s Disease Classification, SPECT Heart, and SPECTF Heart; the retained experiments do not identify the boundary geometry causing them.

The high-dimensional cases mark the boundary of the proposed representation. On Arrhythmia and Parkinson’s Disease Classification, the default policy restricts FuzzyRuleSVM to single-antecedent rules, so the model loses the conjunctive IF-AND structure that motivates the method. The budget analysis in Section 4.10 shows that increasing K_{max} is not a universal fix: Parkinson’s benefits from a larger budget, whereas Arrhythmia peaks at a smaller one. This is why adaptive discovery is future work rather than a central claim.

Against standard classifiers and membership-space baselines, the results warrant a narrower claim. FuzzyRuleSVM is competitive with linear SVM and logistic regression but does not consistently improve over RBF SVM or L1-logistic regression on atomic fuzzy memberships. The observed differences do not isolate whether partitioning, conjunctions, or the learning objective accounts for performance.

When to use FuzzyRuleSVM. If balanced accuracy is the only objective and atomic fuzzy membership features are acceptable audit units, MembershipLogisticL1 is the simpler choice in the current experiments. FuzzyRuleSVM is justified when the audit unit should be a named fuzzy conjunction, such as *glucose is high AND BMI is high*, with a signed contribution to a linear margin. The benefit is representational and audit-oriented; the present study does not prove that users understand the resulting rule lists better than atomic memberships or additive shape functions.

Reduced-budget FARC-HD and IVTURS runs are diagnostic artifacts only, not evidence of superiority over those methods. We do not claim superiority over canonical FARC-HD or IVTURS under their full 20 000-evaluation GA budget. Alcalá-Fdez et al. [2] report substantially higher accuracy in their canonical FARC-HD evaluation; Sanz et al. [32] introduce IVTURS as an extension of FARC-HD with interval-valued type-2 fuzzy partitions, reporting

