Geometric Feature Selection for Interpretable SVM Classifiers in Cyber-Physical Systems

Leonie Hatte
LAAS-CNRS, Université de Toulouse, INSA, UPS, France

Pauline Ribot
LAAS-CNRS, Université de Toulouse, INSA, UPS, France

Elodie Chanthery
LAAS-CNRS, Université de Toulouse, INSA, UPS, France

Abstract

Identifying mode-transition conditions in Cyber-Physical Systems (CPS) is fundamental to model-based diagnosis, but learning them from scarce, multivariate time-series data requires boundary expressions that remain interpretable in physical variables. A central challenge is that some features are actively detrimental: their inclusion distorts the orientation of the learned boundary, degrading robustness to future crossings without affecting training accuracy. We introduce SVM-GRFE (Geometric Recursive Feature Elimination for SVM), a feature selection method that operates on polynomial monomials rather than raw variables, jointly assessing accuracy and geometric criteria that capture each monomial's structural role in boundary orientation; the boundary is back-projected analytically to an explicit polynomial inequality in physical variables. We prove SVM-GRFE returns an interpretable, geometrically stable, and robust boundary. On seven synthetic scenarios, SVM-GRFE achieves 33%–85% monomial reduction, matches or outperforms SVM-RFE in five out of seven scenarios and consistently outperforms SVM-RFE under noise at equal complexity, most notably under heteroscedastic noise.

2012 ACM Subject Classification Computing methodologies → Feature selection; Computing methodologies → Continuous models; Computing methodologies → Representation of mathematical functions; Computing methodologies → Representation of polynomials

Keywords and phrases Feature Selection, SVM, Geometric boundary learning, CPS

Digital Object Identifier 10.4230/OASIcs.DX.2026.6

Supplementary Material *Software (Source Code)*: `https://gitlab.laas.fr/lhatte/dx-26-svm-grfe`, archived at `swh:1:dir:73c6aebeaad3cf0e2b4f896dd0a64ad80d74811b`

Acknowledgements *Use of AI-assisted tools* (Claude, Anthropic) as a writing assistance tool during the drafting and editing of this manuscript. All scientific content, mathematical claims, experimental results, and conclusions were produced, verified, and are the sole responsibility of the authors.

1 Introduction

Cyber-Physical Systems (CPS) combine continuous physical processes with discrete digital control, producing complex multivariate dynamics. In many industrial applications, from manufacturing plants to energy systems, the behavior of a CPS can be decomposed into a succession of operating modes, each characterized by its own continuous dynamics. Accurate identification of these modes and of the conditions governing transitions between them is a prerequisite for reliable model-based fault diagnosis and prognosis [3].

In model-based diagnosis, the behavioral model of the system is used to identify the root cause of observed anomalies by comparing system observations to expected mode behaviors. A central prerequisite is an accurate characterization of the conditions under which the system transitions between modes. These conditions must be *interpretable*: expressed as

explicit inequalities in the original physical variables of the system, directly readable by a domain expert and reusable without further computation in a monitoring, diagnostic, or control pipeline. Interpretability is not merely a convenience in this context, it is a diagnostic requirement. An interpretable condition identifies which physical variables govern the transition, enabling future fault isolation and reducing the diagnostic hypothesis space. A condition that involves spurious variables may incorrectly implicate sensors that are not causally linked to the transition.

Consider an aircraft during climb. If a learning algorithm infers a transition condition involving an irrelevant sensor (the outside temperature for example), the classifier may achieve excellent training accuracy while producing a boundary that is sensitive to sensor noise or drift. Such errors can compromise the reliability of model-based diagnosis.Not all instrumented variables are causally involved in a given transition, and including those that are not introduces a subtler problem than mere redundancy: these variables are *detrimental* (Definition 2) in the sense that their inclusion pulls the learned boundary away from the true transition condition.

This geometric distortion reduces the robustness of the boundary to transition future crossings: a transition observed under noise, measurement drift, or slow variations in the continuous dynamics may no longer be correctly classified, causing spurious fault detections or missed transitions, reducing the resilience of the diagnostic pipeline. Detecting and removing *detrimental* features is therefore central to transition condition learning, and cannot be achieved by accuracy-based criteria alone: a *detrimental* feature may distort the boundary orientation without reducing classification accuracy on the training data, a distinction that purely accuracy-based selection criteria cannot capture.

In practice, a CPS is observed as a multivariate Time Series (TS) $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where each observation $\mathbf{x}_i \in \mathbb{R}^m$ represents the simultaneous measurement of m physical variables at time step i. Recovering an interpretable, exportable, geometrically stable expression under perturbation from data is both a classification problem and a modeling challenge. The available data is scarce: only the two short time segments immediately before and after the transition are informative, typically comprising a few hundred observations, which renders cross-validation (CV) unreliable as a model[1] selection strategy [2]. These conditions are formalized as Assumptions (A1)–(A4) in Section 3. With so few points, each CV fold contains too few samples to be representative of the data distribution, and the resulting accuracy estimates exhibit high variance, making classifier comparison unreliable. Data points are furthermore temporally ordered and locally correlated.

Support Vector Machines (SVMs) are particularly well-suited to this problem: their decision boundary is explicit and can be exported as a polynomial inequality. They can represent complex transition conditions when combined with polynomial feature expansion, and their maximum-margin formulation provides a natural measure of boundary robustness [5]. However, standard SVM feature selection (FS) methods such as SVM-RFE [7] operate on raw input variables and do not account for the geometric properties of the boundary essential for transition condition generalization. Moreover, in a polynomial feature space, the relevant unit of selection is the monomial rather than the physical variable: a monomial may couple several physical variables simultaneously, and its removal eliminates a coupling term from the boundary expression rather than a physical variable.

This paper introduces SVM-GRFE (Geometric Recursive Feature Elimination for SVM), an FS method tailored to transition condition learning from TS data. SVM-GRFE extends the SVM-RFE framework to operate on polynomial monomials rather than on raw physical

[1] We call *model* the learned classifier.

variables, enabling the analytical back-projection of the learned classifier to an explicit polynomial inequality in physical variables. SVM-GRFE has three main proved properties: it returns an interpretable boundary, it is robust to bounded perturbations and ensures geometrical stability. Its central contribution is the detection and removal of *detrimental* monomials, that distort the boundary orientation without accuracy benefit, through geometric criteria that capture the structural role of each monomial in the learned boundary, a distinction that purely accuracy-based selection criteria cannot make. SVM-GRFE can therefore be integrated directly into model learning pipelines. The method is evaluated on synthetic boundary scenarios of increasing complexity against baseline FS methods, and its robustness to five types of measurement noise is assessed at equal model complexity.

The remainder of this paper is organized as follows. Section 2 presents background on SVMs and reviews related work on SVM-based FS methods and their limitations. Section 3 formalizes the problem and details the SVM-GRFE algorithm. Section 4 presents the experimental protocol, results, and noise robustness analysis over the case study. Section 5 concludes and outlines directions for future work.

2 Background and related work

2.1 Support Vector Machines

Support Vector Machines (SVMs) are supervised binary classifiers [5]: given a labeled training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ with $\mathbf{x}_i \in \mathbb{R}^m$ and $y_i \in \{-1, +1\}$, a hard-margin SVM seeks a boundary expression $\mathbf{w}^\top \mathbf{x} + b = 0$ that separates the two classes while maximizing the *margin*, defined as the distance between the boundary and the closest training points of each class. Here, $\mathbf{w} \in \mathbb{R}^m$ is the weight vector, which is normal to the boundary and determines its orientation in the learning feature space, and $b \in \mathbb{R}$ is the bias term, which controls its offset from the origin. These closest points are called *support vectors* and are the only observations that determine the boundary; all other training points are irrelevant to the boundary's position. Maximizing the margin is equivalent to minimizing $\|\mathbf{w}\|^2$, which yields a boundary that is as far as possible from both classes and is therefore robust to small perturbations in the data.

When the two classes are not perfectly separable, the hard-margin formulation has no solution. The soft-margin SVM [5] relaxes this constraint by allowing a controlled number of misclassifications, penalized by a regularization parameter $C \in \mathbb{R}_*^+$. A large value of C forces the classifier to minimize training errors at the risk of overfitting, while a small value of C tolerates more misclassifications in favor of a wider, more generalizable margin.

Linear SVMs produce linear decision boundaries, which is insufficient for the complex, non-linear transition conditions that arise in CPS. Non-linear SVMs with kernels can represent more complex boundaries, but their decision function $f(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$ is an implicit representation defined only via training-point evaluations in a high or infinite-dimensional feature space, not as a closed-form expression in physical variables; it therefore cannot be exported without the full training set and cannot be inspected by a domain expert, making the learned boundary unsuitable for a transition condition on the physical variables of the system. Several approaches have been proposed to recover interpretability from SVM models. ISVMFD (Interpretable SVM for Functional Data) [9], is an interpretable SVM for functional data that produces a function whose shape can be inspected to identify relevant time intervals. However, this framework operates in infinite-dimensional function spaces and relies on the functional structure of the input: it does not extend to multivariate TS and the learned boundary cannot be exported as an inequality in physical variables. pTsm-SVM [4] incorporates data-mined two-dimensional linear rules as inequality constraints into the SVM

optimization problem to enhance interpretability. The result, however, is a set of local rules used to explain individual predictions, not a global analytical boundary expression: the underlying decision function remains a kernel SVM that cannot be exported and reused directly as a boundary expression. In our context, an interpretable expression is a strict requirement: the learned boundary must be a polynomial inequality in physical variables, analytically exportable and directly reusable in a CPS diagnostic or prognostic pipeline.

