Uncertainty-Aware Model-Agnostic Dynamic Feature Selection

Anonymous authors
Paper under double-blind review

Abstract

Dynamic feature selection (DFS) acquires features sequentially under a budget, making it attractive for cost-sensitive decisions. Existing methods train a predictor specifically for sequential acquisition and evaluate it mainly through accuracy–budget curves. This approach leaves two questions open that are significant for deployment: how to know whether the decision under partial observation is truly reliable, and how to deploy DFS in settings where the predictor is already validated or in production. We argue that a pre-trained model can be reused if the acquisition policy explicitly accounts for the subset-dependent uncertainty that partial observation creates. Our central claim is that a native DFS predictor is not always necessary, and that while some decisions with little information are inherently uncertain, it is possible to search for the ones that are more reliable for each budget. We first show that DFS introduces uncertainty sources absent from static prediction: adapting a model across feature subsets induces subset-dependent epistemic uncertainty, imputing missing features biases aleatoric estimates, and predictive entropy can fall even under a poor policy. We then propose an uncertainty-aware, model-agnostic framework that adapts pre-trained neural, tree-based, and rule-based classifiers through efficient subset reparametrisation, and penalises epistemically unstable acquisitions using auxiliary predictors. Experiments show the framework is competitive with and sometimes better than state-of-the-art greedy and reinforcement-learning DFS baselines on tabular and image data. Beyond accuracy, they also show that calibration does not necessarily improve as more features are acquired, and that incorporating epistemic uncertainty provides a useful reliability-control signal for partial-observation decisions.

1 Introduction

Many prediction problems are constrained not by model capacity but by the cost of acquiring new information. In clinical diagnosis, industrial inspection, or interactive data collection, every additional laboratory test, sensor reading, or user query has a monetary, temporal, and human cost. The relevant variables are often instance-specific: two patients may require different tests, and two images may require different regions to be inspected. Dynamic feature selection (DFS) addresses this setting by sequentially acquiring features for each instance under a budget (Melville et al., 2004). Figure 1 illustrates an example of the contrast between static and dynamic feature selection in a medical-testing scenario, where a fixed test panel can be inefficient for some patients while DFS adapts acquisition to each case.

The existing literature on DFS primarily explores two main approaches: reinforcement learning (RL) (Janisch et al., 2019) and greedy optimisation (Chen et al., 2015b). RL-based approaches offer the potential to discover optimal policies for feature acquisition, but they come with the inherent challenges of RL training, including significant computational demands and potential convergence issues (Erion et al., 2022). Greedy optimisation methods maximise a measure that relates the known inputs with the response variable (Chattopadhyay et al., 2023; Takahashi et al., 2025). Usually, this measure is the conditional mutual information (CMI) between input variables and the response variable, which has some desirable qualities in terms of uncertainty minimisation and prediction power (Gadgil et al., 2023), although it has non-differentiability issues (Covert et al., 2023). These approaches can focus on the CMI-estimation task directly (Gadgil et al.,

2023), circumventing the issue of selecting a discrete variable in a neural network, or by using generative models (Ma et al., 2019). Nevertheless, both approaches rely on predictors that operate across a large number of feature subsets. This requirement is closely related to the core problem studied in Invariant Risk Minimisation (Krueger et al., 2021), where a single model must generalise across multiple environments. These subset-adaptive methods allow specialisation while retaining shared knowledge across subsets. Examples of this approach are hierarchical or modular parametrisations (Rebuffi et al., 2018), mixture-of-experts architectures (Shazeer et al., 2017), or meta-learning strategies (Finn et al., 2017).

Despite the benefits of DFS in budget-sensitive scenarios, existing methods have struggled with two major limitations that hinder their application in real-life scenarios:

- **Uncertainty under partial observation.** Existing DFS methods optimise acquisition for predictive performance, but often do not separate aleatoric uncertainty, subset-induced epistemic uncertainty, and calibration during sequential acquisition. This gap prevents practitioners from distinguishing between predictions that are confidently correct versus those that merely appear plausible.
- **Compatibility with deployed predictors.** Existing DFS methods require classifiers specifically designed to handle arbitrary feature subsets, making them incompatible with models already validated and deployed in practice. Training DFS-native methods from scratch can be computationally expensive or infeasible in some cases, e.g., the training data is no longer available. This creates a significant barrier between DFS research and real-world deployment.

Our central claim is that effective DFS does not require a predictor trained specifically for sequential acquisition: an already-validated, pre-trained model can be reused, provided that the acquisition policy explicitly accounts for the subset-dependent uncertainty that partial observation introduces. We develop and support this claim through the following contributions:

- **A model-agnostic DFS framework (main contribution).** We adapt pre-trained classifiers to the sequential setting through subset reparametrisation, and steer acquisition with an epistemic penalty that avoids unstable partial-observation decisions, without retraining the original predictor.
- **Uncertainty analysis of DFS.** We formalise three uncertainty issues that make naive subset reuse unreliable and that each component of the framework is designed to address: a subset adaptation risk gap Δ_S, aleatoric bias induced by feature imputation, and the non-monotonic relationship between predictive entropy and acquisition quality.
- **Evidence across modalities.** We evaluate predictive performance, calibration, and prediction bias across tabular and image datasets, showing both when the framework is competitive with native DFS methods and when a fixed static subset remains preferable.

2 Background and Problem Setup

Notation and Problem Formulation. Let $\mathbf{x} = (x_1, \ldots, x_M)$ be an input feature vector and $y \in \{1, \ldots, C\}$ a target label. The training set is denoted by $\mathcal{D} = \{(\mathbf{x}^n, y^n)\}_{n=1}^N$. For a subset of observed feature indices $S \subseteq \{1, \ldots, M\}$, we write $\mathbf{x}_S = \{x_i : i \in S\}$ and $\bar{S} = \{1, \ldots, M\} \setminus S$ for the unobserved indices. Each feature can have an acquisition cost $c_i > 0$, and a subset is feasible when its total cost $c(S) = \sum_{i \in S} c_i$ does not exceed a budget B.

Dynamic Feature Selection. Dynamic feature selection (DFS) frames feature acquisition as a sequential decision problem: rather than committing to a single global subset before observing a sample, a policy receives the current partial input and chooses which unobserved feature to query next. Formally, at step t the policy selects $i_t \in \bar{S}_t$, observes x_{i_t}, and updates $S_{t+1} = S_t \cup \{i_t\}$ until the budget is exhausted or a stopping rule is met. The final prediction is then made from the acquired subset $\mathbf{x}_{S_T}$. The objective is therefore to obtain both a predictor that performs well for the possible feature subsets queried by the selector, and a selector

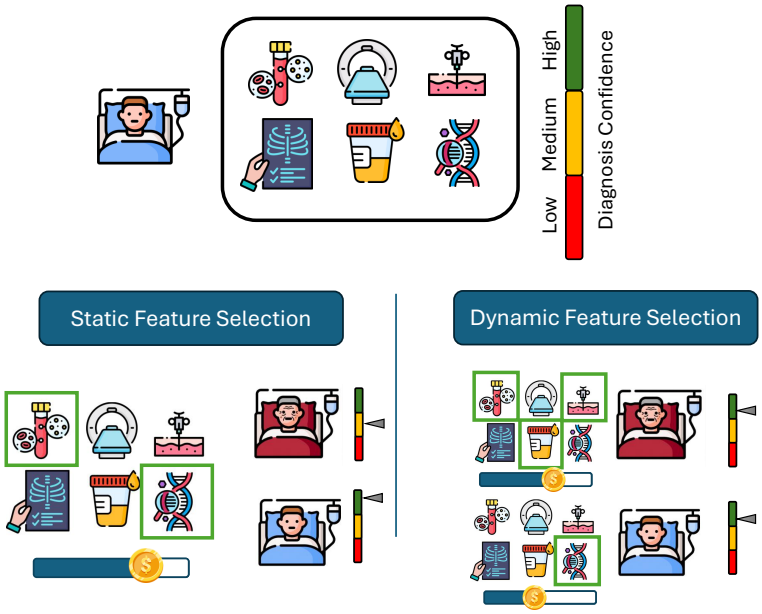


Figure 1: **Comparison of static and dynamic feature selection in a medical setting.** Static feature selection chooses a fixed set of features within the budget, which may be unsuitable for individual patients and may request unnecessary tests. Dynamic feature selection adapts the choice of tests to each patient, producing more consistent confidence while reducing redundant tests.

that chooses the right feature at each moment. This distinguishes DFS from static feature selection, where a single subset is selected once and then used for all test instances, and the predictor does not need to accommodate different feature subsets.

Uncertainty Quantification in Machine Learning. Uncertainty quantification (UQ) measures how much a model's prediction should be trusted. Following the taxonomy given in Hüllermeier & Waegeman (2021), predictive uncertainty is commonly decomposed into aleatoric and epistemic components. Aleatoric uncertainty refers to irreducible ambiguity in the data-generating process, such as class overlap or measurement noise. Epistemic uncertainty refers to uncertainty caused by limited knowledge about the predictor, its parameters, or the region of the input space being evaluated; in principle, it can be reduced by better data, models, or adaptation. Formally, given a distribution Q over plausible parameters $\Theta \sim Q$, the total predictive uncertainty can be decomposed additively as $H(Y \mid \mathbf{x}) = H(Y \mid \mathbf{x}, \Theta) + I(Y; \Theta \mid \mathbf{x})$, where the first term is the aleatoric component and the second the epistemic one. A way to estimate this decomposition is through an ensemble, whose spread across members reflects the parameter uncertainty that the epistemic term is meant to capture. Concretely, given an ensemble of K predictors $\{p_i(\hat{y} \mid \mathbf{x})\}_{i=1}^{K}$ sampled from Q, with mean prediction $\bar{p}(\hat{y} \mid \mathbf{x}) = \frac{1}{K}\sum_{i=1}^{K} p_i(\hat{y} \mid \mathbf{x})$, the decomposition above is approximated by

$$\underbrace{H\big(\bar{p}(\hat{y} \mid \mathbf{x})\big)}_{\text{total}} \approx \underbrace{\frac{1}{K}\sum_{i=1}^{K} H\big(p_i(\hat{y} \mid \mathbf{x})\big)}_{\text{aleatoric}} + \underbrace{\frac{1}{K}\sum_{i=1}^{K} D_{\mathrm{KL}}\big(p_i(\hat{y} \mid \mathbf{x}) \,\big\|\, \bar{p}(\hat{y} \mid \mathbf{x})\big)}_{\text{epistemic}}. \tag{1}$$

The aleatoric term is the average entropy of the individual predictors, whereas the epistemic term is their average disagreement (KL divergence) with the ensemble mean, which estimates the mutual information between the label and the model parameters.

3 Uncertainty Induced by Dynamic Feature Selection

In this section we show that uncertainty in DFS is not simply the uncertainty of a static classifier evaluated on fewer variables. Instead, sequential partial observation changes the statistical target, the optimality of the parameters and the interpretation of predictive confidence according to the features observed. This motivates the model-agnostic procedure introduced in Section 4.

3.1 Subset-Dependent Epistemic Uncertainty

In the DFS literature, the classifier parameters, θ, are usually trained using the following expression (Gadgil et al., 2023):

$$\theta^* = \arg\min_{\theta} \mathbb{E}_S \mathbb{E}_{\mathbf{x},y} l(f_\theta(\mathbf{x}_S), y), \tag{2}$$

where f_θ denotes a model capable of making predictions given any subset of features S. To evaluate $f_\theta(\mathbf{x}_S)$, missing features are typically handled through some form of imputation method, such as zero/mean. Although solving Eq. (2) results in a parameter configuration that minimises the expected loss across all possible subsets, the resulting model may differ from a local model trained exclusively on a particular subset of features with an equivalent training objective:

$$\theta_S^* = \arg\min_{\theta_S} \mathbb{E}_{\mathbf{x},y} l(f_{\theta_S}(\mathbf{x}_S), y). \tag{3}$$

If we measure the risk of a model on subset S as $\mathcal{R}_S(f) = \mathbb{E}_{\mathbf{x},y} l(f(\mathbf{x}_S), y)$, we can expect both models to have different risks. We measure the discrepancy of f_{θ^*} and $f_{\theta_S^*}$ as their difference in risk:

$$\Delta_S = \mathcal{R}_S(f_{\theta^*}) - \mathcal{R}_S(f_{\theta_S^*}). \tag{4}$$

A positive Δ_S indicates that the globally trained model f_{θ^*} performs worse on the specific subset S than a hypothetical local optimum $f_{\theta_S^*}$, reflecting a loss of specialisation caused by averaging across subsets in Eq. (2). Conversely, $\Delta_S \approx 0$ implies that f_{θ^*} generalises well to S, achieving near-optimal performance despite joint training.

The exponential growth of the feature subset space, 2^M, forces θ^* to balance many local optima and unseen feature subsets, which may lead to large Δ_S. Because of this, it is common to train the predictor jointly with a feature selector q_ϕ rather than sampling S uniformly. This changes Eq. (2) into:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{\mathbf{x},y} \mathbb{E}_{S \sim q_\phi(\cdot|\mathbf{x})} l(f_\theta(\mathbf{x}_S), y). \tag{5}$$

However, this approach also has problems. The selector q_ϕ may explore only a small fraction of all possible feature subsets S. This results in a trade-off where predictive power is diminished for rarely selected subsets. Moreover, this scheme is only effective when the predictor guides the feature selection process, but it is vulnerable to changes in external factors. For example, when relevant variables needed by the selector are unavailable for certain samples.

3.2 Aleatoric Bias Under Partial Observation

The ideal DFS classifier should properly estimate $p(y|\mathbf{x}_S)$ for every S in the query process:

$$p(y \mid \mathbf{x}_S) = \int_{\mathbf{x}_{\bar{S}}} p(y \mid \mathbf{x}_S, \mathbf{x}_{\bar{S}}) p(\mathbf{x}_{\bar{S}} \mid \mathbf{x}_S) \, d\mathbf{x}_{\bar{S}}. \tag{6}$$

However, many classical imputation methods for the unobserved features bias this estimation. For example, if we use the classical mean imputation method, we are implicitly assuming a degenerate distribution where the unobserved features are fixed at their expected values, $\mu(\mathbf{x}_{\bar{S}})$. Consequently, Eq. (6) simplifies to $p(y \mid \mathbf{x}_S, \mathbf{x}_{\bar{S}} = \mu(\mathbf{x}_{\bar{S}}))$, which biases the prediction.

In previous DFS literature, there have been two options to solve this issue. One is to use a generative model to estimate $p(\mathbf{x}_{\bar{S}} \mid \mathbf{x}_S)$ (Ma et al., 2019). However, this approach requires a lot of data to work properly

and tends to overfit. The alternative approach consists of imputing the missing values in $\mathbf{x}_{\bar{S}}$ with a fixed value. Then, the input to the classifier is modified by concatenating the observed feature vector with a mask that indicates which features are really observed. This technique allows the model to differentiate between imputed and truly observed features. Nonetheless, the effectiveness of augmenting the input with a missingness mask depends on how informative the missingness pattern is with respect to the response. When missingness carries little additional signal, the mask variables contribute minimally to the prediction, and the resulting model behaves similarly to standard mean imputation (Van Ness et al., 2023). Thus, masks help encode which features are observed, but they do not by themselves guarantee an unbiased estimate of the marginal conditional in Eq. (6).

3.3 Predictive Entropy Does Not Identify Good Policies

A common implicit assumption in dynamic feature selection is that acquiring additional features monotonically improves prediction quality and reduces uncertainty. However, neither prediction uncertainty nor model risk is guaranteed to evolve monotonically. For instance, in the diagnosis of a rare disease (0.1% prevalence), the prior predictive distribution is highly skewed and low-entropy; however, epistemic uncertainty is also very high. After observing a relevant but uncertain test result, the posterior may approach a balanced distribution, increasing predictive entropy and reducing confidence, even though the epistemic uncertainty was reduced.

Epistemic uncertainty is also not guaranteed to decrease when querying additional features, due to model adaptation effects. While it is commonly assumed that predictions made with $S \cup \{i\}$ are more reliable than those made with S, this requires the monotonicity condition $\mathcal{R}_{S\cup\{i\}}(f_{\theta^*}) \leq \mathcal{R}_S(f_{\theta^*})$ to hold. However, this condition is highly problem-dependent and might be violated in practice, particularly when a single model must adapt to many different feature subsets. Existing DFS methods rely on jointly training feature selectors and predictors to mitigate this effect (Chattopadhyay et al., 2023). However, distribution shift and limited exploration in training can reintroduce the non-monotonic behaviour of model risk.

Another related issue is whether total entropy reductions are good signs of good feature selection policies or if it happens even for random querying. Prior work has shown the tendency of classifiers to be overconfident on Out-of-Distribution data (Kendall & Gal, 2017). So, imagine that a random policy guides the feature querying process. As imputed placeholders are replaced by observed values, the input can easily drift towards out-of-distribution samples with overconfident estimates. Distinguishing good from bad policies therefore requires a signal tied to correctness, or calibration, rather than predictive entropy alone, as we confirm in Section 6.

4 Model-Agnostic Dynamic Feature Selection

Directly training models for DFS is not always practical: large pre-trained models are expensive to fine-tune (Kumar et al., 2022), and subset-specific adaptation of interpretable models such as rule-based classifiers incurs prohibitive overhead (Appendix D). Moreover, practitioners often need to incorporate models that have already been validated in deployment, where retraining from scratch is neither feasible nor desirable.

We therefore propose a model-agnostic DFS framework built around a single global predictor f_θ. Missing features are first imputed so that any pre-trained predictor can be evaluated on a complete input vector. When desired, a subset-dependent reparametrisation module can then be trained while keeping the original predictor fixed, allowing the model to adapt some of its parameters to the currently observed subset without full retraining. Our DFS policy also explicitly estimates the epistemic uncertainty induced by working with partial information, which allows it to select features that lead to more reliable decisions. Figure 2 summarises the general scheme of the proposal, while Algorithm 1 shows the pseudocode for training and inference. The procedure is highly modular: the global predictor can be neural, tree-based, or rule-based. The subset adapter is optional as well, and the acquisition policy can be greedy or reinforcement-learning based.