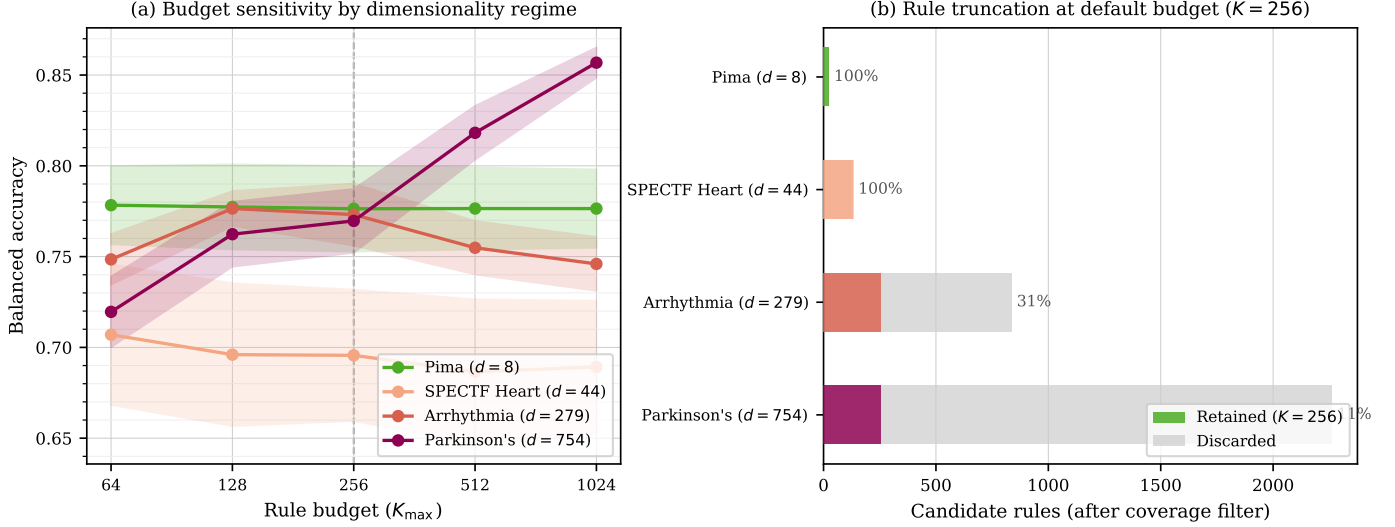


Figure 5: High-dimensional failure mode analysis. (a) Single-antecedent rule-budget sensitivity: balanced accuracy vs. $K_{\max}$ for four datasets spanning low to very-high dimensionality (± 1 -fold standard-error bands; dashed vertical line = $K_{\max} = 256$). (b) Truncation severity at $K_{\max} = 256$ for the forced $L = 1$ candidate pool: coloured bars show candidates retained; grey bars show candidates discarded. Pima and SPECTF Heart are never truncated in this single-antecedent diagnostic; Arrhythmia retains 31%; Parkinson’s Classification retains only 11%.

accuracy improvements over FARC-HD in their evaluation at canonical settings. The defensible comparison is therefore to FURIA, RuleFit, EBM, standard SVM baselines, GBDT baselines, and membership-space ablations under the reported protocol.

5.1 Limitations and future work

The principal limitation is conceptual: exact additive decomposition is algebraic fidelity, not human interpretability. We do not run a user or domain-expert study, so claims about cognitive effort, trust calibration, or decision quality remain future work. Rule coefficients are predictive associations in a regularised model and should not be read causally. The support-rule sets can be unstable across folds, and short top- k displays omit residual contribution mass unless all nonzero terms are included.

The computational boundary is also clear. Length-2 enumeration grows as $3d + 9\binom{d}{2}$, so high-dimensional datasets force screening, atomic rules, or adaptive discovery. The optional column-generation-inspired path needs its own evaluation before it can be treated as a deployable high-dimensional solution. Runtime and memory are higher than for simpler atomic-membership or GBDT baselines on some datasets. FARC-HD and IVTURS remain important fuzzy-rule baselines, but canonical GA-budget reproduction is not completed here. Finally, Steel Plates Faults broadens the domain evidence, but all experiments are still public, offline supervised evaluations; they do not establish industrial deployment performance. APS Failure at Scania Trucks is a natural future stress test, but its size and imbalance make it a separate study rather than a safe add-on to this revision.

6 Conclusion

FuzzyRuleSVM is a modeling reduction, not a new optimiser: it substitutes a linguistic rule activation space for the input features, then applies a standard regularised max-margin classifier. Because each dimension is a named antecedent activation, every prediction decomposes exactly into signed contributions. This guarantees algebraic fidelity when all nonzero terms are included, not human interpretability.

On the regenerated twenty-one-dataset benchmark, FuzzyRuleSVM is evaluated against historical fuzzy-rule, standard SVM/logistic, random-forest, histogram-gradient-boosting, XGBoost, modern glass-box, and membership-space baselines. The exact numerical summaries are generated from checked-in artifacts in Table 6, Table 7, Table 9, and Table 12. Conjunctive semantics apply where length-2 rules are feasible; high-dimensional fixed-policy fits decompose over named atomic memberships.

The strongest defensible claim is therefore narrow: FuzzyRuleSVM offers a faithful, auditable rule-activation representation trained by a standard regularised linear SVM. It should be used when named fuzzy antecedent contributions are valuable, not because the present evidence proves universal predictive dominance, causal rules, or human interpretability. At high dimensionality, candidate growth forces either screening, atomic rules, or a separate adaptive rule-discovery procedure.

Acknowledgements

The authors declare no competing interests. No specific funding was received for this work.