We therefore adopt a linear SVM combined with an explicit polynomial feature expansion defined by the monomial set $\Phi = \{\varphi_1, \ldots, \varphi_{N_{\text{poly}}}\}$ of degree d, where:

$$N_{\text{poly}} = \binom{m+d}{d} - 1 \tag{1}$$

is the number of monomials in the polynomial expansion of total degree at most d, excluding the constant bias term b, and each $\varphi_j : \mathbb{R}^m \to \mathbb{R}$ is a monomial of the polynomial feature expansion. We therefore create non-linear features from the physical variables of the system. The SVM then operates linearly in the expanded polynomial feature space, and the learned boundary in physical variables can be analytically recovered as a polynomial inequality, retaining the ability to represent non-linear transition conditions. Those properties are central to the interpretability objectives of SVM-GRFE (Section 3.3).

From now on, the *variables* refer to the physical variables of the system while *features* or *monomials* refer to the expanded polynomial features in $\mathbb{R}^{N_{\text{poly}}}$. Furthermore, variables in bold refer to vectors.

2.2 Related work on feature selection: methods and limitations

In the context of boundary learning for CPS, FS serves a purpose that goes beyond model simplification: a monomial may distort the orientation or translation of the learned boundary in physical space, yielding a boundary that is not robust to future crossings of the transition. FS is therefore motivated by three complementary objectives: (i) geometric stability: removing monomials whose inclusion rotates or translates the boundary without accuracy benefit; (ii) robustness: choosing the most relevant monomials to make the boundary less sensitive to noise and slow variations in the continuous dynamics; and (iii) interpretability defined as follows.

▶ **Definition 1** (Interpretability of the learned boundary). *The learned boundary $g_\Phi(\mathbf{x}) \geq 0$ is said to be* interpretable *if it can be expressed as a closed-form polynomial inequality in the physical variables $x_0, \ldots, x_{m-1}$, directly readable by a domain expert and analytically exportable for reuse in a CPS modeling, monitoring, or diagnostic pipeline.*

The goal is not to reduce the number of features for its own sake, but to identify and remove monomials that are actively *detrimental* to the geometric quality of the learned boundary.

▶ **Definition 2** (Detrimental monomial). *A monomial $\varphi_j \in \Phi$ is said to be* detrimental *if its removal from the active feature set Φ does not degrade classification accuracy and simultaneously reduces the geometric distortion of the learned boundary.*

FS for SVM has been studied extensively, and two main families of methods are generally distinguished [6, 8]: filter approaches select features independently of the classifier while wrapper approaches evaluate feature subsets by directly measuring the classifier's accuracy on held-out data. Filter approaches, which are agnostic to the classifier, are not considered further; the remainder of this section focuses on wrapper methods.

Wrapper approaches are known to outperform filter methods when the goal is to optimize the performance of a specific classifier rather than to rank features by generic statistical criteria [8]. Importantly, [8] shows that maximizing accuracy and selecting the most relevant features for a given problem are not equivalent objectives and can lead to different chosen subsets. They define as *relevant* a feature for which its removal can degrade the performance of the classifier. Relevance is further distinguished into three levels: a feature is *strongly relevant* if its removal always degrades the optimal classifier; *weakly relevant* if it is not strongly relevant but belongs to some optimal feature subset; and *irrelevant* if its presence never affects the optimal classifier regardless of the subset considered. Their work also establishes that strongly relevant, weakly relevant, and irrelevant features require different treatment, a taxonomy that motivates the multi-criteria categorization adopted in our method SVM-GRFE (see Section 3).

Among wrapper strategies, backward elimination starts from the full set and removes variables recursively, capturing interactions more naturally but at higher computational cost. Forward selection on the other hand starts from an empty feature set and adds variables one at a time, offering computational efficiency at the cost of missing feature interactions. Forward selection is not included as a primary baseline in our work for two reasons. First, the first monomial selected anchors the solution: any subsequent monomial is evaluated conditionally on this initial choice, which introduces a dependence that is particularly problematic when monomials are correlated, as is systematically the case in polynomial expansions. Second, and more fundamentally, forward selection cannot identify *detrimental* monomials since it only adds monomials that improve the current model and never considers removal.

A well-documented limitation of wrapper methods is their susceptibility to overfitting in the feature subset space, caused by repeated use of accuracy estimates [8]. This problem is exacerbated when the number of available samples is small: with so few points, each CV fold contains too few samples to be representative of the data distribution, and the resulting accuracy estimates exhibit high variance, making classification model comparison unreliable [2]. In our setting, the training data consists of a local neighborhood of size $n_{\text{train}} \in \mathbb{N}$, which may comprise as few as a few tens of points. CV is therefore not abandoned for computational reasons, but because it is statistically unreliable with so few data. Unlike a CV fold, which partitions a single observation pool so that train and test share the same recording and local correlations, our train/test split originate from two distinct crossings, recorded under potentially different operating conditions. X_{eval} is therefore an independent realization of the same physical phenomenon, not a held-out subset of the same recording: the resulting accuracy estimate measures generalization to a future crossing (the method's actual operational objective) and its variance reflects the system's natural between-crossing variability rather than the arbitrary choice of partition. SVM-RFE (Recursive Feature Elimination), introduced by [7] for the classification of cancerous tissues from very few samples but a high number of features, addresses FS for linear SVM by recursively eliminating the feature j whose squared weight w_j^2 is smallest, eliminating the feature that contributes least to the margin, and retraining the classifier after each removal. SVM-RFE was shown to automatically reduce redundancy while producing smaller and more accurate subsets. However, SVM-RFE was designed for the case where features coincide with input variables: the weight vector $\mathbf{w}$ has one component per physical variable, so ranking features is equivalent to ranking variables. In SVM-GRFE, by contrast, the SVM operates in an expanded polynomial feature space and $\mathbf{w}$ has one component per monomial of the polynomial feature space. Ranking monomials is therefore not equivalent to ranking physical variables. Furthermore, SVM-RFE relies solely on a weight-based criterion, without accounting for the geometric properties (orientation and translation) of the learned boundary in physical space.

More recent work has explored metaheuristic approaches to FS, combining population-based optimization algorithms (Grey Wolf Optimizer (GWO), Salp Swarm Algorithm (SSA), Particle Swarm Optimization (PSO), and others) with classifiers such as SVM, random forests, or neural networks [10]. These methods identify optimal feature subsets jointly with classifier hyperparameters, evaluated by CV and tuned by Optuna [1]. Their results confirm that FS does not degrade SVM accuracy and yields significant reductions in model complexity and improved interpretability. However, population-based methods require a large number of model evaluations and a reliable CV estimate, making them unsuitable for the small-sample constraint considered here.

Three limitations are shared across the reviewed methods and motivate the design of our new method SVM-GRFE. First, feature is systematically equated with input variable, whereas in a polynomial feature space the relevant unit of selection is the monomial. Second, no method incorporates geometric criteria to assess the structural role of a feature in the learned boundary, relying exclusively on accuracy-based scoring. Third, none of the reviewed methods require producing an exportable expression of the boundary in physical variables, which is a requirement in the context of transition learning for CPS.

3 SVM-GRFE: Geometric Recursive Feature Elimination for SVM

3.1 Problem formulation

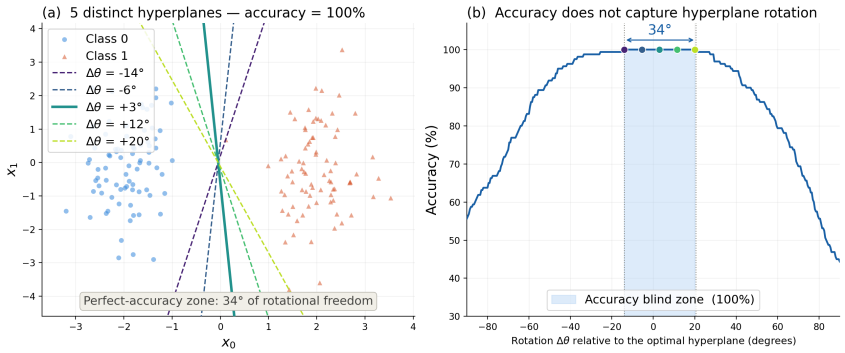

Figure 1 Geometric instability of the decision boundary: five different boundaries achieve perfect classification accuracy (subfig.(a)) while exhibiting substantial mutual rotation (subfig.(b)).

In CPS, the transition conditions between functioning modes are generally complex, non-linear and multivariate.

Accuracy alone is an insufficient criterion for FS when the geometry of the learned boundary matters. As exemplified in Fig. 1, subfig.(a), five distinct boundaries can achieve perfect classification accuracy on the same two-dimensional dataset while exhibiting substantial mutual rotation (subfig.(b)). In the context of boundary learning, such rotational instability yields structurally different expression in physical variables, undermining robustness of the learned boundary. A monomial that induces such rotation without improving accuracy is not merely useless: it is *detrimental* (Definition 2) and should be removed. We are looking for the monomials that ensure what is defined as the geometric stability of the learned boundary.

▶ **Definition 3** (Geometric stability of the learned boundary). *The learned boundary* $g_\Phi(\mathbf{x}) \geq 0$ *is said to be* geometrically stable *with respect to a monomial* $\varphi_j \in \Phi$ *if its removal does not significantly alter the orientation or the position of the boundary in physical space. More*

formally, let τ_θ^{low} and τ_ρ^{low} thresholds defined in Eq. 4. The learned boundary is geometrically stable iff $\theta^{(-j)} \leq \tau_\theta^{low}$ and $\rho^{(-j)} \leq \tau_\rho^{low}$, with $\theta^{(-j)}$ the angle between the orientation of the ablated boundary $g_{\Phi\setminus\{\varphi_j\}}(\mathbf{x})$ and the full boundary $g_\Phi(\mathbf{x})$, and $\rho^{(-j)}$ the normalized plane distance between the two boundaries.

Indeed, the decision boundary must not only classify the training dataset correctly, but also faithfully generalize to future crossings. A boundary that is accurate on the training dataset yet poorly oriented because of detrimental variables that would impact negatively the orientation of the boundary, may fail to generalize to future crossings. Not all features contribute equally to this geometric fidelity: some monomials are irrelevant and can be discarded without consequence, while others actively distort the orientation of the boundary and are therefore detrimental to the robustness of the learned boundary. Identifying and removing such detrimental monomials is the central objective of SVM-GRFE.