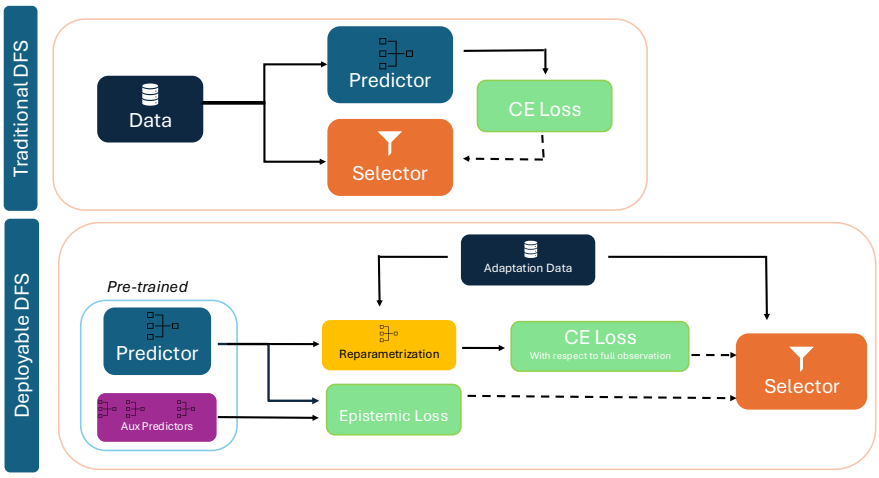


Figure 2: **General scheme of the proposed uncertainty-aware model-agnostic DFS framework.** Existing DFS methods train predictors from scratch for DFS, whereas our proposal adapts an already trained model for DFS use. The selector queries features that make partial-observation decisions match the hypothetical full-observation decisions, while using epistemic uncertainty to favour more reliable acquisitions. An optional subset reparametrisation module adapts the predictor without retraining its original parameters, while epistemic uncertainty is estimated using auxiliary predictors. Dashed arrows mark selector labels.

4.1 Subset Reparametrisation of Pre-Trained Models

The challenges identified in Section 3 become particularly relevant when working with a pre-specified global model f_θ that was not designed for dynamic feature selection. Such models expect inputs in the full feature space and therefore cannot natively process arbitrary observed subsets $\mathbf{x}_S$.

To enable evaluation on any subset S, we complete the partially observed input by imputing the missing features and then evaluate the original model on the imputed representation. This allows the fixed model f_θ to produce predictions for arbitrary subsets. However, these predictions depend on the imputation mechanism and generally differ from the true conditional $p(y \mid \mathbf{x}_S)$, introducing an approximation bias. To reduce this discrepancy, we introduce a subset-dependent reparametrisation of the global model:

$$\theta_S = R_\psi(\theta, S), \tag{7}$$

where R_ψ is a reparametrisation function with parameters ψ, trained to minimise the expected risk across feature subsets:

$$\psi^* = \arg\min_\psi \mathbb{E}_S \, \mathbb{E}_{x,y} l\big(f_{R_\psi(\theta,S)}(\mathbf{x}_S), y\big). \tag{8}$$

This reparametrisation reduces the subset-specific risk gap Δ_S and mitigates the bias induced by imputation, bringing predictions closer to the true conditional $p(y \mid \mathbf{x}_S)$. Compared to full subset-specific training, this approach is computationally efficient, preserves the original parameters of f_θ, and avoids degrading a previously validated model. It also does not require retraining of the original model. For neural networks, R_ψ can adapt only the final layer or consist of low-rank updates (Hu et al., 2022).

For interpretable classifiers, reparametrisation can exploit rule-based and tree-based inference to aim for the optimal target at each time. Given a partially observed subset S, we remove antecedent conditions involving unobserved features, prune rules that no longer contain any observed condition, and re-estimate the class confidence of each reduced rule on the training samples satisfying its reduced antecedent. These reduced-rule confidences can be cached over all sub-rules, making subset adaptation a mere inference-time lookup and restricting acquisition to features that can still influence an active rule. Appendix C gives the full rule-level formulae and caching details. Note that this is a light form of reparametrisation in which neither the structure of the tree nor the rules used are changed, so this limits the optimality of the approximation in some subsets, which is a reasonable trade-off for the speed of this reparametrisation.

Algorithm 1 Model-agnostic DFS with epistemic-aware acquisition

Require: global predictor f_θ, imputation rule m, training data $\mathcal{D}$, auxiliary models $\mathcal{I}$, feature costs c, budget B

Ensure: prediction after acquiring an instance-specific subset S

Training

Define completed inputs $\tilde{\mathbf{x}}_S \leftarrow m(\mathbf{x}_S, S)$ by imputing unobserved features
if a subset adapter is used **then**
  fit $\psi^* \leftarrow \arg\min_\psi \mathbb{E}_S \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \mathcal{L}\big(f_{R_\psi(\theta,S)}(\tilde{\mathbf{x}}_S), y\big)$ with θ fixed
else
  set $R_\psi(\theta, S) \leftarrow \theta$
end if
for all training states $(\tilde{\mathbf{x}}_S, S)$ and candidates $i \in \bar{S}$ **do**
  $\Delta u_i \leftarrow u(\tilde{\mathbf{x}}_S) - u(\tilde{\mathbf{x}}_{S\cup\{i\}})$
  $e_i \leftarrow e(\tilde{\mathbf{x}}_{S\cup\{i\}})$
  set acquisition target (RL reward) $q_i \leftarrow \Delta u_i - \lambda e_i$
  set $q_i \leftarrow q_i / c_i$ for variable feature costs
end for
if greedy acquisition is used **then**
  fit the dual-head estimator $\phi^* \leftarrow \arg\min_\phi \mathcal{L}_{\text{greedy}}(\phi)$
  $\mathcal{L}_{\text{greedy}}(\phi) = \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}} \sum_{i\in\bar{S}} \big[\big(h_{u,i}(\tilde{\mathbf{x}}_S) - \Delta u_i\big)^2 + \big(h_{e,i}(\tilde{\mathbf{x}}_S) - e_i\big)^2\big]$
else
  fit $\phi^* \leftarrow \arg\min_\phi \mathcal{L}_{\text{TD}}(\phi)$ for $Q_\phi(\tilde{\mathbf{x}}_S, i)$
  $\mathcal{L}_{\text{TD}}(\phi) = \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}, i\in\bar{S}} \Big(Q_\phi(\tilde{\mathbf{x}}_S, i) - \big[q_i + \gamma \max_{j\in\overline{S\cup\{i\}}} Q_{\bar{\phi}}(\tilde{\mathbf{x}}_{S\cup\{i\}}, j)\big]\Big)^2.$
end if

Test-time inference for one instance

$S \leftarrow \emptyset$ and $b \leftarrow B$
while $b > 0$ and $\bar{S} \neq \emptyset$ **do**
  impute missing features to form $\tilde{\mathbf{x}}_S \leftarrow m(\mathbf{x}_S, S)$
  for all $i \in \bar{S}$ with $c_i \leq b$ **do**
    score i using $\big(h_{u,i}(\tilde{\mathbf{x}}_S) - \lambda\, h_{e,i}(\tilde{\mathbf{x}}_S)\big)/c_i$ or the learned RL value $Q_\phi(\tilde{\mathbf{x}}_S, i)$
  end for
  $i^\star \leftarrow \arg\max_i \text{score}(i)$
  acquire $x_{i^\star}$
  $S \leftarrow S \cup \{i^\star\}$ and $b \leftarrow b - c_{i^\star}$
end while
if a subset adapter is used **then**
  set $\tilde{\theta} \leftarrow R_\psi(\theta, S)$
else
  set $\tilde{\theta} \leftarrow \theta$
end if
impute missing features to form $\tilde{\mathbf{x}}_S \leftarrow m(\mathbf{x}_S, S)$
return $f_{\tilde{\theta}}(\tilde{\mathbf{x}}_S)$

Patch-wise reparametrisation for image models. For convolutional backbones, imputing unobserved patches and re-running the network feeds the backbone an out-of-distribution input. We instead reparametrise the backbone as a *per-patch encoder*: each patch x_p is upsampled to the backbone's native resolution and encoded independently,

$$z_p \;=\; f_\theta\big(\text{up}(x_p)\big) \in \mathbb{R}^d, \tag{9}$$

so that the embedding z_p depends only on the pixels of patch p and can be cached once. Given an observed subset S, a small Transformer encoder aggregates the variable-size set $\{z_p : p \in S\}$ over a learnable classification token, using learned positional embeddings for each patch location and excluding unobserved patches.

The final classification distribution is read from the classification token (Devlin et al., 2019). Modelling inter-patch structure with set attention allows the module to understand the spatial relationship between the patches, which is particularly relevant for low budgets. We use pre-norm layer normalisation rather than batch normalisation, whose running statistics would otherwise mix the subset sizes seen during training and mismatch each fixed-budget evaluation. The aggregator is trained by cross-entropy on randomly sampled subsets and then frozen, after which the greedy policy is trained.

4.2 Epistemic-Aware Acquisition Objective

Our DFS policy aims to sequentially select features so as to approximate the global model's predictions for each sample. For that, we define a value function, $v_q(i, \mathbf{x}_S)$, that computes the expected added value of adding the feature i given the currently observed $\mathbf{x}_S$. At each step of the DFS, we compute this expression for each unobserved feature.

To compute $v_q(i, \mathbf{x}_S)$ we use a function $q(x_i, \mathbf{x}_S)$ that quantifies the improvement of the prediction quality of $f_\theta(\mathbf{x}_{S\cup\{i\}})$ with respect to the prediction with all features, $f_\theta(\mathbf{x})$:

$$v_q(i, \mathbf{x}_S) = \mathbb{E}_{p(x_i|\mathbf{x}_S)}\left[q(x_i, \mathbf{x}_S)\right]. \tag{10}$$

Function $q(x_i, \mathbf{x}_S)$ combines two components: the predictive improvement from adding feature i, captured by the reduction in discrepancy $u(\mathbf{x}_S) - u(\mathbf{x}_{S\cup\{i\}})$, and the epistemic uncertainty of the resulting model, captured by $e(\mathbf{x}_{S\cup\{i\}})$:

$$q(x_i, \mathbf{x}_S) = u(\mathbf{x}_S) - u(\mathbf{x}_{S\cup\{i\}}) - \lambda e(\mathbf{x}_{S\cup\{i\}}). \tag{11}$$

The parameter λ regulates the relative importance of epistemic uncertainty in the DFS process.

We can opt to have a greedy or an RL approach. If we use a greedy policy, we just choose the feature that maximises v_q. To use RL we formulate the problem as a finite-horizon Markov decision process, where the state at time t is the set of observed features $\mathbf{x}_{S_t}$, the action corresponds to selecting the next feature $i \in \overline{S_t}$ to query, and the immediate reward is given by $v_q(i, \mathbf{x}_{S_t})$. In this setting, the optimal action-value function $Q^*(\mathbf{x}_S, i)$ satisfies the Bellman optimality equation (Sutton et al., 1998):

$$Q^*(\mathbf{x}_{S_t}, i) = v_q(i, \mathbf{x}_{S_t}) + \gamma\, \mathbb{E}_{p(x_i|\mathbf{x}_{S_t})}\Big[\max_{i' \in \overline{S_t \cup \{i\}}} Q^*(\mathbf{x}_{S_t\cup\{i\}}, i')\Big] \tag{12}$$

where $\gamma \in [0, 1]$ is the discount factor, and the expectation accounts for uncertainty over the value of the queried feature x_i. The policy then selects features according to $\pi^*(\mathbf{x}_S) = \arg\max_{i\in\overline{S}} Q^*(\mathbf{x}_S, i)$.

In both approaches, to support variable feature costs, we divide $v_q(i, \mathbf{x}_S)$ by the cost of the feature, so that the policy optimises q per unit cost. This is not generally optimal (Chen et al., 2015b), but it works well in practice and is cheap to compute.

The next step is computing v_q. For that, we need to know: (i) how to compute u using the global prediction (Section 4.2.1); (ii) how to estimate the epistemic uncertainty, e, for a given decision (Section 4.2.2); and (iii) how to estimate $v_q(i, \mathbf{x}_S)$ at test time (Section 4.2.3).

4.2.1 Prediction Discrepancy Objective

For computing the prediction difference of the sub-model defined by S with respect to the full model, we consider the Kullback–Leibler divergence of the model predictions:

$$u(\mathbf{x}_S) = D_{\mathrm{KL}}(p(\hat{y}|\mathbf{x}) \| p(\hat{y}|\mathbf{x}_S)). \tag{13}$$

When an additional feature x_i is queried, this becomes:

$$u(\mathbf{x}_{S\cup\{i\}}) = D_{\mathrm{KL}}(p(\hat{y}|\mathbf{x}) \| p(\hat{y}|\mathbf{x}_{S\cup\{i\}})). \tag{14}$$

As shown in Appendix B, the reduction in Eq. (13) induced by querying x_i admits an information–theoretic interpretation and is related to conditional mutual information (CMI) (Covert et al., 2023; Gadgil et al., 2023;

Ma et al., 2019). Both quantities involve a log-ratio term comparing predictive distributions before and after observing x_i, but differ in the reference measure used for weighting. CMI quantifies the statistical dependence between x_i and $\hat{y}$ conditioned on $\mathbf{x}_S$, whereas our criterion measures how much observing x_i reduces the discrepancy between the subset-based predictor and a fixed global model. Although not equivalent in general, they align when $p(\hat{y} \mid \mathbf{x}_S, x_i)$ closely matches $p(\hat{y} \mid \mathbf{x})$, or when x_i induces a highly concentrated posterior.

4.2.2 Epistemic Uncertainty Estimation

Ensembles are considered state-of-the-art for epistemic uncertainty quantification, as prediction variance across members captures ignorance about a decision (Rahaman et al., 2021; Ovadia et al., 2019). When feasible, this is the preferred solution. For large models where ensembles are computationally prohibitive, alternatives include MC-dropout (Gal & Ghahramani, 2016), provided the global model has dropout layers, or feature-space similarity metrics (Liu et al., 2020). For explainable models, ensembles are also not desirable, as aggregating predictions across ensemble members obscures the decision logic (Bassan et al., 2025). In that case, we propose to keep the fixed primary predictor for decision-making and explanation, and use auxiliary models only to quantify epistemic uncertainty around that predictor's output. Specifically, let $\mathcal{I}$ denote a set of auxiliary classifiers trained using different random initialisations or bootstrap samples. For a given sample with observed features $\mathbf{x}_S$, let $p_\theta(\hat{y}|\mathbf{x}_S)$ denote the prediction of the primary model with parameters θ, and let $p_{\theta_i}(\hat{y}|\mathbf{x}_S)$ denote the prediction of auxiliary model i with parameters θ_i. We define epistemic uncertainty as the average divergence between the primary prediction and the auxiliary predictions:

$$e(\mathbf{x}_S) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} D_{\mathrm{KL}} \left(p_{\theta_i}(\hat{y} \mid \mathbf{x}_S) \,\|\, p_\theta(\hat{y} \mid \mathbf{x}_S) \right). \tag{15}$$

This expression quantifies how atypical the primary model's predictive distribution is relative to a set of alternative plausible predictors. If $\bar{p}(\hat{y} \mid \mathbf{x}_S) = |\mathcal{I}|^{-1} \sum_{i \in \mathcal{I}} p_{\theta_i}(\hat{y} \mid \mathbf{x}_S)$, then Eq. (15) decomposes as

$$e(\mathbf{x}_S) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} D_{\mathrm{KL}}(p_{\theta_i} \| \bar{p}) + D_{\mathrm{KL}}(\bar{p} \| p_\theta), \tag{16}$$

where the first term is the usual ensemble disagreement (the epistemic term of the standard decomposition in Eq. (1)) and the second term measures the mismatch between the fixed primary model used for prediction and the auxiliary-model consensus. Thus, $e(\mathbf{x}_S)$ is sensitive to cases where the deployed predictor is confident but unstable under training perturbations, while still allowing explanations or decisions to be produced by a single primary model. Appendix A gives the derivation and discusses the metric's properties.

4.2.3 Inference-Time Value Estimation

Since unqueried features are unavailable at inference, we train a neural estimator to approximate $v_q(i, \mathbf{x}_S)$ from observed features $\mathbf{x}_S$. The estimator uses a dual-head architecture to jointly predict both the discrepancy reduction $u(\mathbf{x}_S) - u(\mathbf{x}_{S \cup \{i\}})$ and the epistemic term $e(\mathbf{x}_{S \cup \{i\}})$, trained by minimising squared prediction error across all possible feature additions (architecture and training details in Appendix G.2).

5 Experimental Protocol

The experiments are designed to test whether the proposed framework is not only accurate, but also useful for diagnosing the uncertainty issues identified in Section 3.

5.1 Research Questions

Our evaluation is organised around four questions that separate benchmark performance from diagnostic reliability:

- **RQ1:** Does model-agnostic DFS achieve competitive accuracy compared to native DFS methods?

- **RQ2:** Does explicitly including epistemic uncertainty in the feature selection policy affect the performance and epistemic uncertainty of the final decisions?
- **RQ3:** Do existing DFS methods exhibit the subset-adaptation, calibration, aleatoric-bias, and entropy-based pathologies predicted by our analysis?
- **RQ4:** Does the performance gap between native DFS methods and our model-agnostic DFS proposal stem from predictor performance or from successful DFS policy implementation?

5.2 Datasets, Baselines, and Metrics

Our empirical evaluation considers five tabular datasets from both medical and non-medical domains: *Diabetes*, *Heart Disease*, and *Cirrhosis*; and *Wine* and *Yeast*. These datasets range from a few hundred to over 10^5 samples and from 8 to 47 features, capturing substantial variability in both scale and dimensionality. For image datasets, we use CIFAR-10 (Krizhevsky et al., 2009), which contains 60,000 32×32 colour images across 10 classes; MNIST (Lecun et al., 1998), consisting of 70,000 28×28 grayscale images across 10 handwritten-digit classes; and Imagenette (Howard, 2019), a subset of ImageNet with 13,394 images across 10 classes. Full dataset statistics for each tabular dataset are summarised in Appendix G.1. We measure predictive performance using classification accuracy and calibration quality using the Expected Calibration Error (ECE), which evaluates the alignment between predictive confidence and empirical accuracy across confidence bins.