References

- [1] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, et al., *Neural additive models: Interpretable machine learning with neural nets*, In Advances in Neural Information Processing Systems, **34** (2021). <https://doi.org/10.5555/3540261.3540620>
- [2] J. Alcalá-Fdez, R. Alcalá, F. Herrera, *A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning*, IEEE Transactions on Fuzzy Systems, **19**(5) (2011), 857-872. <https://doi.org/10.1109/TFUZZ.2011.2147794>
- [3] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, et al., *Learning certifiably optimal rule lists for categorical data*, Journal of Machine Learning Research, **18**(1) (2017), 1-78. <https://doi.org/10.5555/3122009.3290419>
- [4] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, et al., *Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*, Information Fusion, **58** (2020), 82-115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [5] R. Batuwita, V. Palade, *FSVM-CIL: Fuzzy support vector machines for class imbalance learning*, IEEE Transactions on Fuzzy Systems, **18**(3) (2010), 558-571. <https://doi.org/10.1109/TFUZZ.2010.2042721>
- [6] C. Bénard, G. Biau, S. Da Veiga, E. Scornet, *SIRUS: Stable and interpretable RULe Set for classification*, Electronic Journal of Statistics, **15**(1) (2021), 427-505. <https://doi.org/10.1214/20-EJS1792>
- [7] J. Cao, T. Zhou, S. Zhi, S. Lam, et al., *Fuzzy inference system with interpretable fuzzy rules: Advancing explainable artificial intelligence for disease diagnosis-a comprehensive review*, Information Sciences, **662** (2024). <https://doi.org/10.1016/j.ins.2024.120212>
- [8] R. Caruana, Y. Lou, J. Gehrke, P. Koch, et al., *Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission*, In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2015), 1721-1730. <https://doi.org/10.1145/2783258.2788613>
- [9] J. Cervantes, F. García-Lamont, L. Rodríguez-Mazahua, A. López, *A comprehensive survey on support vector machine classification: Applications, challenges and trends*, Neurocomputing, **408** (2020), 189-215. <https://doi.org/10.1016/j.neucom.2019.10.118>
- [10] C. H. Chang, R. Caruana, A. Goldenberg, *NODE-GAM: Neural generalized additive model for interpretable deep learning*, In International Conference on Learning Representations, (2021). <https://doi.org/10.48550/arXiv.2106.01613>
- [11] Y. Chen, J. Z. Wang, *Support vector learning for fuzzy rule-based classification systems*, IEEE Transactions on Fuzzy Systems, **11**(6) (2003), 716-728. <https://doi.org/10.1109/TFUZZ.2003.819843>
- [12] J. Demšar, *Statistical comparisons of classifiers over multiple data sets*, Journal of Machine Learning Research, **7** (2006), 1-30.
- [13] D. Dua, C. Graff, *UCI machine learning repository*, University of California, Irvine, School of Information and Computer Sciences, (2019). <http://archive.ics.uci.edu/ml>