▶ **Definition 4** (Robustness of the learned boundary). *The boundary condition $g_\Phi(\mathbf{x}) \geq 0$ is said to be* robust *at a point $\mathbf{x}$ if there exists a perturbation bound $\varepsilon(\mathbf{x}) > 0$ such that for any perturbed observation $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$ satisfying $\|\boldsymbol{\delta}\| \leq \varepsilon(\mathbf{x})$, the sign of g_Φ is preserved: (i.e.* $\operatorname{sign}\big(g_\Phi(\tilde{\mathbf{x}})\big) = \operatorname{sign}\big(g_\Phi(\mathbf{x})\big)$*. Such perturbations may include measurement noise, sensor drift, and slow variations of the underlying continuous dynamics.*

The system is observed via a multivariate time series $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with $\mathbf{x}_i \in \mathbb{R}^m$. An observable mode change occurs at a transition instant t^*. We extract from $\mathcal{X}$ a pre-transition training segment and a post-transition training segment. We denote by $\xi^- = \{\mathbf{x}_{-k^-}, \ldots, \mathbf{x}_{-1}\}$ the last $k^- \in \mathbb{N}^*$ observations of $\mathcal{X}$ before t^* and by $\xi^+ = \{\mathbf{x}_1, \ldots, \mathbf{x}_{k^+}\}$ the first $k^+ \in \mathbb{N}^*$ observations of $\mathcal{X}$ after and during t^*, with $n_{\text{train}} = k^- + k^+ = |X_{\text{train}}|$. The training dataset is constructed as:

$$X_{\text{train}} = \{(\mathbf{x}, -1) \mid \mathbf{x} \in \xi^-\} \cup \{(\mathbf{x}, +1) \mid \mathbf{x} \in \xi^+\} \tag{2}$$

so that class -1 corresponds to the pre-transition mode and class $+1$ to the post-transition mode. Once X_{train} is constructed, the temporal ordering of the points is discarded: the SVM treats X_{train} as an unordered set of labeled points in $\mathbb{R}^m$, whose only role is to determine the support vectors and the resulting decision boundary.

The method's design and guarantees are conditioned on the following assumptions, characteristic of transition learning from CPS multivariate time series.

- **(A1) Data scarcity.** The training dataset X_{train} comprises n_{train} labeled observations drawn from a single transition crossing. This assumption renders cross-validation unreliable as a model selection strategy: with so few points, each CV fold contains too few samples to be representative of the data distribution, and the resulting accuracy estimates exhibit high variance, making classifier comparison unreliable [2].
- **(A2) Held-out crossing availability.** At least two crossings of the same transition are available, yielding a segment X_{eval} disjoint from X_{train} and recorded at a different time instant. When no such crossing is available, $X_{\text{eval}} = X_{\text{train}}$ is used as a fallback, at the cost of measuring training fit rather than generalization.
- **(A3) Transition stationarity.** The guard condition governing the transition is stable across crossings: X_{eval} is a representative evaluation of the generalization of the learned boundary to future crossings of the same transition.
- **(A4) Bounded model complexity.** The polynomial degree $d^* \leq d_{\max}$ and the number of physical variables m are moderate, consistent with physically interpretable CPS dynamics.

3.2 Proposed method

SVM-GRFE is a sequential pipeline: first, a full-feature boundary is learned, and then monomials are recursively eliminated according to accuracy and geometric criteria. The central building block throughout is the `train_SVM` function (Alg. 2, Section 3.2.2), which trains a polynomial SVM with hyperparameters (d^*, C^*) (previously identified by grid search, results in the associated git repository[2]), and exports the resulting boundary as an explicit polynomial inequality in physical variables. The objective of SVM-GRFE is therefore not to reduce the number of physical variables, but to identify a reduced subset $\Phi^* \subseteq \Phi$ by eliminating monomials whose presence degrades the geometric stability of the boundary, such that $g_{\Phi^*}(\mathbf{x}) \geq 0$. We first describe the full SVM-GRFE algorithm in its two elimination phases, then the training SVM function.

3.2.1 SVM-GRFE full algorithm

Alg. 1 explains the Geometric Recursive Feature Elimination for SVM (SVM-GRFE). At first, a feature set Φ is initialized as the complete set of N_{poly} (Equation (1)) monomials of degree at most d^* (line 1). A full-feature boundary $g_\Phi(\mathbf{x})$ is then trained via `train_SVM` (Alg. 2, Section 3.2.2) and its accuracy $\text{acc}_{\text{eval}}^{(\text{full})}$ on the evaluation dataset X_{eval} is recorded as the reference accuracy (line 2). SVM-GRFE then proceeds in two phases.

Phase 1: recursive elimination. At each iteration, every remaining monomial $\varphi_j \in \Phi$ is individually removed and an ablated model $g_{\Phi\setminus\{\varphi_j\}}(\mathbf{x})$ is retrained on $\Phi\setminus\{\varphi_j\}$ (lines 5–9). Three scores are computed for each ablated model (line 8, Eq. 3): the signed accuracy variation $\Delta^{(-j)}$, the cosine angle $\theta^{(-j)}$ between the orientation of the ablated boundary $g_{\Phi\setminus\{\varphi_j\}}(\mathbf{x})$ and the full boundary $g_\Phi(\mathbf{x})$, and the normalized plane distance $\rho^{(-j)}$ between the two boundaries, characterizing respectively the accuracy impact, the rotational displacement, and the translational displacement induced by removing φ_j.

We now use the notation $\square^{(\text{ref})}$ when referencing the reference model and $\square^{(\text{ref})}_{\text{proj}(-j)}$ for the element of the reference model projected onto the subspace $\Phi\setminus\{\varphi_j\}$ by discarding the component corresponding to φ_j. The three scores are defined as:

$$\cos(\theta^{(-j)}) = \frac{\mathbf{w}^{(-j)} \cdot \mathbf{w}^{(\text{ref})}_{\text{proj}(-j)}}{\|\mathbf{w}^{(-j)}\| \, \|\mathbf{w}^{(\text{ref})}_{\text{proj}(-j)}\|}, \quad \rho^{(-j)} = \frac{\left|b^{(-j)} - b^{(\text{ref})}_{\text{proj}(-j)}\right|}{\|\mathbf{w}^{(-j)}\|}, \quad \Delta^{(-j)} = \text{acc}^{(-j)}_{\text{eval}} - \text{acc}^{(\text{ref})}_{\text{eval}} \tag{3}$$

A value of $\cos(\theta^{(-j)})$ close to 1 indicates that the boundary orientation is preserved; a small value indicates a significant rotation. The distance $\rho^{(-j)}$ measures the translational displacement of the boundary normalized by $\|\mathbf{w}^{(-j)}\|$.

The three scores are then compared to adaptive thresholds that partition the candidate monomials into elimination categories: ε_{acc} separates accuracy-neutral removals from accuracy-impactful ones, while $\tau_\theta^{\text{low/high}}$ and $\tau_\rho^{\text{low/high}}$ distinguish geometrically negligible, moderate, and significant displacements. All thresholds are computed from the empirical distributions of the scores across all candidate removals at the current iteration (line 10),

Algorithm 1 Geometric Recursive Feature Elimination for SVM (SVM-GRFE).

Input: $C^* \in \mathbb{R}_*^+, \quad d^* \in \mathbb{N}_*$ ▷ SVM hyperparameters
$X_{\text{train}} \in \mathbb{R}^{n_{\text{train}} \times m}, \quad y_{\text{train}} \in \{-1,+1\}^{n_{\text{train}}}$ ▷ train set
$X_{\text{eval}} \in \mathbb{R}^{n_{\text{eval}} \times m}, \quad y_{\text{eval}} \in \{-1,+1\}^{n_{\text{eval}}}$ ▷ evaluation set
Output: $\Phi^* \subseteq \Phi$ ▷ reduced monomial set
$g_{\Phi^*}(\mathbf{x}) \geq 0$ ▷ boundary in physical variables

$\Phi \leftarrow \{\varphi_1, \ldots, \varphi_{N_{\text{poly}}}\}, \quad N_{\text{poly}} = \binom{m+d^*}{d^*} - 1$ ▷ feature set
$g_\Phi(\mathbf{x}) \leftarrow \texttt{train\_SVM}(\Phi)$ ▷ Alg. 2
$\text{acc}_{\text{eval}}^{(\text{ref})} \leftarrow \texttt{accuracy}(g_\Phi(\mathbf{x}), X_{\text{eval}}, y_{\text{eval}})$