We compare both static and dynamic feature selection approaches. For static selection, we include mutual information-based feature scoring with a multilayer perceptron (MLP) classifier. We also include TabNet with budget constraints, which sees all available features and selects relevant ones until the budget is exhausted (Arik & Pfister, 2021). For pure DFS, we evaluate: (i) CMI-based greedy optimisation (DIME) (Covert et al., 2023), (ii) variational greedy information pursuit (VIP) (Chattopadhyay et al., 2023), (iii) Q-learning-based RL (CWCF) (Janisch et al., 2019), and (iv) actor-critic networks (INVASE) (Yoon et al., 2018). We evaluate our DFS adaptation method using different classifiers: a standard MLP and its reparametrised variant (ReMLP), and two rule-based approaches: CART decision trees (Breiman et al., 2017) and fuzzy decision trees (Ftree) (Fumanal-Idocin et al., 2025). For tabular data, we use both greedy and RL-based DFS learned policies. To denote each method, we use the "G-" prefix for those methods using a greedy policy and the "RL-" prefix for those using RL. For image datasets, we only use greedy approaches for computational reasons. Hyperparameter selection and further details for all methods are provided in Appendix G.3.

6 Results

6.1 RQ1: Competitive Accuracy of Model-Agnostic DFS

Tabular variable-cost benchmark. Table 1 reports the main tabular benchmark using randomly generated variable feature costs drawn uniformly from $[1, 4]$. The proposed G-ReMLP variant obtains the best overall accuracy and rank (67.58%, rank 3.4), showing that a model-agnostic DFS policy can be competitive with native DFS methods that train predictors specifically for arbitrary feature subsets. VIP is the strongest native method (66.76%, rank 4.0), followed by the proposed RL-ReMLP (65.57%, rank 5.1) and RL-MLP (64.97%). Static CMI remains a strong baseline (62.33%), outperforming several dynamic methods and highlighting that DFS is not automatically superior when a fixed subset captures most of the predictive signal.

The ranking also shows that competitiveness depends on the backbone–policy pair. The best acquisition strategy is itself backbone-dependent for neural predictors: ReMLP is strongest under greedy selection (G-ReMLP: 67.58% vs. RL-ReMLP: 65.57%), whereas MLP is strongest under RL selection (RL-MLP: 64.97% vs. G-MLP: 61.75%). Reparametrisation improves over the plain MLP backbone on aggregate under both policies, and on every dataset under greedy selection, although under RL selection MLP is stronger on Cirrhosis, Diabetes, and Wine. By contrast, rule-based backbones are more policy-sensitive: G-CART reaches 63.49%, comparable to Static CMI (62.33%) and INVASE (61.50%), whereas RL-CART and RL-FTree are not competitive. Thus, the answer to RQ1 is positive but conditional: model-agnostic

Table 1: Accuracy comparison of static and dynamic feature selection methods for tabular datasets with variable feature costs. Feature costs are randomly generated once per dataset, fixed across train and test splits, shared by all methods, and used with budgets from 5% to 50% of total feature cost in 5% increments. Results show average accuracy across budget levels, with standard deviation indicating variability across budgets. The final rank is the mean of the ranks across the five datasets. Methods marked with * are instances of the proposed model-agnostic framework. Bold marks best method, and underline marks second best. We mark this both per dataset and overall.

		Method	Cirrhosis	Diabetes	Heart	Wine	Yeast	Average (rank)
		Static CMI	**72.62 ± 1.68**	**62.63 ± 0.13**	65.58 ± 9.27	79.17 ± 1.97	31.65 ± 3.52	62.33 (6.7)
		TabNet	42.86 ± 7.01	53.15 ± 4.02	62.30 ± 8.75	61.11 ± 10.67	31.65 ± 4.13	50.21 (12.3)
Dynamic	RL	ReMLP*	69.18 ± 3.28	53.85 ± 1.93	78.33 ± 5.24	93.83 ± 2.70	32.66 ± 4.68	65.57 (5.1)
		MLP*	71.56 ± 3.49	54.78 ± 1.76	70.67 ± 2.77	96.29 ± 1.39	31.54 ± 5.55	64.97 (6.4)
		CWCF	68.65 ± 12.21	54.81 ± 1.07	66.85 ± 4.00	93.52 ± 1.96	34.04 ± 5.27	63.57 (6.8)
		INVASE	69.18 ± 1.83	46.85 ± 1.72	**79.78 ± 2.59**	78.71 ± 5.55	33.00 ± 2.23	61.50 (7.3)
		FTree*	55.95 ± 0.00	53.41 ± 1.66	73.59 ± 3.33	79.63 ± 10.11	28.96 ± 0.00	58.31 (10.0)
		CART*	66.54 ± 2.94	50.85 ± 2.09	67.94 ± 6.91	76.54 ± 7.36	28.36 ± 3.74	58.05 (11.6)
	Greedy	ReMLP*	66.93 ± 3.03	56.48 ± 2.43	76.69 ± 3.91	**96.60 ± 4.11**	**41.19 ± 7.86**	**67.58** (3.4)
		VIP	70.77 ± 1.98	60.89 ± 1.00	75.78 ± 12.01	93.21 ± 2.45	33.14 ± 1.13	66.76 (4.0)
		DIME	63.89 ± 2.73	58.07 ± 1.15	71.95 ± 3.42	89.20 ± 3.53	35.50 ± 3.62	63.72 (6.4)
		CART*	60.85 ± 7.63	59.41 ± 1.27	72.13 ± 6.19	93.52 ± 14.50	31.54 ± 3.48	63.49 (7.0)
		MLP*	65.74 ± 5.61	52.15 ± 1.17	69.76 ± 2.61	83.95 ± 9.29	37.15 ± 3.95	61.75 (8.2)
		FTree*	51.98 ± 6.94	55.15 ± 1.17	75.41 ± 0.00	73.15 ± 17.18	29.97 ± 1.81	57.13 (9.8)

Table 2: Accuracy Comparison of dynamic feature selection methods for image datasets. Results show average accuracy up to 50% of the total patches acquired. Best result in bold.

Method	MNIST	CIFAR-10	Imagenette
DIME	70.10 ± 13.64	51.18 ± 17.60	67.71 ± 13.35
VIP	75.76 ± 19.35	48.76 ± 18.00	**71.41 ± 8.90**
Greedy-CNN	88.61 ± 8.97	33.37 ± 12.29	46.58 ± 21.76
Greedy-ReCNN	**89.25 ± 15.62**	**51.61 ± 12.94**	66.40 ± 17.17

DFS can match or exceed native DFS methods, but only when the predictor and acquisition strategy are well matched. The uniform-cost results in Appendix E reinforce the same idea, with static CMI and VIP remaining particularly strong baselines there.

Statistical comparison. The Friedman test over the five datasets rejects equal performance across the DFS methods ($\chi^2 = 25.9$, $p = 0.011$). However, all-pairs rank-based post-hoc tests (Nemenyi) are severely underpowered at five datasets. We therefore follow the Bayesian comparison protocol of Benavoli et al. (2017), recommended over null-hypothesis testing for small numbers of datasets (Demšar, 2006). Using the per-dataset normalised accuracy-curve AUC with a region of practical equivalence of ±0.01, Table 3 shows that G-ReMLP outperforms DIME, INVASE, and CWCF with posterior probability 0.98, 0.97, and 0.94, respectively, and is statistically indistinguishable from the strongest native baseline VIP ($P(\text{better}) = 0.39$, $P(\text{rope}) = 0.32$, $P(\text{worse}) = 0.29$). Therefore, we can safely report competitiveness rather than dominance: the model-agnostic predictor matches the best native DFS method while additionally supporting pre-trained and interpretable backbones.

Image variable-cost benchmark. Table 2 shows that native DFS methods dominate on Imagenette, where VIP reaches 71.41%. On CIFAR-10 and MNIST, however, the best results are obtained by the model-agnostic CNN variant: G-ReCNN reaches 51.61% on CIFAR-10 (versus DIME's 51.18%) and 89.25% on MNIST, the best result on both datasets, while remaining competitive (66.40%) on Imagenette. Reparametrisation significantly improved the results in CIFAR-10 and Imagenette compared to the non-reparametrised baseline. On MNIST, the easiest of the three datasets, the model-agnostic DFS-policy was good enough

Table 3: Bayesian signed-rank comparison of the proposed G-ReMLP against native DFS baselines, on the per-dataset normalised accuracy-curve AUC (variable costs). P(better) and P(worse) are the posterior probabilities that G-ReMLP out-performs the baseline beyond a region of practical equivalence of ± 0.01 AUC, and P(rope) is the probability of practical equivalence.

Baseline	P(better)	P(rope)	P(worse)
DIME	0.981	0.010	0.009
INVASE	0.965	0.026	0.009
CWCF	0.937	0.051	0.013
VIP	0.391	0.318	0.291

Table 4: Effect of epistemic uncertainty weight λ for G-CART and G-ReMLP backbones. Results are averaged across 5 datasets and 7 budget levels (20%–80%). The best accuracy is underlined for G-CART and bolded for G-ReMLP.

λ	G-CART		G-ReMLP	
	Accuracy	Epistemic Unc.	Accuracy	Epistemic Unc.
-1.0	67.34 ± 15.46	1.803 ± 1.221	68.01 ± 16.07	0.177 ± 0.109
-0.5	67.83 ± 17.65	1.540 ± 1.000	68.81 ± 16.17	0.179 ± 0.111
-0.1	67.00 ± 16.81	1.384 ± 0.807	69.07 ± 16.37	0.193 ± 0.105
0.0	67.56 ± 18.29	1.426 ± 0.851	69.05 ± 17.24	0.188 ± 0.107
0.1	67.58 ± 18.40	1.365 ± 0.788	**70.86 ± 17.68**	0.185 ± 0.096
0.5	<u>68.57 ± 18.92</u>	1.358 ± 0.743	69.31 ± 16.91	0.194 ± 0.095
1.0	67.69 ± 18.96	1.336 ± 0.734	69.33 ± 16.12	0.187 ± 0.088

to obtain very good results, which suggests that most decisions in MNIST require only a small part of the image.

6.2 RQ2: Effect of Epistemic Uncertainty in the DFS Policy

We investigate incorporating different weights of the epistemic uncertainty penalty in feature selection policy in Eq. (11), evaluating λ across the $[-1.0, 1.0]$ range for G-CART and G-ReMLP backbones using uniform costs. Table 4 shows results for tabular data. For G-CART, positive λ values reduce both epistemic uncertainty and improve performance relative to the unpenalised baseline (1.426 at $\lambda = 0$ versus 1.336 at $\lambda = 1.0$), and the best accuracy is obtained at $\lambda = 0.5$ (68.57%). For G-ReMLP, epistemic uncertainty is already much lower and varies little across λ, but a mild positive penalty gives the best accuracy (70.86% at $\lambda = 0.1$ versus 69.05% at $\lambda = 0$). Overall, RQ2 suggests that epistemic uncertainty is most useful as a reliability-control signal, but can also result in accuracy gains when the penalty is tuned to the backbone. Appendix F shows how the epistemic signal correlates with distribution shift, which reinforces the idea that this mechanism helps the predictor stay in-distribution and make reliable predictions. The same qualitative trend holds on image data with smaller magnitude (Appendix F.5).

6.3 RQ3: Evidence for DFS Uncertainty Issues

Prediction bias under partial observation. A central problem in DFS is that a method may make accurate predictions without approximating the intended conditional distribution $p(y \mid \mathbf{x}_S)$. To test this, we generate feature subsets of varying cardinalities, and train a G-MLP exclusively on $\mathbf{x}_S$ as an empirical proxy for the subset-optimal conditional that serves as a local oracle. Then, we compare DFS predictions against both this reference and a full-feature model using KL divergence. Larger divergence from the subset-specific reference, accompanied by stronger agreement with the full-feature model and no corresponding accuracy or calibration gain, suggests that the method may be relying on imputation artefacts or feature correlations rather than learning proper subset conditionals.

Table 5: Average prediction bias metrics across datasets. $\mathrm{KL}_{\mathrm{Ref}}$ captures distance to subset optimal predictions, while $\mathrm{KL}_{\mathrm{Full}}$ captures distance to whole feature set predictions. KL Ratio = $\mathrm{KL}_{\mathrm{Ref}}/\mathrm{KL}_{\mathrm{Full}}$ (values > 1 indicate predictions closer to the full-feature model than to the subset reference); Agreement is the top-1 agreement rate with the full-feature model; $\mathrm{L2}_{\mathrm{Ref}}$ is the mean ℓ_2 distance to the subset reference; and Sig. Ref. counts the datasets (out of five) on which the divergence from the subset reference is statistically significant.

Method	$\mathrm{KL}_{\mathrm{Ref}}$	$\mathrm{KL}_{\mathrm{Full}}$	KL Ratio	Agreement	$\mathrm{L2}_{\mathrm{Ref}}$	Sig. Ref.
G-ReMLP-DFS	1.03 ± 1.48	0.83 ± 0.83	1.23 ± 0.51	$\mathbf{0.81 \pm 0.08}$	0.28 ± 0.14	3/5
RL-MLP-DFS	0.37 ± 0.25	0.38 ± 0.18	0.97 ± 0.39	0.75 ± 0.09	0.27 ± 0.06	3/5
VIP	0.39 ± 0.46	0.37 ± 0.25	1.05 ± 0.50	0.81 ± 0.09	0.24 ± 0.12	3/5
DIME	$\mathbf{0.15 \pm 0.10}$	$\mathbf{0.21 \pm 0.14}$	$\mathbf{0.74 \pm 0.32}$	0.76 ± 0.08	$\mathbf{0.20 \pm 0.06}$	**4/5**
CWCF	0.79 ± 1.03	0.56 ± 0.58	1.43 ± 0.59	0.74 ± 0.09	0.24 ± 0.07	3/5
G-CART	1.71 ± 1.74	1.44 ± 1.10	1.19 ± 0.39	0.68 ± 0.06	0.28 ± 0.11	3/5

Table 6: Correlation between average epistemic uncertainty and the performance gap $\Delta_S = R_S(f_{\theta^*}) - R_S(f_{\theta^*_S})$, per dataset and pooled across all random subsets and budget levels (10–90%). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Dataset	G-CART		G-ReMLP		n
	Pearson r	Spearman ρ	Pearson r	Spearman ρ	
Cirrhosis	0.452^{***}	0.450^{***}	-0.020	-0.037	150
Diabetes	-0.070	-0.124	-0.337^{***}	-0.378^{***}	150
Heart	0.305^{***}	0.296^{***}	0.064	0.154	150
Wine	0.701^{***}	0.722^{***}	0.216^{**}	0.185^{*}	150
Yeast	-0.186^{*}	-0.183^{*}	0.211^{**}	0.189^{*}	150
Avg.	0.664^{***}	0.490^{***}	0.020	0.054	750

Table 5 shows that this problem is method-dependent. DIME has the lowest divergence from the subset reference (0.153) and a KL ratio below one, indicating effective adaptation with limited full-feature leakage, but also significantly lower accuracy compared to the best DFS methods tested. In contrast, G-ReMLP and VIP show larger relative agreement with the full model, suggesting that some high-performing methods partly preserve full-information behaviour under partial observation. CWCF and G-CART also showed ratios significantly larger than 1, which shows that although closeness to the reference oracle does not guarantee good results, large deviations from this reference can be problematic as well.

Subset adaptation and epistemic uncertainty. To measure this, we sample random subsets across budget levels (10–90%) and compute the subset risk gap Δ_S together with the mean epistemic uncertainty. For G-CART, the correlation between both is clearly positive on Wine, Heart, and Cirrhosis ($r = 0.70$, 0.31, 0.45; all $p < 0.001$), near zero on Diabetes, and negative on Yeast, with an overall Pearson $r = 0.66$ and Spearman $\rho = 0.49$ pooled across subsets and budgets (Table 6 and Figure 3). Note that the pooled coefficients exceed most per-dataset values because pooling also reflects between-dataset differences in the scales of Δ_S and epistemic uncertainty, so the per-dataset correlations are the more conservative estimates. For G-ReMLP, the correlation is essentially absent overall ($r = 0.02$, $\rho = 0.05$), which is consistent with its near-zero adaptation gap: once reparametrisation removes most of Δ_S, there is little residual subset-specific error for an uncertainty proxy to track. We also note that this epistemic–Δ_S relationship is backbone and dataset-dependent: it is clear for CART on Wine, Heart, and Cirrhosis but weak or absent on Diabetes and Yeast, and largely vanishes once reparametrisation closes the gap. Epistemic uncertainty should therefore be read as a useful but not universal indicator of subset-adaptation risk.

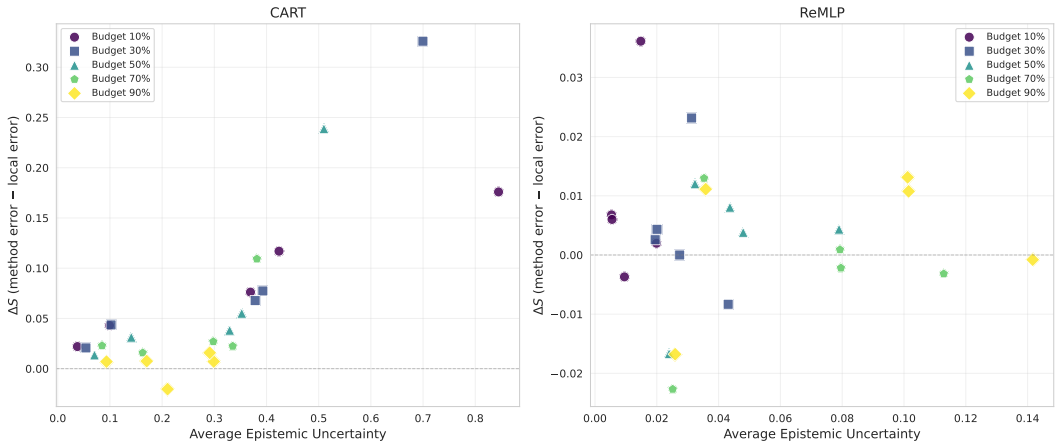


Figure 3: Scatter plot of Δ_S versus epistemic uncertainty. Each point averages a (dataset, budget) cell for visualisation; per-dataset and pooled per-subset correlations are reported in Table 6. Left: CART shows a clear positive association (overall per-subset r=0.66, p<0.001): subsets where the global model incurs a larger gap over the local oracle tend to have higher epistemic uncertainty. Right: G-ReMLP exhibits essentially no association (overall r=0.02), consistent with its near-zero Δ_S; the reparametrised model leaves little adaptation gap for the uncertainty signal to track.