- [14] M. Elkano, M. Galar, J. A. Sanz, A. Fernandez, et al., *Enhancing multiclass classification in FARC-HD fuzzy classifier: On the synergy between n -dimensional overlap functions and decomposition strategies*, IEEE Transactions on Fuzzy Systems, **23**(5) (2015), 1562-1580. <https://doi.org/10.1109/TFUZZ.2014.2370677>
- [15] M. Friedman, *A comparison of alternative tests of significance for the problem of m rankings*, The Annals of Mathematical Statistics, **11**(1) (1940), 86-92. <https://doi.org/10.1214/aoms/1177731944>
- [16] J. H. Friedman, B. E. Popescu, *Predictive learning via rule ensembles*, The Annals of Applied Statistics, **2**(3) (2008), 916-954. <https://doi.org/10.1214/07-AOAS148>
- [17] X. Gu, J. Han, Q. Shen, P. P. Angelov, *Autonomous learning for fuzzy systems: A review*, Artificial Intelligence Review, **56**(8) (2023), 7549-7595. <https://doi.org/10.1007/s10462-022-10355-6>
- [18] S. Holm, *A simple sequentially rejective multiple test procedure*, Scandinavian Journal of Statistics, **6**(2) (1979), 65-70. <https://www.jstor.org/stable/4615733>
- [19] X. Hu, C. Rudin, M. Seltzer, *Optimal sparse decision trees*, In Advances in Neural Information Processing Systems, (2019), 7267-7275. <https://doi.org/10.48550/arXiv.1904.12847>
- [20] J. C. Hühn, E. Hüllermeier, *FURIA: An algorithm for unordered fuzzy rule induction*, Data Mining and Knowledge Discovery, **19**(3) (2009), 293-319. <https://doi.org/10.1007/s10618-009-0131-8>
- [21] J. S. R. Jang, *ANFIS: Adaptive-network-based fuzzy inference system*, IEEE Transactions on Systems, Man, and Cybernetics, **23**(3) (1993), 665-685. <https://doi.org/10.1109/21.256541>
- [22] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, *Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model*, The Annals of Applied Statistics, **9**(3) (2015), 1350-1371. <https://doi.org/10.1214/15-AOAS848>
- [23] C. F. Lin, S. D. Wang, *Fuzzy support vector machines*, IEEE Transactions on Neural Networks, **13**(2) (2002), 464-471. <https://doi.org/10.1109/72.991432>
- [24] C. T. Lin, C. M. Yeh, S. F. Liang, J. F. Chung, et al., *Support-vector-based fuzzy neural network for pattern classification*, IEEE Transactions on Fuzzy Systems, **14**(1) (2006), 31-41. <https://doi.org/10.1109/TFUZZ.2005.861604>
- [25] Y. Lou, R. Caruana, J. Gehrke, *Intelligible models for classification and regression*, In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2012), 150-158. <https://doi.org/10.1145/2339530.2339556>
- [26] H. Nori, S. Jenkins, P. Koch, R. Caruana, *InterpretML: A unified framework for machine learning interpretability*, Arxiv, (2019). <https://doi.org/10.48550/arXiv.1909.09223>
- [27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, et al., *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, **12** (2011), 2825-2830. <https://doi.org/10.5555/1953048.2078195>
- [28] O. Rainio, J. Teuho, R. Klén, *Evaluation metrics and statistical tests for machine learning*, Scientific Reports, **14**(1) (2024), 6086. <https://doi.org/10.1038/s41598-024-56706-x>
- [29] C. O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger, et al., *Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists*, Cognitive Systems Research, **86** (2024), 101243. <https://doi.org/10.1016/j.cogsys.2024.101243>
- [30] C. Rudin, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, Nature Machine Intelligence, **1**(5) (2019), 206-215. <https://doi.org/10.1038/s42256-019-0048-x>
- [31] C. Rudin, C. Chen, Z. Chen, H. Huang, et al., *Interpretable machine learning: Fundamental principles and 10 grand challenges*, Statistics Surveys, **16** (2022), 1-85. <https://doi.org/10.1214/21-SS133>
- [32] J. A. Sanz, A. Fernández, H. Bustince, F. Herrera, *IVTURS: A linguistic fuzzy rule-based classification system based on a new interval-valued fuzzy reasoning method with tuning and rule selection*, IEEE Transactions on Fuzzy Systems, **21**(3) (2013), 399-411. <https://doi.org/10.1109/TFUZZ.2013.2243153>

- [33] D. Slack, S. Hilgard, E. Jia, S. Singh, et al., *Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods*, In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), (2020), 180-186. <https://doi.org/10.1145/3375627.3375830>
- [34] K. Sokol, P. Flach, *Interpretable representations in explainable AI: From theory to practice*, Data Mining and Knowledge Discovery, **38**(5) (2024), 3102-3140. <https://doi.org/10.1007/s10618-024-01010-5>
- [35] F. Wilcoxon, *Individual comparisons by ranking methods*, Biometrics Bulletin, **1**(6) (1945), 80-83. <https://doi.org/10.2307/3001968>
- [36] X. Yang, G. Zhang, J. Lu, J. Ma, *A kernel fuzzy c-means clustering-based fuzzy support vector machine algorithm for classification problems with outliers or noises*, IEEE Transactions on Fuzzy Systems, **19**(1) (2011), 105-115. <https://doi.org/10.1109/TFUZZ.2010.2087382>
- [37] S. M. Zhou, J. Q. Gan, *Constructing L2-SVM-based fuzzy classifiers in high-dimensional space with automatic model selection and fuzzy rule ranking*, IEEE Transactions on Fuzzy Systems, **15**(3) (2007), 398-409. <https://doi.org/10.1109/TFUZZ.2006.882464>