Phase 1: recursive elimination
while $|\Phi| > 1$ **do**
  for each $\varphi_j \in \Phi$ **do**
    $g_{\Phi\setminus\{\varphi_j\}}(\mathbf{x}) \leftarrow \texttt{train\_SVM}(\Phi\setminus\{\varphi_j\})$ ▷ Alg. 2
    $\text{acc}_{\text{eval}}^{(-j)} \leftarrow \texttt{accuracy}(g_{\Phi\setminus\{\varphi_j\}}(\mathbf{x}), X_{\text{eval}}, y_{\text{eval}})$
    Compute $\Delta^{(-j)}, \theta^{(-j)}, \rho^{(-j)}$ ▷ Eq. 3
  end for
  Compute $\varepsilon_{\text{acc}}, \tau_\theta^{\text{low}}, \tau_\theta^{\text{high}}, \tau_\rho^{\text{low}}, \tau_\rho^{\text{high}}$ ▷ Eq. 4
  $\mathcal{D}_{\text{detrimental}} \leftarrow \{\varphi_j \in \Phi : \Delta^{(-j)} \geq +\varepsilon_{\text{acc}}\}$ ▷ set of detrimental features
  $\mathcal{D}_{\text{geom}} \leftarrow \left\{\varphi_j \in \Phi\setminus\mathcal{D}_{\text{detrimental}} : \left(\theta^{(-j)} \geq \tau_\theta^{\text{high}} \text{ or } \rho^{(-j)} \geq \tau_\rho^{\text{high}}\right) \text{ and } \Delta^{(-j)} \geq -\varepsilon_{\text{acc}}\right\}$ ▷ set of geometry destabilizing features
  if $\mathcal{D}_{\text{detrimental}} \cup \mathcal{D}_{\text{geom}} = \emptyset$ **then break**
  $\varphi^* \leftarrow \begin{cases} \arg\max_{\varphi_j \in \mathcal{D}_{\text{detrimental}}} \Delta^{(-j)} & \text{if } \mathcal{D}_{\text{detrimental}} \neq \emptyset \\ \arg\max_{\varphi_j \in \mathcal{D}_{\text{geom}}} \left(\tilde{\theta}^{(-j)} + \tilde{\rho}^{(-j)}\right) & \text{otherwise} \end{cases}$ ▷ with normalized scores $\tilde{\theta}, \tilde{\rho}$
  $\Phi \leftarrow \Phi\setminus\{\varphi^*\}$
  $g_\Phi(\mathbf{x}) \leftarrow \texttt{train\_SVM}(\Phi)$ ▷ Alg. 2
  $\text{acc}_{\text{eval}}^{(\text{ref})} \leftarrow \texttt{accuracy}(g_\Phi(\mathbf{x}), X_{\text{eval}}, y_{\text{eval}})$
end while

Phase 2: batch removal of useless monomials
if $|\Phi| > 1$ **then**
  $\mathcal{D}_{\text{useless}} \leftarrow \left\{\varphi_j \in \Phi : \theta^{(-j)} \leq \tau_\theta^{\text{low}} \text{ and } \rho^{(-j)} \leq \tau_\rho^{\text{low}} \text{ and } \Delta^{(-j)} \geq -\varepsilon_{\text{acc}}\right\}$ ▷ reuse last Phase 1 scores
  if $\mathcal{D}_{\text{useless}} \neq \emptyset$ **then**
    $\Phi \leftarrow \Phi\setminus\mathcal{D}_{\text{useless}}$
    $g_\Phi(\mathbf{x}) \leftarrow \texttt{train\_SVM}(\Phi)$ ▷ Alg. 2
  end if
end if
return $\Phi^* \leftarrow \Phi, \quad g_{\Phi^*}(\mathbf{x}) \geq 0$

using quantiles to avoid introducing additional hyperparameters:

$$\begin{aligned} \varepsilon_{\text{acc}} &= \max\left(\tfrac{1}{n_{\text{eval}}},\ Q_{75}\left(\left\{\left|\Delta^{(-j)}\right|\right\}_{\varphi_j\in\Phi}\right)\right) \\ \tau_\theta^{\text{low}} &= Q_{25}\left(\left\{\theta^{(-j)}\right\}_{\varphi_j\in\Phi}\right) & \tau_\theta^{\text{high}} &= Q_{75}\left(\left\{\theta^{(-j)}\right\}_{\varphi_j\in\Phi}\right) \\ \tau_\rho^{\text{low}} &= Q_{25}\left(\left\{\rho^{(-j)}\right\}_{\varphi_j\in\Phi}\right) & \tau_\rho^{\text{high}} &= Q_{75}\left(\left\{\rho^{(-j)}\right\}_{\varphi_j\in\Phi}\right) \end{aligned} \tag{4}$$

The accuracy tolerance ε_{acc} is floored at $1/n_{\text{eval}}$, with n_{eval} the number of points of the evaluation dataset X_{eval}, the smallest observable accuracy difference on X_{eval}, to prevent a single misclassified point from being interpreted as a meaningful accuracy change and triggering elimination. All thresholds are data-driven and require no tuning.

Two elimination candidate sets are then constructed (lines 11–12).

▶ **Definition 5** (Accuracy-detrimental monomial). *A monomial $\varphi_j \in \Phi$ is* accuracy-detrimental *if its removal improves classification accuracy beyond the tolerance threshold: $\Delta^{(-j)} \geq +\varepsilon_{acc}$. The set of accuracy-detrimental monomials is $\mathcal{D}_{detrimental} = \{\varphi_j \in \Phi : \Delta^{(-j)} \geq +\varepsilon_{acc}\}$.*

▶ **Definition 6** (Geometrically destabilizing monomial). *A monomial $\varphi_j \in \Phi \setminus \mathcal{D}_{detrimental}$ is* geometrically destabilizing *if its presence distorts the boundary orientation without contributing to separability: its removal causes a significant displacement of the boundary, in orientation ($\theta^{(-j)} \geq \tau_\theta^{high}$) or in position ($\rho^{(-j)} \geq \tau_\rho^{high}$), without degrading classification accuracy ($\Delta^{(-j)} \geq -\varepsilon_{acc}$). Its elimination therefore improves the geometric stability of the learned boundary.*

$$\mathcal{D}_{geom} = \left\{ \varphi_j \in \Phi \setminus \mathcal{D}_{detrimental} : \left(\theta^{(-j)} \geq \tau_\theta^{high} \text{ or } \rho^{(-j)} \geq \tau_\rho^{high} \right) \text{ and } \Delta^{(-j)} \geq -\varepsilon_{acc} \right\}$$

If both $\mathcal{D}_{\text{detrimental}}$ and $\mathcal{D}_{\text{geom}}$ sets are empty, Phase 1 terminates (line 13). Otherwise, the most detrimental or destabilizing monomial $\varphi^* \in \mathcal{D}_{\text{detrimental}} \cup \mathcal{D}_{\text{geom}}$ is selected for elimination (line 14): accuracy-detrimental monomials are prioritized by $\Delta^{(-j)}$, and geometrically destabilizing monomials by the sum of their normalized scores $\tilde{\theta}^{(-j)} + \tilde{\rho}^{(-j)}$. The selected monomial φ^* is dropped (line 15), the SVM model is retrained (line 16), and the reference accuracy is updated (line 17).

X_{eval} denotes a held-out evaluation set used to assess accuracy during elimination. Two choices are possible depending on data availability:

- **Test set (recommended):** $X_{\text{eval}} = X_{\text{test}}$, constructed from one or more future crossings of the same transition: segments $\tilde{\xi}^-$ and $\tilde{\xi}^+$ recorded at a different time than those used to build X_{train}, such that $X_{\text{train}} \cap X_{\text{test}} = \emptyset$. This provides an unbiased estimate of the generalization of the learned condition to unseen crossings.
- **Training set (fallback):** $X_{\text{eval}} = X_{\text{train}}$ when no held-out future crossing is available. Accuracy score $\Delta^{(-j)}$ measures training fit rather than generalization, which may lead to less reliable elimination decisions.

In the experiments of Section 4.2, $X_{\text{eval}} = X_{\text{test}}$, using a fixed train/test split in place of CV, which is unreliable in the small-sample constraint considered here [2].

Phase 2: batch removal of useless monomials. Upon termination of Phase 1, the scores from its last iteration are reused directly to identify the set of useless monomials $\mathcal{D}_{\text{useless}}$ (line 20): monomials for which both $\theta^{(-j)}$ and $\rho^{(-j)}$ are below their respective low thresholds and whose removal does not degrade accuracy. These monomials contribute neither to the boundary orientation, its position, nor its accuracy, and are removed in a single batch (lines 21–24), followed by a single retraining (line 23).

This two-phase design eliminates, in Phase 1, monomials that are either accuracy-detrimental (Definition 5) or geometrically destabilizing (Definition 6), and removes in batch in Phase 2 those that are useless (neither accuracy-detrimental nor geometrically destabilizing). These three populations are necessarily disjoint (Proposition 12), ensuring that each monomial is processed by at most one elimination criterion.

The final reduced monomial set Φ^* and the corresponding boundary $g_{\Phi^*}(\mathbf{x}) \geq 0$ are returned (line 26). To ensure that all SVM models trained during SVM-GRFE are consistently oriented and therefore geometrically comparable, the sign of g_{Φ^*} is fixed by evaluating it at the centroid $\boldsymbol{\mu}$ of X_{train} in physical space (details in Section 3.2.2 and Algorithm 2).

The monomials retained in Φ^* at the end of Phase 2 are those for which no elimination criterion was triggered: they are neither accuracy-detrimental, nor geometrically destabilizing, nor useless under the adaptive thresholds computed at their respective iteration. This does

not imply that retained monomials are *necessary* in a strict sense (that is, that their removal would necessarily degrade accuracy or geometry). Two sources prevent this stronger claim. First, the adaptive thresholds are defined relative to the score distributions of the current iteration: a monomial may be retained simply because other monomials exhibit larger geometric effects at that step, keeping it below the elimination threshold. Second, the sequential nature of the elimination introduces a path dependence: the scores $\theta^{(-j)}$ and $\rho^{(-j)}$ are recomputed at each iteration on the current model, so a monomial that was not eliminated may have been eliminated under a different elimination order. The set Φ^* should therefore be interpreted as a *sufficient* reduced feature set: one for which no monomial was identified as removable without violating the accuracy or geometry constraints of the method, rather than as a minimal or necessary one.

3.2.2 SVM training function

Algorithm 2 `train_SVM`(S).