Calibration does not improve with budget. A natural expectation is that acquiring more features should make predictions better calibrated, since the model conditions on more evidence. Figure 4 shows that this does not happen: the Expected Calibration Error averaged across tabular datasets remains within a narrow 0.10–0.18 band and fluctuates non-monotonically as the budget grows, with no method exhibiting a clear downward trend. Several methods instead drift *towards* the full-feature baselines (dashed lines) as more features are acquired, so the additional budget that improves accuracy does not translate into more reliable confidence.

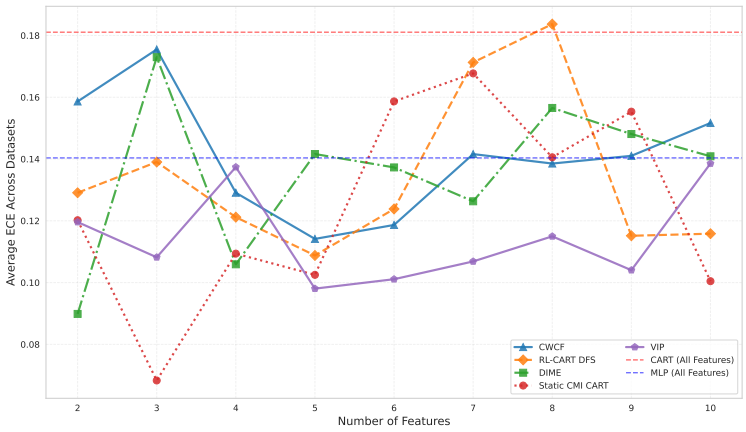


Figure 4: **Expected Calibration Error averaged across the tabular datasets as a function of the acquisition budget.** A selection of methods is shown here. Calibration does not necessarily improve in any of them as more features are acquired: the curves stay within a narrow band and fluctuate non-monotonically. The full-feature baselines (dashed lines) show that the calibration error remains in the same scale up until full observation. Per-dataset curves are reported in Appendix E (Figure 12).

Selective prediction. We further evaluate calibration through accuracy-rejection curves, progressively rejecting the instances assigned the highest uncertainty and measuring the accuracy on the retained set. Well-calibrated uncertainty estimates should produce monotonic improvements as rejection increases; flat

or declining curves indicate that the uncertainty score does not reliably identify difficult partial-observation predictions. Figure 5 shows that total predictive uncertainty and $1 - \max$ confidence scores produce the steepest and most consistent improvements, while isolated aleatoric and epistemic components are flatter, especially for DIME and G-CART.

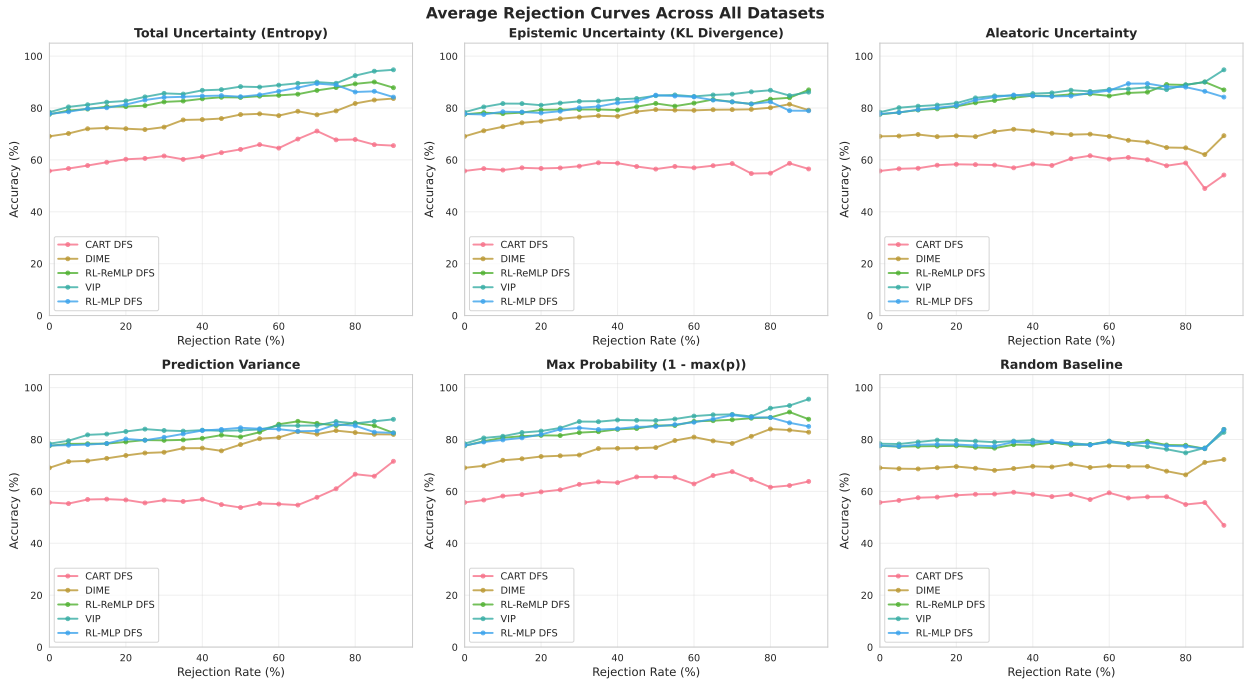


Figure 5: Evolution of accuracy-rejection curves based on different uncertainty measures. Results are shown for 50% of the budget and averaged across all tabular datasets.

Table 7: Selective prediction metrics for tabular data, averaged across budgets. AURC, E-AURC and risk are better when lower; AUROC is better when higher.

	AURC	E-AURC	AUROC	Risk@80%	Risk@90%
G-ReMLP	**21.77 ± 0.46**	**15.65 ± 0.04**	71.77 ± 0.88	27.54 ± 2.93	**29.50 ± 2.05**
VIP	24.19 ± 0.72	17.00 ± 0.95	**77.70 ± 12.11**	**27.34 ± 1.33**	29.56 ± 1.47
G-CART	25.46 ± 1.72	18.92 ± 1.23	63.63 ± 0.28	29.26 ± 1.20	30.14 ± 0.82
RL-MLP	26.18 ± 2.37	17.46 ± 1.89	67.40 ± 2.46	31.92 ± 4.50	33.21 ± 3.31
DIME	26.65 ± 2.39	19.08 ± 2.86	69.53 ± 3.82	32.43 ± 1.20	33.38 ± 0.34
RL-ReMLP	26.86 ± 1.00	18.69 ± 0.98	71.60 ± 0.57	32.91 ± 0.34	33.85 ± 0.81
RL-CART	32.27 ± 1.43	21.58 ± 0.95	68.88 ± 2.68	39.66 ± 2.27	41.75 ± 1.96

Table 7 confirms that calibration quality differs substantially across DFS methods. G-ReMLP obtains the best AURC and E-AURC (21.77/15.65) and the lowest risk at 90% coverage (29.50), while VIP obtains the best AUROC (77.70) and the lowest risk at 80% coverage (27.34), with both methods essentially tied at the highest coverage level. RL-ReMLP improves AUROC over RL-MLP (71.60 vs. 67.40), supporting the role of reparametrisation in preserving reliable uncertainty estimates. The G-CART backbones have the weakest selective-prediction metrics (G-CART AUROC 63.63, and RL-CART the worst AURC at 32.27), suggesting that an interpretable backbone also comes at the cost of uncertainty quality under partial observation.

Taken together, these results answer RQ3 affirmatively: the predicted pathologies appear empirically, but each method is affected in different ways.

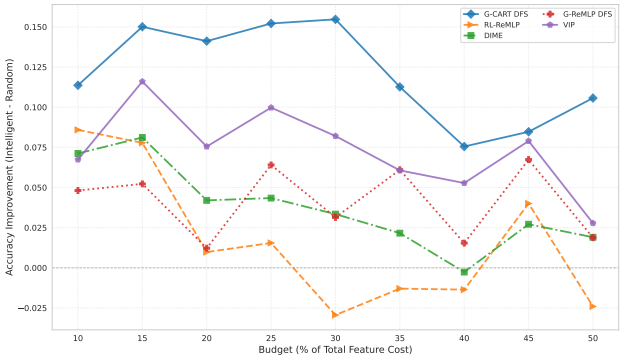


Figure 6: Evolution of average accuracy difference between feature selected and random masks for increasing budgets in the tabular datasets studied.

6.4 RQ4: Separating Predictor and Policy Effects

Entropy and uncertainty decomposition. Predictive entropy generally decreases as more features or patches are acquired, but this trend can occur for both learned and random acquisition policies, so entropy alone does not identify whether the selected features are useful. On image data, methods can have similar entropy levels while exhibiting visibly different calibration errors (Figure 8).

Policy contribution analysis. Figure 6 compares trained feature-selection policies against random masks, using the same predictors, for tabular data. This isolates the acquisition policy from the predictor's ability to operate under partial observation. G-CART benefits most from a learned policy (up to ∼15 percentage points over random acquisition), and VIP and G-ReMLP show consistent positive gains across budgets, while DIME shows more moderate gains. RL-ReMLP is the exception: its advantage over random acquisition is small and occasionally negative at mid budgets, indicating that once reparametrisation makes the predictor robust to partial observation, the additional value of the learned RL policy is limited. The large gains for G-CART show that a learned policy matters most for the weaker backbone, but cannot by itself lift it to the accuracy of stronger predictors. So, even after addressing subset adaptation through reparametrisation, a successful policy remains important.

For image data, we can see in Figure 7 that the learned acquisition policy is most useful for the pre-trained predictor based approaches, while VIP exhibits little gain over random acquisition. This shows that a strong predictor can compensate for a weak patch-selection policy and that indeed the backbone performance is the main factor in the image results reported in Table 2.

The answer to RQ4 depends then on the modality of the data: on tabular data, policy quality and subset reparametrisation surpass native DFS methods; on image data, the predictor dominates, and while the proposed policy significantly overperforms random selection, it cannot close the gap in the most difficult image tasks.

7 Related Work

Dynamic Feature Selection. The dominant view of DFS frames each acquisition as the feature that maximally reduces uncertainty about the label, formalised through the conditional mutual information (CMI) $I(y; x_i \mid \mathbf{x}_S)$ (Melville et al., 2004; Gadgil et al., 2023). Because CMI is intractable for arbitrary subsets, recent work learns to amortise it. Covert et al. (2023) train a predictor network jointly with a policy network under random feature masking, so that a single greedy estimator approximates the CMI-maximising feature at every step; Chattopadhyay et al. (2023) learn query chains that maximise the information gain. More recent work has criticised the use of CMI for DFS (Norcliffe et al., 2025), and has opted for using an ad hoc metric instead. A second family models the unobserved features explicitly. EDDI (Ma et al., 2019) fits a Partial VAE to $p(\mathbf{x}_{\bar{S}} \mid \mathbf{x}_S)$ and acquires the feature with the largest expected information reward,

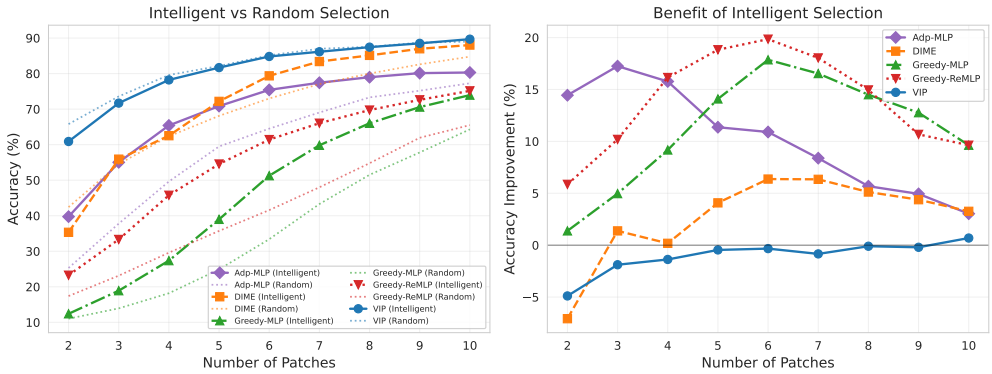


Figure 7: Comparison between intelligent and random feature acquisition on *Imagenette*. Left: Classification accuracy for intelligent feature selection policies (solid lines) and random acquisition baselines (dotted lines). Right: Accuracy improvement of intelligent feature selection over random acquisition.

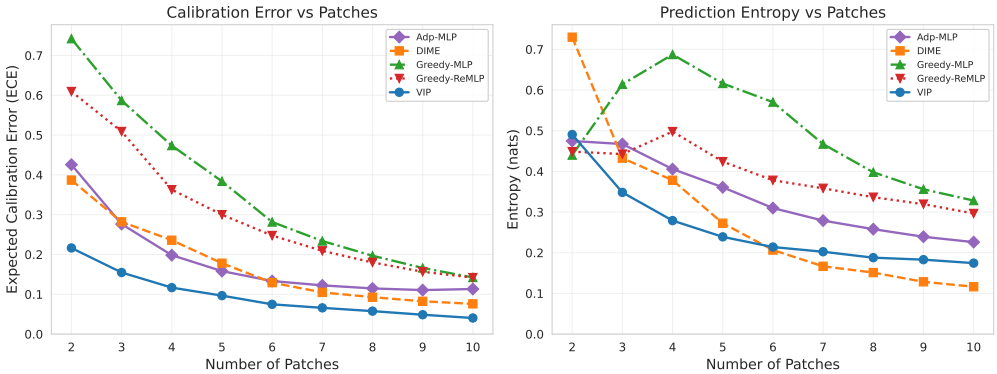


Figure 8: Calibration behaviour on *Imagenette* under dynamic feature acquisition. Left: ECE as a function of the number of acquired image patches. Right: predictive entropy evolution across acquisition steps.

while Li & Oliva (2021) pair a generative surrogate with model-based RL to amortise the same lookahead. Model-free RL formulations such as Janisch et al. (2019) and Yin et al. (2020) instead learn an acquisition-and-stop policy end to end by treating cost as a negative reward. Earlier cost-sensitive formulations framed acquisition as budgeted classification (Karayev et al., 2013; Kachuee et al., 2018) or as maximising the value of information at each step (Chen et al., 2015a), and the idea has been extended to images through probabilistic hard-attention policies that acquire informative glimpses (Rangrej & Clark, 2021). Across all these families the predictor is necessarily trained from scratch for the feature acquisition setting, in our work we instead decouple the predictor from the policy so that an already-trained model can be reused.

Static feature selection and interpretable predictors. Classical feature selection chooses a single global subset for all instances through filter, wrapper, or embedded criteria (Guyon et al., 2008), including differentiable relaxations such as concrete autoencoders (Balın et al., 2019). Such methods do not adapt the subset per instance, but provide strong, low-cost baselines that DFS must beat to justify its overhead, as our experiments confirm. A separate motivation for keeping a single fixed predictor is interpretability: rule- and tree-based models are valued because their decision logic is inspectable (Rudin, 2019; Molnar, 2020), and pooling them into ensembles for uncertainty estimation would obscure that logic. Our framework therefore estimates epistemic uncertainty with auxiliary models while leaving one interpretable predictor in charge of the decision, complementing concept-based routes to interpretable acquisition (Koh et al., 2020).

Uncertainty quantification. The aleatoric/epistemic decomposition (Hüllermeier & Waegeman, 2021; Valdenegro-Toro & Mori, 2022) splits the predictive uncertainty of a model into an irreducible part, originating from noise and class overlap in the data, and a reducible part, originating from the model's limited knowledge of the parameters or the input region being evaluated. In practice, total uncertainty is computed as the entropy of the prediction, while the epistemic component is estimated from the disagreement among several predictors obtained through MC-Dropout or deep ensembles. Other methods for uncertainty quantification include post-hoc calibration, such as temperature scaling, and deterministic single-pass estimators (Gal & Ghahramani, 2016; Rahaman et al., 2021; Guo et al., 2017; Naeini et al., 2015; Liu et al., 2020; Mukhoti et al., 2023). These estimators are almost always studied for a fixed input, where the aleatoric uncertainty is constant and only the epistemic term is expected to reduce with more data (Hoarau et al., 2026). DFS violates that premise: the conditioning set changes at every acquisition step, so both components become subset-dependent, and a predictor that is well calibrated on the full feature set need not remain calibrated on partial subsets. We make this dependence explicit by quantifying the gap between a globally trained predictor and its subset-specific counterpart (the risk gap Δ_S, Eq. 4) and by tracking how these estimators behave as features are sequentially revealed, a characterization that to our knowledge is absent from the uncertainty and DFS literature.

8 Limitations

Our results should be interpreted with several limitations in mind. First, the proposed epistemic uncertainty estimates depend on auxiliary models and therefore inherit the approximation quality and computational cost of the chosen ensemble or perturbation scheme: with K auxiliary models, uncertainty estimation multiplies inference cost by roughly K (or by the number of MC-dropout samples), although these evaluations are independent and trivially parallelisable. Second, although the framework is model-agnostic, the DFS performance still depends on the performance of the backbone method. Strong reparametrisations are also needed to match native DFS methods for very high-dimensional or strongly structured domains. In particular, reparametrisation of a fixed convolutional backbone was not enough to match the performance of native DFS methods, even though it was able to obtain competitive results. Finally, the epistemic penalty has a modest effect on average accuracy in our experiments, so its main role should be understood as reliability control rather than as a significant accuracy-improving term.

9 Conclusions

In this work, we formalised how DFS introduces distinct sources of uncertainty beyond the static setting: model adaptation to different feature subsets creates non-uniform epistemic uncertainty across the exponential feature space, standard imputation techniques bias intermediate estimates, and confidence-based stopping criteria cannot reliably discriminate between good and bad selection policies. These findings have direct implications for deployment in high-stakes domains, where systems must reliably quantify their uncertainty. Beyond this analysis, we introduced a model-agnostic DFS framework that adapts pre-trained classifiers, including interpretable-by-design models, to the sequential acquisition setting through subset reparametrisation, guiding the acquisition with an epistemic penalty and without retraining the original predictor. Results showed that reparametrisation was able to obtain better results than native DFS methods in tabular data and obtain competitive results in image as well, while additionally supporting deployed and interpretable backbones that native methods cannot reuse.

Future work should explore tailored adaptation strategies for interpretable classifiers, including extensions to non-tabular domains through concept-based architectures (Koh et al., 2020), as well as more expressive reparametrisation techniques to further close the subset adaptation gap.