Input: $S = \{\gamma_1, \dots, \gamma_{|S|}\}$ ▷ feature set
$X_{\text{train}} = \{\mathbf{x}_1, \dots, \mathbf{x}_{n_{\text{train}}}\}$, $X_{\text{train}} \in \mathbb{R}^{n_{\text{train}} \times m}$ ▷ train set
$y_{\text{train}} \in \{-1, +1\}^{n_{\text{train}}}$ ▷ train classes
$C^* \in \mathbb{R}_*^+$, $d^* \in \mathbb{N}_*$ ▷ SVM hyperparameters
Output: $g_S(\mathbf{x}) \geq 0$ ▷ boundary in physical variables on input feature set S

1: Compute $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^m$ ▷ mean and standard deviation of X_{train}
2: Compute $Z_{\text{train}} = \{\mathbf{z}_1, \dots, \mathbf{z}_{n_{\text{train}}}\}$, $Z_{\text{train}} \in \mathbb{R}^{n_{\text{train}} \times m}$ ▷ standardize X_{train} by $\boldsymbol{\mu}, \boldsymbol{\sigma}$
3: $\Psi_{\text{train}} \leftarrow [\gamma_j(\mathbf{z}_i)]_{i,j} \in \mathbb{R}^{n_{\text{train}} \times |S|}, i \in [1, m], j \in [1, |S|]$ ▷ matrix of polynomial expansion

Compute decision function: $f(\mathbf{z}) = \mathbf{w}_f^\top \Psi(\mathbf{z}) + b_f$ in standardized space

4: $(\mathbf{w}_f, b_f) \leftarrow$ `LinearSVC`$(C^*, \Psi_{\text{train}}, y_{\text{train}})$, $\mathbf{w}_f \in \mathbb{R}^{|S|}, b_f \in \mathbb{R}$ ▷ learn boundary

Back-project to decision function $g_S(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ in physical variables

5: $g_S(\mathbf{x}) \leftarrow$ `backproject`$(\mathbf{w}_f, b_f, \boldsymbol{\mu}, \boldsymbol{\sigma}, S)$ ▷ Eq. 5
6: **if** $g_S(\boldsymbol{\mu}) < 0$ **then** ▷ $\boldsymbol{\mu}$ is the centroid of X_{train} in physical space
7: $g_S(\mathbf{x}) \leftarrow -g_S(\mathbf{x})$ ▷ orient so that $g_S(\boldsymbol{\mu}) \geq 0$
8: **end if**
9: **return** $g_S(\mathbf{x}) \geq 0$

Algorithm 2 details the `train_SVM` function, which is called throughout SVM-GRFE. It takes as fixed inputs any feature set S, training dataset X_{train} and its associated classes $y_{\text{train}} \in \{-1, +1\}^{n_{\text{train}}}$, degree d^*, regularization parameter C^*; and returns an oriented boundary $g_S(\mathbf{x}) \geq 0$ in physical variables.

The variable-wise mean $\boldsymbol{\mu} \in \mathbb{R}^m$ and standard deviation $\boldsymbol{\sigma} \in \mathbb{R}^m$ of X_{train} are first computed (line 1), and the dataset is standardized to zero mean and unit variance, producing $Z_{\text{train}} \in \mathbb{R}^{n_{\text{train}} \times m}$ (line 2). Standardization is necessary because the polynomial expansion would otherwise produce monomials with very different numerical magnitudes depending on the physical units of the variables, which would distort the SVM weight vector $\mathbf{w}$. The polynomial feature matrix $\Psi_{\text{train}} \in \mathbb{R}^{n_{\text{train}} \times |S|}$ is then constructed by evaluating each monomial $\gamma_j \in S$ on each standardized observation $\mathbf{z}_i \in Z_{\text{train}}$ (line 3): entry (i, j) of Ψ_{train} is $\gamma_j(\mathbf{z}_i)$. A `LinearSVC` (Linear Support Vector Classifier) with regularization parameter C^* is fitted on $(\Psi_{\text{train}}, y_{\text{train}})$, producing the weight vector $\mathbf{w}_f \in \mathbb{R}^{|S|}$ and intercept $b_f \in \mathbb{R}$ of the decision function $f(\mathbf{z}) = \mathbf{w}_f^\top \Psi(\mathbf{z}) + b_f$ in standardized space (line 4).

The `backproject`$(\mathbf{w}_f, b_f, \boldsymbol{\mu}, \boldsymbol{\sigma}, S)$ function (line 5) converts the SVM decision function $f(\mathbf{z}) = \mathbf{w}_f^\top \Psi(\mathbf{z}) + b_f$ defined in standardized space, into an explicit polynomial $g_S(\mathbf{x})$ in physical variables by applying the substitution $z_i = (x_i - \mu_i)/\sigma_i$ analytically to each monomial $\gamma_j \in S$:

$$g_S(\mathbf{x}) = \sum_{j=1}^{|S|} w_{f,j}\, \gamma_j\left(\frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}\right) + b_f \;=\; \mathbf{w}^\top \mathbf{x} + b \tag{5}$$

where $\mathbf{w} \in \mathbb{R}^{|S|}$ and $b \in \mathbb{R}$ are the coefficients of the resulting polynomial in $\mathbf{x}$, obtained by expanding and collecting the terms of $\sum_j w_{f,j}\, \gamma_j\left(\frac{\mathbf{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right) + b_f$, and are distinct from $\mathbf{w}_f \in \mathbb{R}^{|S|}$ and $b_f \in \mathbb{R}$, which are the SVM weights in the standardized polynomial feature space.

To ensure that all returned SVM models are consistently oriented and therefore geometrically comparable, the sign of g_S is fixed by evaluating it at the centroid $\boldsymbol{\mu}$ of X_{train} in physical space (Algorithm 2). If $g_S(\boldsymbol{\mu}) < 0$, the sign of g_S is flipped (line 7), so that class $+1$ always contains the centroid of X_{train}. The oriented boundary $g_S(\mathbf{x}) \geq 0$ is returned (line 9).

The resulting inequality $g_S(\mathbf{x}) \geq 0$ is a polynomial inequality in the original state variables that can be directly reused as a transition condition in a CPS modeling, monitoring, diagnostic, or control pipeline, and the retained monomials explicitly identify which physical variables and their interactions govern the mode transition, allowing a diagnostic expert to determine which sensors are causally involved and which are irrelevant, reducing the diagnostic hypothesis space.

3.3 Properties of SVM-GRFE

SVM-GRFE satisfies two key properties that jointly motivate its design: the learned boundary is *interpretable*, in that it is expressed as a closed-form polynomial inequality in physical variables, and *robust*, in that the maximum-margin formulation provides a formal guarantee of classification stability under bounded perturbations. These two properties are stated and proved below, followed by a discussion of how geometric stability criteria connect them in the feature selection procedure.

▶ **Proposition 7** (SVM-GRFE returns an interpretable learned boundary)**.** *Let $\Phi^* \subseteq \Phi$ be the reduced monomial set returned by SVM-GRFE. The learned boundary $g_{\Phi^*}(\mathbf{x}) \geq 0$ is a polynomial inequality in the physical variables $x_0, \ldots, x_{m-1}$, whose coefficients are recovered analytically from the SVM weight vector $\mathbf{w}$ via the back-projection procedure of Alg. 2 (line 5). It is directly human-readable and analytically exportable for reuse in a CPS model for diagnosis and prognosis.*

Proof. By construction of `backproject` (Eq. 5), the decision function $f(\mathbf{z}) = \mathbf{w}_f^\top \Psi(\mathbf{z}) + b_f$ defined in the standardized polynomial feature space is converted into an explicit polynomial $g_{\Phi^*}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ in physical variables by substituting $z_i = (x_i - \mu_i)/\sigma_i$ analytically into each monomial $\varphi_j \in \Phi^*$ and collecting terms. The result is a finite sum of monomials in $x_0, \ldots, x_{m-1}$ with explicit real coefficients, readable without computational evaluation of kernel functions. ◀

▶ **Definition 8** (Physical-space margin)**.** *Let $g_\Phi(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ be the decision function of the SVM in physical space. The* margin *at a point $\mathbf{x} \in \mathbb{R}^m$ is defined as the perpendicular distance from $\mathbf{x}$ to the decision boundary $\{\mathbf{x}' \in \mathbf{x} \mid g_\Phi(\mathbf{x}') = 0\}$:*

$$d(\mathbf{x}) = \frac{|g_\Phi(\mathbf{x})|}{\|\mathbf{w}\|} \tag{6}$$

▶ **Proposition 9** (SVM-GRFE returns a boundary robust to bounded perturbations)**.** *Let* $\mathbf{x} \in \mathbb{R}^m$ *be a correctly classified point. The boundary* $g_\Phi(\mathbf{x}) \geq 0$ *is robust at* $\mathbf{x}$ *with* $\varepsilon(\mathbf{x}) = d(\mathbf{x}) = |g_\Phi(\mathbf{x})|/\|\mathbf{w}\|$.

Proof. The perturbed decision function reads:

$$g_\Phi(\mathbf{x} + \boldsymbol{\delta}) = \mathbf{w}^\top(\mathbf{x} + \boldsymbol{\delta}) + b = g_\Phi(\mathbf{x}) + \mathbf{w}^\top\boldsymbol{\delta} \tag{7}$$

The predicted class is preserved if and only if $g_\Phi(\mathbf{x})$ and $g_\Phi(\mathbf{x} + \boldsymbol{\delta})$ have the same sign. By the Cauchy–Schwarz inequality, $|\mathbf{w}^\top\boldsymbol{\delta}| \leq \|\mathbf{w}\| \cdot \|\boldsymbol{\delta}\|$. If $\|\boldsymbol{\delta}\| < \gamma(\mathbf{x}) = |g_\Phi(\mathbf{x})|/\|\mathbf{w}\|$, then:

$$|\mathbf{w}^\top\boldsymbol{\delta}| \leq \|\mathbf{w}\| \cdot \|\boldsymbol{\delta}\| < \|\mathbf{w}\| \cdot \frac{|g_\Phi(\mathbf{x})|}{\|\mathbf{w}\|} = |g_\Phi(\mathbf{x})| \tag{8}$$

so $g_\Phi(\mathbf{x} + \boldsymbol{\delta})$ cannot change sign, and the predicted class is preserved. ◀

▶ Remark 10. Proposition 9 is stated in physical space and $\boldsymbol{\delta} \in \mathbb{R}^m$ is a perturbation in physical space representing measurement noise, sensor drift, or slow variations in the continuous dynamics of the system. The perturbation tolerance (the margin in physical space) $d(\mathbf{x})$ increases with the distance of $\mathbf{x}$ to the boundary and decreases with $\|\mathbf{w}\|$: a smaller weight vector norm yields a wider margin and greater robustness. Note that the SVM maximum-margin criterion maximizes the margin in the standardized polynomial feature space, not in physical space; the physical-space margin $d(\mathbf{x})$ is therefore an indicator of robustness in the physical variables, evaluated empirically in Section 4.4.