References

Stefan Aeberhard and M. Forina. Wine. UCI Machine Learning Repository, 1992. DOI: https://doi.org/10.24432/C5PC7J.

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

Muhammed Fatih Balın, Abubakar Abid, and James Zou. Concrete autoencoders: Differentiable feature selection and reconstruction. In *International conference on machine learning*, pp. 444–453. PMLR, 2019.

Shahaf Bassan, Guy Amir, Meirav Zehavi, and Guy Katz. What makes an ensemble (un) interpretable? *arXiv preprint arXiv:2506.08216*, 2025.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18(77): 1–36, 2017.

Leo Breiman, Jerome Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Chapman and Hall/CRC, 2017.

Aditya Chattopadhyay, Kwan Ho Ryan Chan, Benjamin David Haeffele, Donald Geman, and Rene Vidal. Variational information pursuit for interpretable predictions. In *The Eleventh International Conference on Learning Representations*, 2023.

Suming Chen, Arthur Choi, and Adnan Darwiche. Value of information based on decision robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015a.

Yuxin Chen, S Hamed Hassani, Amin Karbasi, and Andreas Krause. Sequential information maximization: When is greedy near-optimal? In *Conference on Learning Theory*, pp. 338–363. PMLR, 2015b.

Jesse Clifton and Eric Laber. Q-learning: Theory and applications. *Annual Review of Statistics and Its Application*, 7(1):279–301, 2020.

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J White, and Su-In Lee. Learning to maximize mutual information for dynamic feature selection. In *International Conference on Machine Learning*, pp. 6424–6447. PMLR, 2023.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Robert C. Detrano, András Jánosi, Walter Steinbrunn, Matthias Emil Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64 5:304–10, 1989. URL https://api.semanticscholar.org/CorpusID:23545303.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Gabriel Erion, Joseph D Janizek, Carly Hudelson, Richard B Utarnachitt, Andrew M McCoy, Michael R Sayre, Nathan J White, and Su-In Lee. A cost-aware framework for the development of ai models for healthcare applications. *Nature Biomedical Engineering*, 6(12):1384–1398, 2022.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 2017.

Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*. John Wiley & Sons, 2013.

Javier Fumanal-Idocin, Raquel Fernandez-Peralta, and Javier Andreu-Perez. A fast interpretable fuzzy tree learner. *arXiv preprint arXiv:2512.11616*, 2025.

Soham Gadgil, Ian Connick Covert, and Su-In Lee. Estimating conditional mutual information for dynamic feature selection. In *The Twelfth International Conference on Learning Representations*, 2023.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

Arthur Hoarau, Benjamin Quost, Sébastien Destercke, and Willem Waegeman. Reducing aleatoric and epistemic uncertainty through multi-modal data acquisition. *Machine Learning*, 115(5):121, 2026.

Jeremy Howard. Imagenette. https://github.com/fastai/imagenette, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. Classification with costly features using deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3959–3966, 2019.

Mohammad Kachuee, Sajad Darabi, Babak Moatamed, and Majid Sarrafzadeh. Dynamic feature acquisition using denoising autoencoders. *IEEE transactions on neural networks and learning systems*, 30(8):2252–2262, 2018.

Sergey Karayev, Mario J Fritz, and Trevor Darrell. Dynamic feature selection for classification on a budget. In *International conference on machine learning (ICML): Workshop on prediction with sequential models*, volume 95, 2013.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5815–5826. PMLR, 2021.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *International conference on machine learning*, pp. 6450–6459. PMLR, 2021.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems*, 33:7498–7512, 2020.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *International Conference on Machine Learning*, pp. 4234–4243, 2019.

Prem Melville, Maytal Saar-Tsechansky, Foster Provost, and Raymond Mooney. Active feature-value acquisition for classifier induction. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pp. 483–486. IEEE, 2004.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Jishnu Mukhoti, Andreas Kirsch, Joost Van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24384–24394, 2023.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Kenta Nakai. Yeast. UCI Machine Learning Repository, 1991.

Alexander Luke Ian Norcliffe, Changhee Lee, Fergus Imrie, Mihaela Van Der Schaar, and Pietro Lio. Stochastic Encodings for Active Feature Acquisition. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 46784–46814. PMLR, 2025.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Rahul Rahaman et al. Uncertainty quantification and deep ensembles. *Advances in neural information processing systems*, 34:20063–20075, 2021.

Samrudhdhi B Rangrej and James J Clark. A probabilistic hard attention model for sequentially observed scenes. *arXiv preprint arXiv:2111.07534*, 2021.

Sylvestre-Alvise Rebuffi, Andrea Vedaldi, and Hakan Bilen. Efficient parametrization of multi-domain deep neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8119–8127, 2018.

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

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014(1):781670, 2014.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

Katsumi Takahashi, Koh Takeuchi, and Hisashi Kashima. Dynamic feature selection from variable feature sets using features of features. *arXiv preprint arXiv:2503.09181*, 2025.

Matias Valdenegro-Toro and Daniel Saromo Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1508–1516. IEEE, 2022.

Mike Van Ness, Tomas M. Bosschieter, Roberto Halpin-Gregorio, and Madeleine Udell. The missing indicator method: From low to high dimensions. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5004–5015. Association for Computing Machinery, 2023.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In Robin J. Evans and Ilya Shpitser (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2282–2292. PMLR, 2023.

Haiyan Yin, Yingzhen Li, Sinno Jialin Pan, Cheng Zhang, and Sebastian Tschiatschek. Reinforcement learning with efficient active feature acquisition. *arXiv preprint arXiv:2011.00825*, 2020.

Jinsung Yoon, James Jordon, and Mihaela Van der Schaar. Invase: Instance-wise variable selection using neural networks. In *International conference on learning representations*, 2018.

Appendix Overview

The appendix is organised into three parts.

Theoretical Analysis. Appendices A–D provide additional theoretical results and methodological details:

- Appendix A: Properties of the proposed epistemic uncertainty metric.
- Appendix B: Formal relation between prediction similarity minimisation and conditional mutual information.
- Appendix C: Efficient adaptation of rule-based classifiers to arbitrary feature subsets.
- Appendix D: Adaptation of classical rule-learning heuristics to the DFS setting.

Extended Empirical Results. Appendix E reports additional experimental analyses:

- Performance under uniform feature costs.
- Per-dataset accuracy and calibration evolution.
- Selective prediction and accuracy–rejection curves.
- Bias decomposition of sequential predictions.

Targeted Validation Experiments. Appendix F reports additional experiments designed to empirically validate additional phenomena of interest:

- Robustness under distribution shift.
- Correlation between epistemic uncertainty and subset risk gap Δ_S.
- Effect of reparametrisation on reducing subset risk.
- Correlation between aleatoric and epistemic uncertainties.
- Effect of epistemic penalty λ on image datasets.

Implementation Details. Appendix G provides reproducibility information, including dataset statistics, model architectures, hyperparameters, and training configurations.

A Properties of the Proposed Epistemic Uncertainty Metric

For uncertainty decomposition, a standard information-theoretic identity is

$$H(Y \mid \mathbf{x}_S) = H(Y \mid \mathbf{x}_S, \Theta) + I(Y; \Theta \mid \mathbf{x}_S), \tag{17}$$

where Θ is a random variable of model's parameters $\theta \sim Q$ (Hüllermeier & Waegeman, 2021). The first term is commonly interpreted as aleatoric uncertainty (AU), since it captures the irreducible uncertainty in Y that remains even if the model parameters were known. The second term, $I(Y; \Theta \mid \mathbf{x}_S)$, is interpreted as epistemic uncertainty (EU), since it quantifies the expected reduction in uncertainty about Y that would be obtained by observing Θ. In practice, these quantities are approximated via an ensemble of $K \in \mathbb{N}$ predictors $\{\theta^{(i)}\}_{i=1}^{K}$ sampled from Q. Denoting

$$p_i(\hat{y}) \equiv p(\hat{y} \mid \mathbf{x}_S, \theta^{(i)}), \qquad \bar{p}(\hat{y}) \equiv \frac{1}{K} \sum_{i=1}^{K} p_i(\hat{y}), \tag{18}$$

we obtain the Monte Carlo estimators

$$\mathrm{AU}(\mathbf{x}_S) \;\approx\; \frac{1}{K}\sum_{i=1}^{K} H(p_i), \tag{19}$$

$$\mathrm{EU}(\mathbf{x}_S) \;\approx\; H(\bar{p}) \;-\; \frac{1}{K}\sum_{i=1}^{K} H(p_i) \;=\; \frac{1}{K}\sum_{i=1}^{K} D_{\mathrm{KL}}(p_i \| \bar{p}). \tag{20}$$

However, our setting differs from standard ensemble inference: predictions and explanations are produced by a single fixed model, denoted by

$$q(\hat{y}) \equiv p(\hat{y} \mid \mathbf{x}_S, \theta), \tag{21}$$

while auxiliary ensemble members are used only for uncertainty quantification. To measure epistemic uncertainty around the selected model, we use the quantity Eq. (15), which measures how much other reasonable models would disagree with the predictor we actually use at $\mathbf{x}_S$ and thus reflects how stable that decision is under small changes in training. It is straightforward to prove that this formula has the following relationship with standard ensemble epistemic uncertainty

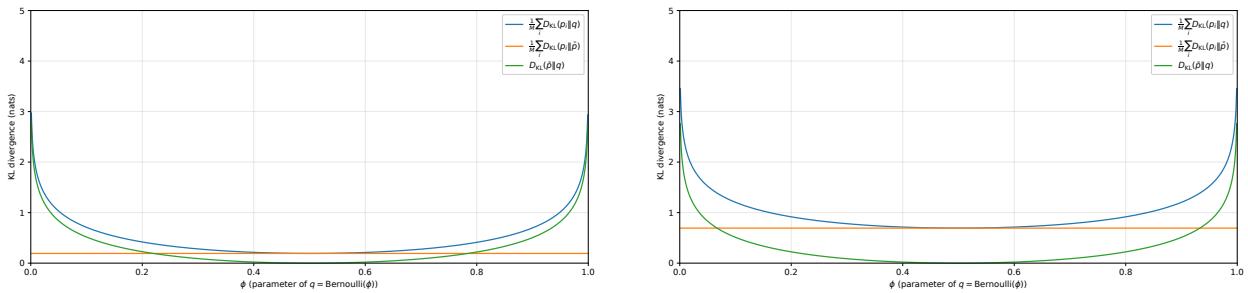
$$e(\mathbf{x}_S) = \frac{1}{K}\sum_{i=1}^{K} D_{\mathrm{KL}}(p_i \| q) = \frac{1}{K}\sum_{i=1}^{K} D_{\mathrm{KL}}(p_i \| \bar{p}) + D_{\mathrm{KL}}(\bar{p} \| q) = \mathrm{EU}(\mathbf{x}_S) + D_{\mathrm{KL}}(\bar{p} \| q). \tag{22}$$

By using the definition in Eq. (15) for epistemic uncertainty, we capture both the usual disagreement among ensemble members around their average prediction (left term) and the additional discrepancy introduced by relying on the single fixed model q at inference time, rather than the ensemble mean $\bar{p}$ (right term).

Importantly, while the left term is bounded by $\ln C$ (where C is the number of classes), the right term $D_{\mathrm{KL}}(\bar{p} \| q)$ is unbounded and can grow arbitrarily large when q assigns near-zero probability to regions where $\bar{p}$ has significant mass. This asymmetry is particularly beneficial for uncertainty quantification: when the selected model q significantly deviates from the ensemble consensus $\bar{p}$, the unbounded right term dominates the decomposition, effectively raising a strong alarm about potential model misspecification. In contrast, the standard epistemic uncertainty (left term alone) could remain relatively small even when q is a poor representative of the ensemble, failing to capture this critical aspect on explainable models when ensemble inference is not adequate.

Now, we discuss the properties of $e(\mathbf{x}_S)$ as a measure of epistemic uncertainty. Since it is defined via the Kullback–Leibler divergence, $e(\mathbf{x}_S)$ is non-negative and vanishes if and only if all auxiliary predictive distributions coincide with the primary one at $\mathbf{x}_S$, i.e. $p_i = \bar{p} = q$; this holds in particular when Q reduces to a Dirac measure $Q = \delta_\theta$, though predictive agreement does not require parameter identity. Thus, zero epistemic uncertainty characterises the local stability of the selected model's prediction. Nonetheless, like the mutual information approach (Wimmer et al., 2023), our measure $e(\mathbf{x}_S)$ also exhibits some non-standard behaviours: mean-preserving spreads may not increase $e(\mathbf{x}_S)$, location shifts may not preserve it, and it is not maximal when the second-order distribution over predictions is uniform. However, these properties are less relevant in our framework, as we aim to capture epistemic uncertainty from a different perspective. Indeed, in our case, we are not interested in minimising $\mathrm{EU}(\mathbf{x}_S)$ alone, but rather in balancing this term with $D_{\mathrm{KL}}(\bar{p} \| q)$ to assess whether the selected model is coherent with the ensemble.

To illustrate the behaviour of our measure, we consider the following experiment. We study an ensemble of Bernoulli models $p_i = \mathrm{Bernoulli}(\theta_i)$ and compute $e(\mathbf{x}_S)$ for different reference models $q = \mathrm{Bernoulli}(\phi)$. We compare two choices for the distribution Q of the parameters θ_i: (i) a uniform distribution on $[0, 1]$, and (ii) an atomic distribution $Q = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. Fig. 9 compares the behaviour of the $e(\mathbf{x}_S)$ decomposition for these two choices of the second-order distribution Q. As shown in Wimmer et al. (2023), the term $\mathrm{EU}(\mathbf{x}_S)$ is close to its theoretical maximum $\ln(2) \approx 0.69$ for the mixture of Dirac distributions, while for the uniform distribution this term is much smaller, around 0.19. For the second term, $D_{\mathrm{KL}}(\bar{p} \| q)$, we observe that it becomes large at the extremes, where the selected model is very different from the ensemble average. When the selected model is more consistent with the ensemble, this term is close to zero, and only the ensemble epistemic uncertainty remains. This behaviour is aligned with our objective, as the measure naturally penalises incoherent model choices while remaining stable when the selected model agrees with the ensemble.



(a) $\theta_i \sim \text{Uniform}(0, 1)$

(b) $\theta_i \sim \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$

Figure 9: Epistemic uncertainty decomposition for Bernoulli ensembles.

B Relation between Prediction Similarity minimization and CMI

In this section, we study the relationship between the conditional mutual information maximisation and the prediction difference minimisation as DFS policies.

First, we rewrite the decrease in prediction discrepancy obtained by adding a candidate feature x_i to the currently observed set S.

Lemma 1. *Let* $u(\mathbf{x}_S) = D_{\mathrm{KL}}(p(\hat{y} \mid \mathbf{x}) \,\|\, p(\hat{y} \mid \mathbf{x}_S))$ *and define*

$$\Delta u_i = u(\mathbf{x}_S) - u(\mathbf{x}_{S\cup\{i\}}).$$

Then

$$\Delta u_i = \sum_{\hat{y}} p(\hat{y} \mid \mathbf{x}) \log \left(\frac{p(\hat{y} \mid \mathbf{x}_S, x_i)}{p(\hat{y} \mid \mathbf{x}_S)} \right).$$

Proof. By definition,

$$u(\mathbf{x}_S) = \sum_{\hat{y}} p(\hat{y} \mid \mathbf{x}) \log \frac{p(\hat{y} \mid \mathbf{x})}{p(\hat{y} \mid \mathbf{x}_S)},$$

$$u(\mathbf{x}_{S\cup\{i\}}) = \sum_{\hat{y}} p(\hat{y} \mid \mathbf{x}) \log \frac{p(\hat{y} \mid \mathbf{x})}{p(\hat{y} \mid \mathbf{x}_S, x_i)}.$$

Subtracting yields

$$\Delta u_i = u(\mathbf{x}_S) - u(\mathbf{x}_{S\cup\{i\}}) = \sum_{\hat{y}} p(\hat{y} \mid \mathbf{x}) \log \frac{p(\hat{y} \mid \mathbf{x}_S, x_i)}{p(\hat{y} \mid \mathbf{x}_S)}.$$

□

Next, we compare the conditional mutual information score $I(x_i; \hat{y}|\mathbf{x}_S)$ with the expected decrease in discrepancy $\mathbb{E}_{p(x_i|\mathbf{x}_S)}[\Delta u_i]$. The lemma below gives an exact identity, disclosing that their differences depend on how close the partial predictor $p(\hat{y} \mid \mathbf{x}_S, x_i)$ is to the global model $p(\hat{y} \mid \mathbf{x})$.

Lemma 2.

$$I(x_i; \hat{y} \mid \mathbf{x}_S) - \mathbb{E}_{p(x_i|\mathbf{x}_S)}[\Delta u_i] = \mathbb{E}_{p(x_i|\mathbf{x}_S)} \left[\sum_{\hat{y}} \big(p(\hat{y} \mid \mathbf{x}_S, x_i) - p(\hat{y} \mid \mathbf{x}) \big) \log \frac{p(\hat{y} \mid \mathbf{x}_S, x_i)}{p(\hat{y} \mid \mathbf{x}_S)} \right].$$

Proof. From Lemma 1,

$$\Delta u_i = \sum_{\hat{y}} p(\hat{y} \mid \mathbf{x}) \log \frac{p(\hat{y} \mid \mathbf{x}_S, x_i)}{p(\hat{y} \mid \mathbf{x}_S)}.$$

The conditional mutual information can be written as

$$I(x_i; \hat{y} \mid \mathbf{x}_S) = \sum_{x_i, \hat{y}} p(x_i, \hat{y} \mid \mathbf{x}_S) \log \frac{p(\hat{y} \mid \mathbf{x}_S, x_i)}{p(\hat{y} \mid \mathbf{x}_S)} = \mathbb{E}_{p(x_i \mid \mathbf{x}_S)} \left[\sum_{\hat{y}} p(\hat{y} \mid \mathbf{x}_S, x_i) \log \frac{p(\hat{y} \mid \mathbf{x}_S, x_i)}{p(\hat{y} \mid \mathbf{x}_S)} \right].$$

Therefore,

$$\mathbb{E}_{p(x_i \mid \mathbf{x}_S)}[\Delta u_i] = \mathbb{E}_{p(x_i \mid \mathbf{x}_S)} \left[\sum_{\hat{y}} p(\hat{y} \mid \mathbf{x}) \log \frac{p(\hat{y} \mid \mathbf{x}_S, x_i)}{p(\hat{y} \mid \mathbf{x}_S)} \right],$$

and taking the difference yields the result. □

Notice that both criteria involve the same log-ratio term $\log \frac{p(\hat{y} \mid \mathbf{x}_S, x_i)}{p(\hat{y} \mid \mathbf{x}_S)}$, but they differ in the weighting distribution over $\hat{y}$: mutual information weights by $p(\hat{y} \mid \mathbf{x}_S, x_i)$, whereas the prediction-difference gain weights by the global predictor $p(\hat{y} \mid \mathbf{x})$. Therefore, prediction-difference minimisation can be interpreted as a targeted version of conditional mutual information, where feature selection is guided by alignment with a reference (global) predictive model.

The two DFS policies coincide whenever $p(\hat{y} \mid \mathbf{x}) \approx p(\hat{y} \mid \mathbf{x}_S, x_i)$, in which case the term in Lemma 2 vanishes. This situation arises, for instance, when $S \cup \{i\}$ captures nearly all predictive information in $\mathbf{x}$, so that the partial and global predictors agree. More generally, the two criteria need not coincide; however, when observing x_i given $\mathbf{x}_S$ yields a low-entropy (nearly deterministic) posterior over $\hat{y}$, both scores typically rank such a feature highly, since the log-ratio term becomes large in magnitude and is dominated by a single outcome.

C Rule-Based Model Adaptation for Arbitrary Feature Subsets

Rule-based classifiers offer a uniquely efficient mechanism for adaptation to arbitrary feature subsets. Unlike neural networks that require retraining or forward passes through modified architectures, rule-based models can be reparametrised through structural manipulation and confidence re-estimation. Moreover, the explicit feature dependencies encoded in rule antecedents allow us to constrain the feature search space to only those features that can influence the final prediction.

C.1 Rule-Based Classifier Structure

A rule-based classifier consists of a set of rules $\mathcal{R} = \{r_1, \ldots, r_L\}$, where each rule r has the form:

$$r : r_A \Rightarrow c, \tag{23}$$

with r_A being a conjunction of conditions on features (the antecedent) and $c \in \{1, \ldots, C\}$ being the predicted class (the consequent). For a sample $\mathbf{x}$, a rule fires if all conditions in r_A are satisfied. The classifier's prediction is typically determined by the rule with the highest confidence or firing strength among all activated rules.

C.2 Adaptation Procedure for Subset S

Given a global rule-based model trained on the complete feature set and a subset S of observed features, we adapt the model through the following steps:

1. Condition Removal: for each rule $r \in \mathcal{R}$, we remove all conditions in r_A that involve features in $\bar{S}$ (the unobserved features). This yields a reduced antecedent r_A^S containing only conditions on observed features.

2. Rule Pruning: Remove any rules where $r_A^S = \emptyset$ (i.e., no conditions remain after feature removal).

3. Confidence Re-estimation: For each remaining rule r with reduced antecedent r_A^S, re-estimate its class prediction confidence. For crisp rules, the confidence is:

$$\text{Confidence}(r^S \Rightarrow c) = \frac{\sum_{i:\, \mathbf{x}^{(i)} \in X_{r_A^S}} \mathbb{I}\big(y^{(i)} = c\big)}{\left|X_{r_A^S}\right|}, \tag{24}$$

where $X_{r_A^S}$ is the set of training samples satisfying the reduced antecedent r_A^S. For fuzzy rules with membership function μ_r, the confidence becomes:

$$\text{Confidence}(r^S \Rightarrow c) = \frac{\sum_{i:\, \mathbf{x}^{(i)} \in X} \mu_{r^S}(\mathbf{x}_S^{(i)}) \cdot \mathbb{I}(y^{(i)} = c)}{\sum_{i:\, \mathbf{x}^{(i)} \in X} \mu_{r^S}(\mathbf{x}_S^{(i)})}. \tag{25}$$

This re-estimation ensures predictions are unbiased estimates of $p(y|\mathbf{x}_S)$ rather than incorrectly treating missing conditions as satisfied.

Computational Efficiency Through Caching. The key advantage of this approach is that all confidence values can be precomputed during training. For a rule with k conditions, there are 2^k possible sub-rules, corresponding to all subsets of its antecedent conditions. If we restrict rules to contain at most $k_{\max}$ conditions, the number of sub-rules per rule is therefore bounded by $2^{k_{\max}}$.

For $k_{\max} = 4$ (a reasonable constraint for interpretability), each rule generates at most 16 sub-rules. Even for a rule base with 100 rules, this yields only 1,600 confidence values to cache.

At inference time, adapting to any subset S requires only:

1. Identifying which conditions remain in each rule (constant time with proper indexing).
2. Looking up the precomputed confidence for each reduced rule (constant time hash table lookup).

C.3 Discarding Features in the Search Space

The output of a rule-based classifier is determined by

$$\hat{y} = c, \quad \text{where } r^* = \arg\max_{r \in \mathcal{R}} r(\mathbf{x}),$$

where $r(\mathbf{x})$ denotes the evaluation of rule r, and $r_A(\mathbf{x})$ denotes the evaluation of its antecedent. For crisp rules, $r_A(\mathbf{x}) \in \{0, 1\}$ indicates whether all conditions are satisfied, while for fuzzy rules $r_A(\mathbf{x}) \in [0, 1]$ is obtained by aggregating condition memberships using a t-norm.

Due to the conjunctive structure of rule antecedents, if any condition evaluates to zero, the antecedent evaluation becomes zero, i.e.,

$$r_A(\mathbf{x}) = 0 \quad \text{if } \exists\, t \in r_A \text{ such that } t(\mathbf{x}) = 0.$$

Under partial feature observation, after removing conditions involving unobserved features, the same property applies to the reduced antecedent r_A^S and its evaluation $r_A^S(\mathbf{x}_S)$.

This property enables efficient search space reduction through two mechanisms:

1. Early termination: if any evaluated condition returns 0, the antecedent evaluation becomes 0, allowing us to discard rule r without evaluating the remaining conditions. More generally, a pruning threshold $\tau > 0$ can be used, which is particularly useful in fuzzy rule systems.
2. Feature selection: the active feature space can be reduced to

$$S_{\text{active}} = \bigcup_{\substack{r \in \mathcal{R} \\ r_A^S(\mathbf{x}_S) > \tau}} \bigcup_{t \in r_A^S} \text{feature}(t),$$

where feature(t) denotes the index of the feature involved in condition t.

This optimisation reduces computational complexity while preserving exact rule evaluation when $\tau = 0$, and introduces a controllable approximation otherwise.

C.4 Example: Decision Tree Adaptation

Consider a decision tree (CART) as a special case of a rule-based classifier, where each leaf node corresponds to a rule. A path from root to leaf defines the rule antecedent as a conjunction of all split conditions along that path.

For a tree with maximum depth d, each leaf has at most d conditions. The number of possible reduced rules per leaf is therefore at most 2^d. For $d = 5$, this is 32 sub-rules per leaf. A tree with 50 leaves thus requires caching at most 1,600 confidence values. Additionally, sub-rules can be shared across multiple original rules, allowing for additional space optimisation.

D Adapting classical heuristics in rule-learning for DFS

Training rule-based classifiers in the DFS setting requires modifying standard split selection criteria to account for feature availability uncertainty. We describe how to adapt the greedy rule-learning algorithms used in this paper.

The aim in CART for classification is to find the split ζ (on feature j and threshold t) that maximises the purity gain, typically measured by the decrease in Gini Impurity. At a node, the Gini Impurity is defined as:

$$I_G(X) = 1 - \sum_{k=1}^{C} p_k^2 \tag{26}$$

where p_k is the proportion of class k in X. The quality of a split ζ that partitions X into X_{left} and X_{right} is measured by the weighted impurity reduction:

$$\Delta I_G(X, \zeta) = I_G(X) - \left(\frac{|X_{\text{left}}|}{|X|} I_G(X_{\text{left}}) + \frac{|X_{\text{right}}|}{|X|} I_G(X_{\text{right}}) \right). \tag{27}$$

The optimal split is then selected as:

$$\zeta^* = \arg\max_{\zeta} \Delta I_G(X, \zeta). \tag{28}$$

In the DFS setting, we must account for the fact that at inference time, only a subset S of features may be observed. To train a tree that performs well across arbitrary feature subsets, we modify the split criterion to maximise the expected impurity reduction over the distribution of feature subsets that will include the splitting feature. Specifically, when considering a split on feature j, we compute:

$$\zeta^*_{\text{DFS}} = \underset{\zeta:\, \text{feature}(\zeta)=j}{\arg\max} \; \mathbb{E}_{S \sim p(S \mid j \in S)} \left[\Delta I_G(X_S, \zeta) \right], \tag{29}$$

where $p(S \mid j \in S)$ is the distribution over feature subsets conditioned on j being observed, and $X_S = \{(\mathbf{x}_S^{(i)}, y^{(i)})\}_{i=1}^N$ denotes the dataset where only features in S are visible (other features imputed or masked). In practice, this expectation is approximated by sampling T feature subsets $\{S_1, \ldots, S_T\}$ such that $j \in S_i$ for all i, computing $\Delta I_G(X_{S_i}, \zeta)$ for each subset, and taking the empirical average $\frac{1}{T} \sum_{i=1}^{T} \Delta I_G(X_{S_i}, \zeta)$. This approach ensures that the selected splits remain effective even when only partial information is available at test time.

For Fuzzy Tree, the adaptation is analogous.

E Extended Empirical Results

In this section, we display additional results regarding tabular data, disaggregated by individual datasets for:

- Uniform feature costs.
- Accuracy and calibration error.
- Uncertainty evolution with increasing budgets.
- Prediction bias during feature acquisition steps.

To measure calibration, we analyse the Expected Calibration Error (ECE) for different budgets. ECE quantifies the alignment between predicted confidence and actual accuracy by partitioning predictions into equally-spaced confidence bins and measuring discrepancies:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|, \tag{30}$$

where B_m represents the m-th confidence bin, n is the total number of samples, $\text{acc}(B_m)$ is the empirical accuracy within the bin, and $\text{conf}(B_m)$ is the average predicted confidence (Naeini et al., 2015).

E.1 Results for uniform feature costs

Table 8: Accuracy comparison of Static and Dynamic Feature Selection methods for tabular datasets. Results show the average accuracy across all budget levels (1-10 features), with standard deviation indicating variability across budget levels. The final rank is computed as the mean of the ranks obtained across the five individual datasets. Methods marked with * are our proposed approaches.

		Method	Cirrhosis	Diabetes	Heart	Wine	Yeast	Average (rank)
		Static CMI	72.22 ± 1.03	57.81 ± 5.48	83.97 ± 2.69	93.82 ± 3.04	54.47 ± 7.94	72.46 (3.6)
		TabNet	68.78 ± 4.28	55.96 ± 3.89	79.05 ± 2.43	93.21 ± 5.22	56.23 ± 4.64	70.65 (5.2)
Dynamic	RL	MLP*	72.09 ± 3.26	52.04 ± 1.71	81.24 ± 2.03	96.91 ± 2.93	54.55 ± 4.22	71.37 (4.8)
		ReMLP*	68.65 ± 3.31	53.18 ± 1.53	79.78 ± 2.46	$\mathbf{97.84 \pm 3.04}$	54.77 ± 1.91	70.85 (5.1)
		CART*	69.71 ± 1.69	51.15 ± 2.21	77.41 ± 2.56	92.90 ± 3.95	$\mathbf{57.09 \pm 2.18}$	69.65 (7.2)
		CWCF	$\mathbf{74.60 \pm 5.19}$	54.33 ± 0.47	76.14 ± 13.86	83.33 ± 13.82	51.37 ± 11.05	67.95 (6.9)
		INVASE	70.11 ± 2.62	47.74 ± 2.07	79.78 ± 3.93	81.79 ± 6.53	32.89 ± 2.43	62.46 (9.5)
		FTree*	55.95 ± 0.00	51.34 ± 1.47	67.94 ± 1.19	91.05 ± 3.34	31.05 ± 2.77	59.47 (12.2)
	Greedy	VIP	72.75 ± 1.10	$\mathbf{60.70 \pm 0.61}$	$\mathbf{84.70 \pm 4.02}$	96.60 ± 3.04	56.01 ± 2.37	**74.15** (2.0)
		ReMLP*	62.70 ± 2.92	59.59 ± 1.21	79.24 ± 4.26	89.51 ± 4.11	48.15 ± 7.45	67.84 (7.6)
		CART*	62.57 ± 7.83	52.22 ± 3.62	78.14 ± 4.10	93.21 ± 8.35	49.64 ± 9.91	67.16 (8.6)
		MLP*	63.76 ± 7.41	53.33 ± 3.61	78.69 ± 3.38	79.32 ± 18.01	48.56 ± 8.65	64.73 (9.4)
		DIME	67.07 ± 2.23	54.33 ± 0.41	73.59 ± 9.00	79.01 ± 9.33	45.94 ± 3.05	63.99 (10.5)
		FTree*	58.47 ± 11.14	46.15 ± 0.79	75.41 ± 0.00	82.10 ± 14.51	39.17 ± 1.53	60.26 (12.4)

Table 8 presents the average performance and standard deviation across all budget levels (1-10 features) for all methods and datasets for uniform feature costs. RL-based methods show strong overall performance. Among RL-based approaches, RL-MLP and RL-ReMLP achieve the best results (71.37% and 70.85%). The rule-based RL-CART (69.65%) outperformed CWCF (67.95%, rank 6.9) in this case, which is contrary to the results obtained with variable feature costs. Greedy optimisation approaches show surprisingly variable performance. VIP achieves the best overall results (74.15%, rank 2.0), and was the only dynamic method outperforming static CMI selection in terms of average accuracy (72.46%, rank 3.6). However, the rest of the greedy strategies usually perform worse than RL approaches, which is particularly visible in methods tested using both approaches.

Despite the theoretical advantages of DFS, static CMI-based feature selection (72.46%, rank 3.6) outperforms eleven of the twelve dynamic methods tested (all except VIP), including sophisticated approaches like DIME and INVASE. This result underscores the difficulty of learning effective DFS systems, which can outweigh the potential benefits of sample-specific feature selection. TabNet (70.65%, rank 5.2) also falls below Static CMI, showing the importance of model adaptability even when all features are visible.

All methods show consistently better results in the uniform-cost case than in the variable-cost case, as the total number of features queried within the budget is larger. However, the different rank results also suggest that the heuristic of dividing the usefulness of a feature by its cost is more successful in some methods than others.

E.2 Performance evolution for Individual Datasets

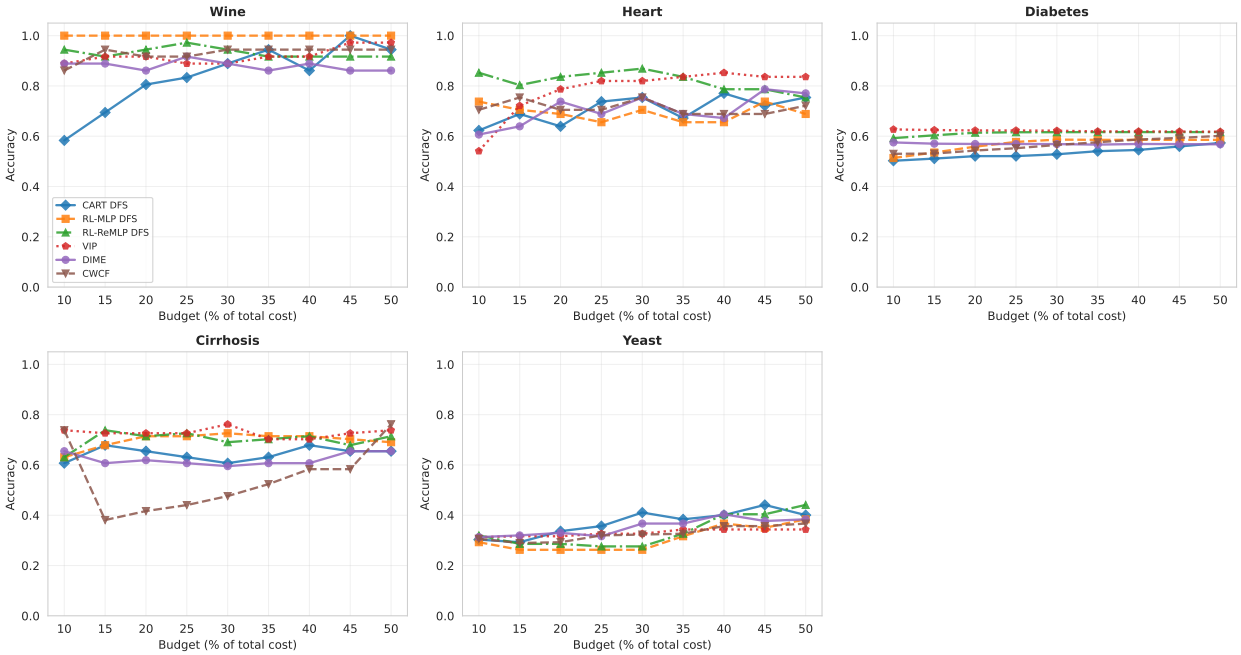


Figure 10: Accuracy evolution according to budget for all tabular datasets tested for different DFS selection methods using variable feature costs.

Here, we analyse accuracy progression (Figures 10 and 11) and calibration quality via ECE (Figure 12) individually per dataset. Because of visibility limitations, the figures only show the following DFS methods: G-CART, RL-MLP, RL-ReMLP, VIP, DIME and CWCF.

Wine dataset. It represents a well-structured classification problem where features have clear discriminative power. All methods show strong convergence to high accuracy: most methods achieve above 95% accuracy by 7–8 features (Figure 11), with VIP, RL-ReMLP, and RL-MLP reaching near-perfect performance (98–100%). This suggests the dataset has highly informative features that enable reliable classification with relatively small feature subsets. However, G-CART shows notable volatility in the 2–4 feature range (60–85% accuracy), only stabilising after acquiring at least 5 features. This reflects the challenge of adapting decision tree structures trained for the whole set of features to very small feature subsets, a pattern not observed for RL-MLP despite its use of a global classifier backbone. DIME exhibits the poorest performance in the uniform setting, but performs substantially better with variable feature costs. The ECE analysis reveals that high accuracy does not translate to reliable confidence estimates. Most methods maintain ECE of 0.15-0.26, indicating systematic overconfidence. VIP shows particularly erratic calibration (ECE ranging 0.05-0.38), with the worst calibration occurring at 4 features (ECE = 0.38) despite reasonable accuracy at that point. RL-ReMLP maintains the most stable and best calibration (ECE $\sim$0.05 throughout). However, these results vary significantly across datasets.