Proposition 9 shows that robustness is directly controlled by the physical-space margin $d(\mathbf{x}) = |g_\Phi(\mathbf{x})|/\|\mathbf{w}\|$. The maximum-margin formulation of the SVM maximizes the minimum margin over the training set, providing the largest possible robustness guarantee on the training dataset. Rather than maximizing the margin directly during FS, which would require solving a joint optimization problem, SVM-GRFE uses the angular deviation $\theta^{(-j)}$ and the normalized plane distance $\rho^{(-j)}$ as proxies for geometric stability: a monomial whose removal significantly rotates or translates the boundary is deemed structurally relevant, and its elimination is avoided unless it is accuracy-detrimental (Section 3.2).

Accuracy and robustness are therefore complementary but not equivalent objectives: a boundary that correctly classifies all points may nonetheless exhibit a small physical-space margin $d(\mathbf{x})$, and remain fragile to perturbations of moderate magnitude. This distinction motivates the joint accuracy-and-geometry criterion of SVM-GRFE, which preserves the margin by avoiding geometrically destabilizing eliminations even when they are accuracy-neutral.

▶ **Proposition 11** (SVM-GRFE per-step elimination guarantee)**.** *At each iteration of Phase 1, the eliminated monomial* φ^* *satisfies at least one of the following conditions:*

1. $\Delta^{(-j^*)} \geq +\varepsilon_{acc}$*: accuracy does not decrease,* *(accuracy-detrimental case)*
2. $\Delta^{(-j^*)} \geq -\varepsilon_{acc}$*: accuracy decreases by at most* ε_{acc}*.* *(geometrically destabilizing case)*

In particular, no monomial whose removal simultaneously degrades accuracy beyond ε_{acc} *and causes a significant geometric displacement is ever eliminated.*

Proof. By construction, $\varphi^* \in \mathcal{D}_{\text{detrimental}} \cup \mathcal{D}_{\text{geom}}$. If $\varphi^* \in \mathcal{D}_{\text{detrimental}}$, then $\Delta^{(-j^*)} \geq +\varepsilon_{\text{acc}}$ by definition, so 1. holds. If $\varphi^* \in \mathcal{D}_{\text{geom}} \setminus \mathcal{D}_{\text{detrimental}}$, then $\Delta^{(-j^*)} \geq -\varepsilon_{\text{acc}}$ by the membership condition of $\mathcal{D}_{\text{geom}}$, so 2. holds. Since any eliminated monomial belongs to one of these two sets, no monomial outside $\mathcal{D}_{\text{detrimental}} \cup \mathcal{D}_{\text{geom}}$ is ever removed. ◀

▶ **Proposition 12** (Disjointness of elimination sets). *The three elimination sets $\mathcal{D}_{detrimental}$, $\mathcal{D}_{geom}$, and $\mathcal{D}_{useless}$ are pairwise disjoint:*

$$\mathcal{D}_{detrimental} \cap \mathcal{D}_{geom} = \emptyset, \qquad \mathcal{D}_{detrimental} \cap \mathcal{D}_{useless} = \emptyset, \qquad \mathcal{D}_{geom} \cap \mathcal{D}_{useless} = \emptyset$$

Proof. $\mathcal{D}_{\text{detrimental}} \cap \mathcal{D}_{\text{geom}} = \emptyset$ follows directly from the definition of $\mathcal{D}_{\text{geom}}$, which requires $\varphi_j \in \Phi \setminus \mathcal{D}_{\text{detrimental}}$.$\mathcal{D}_{\text{geom}} \cap \mathcal{D}_{\text{useless}} = \emptyset$: a monomial in $\mathcal{D}_{\text{useless}}$ satisfies $\theta^{(-j)} \leq \tau_\theta^{\text{low}}$ and $\rho^{(-j)} \leq \tau_\rho^{\text{low}}$, while membership in $\mathcal{D}_{\text{geom}}$ requires $\theta^{(-j)} \geq \tau_\theta^{\text{high}}$ or $\rho^{(-j)} \geq \tau_\rho^{\text{high}}$. Since $\tau_\theta^{\text{low}} < \tau_\theta^{\text{high}}$ and $\tau_\rho^{\text{low}} < \tau_\rho^{\text{high}}$ by construction ($Q_{25} < Q_{75}$), no monomial can satisfy both conditions simultaneously.$\mathcal{D}_{\text{detrimental}} \cap \mathcal{D}_{\text{useless}} = \emptyset$: $\mathcal{D}_{\text{useless}}$ requires $\Delta^{(-j)} \geq -\varepsilon_{\text{acc}}$ while $\mathcal{D}_{\text{detrimental}}$ requires $\Delta^{(-j)} \geq +\varepsilon_{\text{acc}}$; any monomial satisfying the latter is eliminated in Phase 1 before Phase 2 is reached. ◀

▶ **Proposition 13** (SVM-GRFE global guarantee of accuracy). *The per-step guarantee ensures that no single elimination degrades accuracy by more than ε_{acc}, but does not bound the cumulative loss across all iterations, since ε_{acc} is recomputed at each step and individual losses may accumulate. The total accuracy variation is therefore evaluated empirically in Section 4.2.*

3.4 Computational complexity of SVM-GRFE

We derive the computational complexity of Alg. 1 in terms of the number of `LinearSVC` training calls, which dominate the total cost. Let m denote the number of physical variables, d^* the polynomial degree, n_{train} the training dataset size, and N_{poly} the initial number of monomials in Φ. A single `LinearSVC` training on n_{train} points with p features costs $\mathcal{O}(n_{\text{train}} \cdot p)$.

Phase 1: cost per step. At elimination step $s \geq 1$, the current monomial set Φ contains $p_s = N_{\text{poly}} - (s-1)$ monomials. The inner loop (lines 5–9) trains one ablated model per candidate monomial: for each $\varphi_j \in \Phi$, `train_SVM` is called on $p_s - 1$ features, at cost $\mathcal{O}(n_{\text{train}} \cdot p_s)$. Since there are p_s candidates, the total cost of the inner loop at step s is:

$$C_s^{\text{inner}} = \mathcal{O}(n_{\text{train}} \cdot p_s) \times p_s = \mathcal{O}(n_{\text{train}} \cdot p_s^2) \tag{9}$$

After the inner loop, the selected monomial φ^* is eliminated and the model is retrained once on $p_s - 1$ features (line 16), at cost $\mathcal{O}(n_{\text{train}} \cdot p_s)$ which is dominated by C_s^{inner}. The total cost of step s is therefore $\mathcal{O}(n_{\text{train}} \cdot p_s^2)$.

Phase 1: total cost. Let $T \leq N_{\text{poly}} - 1$ be the number of elimination steps before the break condition at line 13. Summing the per-step costs:

$$C^{\text{Phase 1}} = \sum_{s=1}^{T} \mathcal{O}(n_{\text{train}} \cdot p_s^2) = \mathcal{O}\left(n_{\text{train}} \cdot \sum_{s=1}^{T} p_s^2 \right) \tag{10}$$

Since $p_s \leq N_{\text{poly}}$ for all s, each term satisfies $p_s^2 \leq N_{\text{poly}}^2$, giving the upper bound:

$$C^{\text{Phase 1}} \leq \mathcal{O}(n_{\text{train}} \cdot N_{\text{poly}}^2 \cdot T) \tag{11}$$

Worst case. When $T = N_{\text{poly}} - 1$, every monomial is eliminated one by one. Substituting $p_s = N_{\text{poly}} - (s-1)$:

$$\sum_{s=1}^{N_{\text{poly}}-1} p_s^2 = \sum_{p=2}^{N_{\text{poly}}} p^2 = \frac{N_{\text{poly}}(N_{\text{poly}}+1)(2N_{\text{poly}}+1)}{6} - 1 = \Theta(N_{\text{poly}}^3) \tag{12}$$

The worst-case complexity of Phase 1 is therefore $\Theta(n_{\text{train}} \cdot N_{\text{poly}}^3)$.

Phase 2: negligible cost. Phase 2 (lines 19–24) reuses the scores $\Delta^{(-j)}$, $\theta^{(-j)}$, $\rho^{(-j)}$ computed during the last iteration of Phase 1, without any additional `train_SVM` call. It incurs at most one retraining on the final reduced set $\Phi \setminus \mathcal{D}_{\text{useless}}$ at line 23, at cost $\mathcal{O}(n_{\text{train}} \cdot p_T)$. This is dominated by $C^{\text{Phase 1}}$ and does not affect the overall complexity.

Effective complexity. In practice, early stopping at line 13 terminates Phase 1 after a small number of steps $T \ll N_{\text{poly}}$, giving an effective cost of $\mathcal{O}(n_{\text{train}} \cdot N_{\text{poly}}^2)$. Since $N_{\text{poly}} = \mathcal{O}(m^{d^*})$ for fixed d^*, the method scales polynomially in the number of physical variables and remains tractable for the moderate dimensions ($m \leq 10$, $d^* \leq 3$) typical of physical system dynamics.

4 Case study

We evaluate SVM-GRFE on seven synthetic boundary scenarios[2] $B1$ to $B7$ of increasing complexity, ranging from univariate linear thresholds to multivariate polynomial boundaries in up to five physical variables.