Heart dataset. It presents a binary classification problem with moderate sample size and mixed feature informativeness. Unlike Wine, the Heart dataset shows a smoother performance evolution throughout the acquisition process. However, multiple methods exhibit non-monotonic accuracy curves in the uniform setting, which highlights the challenges discussed in Section 3. The results also show the importance of the back-

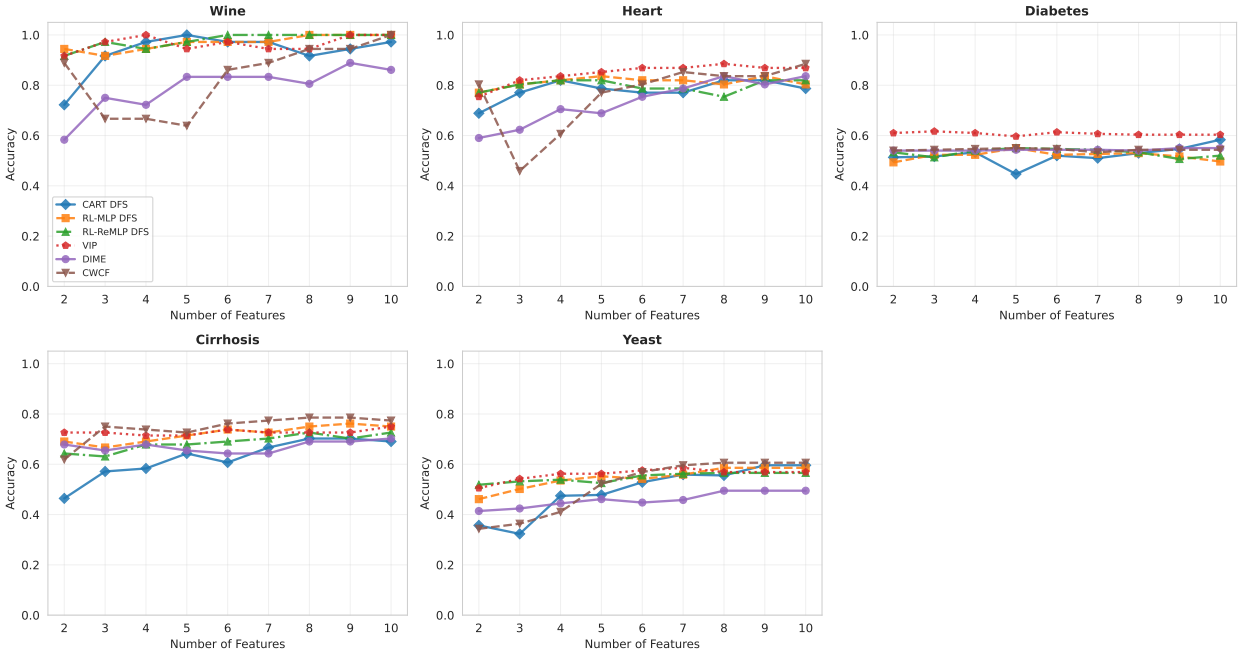
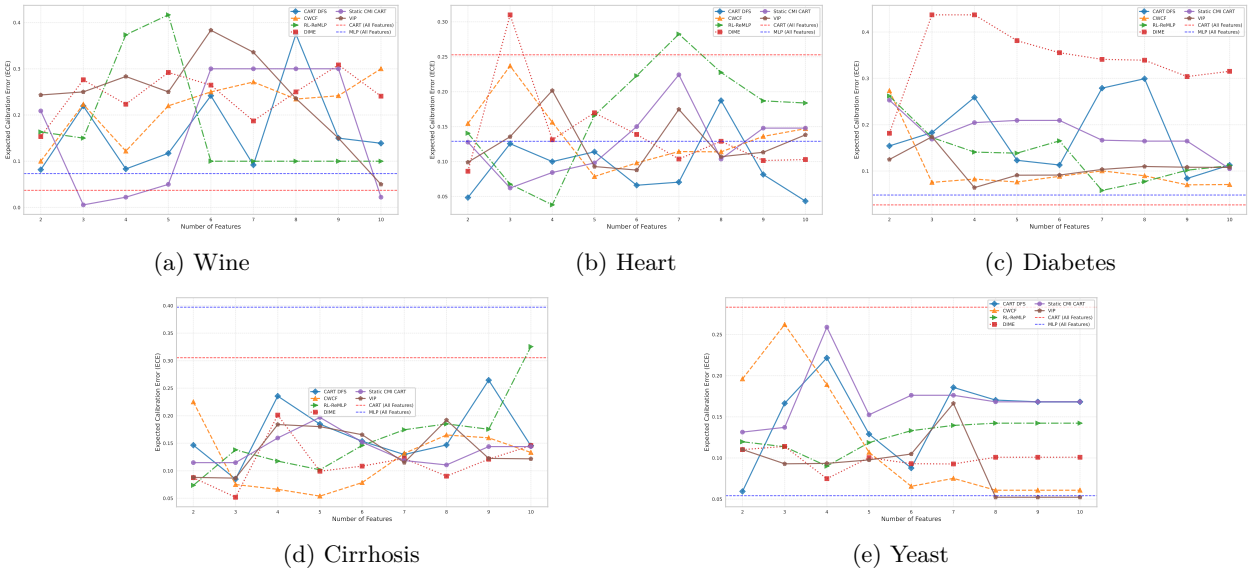


Figure 11: Accuracy evolution according to budget for all tabular datasets tested for different DFS selection methods using uniform feature costs.



(a) Wine (b) Heart (c) Diabetes

(d) Cirrhosis (e) Yeast

Figure 12: ECE evolution according to budget for all tabular datasets tested for different DFS selection methods using variable feature costs.

bone classifier: most neural methods converge to similar accuracy at the maximum budget, while G-CART remains substantially lower. ECE evolution also shows extreme volatility on Heart. No method maintains consistently good calibration, although average ECE at larger budgets is lower than at the beginning, suggesting that calibration can be affected by the available budget. This tendency, however, is not present in the remaining datasets.

Diabetes dataset is the largest in our evaluation, with high dimensionality and substantial class imbalance. Regarding performance, all methods converge to remarkably similar accuracy values (60–63%) regardless of feature budget, with minimal improvement beyond 4–5 features. This flat performance landscape suggests that the most discriminative features provide limited predictive power, and additional features offer marginal information gain. The backbone classifier is again important: G-CART is the weakest performer, illustrating the trade-off between explainability and performance. The uniform-cost setting is also substantially more successful than the variable-cost setting, highlighting the importance of the acquisition heuristic used in each situation. Unlike other datasets, Diabetes shows some methods maintaining remarkably stable ECE values across all budgets. VIP and RL-ReMLP hold ECE around 0.02, representing the best calibration observed across all datasets and methods. This is particularly notable as both are also the most accurate DFS methods overall, ruling out the possibility that good calibration merely reflects overly conservative base-rate estimation. This stability suggests that well-adapted methods on uniformly weak features can learn to consistently express the dataset's irreducible aleatoric uncertainty regardless of acquisition budget.

Cirrhosis dataset. It represents a three-class clinical problem with moderate dimensionality and limited training data. This is another case where the G-CART method works well for uniform feature costs but lags behind other methods in variable feature costs. We also see that the methods are not always monotonic in their performance. CWCF shows an unusually poor performance in the uniform setting, which we attribute to the RL-trained policy struggling on smaller datasets with limited training samples. Other RL approaches used a retrained backbone, which protects them from this instability in the predictor training phase. Regarding ECE, this is completely erratic in this dataset, as all methods exhibit substantial ECE fluctuations.

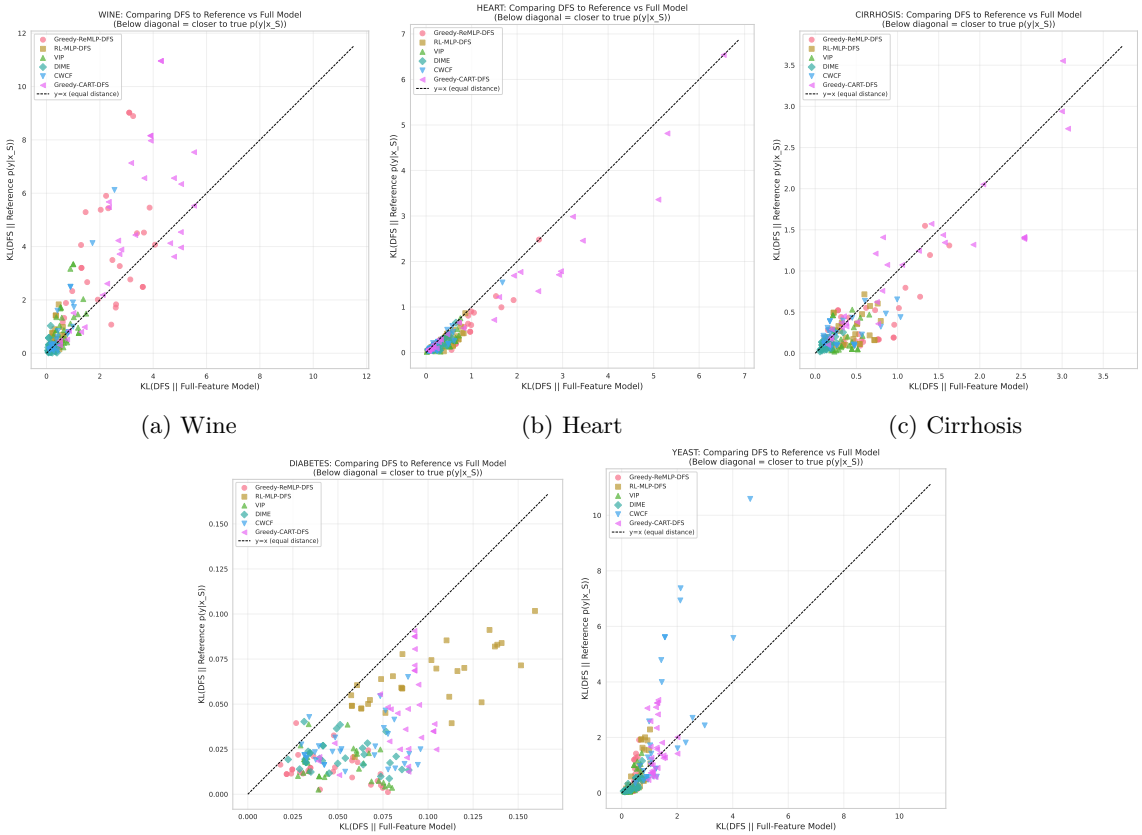
Yeast dataset. It presents the most challenging classification problem with 10 classes and minimal feature dimensionality. In this case, querying for new features is less important than the base classifier. This, for example, leaves a performance gap of 10 points between DIME and CWCF when all the features are queried. ECE here is also quite erratic.

E.3 Prediction Bias Under Partial Feature Observation

To assess how DFS methods approximate the true conditional distribution under partial observation, we compare their predictions against two references: (i) a subset-adaptive classifier trained exclusively on x_S, used as a proxy for $p(y \mid x_S)$, and (ii) the full-information model trained on all features. For each method and dataset, we compute $D_{\mathrm{KL}}\left(f_{\theta^*}(\mathbf{x}_S) \| \, f_{\theta^*_S}(\mathbf{x}_S)\right)$ and $D_{\mathrm{KL}}\left(f_{\theta^*}(\mathbf{x}_S) \| \, f_{\theta^*}(\mathbf{x})\right)$. The first term quantifies subset-adaptation error, while the second captures bias toward the full-information model.

Figure 13 plots these quantities against each other. Points below the diagonal indicate closer agreement with the subset-adaptive reference, whereas points near or above the diagonal suggest imperfect subset adaptation or excessive similarity to the full model. Across datasets, the relative position of methods with respect to the diagonal varies substantially, confirming that subset adaptation and approximation to the full model are distinct phenomena that depend on both dataset structure and DFS strategy. On Wine and Cirrhosis, several methods lie predominantly above the diagonal, indicating that imperfect subset adaptation is the dominant source of error. In contrast, on Diabetes, most methods cluster near the origin, suggesting that small feature subsets already recover much of the predictive signal, leaving limited room for either adaptation error or leakage. Yeast exhibits the highest dispersion, with both subset adaptation and approximation errors coexisting and varying across methods.

Comparing the relative distance of points to the diagonal reveals systematic differences in how DFS methods handle missing information. Methods such as DIME consistently appear below or near the diagonal across datasets, indicating consistency with subset-adaptive reference predictions. In contrast, several RL-based and variational approaches exhibit points closer to the diagonal or below it in datasets where subset adaptation should induce stronger deviations, suggesting partial reconstruction of missing information through learned correlations. Importantly, large deviations from both references, as observed for G-CART on Wine and Yeast, are not indicative of leakage but rather of unstable subset adaptation, emphasising that low agreement with the full model alone is insufficient to diagnose correct DFS behaviour. Overall, these results reinforce the need to jointly evaluate divergence to both subset-adaptive and full-information references when assessing DFS



(a) Wine (b) Heart (c) Cirrhosis

(d) Diabetes (e) Yeast

Figure 13: Bias decomposition of DFS predictions across datasets. Each subfigure plots $D_{\mathrm{KL}}(f_{\theta^*}(\mathbf{x}_S) \,\|\, f_{\theta^*}(\mathbf{x}))$ (x-axis) against $D_{\mathrm{KL}}\big(f_{\theta^*}(\mathbf{x}_S) \,\|\, f_{\theta^*_S}(\mathbf{x}_S)\big)$ (y-axis). Points below the diagonal indicate good subset adaptation, while points above indicate a stronger bias toward the full-information model.

methods, as reliance on either comparison in isolation can lead to misleading conclusions about adaptation quality and information usage.

F Targeted Validation Experiments

We provide five additional experiments to further validate robustness under distribution shift, the relationship between epistemic uncertainty and subset adaptation risk, the effect of last-layer reparametrisation on reducing adaptation error, the correlation between aleatoric and epistemic uncertainty, and the effect of the epistemic penalty λ on image datasets. Unless stated otherwise, these tests use the G-CART, G-MLP, and G-ReMLP backbones with two budget levels: up to 30% and up to 70% of the total features.

F.1 Robustness Under Distribution Shift

Here, we test if epistemic uncertainty increases under different distributional perturbations and how much predictive performance and epistemic uncertainty are affected by them. Perturbations are applied only to the test set to simulate distribution shift: for 50% of features, additive Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is applied with $\sigma \in \{0.1, 0.2, 0.8, 2.0\}$. For each model and budget level, we compute accuracy, ECE and mean predictive entropy. A consistent increase in entropy and selective prediction performance under perturbation indicates that uncertainty estimates respond appropriately to distribution shift.

Table 9: Performance under distribution shift, averaged over all tabular datasets. B: budget level (% of features); e: mean epistemic uncertainty (only for methods with auxiliary models).

		No shift			σ=0.1			σ=0.2			σ=0.8			σ=2.0		
Method	B	Acc	ECE	e	Acc	ECE	e	Acc	ECE	e	Acc	ECE	e	Acc	ECE	e
G-CART	30%	.69	.08	.043	.67	.08	.045	.68	.08	.046	.60	.14	.051	.54	.17	.057
	70%	.73	.08	.046	.73	.08	.047	.71	.10	.050	.63	.16	.058	.56	.22	.069
RL-ReMLP	30%	.72	.06	.006	.71	.07	.006	.70	.08	.006	.64	.13	.008	.59	.22	.017
	70%	.76	.10	.015	.76	.08	.015	.74	.10	.015	.67	.18	.024	.54	.35	.042
DIME	30%	.68	.08	—	.67	.08	—	.67	.08	—	.64	.15	—	.59	.23	—
	70%	.74	.09	—	.74	.08	—	.74	.09	—	.71	.12	—	.62	.24	—
VIP	30%	.66	.18	—	.67	.19	—	.66	.18	—	.66	.18	—	.60	.12	—
	70%	.72	.19	—	.72	.19	—	.72	.19	—	.70	.17	—	.65	.16	—

Table 9 reports the averaged results across datasets. Under mild covariate noise ($\sigma = 0.1$), all methods exhibit negligible accuracy loss (below 3 points on average), confirming that DFS predictions are robust to small input perturbations. At $\sigma = 0.2$, the drops remain moderate, around 2 points for G-CART at the 70% budget, and no method degrades catastrophically. A stronger perturbation, $\sigma = 0.8$, reveals clearer separation: G-CART drops to .60 and .63 at the 30% and 70% budgets, respectively, while neural methods' performance remains less affected. Under the largest perturbation ($\sigma = 2.0$), all methods experience significant accuracy degradation. The sharpest drops occur in G-CART and RL-ReMLP. Methods whose architectures are jointly trained to handle partial observations are more resilient here. The epistemic uncertainty is considerably higher than in previous σ values as well, further confirming that epistemic uncertainty reliably grows with severe distributional mismatch.

For the methods that provide epistemic uncertainty through auxiliary models, even the mildest perturbation ($\sigma = 0.1$) produces a measurable increase in e. At $\sigma = 0.8$ G-CART epistemic uncertainty rises by 19% (from .043 to .051) at the 30% budget and by 26% (from .046 to .058) at 70%, while RL-ReMLP increases by 33% and 60%, respectively. Under the most severe perturbation ($\sigma = 2.0$), the epistemic response is even more pronounced: RL-ReMLP epistemic uncertainty at 70% reaches .042, nearly triple its unperturbed value, while G-CART rises to .069 (+50%). ECE follows a similar upward trend for G-CART, RL-ReMLP, and DIME, confirming progressive calibration degradation under shift; VIP is the exception, as its ECE remains flat and even decreases slightly under the strongest perturbation, although it starts from the highest baseline calibration error.

These results support two conclusions. First, the DFS predictions are robust to the levels of covariate noise that might arise from measurement error, with accuracy losses below 4 points up to $\sigma = 0.2$. Second, and more importantly for the uncertainty framework, the epistemic estimates increase monotonically with perturbation severity, confirming that they capture genuine distributional mismatch rather than being mere artefacts of the ensemble.