Synthetic data are used deliberately: standard benchmark datasets do not satisfy the constraints specific to our problem (few samples, physically meaningful variables, complex continuous dynamics, and a decision boundary whose expression matters as much as its accuracy). Each scenario generates labeled trajectory data of $n_{\text{train}} \approx 300$ training points and $n_{\text{test}} \approx 200$ test points per transition, reflecting the small-data constraint that motivates our design choices. In all experiments, $X_{\text{eval}} = X_{\text{test}}$, a fixed held-out split disjoint from X_{train}, used in place of CV.

From a diagnostic perspective, the seven scenarios represent transition conditions of increasing complexity between operating modes of a CPS.

4.1 Baselines

Three methods are compared throughout this section:

- FULL: the unpruned SVM model trained on the full monomial set Φ, serving as the accuracy reference;
- SVM-RFE: standard SVM-RFE applied in the polynomial feature space, which eliminates monomials one by one according to their squared SVM weight w_j^2, $j \in |\Phi|$;
- SVM-GRFE: the proposed method.

FULL retains all N_{poly} monomials and serves as the accuracy reference. SVM-RFE baseline is truncated to the same cardinality $|\Phi^*|$ as SVM-GRFE, enabling an iso-complexity comparison that isolates the effect of the selection criterion from the effect of model complexity.

[2] For each scenario, the associated git repository https://gitlab.laas.fr/lhatte/dx-26-svm-grfe provides the full true boundary expression, the number of physical variables, the true polynomial degree, the selected hyperparameters, and the dataset sizes; we summarize only the aggregate complexity trends in this article.

Table 1 Comparison of FULL, SVM-GRFE, and SVM-RFE across all seven scenarios; best accuracy per scenario in **bold**.

			FULL	SVM-GRFE		SVM-RFE	
Boundary	N_{poly}	True monomials	Accuracy	Accuracy	Φ^*	Accuracy	Φ^*
$B1$	2	x_0, x_1	**0.995**	**0.995**	x_0	**0.995**	x_0
$B2$	5	x_0^2, x_1^2	**0.990**	0.930	x_1, x_0^2, x_1^2	**0.990**	x_0, x_1, x_0^2
$B3$	2	x_0x_1	**0.990**	**0.990**	x_0	**0.990**	x_0
$B4$	3	x_0^2, x_1, x_2	**0.994**	0.982	x_0	0.988	x_1
$B5$	9	x_0^3, $x_0x_1^2$	0.975	**0.990**	x_0, x_1^2, x_0^3	0.980	x_0, x_1, x_0^3
$B6$	3	x_0x_1, x_1x_2, x_0^2	**0.994**	**0.994**	x_1, x_2	**0.994**	x_1, x_2
$B7$	20	x_2	0.950	**0.995**	x_2, x_2x_3, x_3^2	0.920	x_0, x_2, x_4

Three stopping criteria are commonly used in the literature for SVM-RFE: stopping at a fixed cardinality specified a priori; stopping when cross-validated accuracy begins to degrade; eliminating all but one feature to produce a complete ranking [7]. The second criterion requires reliable CV estimates, which are unavailable in our small-sample constraint. The third produces a single-monomial output that is systematically too aggressive for polynomial boundaries. We therefore adopt the first criterion, stopping SVM-RFE at the same monomial cardinality $|\Phi^*|$ as SVM-GRFE, which enables a fair comparison that isolates the effect of the selection criterion (geometric vs. weight-based) from the effect of model complexity. All methods use the same hyperparameters (d^*, C^*) selected by grid search on a fixed train/test split, ensuring that observed differences in accuracy and geometry are attributable solely to the feature selection strategy (see results in the git repository[2]).

4.2 SVM-GRFE results and comparison to baseline methods

Table 1 summarizes the results for all three methods (FULL, SVM-RFE, SVM-GRFE) across the seven scenarios. For each method we report the selected monomial set Φ^* and the test accuracy. The best accuracy per scenario is in **bold**.

For $B1$, $B3$, and $B6$, (scenarios with $d^* = 1$) SVM-GRFE and SVM-RFE converge to the same monomial set and produce identical boundaries with no accuracy loss relative to FULL. For $B3$ and $B6$, the true boundary involves degree-2 monomials absent from the candidate set since $d^* = 1$: both methods identify the best available linear approximation within the constrained feature space. These low-complexity scenarios confirm that the geometric criterion of SVM-GRFE does not introduce unnecessary eliminations when the feature space is small and well-conditioned.

For $B2$ and $B4$, SVM-RFE outperforms SVM-GRFE. In $B4$, SVM-GRFE retains x_0 (Accuracy = 0.982) while SVM-RFE retains x_1 (Accuracy = 0.988, boundary $8.738\,x_1 - 4.247 \geq 0$), which is the more discriminative variable according to the unpruned model weights $(6.142\,x_1 + 2.080\,x_2 + 1.142\,x_0 - 3.878 \geq 0)$. In $B2$, SVM-GRFE incurs a difference in accuracy of -0.060 by eliminating x_0, whose small but non-negligible weight $(0.307\,x_1 - 0.089\,x_0^2 - 0.032\,x_0 + \cdots \geq 0)$ is underestimated by the geometric criterion. Both cases correspond to near-tie configurations where variables are individually informative and geometrically comparable: the weight-based criterion of SVM-RFE is better suited to these settings because it directly reflects discriminative power, whereas the geometric criterion of SVM-GRFE is designed to detect structural irrelevance rather than marginal accuracy differences.

For $B5$ and $B7$, the full model is severely degraded by detrimental monomials that carry large SVM weights, misleading the criterion of SVM-RFE. In $B5$, SVM-GRFE converges to $\{x_0, x_1^2, x_0^3\}$ (Accuracy = 0.990, +1.5% over FULL) by correctly identifying x_1

Table 2 Geometric indicators θ, ρ, $1/\|\mathbf{w}\|$ for SVM-GRFE and SVM-RFE relative to FULL.

	FULL	SVM-GRFE			SVM-RFE		
Boundary	$\frac{1}{\|\mathbf{w}\|}$	θ	ρ	$\frac{1}{\|\mathbf{w}\|}$	θ	ρ	$\frac{1}{\|\mathbf{w}\|}$
$B1$	0.501	49.1°	1.26	0.726	49.1°	1.26	0.726
$B2$	3.113	24.3°	0.99	2.150	1.2°	0.08	3.217
$B3$	1.001	53.7°	0.70	1.093	53.7°	0.70	1.093
$B4$	0.152	80.0°	0.13	0.219	21.1°	0.04	0.114
$B5$	0.007	58.7°	0.89	0.074	72.6°	1.31	0.103
$B6$	0.234	26.5°	0.10	0.203	26.5°	0.10	0.203
$B7$	0.020	90.0°	6.21	1.412	65.5°	0.12	0.088

as detrimental. SVM-RFE retains x_1 instead of x_1^2, achieving only Accuracy $= 0.980$. In $B7$, SVM-GRFE converges to $\{x_2, x_2x_3, x_3^2\}$, correctly discarding all x_4-related monomials. SVM-RFE retains x_4 (Accuracy $= 0.920$) because its large weight $|w_{x_4^2}| = 45.37$ reflects its dominance in the distorted full model rather than its relevance to the true boundary. In both cases, the geometric criterion of SVM-GRFE detects that removing detrimental monomials simultaneously stabilizes the boundary orientation and improves accuracy, a distinction that the weight ranking of SVM-RFE cannot capture.

These results must be interpreted in the context of the selection strategy of d^* and C^* that are selected by grid search that maximize F_1 score on a fixed split, a criterion that is agnostic to boundary geometry and trajectory dynamics. This means that the selected hyperparameters are optimized for classification accuracy, not for the geometric robustness required.

Overall, SVM-GRFE matches or outperforms SVM-RFE in five out of seven scenarios, strictly outperforms the full model in two ($B5$, $B7$), and identifies compact, physically interpretable boundaries in the high-complexity constraint. This is the expected trade-off for a criterion designed to detect structural irrelevance rather than marginal accuracy differences. Its advantage is most pronounced when the monomial space is large, and accuracy-based criteria are misled by detrimental features with large but spurious SVM weights.

4.3 Geometric analysis of learned boundaries

Table 2 reports the geometric indicators of SVM-GRFE and SVM-RFE relative to FULL: the angular deviation θ (degrees), the scaled plane distance ρ which measures the translational displacement between the two boundaries in units of $\|\mathbf{w}\|$, and the SVM margin $1/\|\mathbf{w}\|$ in physical space. The margin quantifies the robustness of the learned boundary to perturbations in the feature space, as formalized in Section 3.3, Proposition 9: a larger margin implies a wider robustness guarantee around the boundary. θ and ρ are computed between the final selected boundary $g_{\Phi^*}(\mathbf{x})$ and the unpruned reference $g_{\Phi}(\mathbf{x})$ using Eq. 3, with $\Phi \setminus \{\varphi_j\}$ replaced by Φ^*.

For $B1$, $B3$, and $B6$, both methods produce geometrically identical boundaries (θ and ρ equal), consistent with their identical monomial selections. For $B4$, SVM-GRFE diverges strongly from FULL ($\theta = 80.0°$) while SVM-RFE stays close ($\theta = 21.1°$), consistent with their respective monomial selections. For $B2$, SVM-RFE stays very close to the full model ($\theta = 1.2°$, $\rho = 0.08$) while SVM-GRFE diverges ($\theta = 24.3°$, $\rho = 0.99$), consistent with SVM-GRFE's erroneous elimination of x_0 in this near-tie configuration. For $B5$, SVM-GRFE ($\theta = 58.7°$) is closer to FULL than SVM-RFE ($\theta = 72.6°$), yet achieves higher accuracy. For $B7$, SVM-GRFE reaches $\theta = 90°$ and $\rho = 6.21$: the boundary is not only rotated but

Table 3 Accuracy at $\varsigma = 0$ and $\varsigma = 0.50$ (averaged over all noise types).

Boundary	$\varsigma = 0$		$\varsigma = 0.50$		Δ_{acc}
	SVM-GRFE	SVM-RFE	SVM-GRFE	SVM-RFE	
	Accuracy	Accuracy	Accuracy	Accuracy	
$B1$	0.995	0.995	0.858	0.849	+0.009
$B2$	0.930	0.990	0.858	0.895	−0.037
$B3$	0.990	0.990	0.875	0.870	+0.005
$B4$	0.982	0.988	0.909	0.902	+0.007
$B5$	0.990	0.965	0.873	0.858	+0.015
$B6$	0.994	0.994	0.935	0.847	+0.088
$B7$	0.995	0.920	0.920	0.815	+0.105
Mean	0.982	0.977	0.890	0.862	+0.028

also substantially translated away from the distorted full model. SVM-RFE diverges less ($\theta = 65.5°$, $\rho = 0.12$) but retains x_4, achieving only Accuracy = 0.920. This confirms that proximity to the full model is not a reliable indicator of boundary quality in high-dimensional monomial spaces with detrimental features.

The margin column reveals a pattern complementary to accuracy. For FULL in $B5$ and $B7$, the margin collapses to extremely small values (0.007 and 0.020 respectively), indicating that the boundary is nearly tangent to the training points in the distorted polynomial space: any small perturbation of the physical trajectory is likely to cause a misclassification. SVM-GRFE in $B5$ obtains margins of 0.074 (10.6 times the margin of FULL) and 1.412 in $B7$ (70.6 times the margin of the full model), a dramatic improvement that is directly attributable to the elimination of detrimental monomials that dominated the weight vector $\mathbf{w}$ and inflated $\|\mathbf{w}\|$. This confirms the theoretical connection established in Section 3.1: a boundary that is accurate on training data but with a small margin will fail under noise or sensor drift, whereas SVM-GRFE's geometric criterion explicitly favors boundaries with larger margins by penalizing the elimination of monomials that would destabilize the weight vector. SVM-RFE, by contrast, retains x_4 in $B7$ and achieves a margin of only 0.088, only marginally better than the full model (0.020). For $B4$, SVM-GRFE achieves a margin of 0.219 versus 0.114 for SVM-RFE and 0.152 for the full model, suggesting that despite its lower accuracy, the SVM-GRFE boundary in $B4$ is more robust to future perturbations. For $B1$, $B3$, and $B6$, margins are comparable across methods, consistent with the identical boundaries produced.

4.4 Robustness to measurement noise

To evaluate whether the geometric criterion of SVM-GRFE produces boundaries that are more robust than those of SVM-RFE under realistic perturbations, we evaluate all three methods on corrupted versions of X_{test} while keeping X_{train} clean. Five noise types are tested: additive Gaussian, constant bias, linear drift, salt-and-pepper (random outliers), and heteroscedastic (noise proportional to $|x_{ij}|$). Each is applied at six intensity levels $\varsigma \in \{0, 0.05, 0.10, 0.20, 0.30, 0.50\}$, expressed as a fraction of the per-variable standard deviation of X_{train}. Since FULL uses all N_{poly} monomials while SVM-GRFE and SVM-RFE use a reduced subset, the primary comparison is between SVM-GRFE and SVM-RFE at equal monomial cardinality $|\Phi^*|$.

Table 3 reports the accuracy of SVM-GRFE and SVM-RFE without noise ($\varsigma = 0$) and at the highest noise level ($\varsigma = 0.50$), averaged over all five noise types. Since the five noise types are qualitatively different and may not be equally likely in practice, this average

Table 4 Mean accuracy at $\varsigma = 0.50$ by noise type, averaged over the seven scenarios.

Noise type	Full	SVM-GRFE	SVM-RFE	Δ_{acc}
	Accuracy	Accuracy	Accuracy	
Gaussian	0.941	0.916	0.911	+0.005
Bias	0.964	0.923	0.927	−0.004
Drift	0.976	0.965	0.946	+0.019
Heteroscedastic	0.840	0.870	0.769	+0.101
Salt-and-pepper	0.794	0.775	0.760	+0.015
Mean	0.903	0.890	0.863	+0.027

should be interpreted as a summary ranking indicator rather than a physically meaningful quantity. The difference $\Delta_{\text{acc}} = \text{Accuracy}_{\text{SVM-GRFE}} - \text{Accuracy}_{\text{SVM-RFE}}$ at $\varsigma = 0.50$ quantifies the robustness advantage of SVM-GRFE over SVM-RFE at equal complexity. SVM-GRFE outperforms SVM-RFE in six out of seven scenarios, with a mean advantage of $\Delta_{\text{acc}} = +0.028$. The advantage is most pronounced in $B6$ (+0.088) and $B7$ (+0.105): in $B6$, both methods select $\{x_1, x_2\}$ but SVM-GRFE produces a better-oriented boundary; in $B7$, SVM-GRFE correctly discards the detrimental x_4-related monomials while SVM-RFE retains them. The only scenario where SVM-RFE is more robust is $B2$ ($\Delta_{\text{acc}} = -0.037$) with or without noise.

Table 4 breaks down the accuracy at $\varsigma = 0.50$ by noise type, averaged over the seven scenarios. SVM-GRFE outperforms SVM-RFE in four out of five noise types, with the largest advantage for heteroscedastic noise ($\Delta_{\text{acc}} = +0.101$): since noise amplitude scales with measurement magnitude, the detrimental monomials retained by SVM-RFE (such as x_4-related terms in $B7$) are amplified, while SVM-GRFE has already discarded them. For drift, SVM-GRFE also shows a clear advantage ($\Delta_{\text{acc}} = +0.019$). Bias is the only noise type where SVM-RFE marginally outperforms SVM-GRFE ($\Delta_{\text{acc}} = -0.004$), a difference that falls within the resolution of the accuracy metric at the test set sizes considered. Full achieves the highest accuracy under all noise types except heteroscedastic, where SVM-GRFE surpasses it (0.870 vs. 0.840), confirming that the elimination of detrimental monomials can improve noise robustness even relative to the unpruned model. Table 5 reports, for each method and noise type, the mean over the seven scenarios of the lowest tested noise level $\varsigma \in \{0.05, 0.10, 0.20, 0.30, 0.50\}$ at which the accuracy first drops by more than 5% relative to the baseline without noise of each method. Since this threshold is averaged over scenarios, the resulting values may fall between two grid points and do not correspond to a single tested level. A value of 0.50 indicates that the 5% threshold was never reached within the tested range. Cells marked with † (see the note in the table) indicate that the threshold was never reached within the range tested. Drift is the most benign perturbation: Full and SVM-GRFE never reach the 5% threshold, while SVM-RFE degrades at $\varsigma = 0.407$. Salt-and-pepper is the most damaging, with SVM-RFE degrading at $\varsigma = 0.121$ versus 0.143 for SVM-GRFE. For all noise types, SVM-GRFE either matches or outperforms SVM-RFE.

5 Conclusion

This paper introduced SVM-GRFE, a FS method for learning compact, interpretable, and robust transition conditions in CPS from multivariate TS data. By operating on polynomial monomials rather than raw physical variables and combining classification accuracy with geometric criteria, SVM-GRFE identifies detrimental monomials while preserving an explicit

Table 5 Mean noise level at which accuracy first drops by more than 5% relative to the baseline without noise, by noise type.

Noise type	FULL	SVM-GRFE	SVM-RFE
Gaussian	0.471	0.379	0.364
Bias	0.471	0.314	0.343
Drift	$\geq 0.5^{\dagger}$	$\geq 0.5^{\dagger}$	0.407
Heteroscedastic	0.300	0.293	0.257
Salt-and-pepper	0.200	0.143	0.121

†The 5% threshold was never reached within the tested range $\varsigma \in [0, 0.50]$; the reported value is censored at the upper bound and is not a measured degradation level.

polynomial boundary directly reusable in diagnosis and control. Experiments on seven synthetic scenarios show that SVM-GRFE reduces model complexity while matching or outperforming the full model and standard SVM-RFE in most cases. Robustness evaluations under multiple noise types further demonstrate improved resilience to measurement uncertainty. Future work will focus on geometry-aware hyperparameter selection and validation on real CPS datasets. A case study assessment focusing on a large city will be conducted to examine scaling up. From a diagnostic standpoint, Beyond feature selection, the interpretable polynomial boundary expression produced by SVM-GRFE provides interpretable transition conditions that directly support fault isolation by explicitly identifying the physical variables involved in each mode transition.

References

1 T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *KDD*. ACM, 2019.

2 U. M Braga-Neto and E. R Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380, 2004. doi:10.1093/BIOINFORMATICS/BTG419.

3 C. G Cassandras and S. Lafortune. *Introduction to discrete event systems*. Springer, 3rd edition, 2021.

4 S. Chen, C. Gao, and P. Zhang. Incorporation of data-mined knowledge into black-box svm for interpretability. *ACM Trans. on Int. Syst. and Tech.*, 14(1):1–22, 2022. doi:10.1145/3548775.

5 C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. doi:10.1007/BF00994018.

6 H. Frohlich, O. Chapelle, and B. Scholkopf. Feature selection for support vector machines by means of genetic algorithm. In *ICTAI*, pages 142–148. IEEE, 2003.

7 I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3):389–422, 2002. doi:10.1023/A:1012487302797.

8 R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2):273–324, 1997. doi:10.1016/S0004-3702(97)00043-X.

9 B. Martin-Barragan, R. Lillo, and J. Romo. Interpretable support vector machines for functional data. *European Journal of Operational Research*, 232(1):146–155, 2014. doi:10.1016/J.EJOR.2012.08.017.

10 A. Salhi, R. Alshamrani, A. Althbiti, A. Ismail, M. Abd-ElRahman, and B. M Hassan. Optimizing high dimensional data classification with a hybrid ai driven feature selection framework and machine learning schema. *Scientific Reports*, 15(1):35038, 2025.