F.2 Epistemic Uncertainty and Subset Risk Gap

Here, we empirically assess whether epistemic uncertainty correlates with the subset adaptation risk gap $\Delta_S = R_S(f_{\theta^*}) - R_S\big(f_{\theta^*_S}\big)$ introduced in Section 3.1. For a subset S with $|S| = b$, we denote $R_S(f_{\theta^*})$ as the test error of the globally trained model evaluated under mask S, $R_S\big(f_{\theta^*_S}\big)$ as the test error of a model trained exclusively on features S, and $\bar{e}_S$ as the mean epistemic uncertainty across test samples under mask S. Then, for each of the five budget levels (10–90%), we sample 30 random subsets of size b ($n = 150$ per dataset). For each subset, we compute $(\Delta_S, \bar{e}_S)$ and report Pearson and Spearman correlations across all sampled subsets. A positive correlation indicates that epistemic uncertainty reliably identifies subsets for which global masking induces substantial performance degradation relative to a locally adapted oracle.

Table 6 and Figure 3 report the results. For G-CART, the correlation between Δ_S and $\bar{e}_S$ is strongly positive on Wine, Cirrhosis, and Heart (wine: r=0.70; cirrhosis: r=0.45; heart: r=0.31; all p<0.001), near

Table 10: Reparametrisation effect on the subset risk gap. Δ_S is the signed gap (positive = global worse than local); $|\Delta_S|$ is the absolute gap. "Wins" reports the fraction of subsets where $|\Delta_S^{\text{G-ReMLP}}| < |\Delta_S^{\text{G-MLP}}|$. Wilcoxon p-values test whether $|\Delta_S|$ differs between the two models.

	Mean Δ_S		Mean $\lvert\Delta_S\rvert$		Wins	Wilcoxon
Dataset	G-MLP	G-ReMLP	G-MLP	G-ReMLP	(%)	p
Wine	+.028	+.003	.051	.040	39.5	.008
Heart	−.040	−.030	.053	.047	47.0	.048
Cirrhosis	−.025	−.023	.041	.042	40.0	.706
Diabetes	+.002	−.002	.025	.024	46.5	.871
Yeast	+.012	−.002	.019	.018	39.5	.734
All	−.005	−.011	.038	.034	42.5	.022

zero on Diabetes ($r = -0.07$), and negative on Yeast ($r = -0.19$), with an overall Pearson r of 0.66 and Spearman ρ of 0.49 pooled across subsets and budgets. As noted in Section 6, the pooled coefficients exceed most per-dataset values because pooling additionally captures between-dataset differences in the scales of Δ_S and epistemic uncertainty. For G-ReMLP, correlations are much weaker (overall r=0.02, ρ=0.05), with mixed signs across datasets. This is explained by the method's characteristically small Δ_S: averaged across datasets, $\bar{\Delta}_S = -0.01$ at both budget levels, indicating that the reparametrised global model matches or slightly outperforms locally trained oracles. When the adaptation gap is near zero, there is little signal for any uncertainty proxy to track.

F.3 Effect of Reparametrisation on Subset Risk

In this experiment, we test how much last-layer reparametrisation reduces the subset risk gap Δ_S. For each sampled subset S, we compute:

$$\Delta_S^{\text{G-MLP}} = R_S\big(f_{\theta^*}^{\text{G-MLP}}\big) - R_S\Big(f_{\theta_S^*}^{\text{G-MLP}}\Big), \tag{31}$$

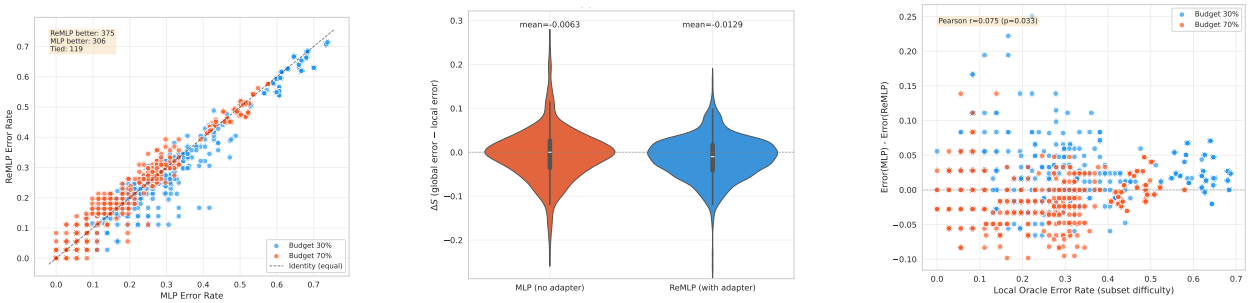
$$\Delta_S^{\text{G-ReMLP}} = R_S\big(f_{\theta^*}^{\text{G-ReMLP}}\big) - R_S\Big(f_{\theta_S^*}^{\text{G-ReMLP}}\Big). \tag{32}$$

We compare the distributions of $\Delta_S^{\text{G-MLP}}$ and $\Delta_S^{\text{G-ReMLP}}$ across all subsets and budgets using the Wilcoxon signed-rank test. We additionally report mean and median risk gaps and the fraction of subsets for which reparametrisation reduces $|\Delta_S|$. A systematic reduction of Δ_S confirms that the adapter mitigates the bias induced by mean imputation under arbitrary subsets.

Table 10 reports the signed and absolute risk gaps for each dataset. Across all 800 sampled subsets (5 datasets $\times$ 2 budgets $\times$ 80 subsets), the reparametrised model has the smaller absolute risk gap $|\Delta_S|$ on 42.5% of subsets, yet attains a lower mean $|\Delta_S|$ than the standard MLP overall (0.034 vs. 0.038, Wilcoxon signed-rank $p = 0.022$). That is, the adapter is closer to the local oracle in fewer than half of the individual subsets, but its improvements are larger in magnitude than its degradations, so it narrows the gap between global and locally trained models on average rather than uniformly across subsets.

The effect is most pronounced on *wine*, where the adapter nearly eliminates the positive gap ($\Delta_S^{\text{G-MLP}} = +0.028 \rightarrow \Delta_S^{\text{G-ReMLP}} = +0.003$; Wilcoxon $p = 0.008$), and is also significant on *heart* ($p = 0.048$). On *cirrhosis*, *diabetes*, and *yeast*, the reduction is not statistically significant, which we attribute to both models already achieving near-local performance on these datasets (mean $|\Delta_S| \leq 0.025$ on diabetes and yeast).

The violin plot (Figure 14b) further illustrates the narrower Δ_S distribution of the reparametrised model: while the G-MLP distribution is centred near zero ($\overline{\Delta_S} = -0.005$), the G-ReMLP distribution is shifted further below zero ($\overline{\Delta_S} = -0.011$), indicating that the adapter not only reduces the gap but systematically improves upon local specialists. This could indicate that global behaviour plus local reparametrisation can achieve a regularisation effect, or that some leakage with respect to the full prediction could be happening.



(a) (b) (c)

Figure 14: Reparametrisation analysis across 800 sampled subsets (5 datasets × 2 budgets × 80 subsets). **(a)** Scatter of per-subset error rates: points below the identity line indicate G-ReMLP outperforms G-MLP. **(b)** Violin plots of Δ_S for each model; the G-ReMLP distribution is narrower and shifted further below zero. **(c)** G-ReMLP advantage as a function of subset difficulty (local oracle error).

Table 11: Effect of epistemic uncertainty weight λ on image datasets (MC Dropout rate 0.4) when the budget is 50% of the total budget. Mean accuracy (in %) and standard deviation (±) across 5 different seeds. Best accuracy per dataset in **bold**.

	MNIST		CIFAR-10		Imagenette	
λ	Acc.	Epistemic Unc.	Acc.	Epistemic Unc.	Acc.	Epistemic Unc.
−1.0	98.24 ± 1.14	0.150 ± 0.047	72.50 ± 5.26	0.129 ± 0.040	77.84 ± 3.56	0.214 ± 0.044
−0.5	98.42 ± 0.97	0.138 ± 0.041	72.41 ± 5.34	0.129 ± 0.039	77.59 ± 3.41	0.210 ± 0.049
−0.1	98.51 ± 0.89	0.133 ± 0.039	72.58 ± 5.47	0.130 ± 0.040	77.69 ± 3.56	0.208 ± 0.049
0.0	98.52 ± 0.88	0.132 ± 0.039	72.63 ± 5.39	0.129 ± 0.039	77.67 ± 3.64	0.208 ± 0.048
0.1	98.52 ± 0.88	0.132 ± 0.039	72.67 ± 5.35	0.130 ± 0.040	77.66 ± 3.66	0.207 ± 0.050
0.5	98.53 ± 0.85	0.131 ± 0.037	**72.86 ± 5.14**	0.129 ± 0.040	77.73 ± 3.52	0.207 ± 0.049
1.0	**98.56 ± 0.80**	0.129 ± 0.035	72.77 ± 5.12	0.128 ± 0.040	**77.92 ± 3.55**	0.207 ± 0.050

F.4 Relationship between Epistemic and Aleatoric Uncertainty

We examine how aleatoric and epistemic uncertainty jointly distribute across instances, computed according to Eqs. (19) and (15). Figure 15 displays scatter plots of both quantities across all experiments and DFS methods. For most models, epistemic uncertainty remains predominantly low, while aleatoric uncertainty is more broadly spread across instances. Notably, instances with the highest aleatoric uncertainty tend to exhibit low epistemic uncertainty, and vice versa. This result differs from findings in prior work (Valdenegro-Toro & Mori, 2022), but is not entirely surprising. In high aleatoric cases, significant class overlap causes auxiliary models to quickly converge to conservative and similar predictions. Conversely, when epistemic uncertainty is high, aleatoric uncertainty appears difficult to estimate properly. This mutual exclusivity is particularly striking for G-CART, whose instances form a clear L-shape with almost no points combining high epistemic and high aleatoric uncertainty.

F.5 Effect of Epistemic Penalty on Image Datasets

The pattern is analogous to the tabular case: positive λ values produce small accuracy gains on MNIST, CIFAR-10 and Imagenette, but epistemic estimate reductions are smaller. Note that this is computed with MC-Dropout instead of real ensembles, which may explain the more conservative epistemic uncertainty values (Rahaman et al., 2021).

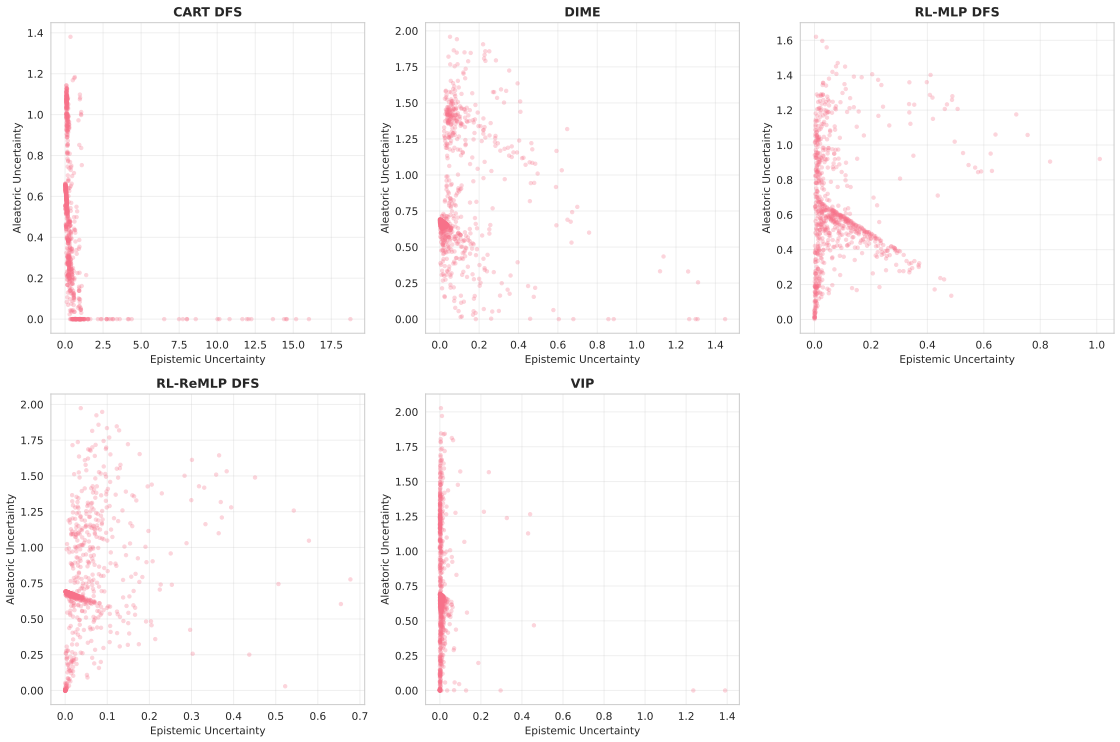


Figure 15: Relationship between aleatoric and epistemic uncertainty, averaged across all datasets and budgets, for the different methods shown.

G Implementation Details and Reproducibility

G.1 Dataset Statistics

Table 12: Dataset Summary

Dataset	Samples	Features	Classes	Acc. LR
Diabetes	101,766	47	2	0.61
Heart	303	13	2	0.86
Cirrhosis	520	17	3	0.72
Wine	178	13	3	0.98
Yeast	1,484	8	10	0.57

A summary of the classification datasets' size and properties is in Table 12. We also included the performance of a logistic regressor for each one as a baseline reference.

- *Diabetes* dataset (Strack et al., 2014): has 101766 samples and 47 features regarding patient condition. The target value is if the patient is readmitted to the hospital in a 30 day period after the last discharge.

- *Heart* disease (Detrano et al., 1989): contains 13 clinical features for each of the 303 patients. The target field refers to the presence of heart disease in the patient, which has been binarised into presence and absence.

- *Cirrhosis* Patient Survival Prediction (Fleming & Harrington, 2013): uses 17 clinical features for predicting the survival state of 520 patients with liver cirrhosis. These include death, alive, or alive due to liver transplantation.
- *Wine*: the classical dataset containing chemical analysis results of wines (Aeberhard & Forina, 1992). It includes 178 samples with 13 features, used for classification into wine types.
- *Yeast*: protein localization sites in yeast cells (Nakai, 1991). It consists of 1,484 samples with 8 features, predicting the cellular compartment where a protein resides.

G.2 Value Function Estimator Architecture

At inference time, the value function $v_q(i, \mathbf{x}_S)$ defined in Eq. (10) cannot be directly computed because the unqueried features x_i are not observed. To address this, we train a neural network estimator h_ϕ that takes the observed feature subset $\mathbf{x}_S$ as input and predicts the expected value of querying each unobserved feature.

Architecture. We use a single neural network with two output heads:

- $h_{u,i}(\mathbf{x}_S)$: predicts the reduction in prediction discrepancy $u(\mathbf{x}_S) - u(\mathbf{x}_{S\cup\{i\}})$
- $h_{e,i}(\mathbf{x}_S)$: predicts the epistemic term $e(\mathbf{x}_{S\cup\{i\}})$

Each head has M outputs (one per feature), allowing the network to predict the value of adding any unobserved feature given the current observation $\mathbf{x}_S$.

Network Architecture Details. For tabular experiments, we use:

- Shared encoder: 2-layer MLP with hidden size 64.
- Activation: SELU with dropout rate 0.5.
- Two separate output heads, each producing M values.
- Training: 100 epochs (comparison experiments), 10 epochs (individual experiments).
- Learning rate: 0.01, Adam optimiser.
- L1 regularisation: tested weights $\in \{0.0, 0.001, 0.01, 0.1, 0.5, 1.0, 2.0\}$

G.3 Hyperparameters

Tabular Experiments. The architecture and training parameters of the value-function estimator are detailed in Appendix G.2.

The hyperparameters for the backbone classifiers for our DFS proposal are:

- **Decision Tree**: minimum of 5 samples for a split, a minimum of 2 samples to generate a leaf, no maximum depth constraint
- **MLP**: one hidden layer of size 100 with ReLU activation, Adam optimiser with 200 iterations

Reinforcement Learning Policy. RL-based feature selection is trained via Q-learning (Clifton & Laber, 2020). The state consists of the binary mask of observed features concatenated with their current values. The Q-network is a 3-layer MLP (hidden size 128, ReLU activations) with a duelling architecture, outputting Q-values for all candidate feature queries. Training uses ϵ-greedy exploration (ϵ decayed from 1.0 to 0.01), discount factor $\gamma = 0.99$, replay buffer size 10,000, batch size 32, target network updates every 100 steps, and Adam optimiser (learning rate 10^{-4}). Rewards are defined as $R_t = q(x_i, \mathbf{x}_{S_t})$ (Eq. 11), where x_i is the observed value of the queried feature, and policies are trained for 1,000 episodes per dataset.

Deep Learning Models. All deep learning methods are trained for 100 total epochs to ensure comparability, with method-specific stage allocation: VIP (100 joint), DIME (25 pretraining + 75 joint), Greedy-MLP (100 policy), Greedy-ReMLP (60 adapter + 40 policy).

Training uses a batch size of 64 and the Adam optimiser. Learning rates are 10^{-3} for pretraining/adapter phases and 10^{-4} for policy networks. We use random seed 42 and 30 MC-Dropout samples for uncertainty estimation. Images are divided into a 4×4 grid (16 patches) with a budget range of 2–10 patches.

All methods use a ResNet-18 backbone, pretrained on ImageNet for RGB datasets (CIFAR-10, Imagenette) and trained from scratch for MNIST. The dropout rate is set to 0.4 for MC-Dropout inference.

Method-Specific Architectures.

- **VIP Query Network**: Fully connected layers with hidden sizes 1000 and 500, dropout rate 0.3.
- **DIME Selector Network**: Fully connected layers with hidden sizes 512 and 256, dropout rate 0.3.
- **Greedy Policy Estimator**: Hidden dimension 256, dropout rate 0.3, L1 regularisation weight 0.1.
- **Subset Adapter (ReMLP methods)**: hidden dimension 64, generating scale and shift parameters for the final classification layer.

G.4 Code Availability

All code, including custom implementations of dynamic feature selection methods, neural estimators, and evaluation frameworks, will be made publicly available upon publication.