ELEKTROTEHNIŠKI VESTNIK 93(1-2): 14–31, 2026
ORIGINAL SCIENTIFIC PAPER

Active feature acquisition by prediction explanations

Matjaž Kukar

Machine Learning and Language Technologies Laboratory, University of Ljubljana,
Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenia,

E-mail: matjaz.kukar@fri.uni-lj.si

Abstract. In many real-world machine learning applications, particularly in resource-constrained domains such as medical diagnostics, acquiring feature values is a costly and often sequential process. Active Feature Acquisition (AFA) addresses this by selecting the most informative subset of features to acquire for a given instance, balancing predictive accuracy against acquisition costs. Conventional AFA strategies often rely on static, global feature importance metrics, which are instance-agnostic and can be suboptimal when feature relevance is context-dependent. We investigate a practically relevant two-stage acquisition scenario, where an initial subset of features is observed, and a subsequent non-overlapping subset is selected for acquisition. We propose a novel AFA strategy that leverages instance-specific model explanations, specifically Shapley additive explanations (SHAP), to guide the selection process. We conduct a systematic comparative study, evaluating three distinct acquisition strategies: random acquisition (as a baseline), acquisition guided by static global feature importance lists, and the proposed SHAP-based approach. Using XGBoost-trained models, these methods are evaluated across ten benchmark datasets under two missingness scenarios: Missing Completely at Random (MCAR) and Missing Not at Random (MNAR). The empirical results demonstrate that the SHAP-based strategy significantly outperforms static global feature importance methods in complex settings, particularly under MNAR missingness, where the context of observed features is critical for effective decision-making. While it also performs strongly in simpler MCAR scenarios (especially at high missingness rates), its robustness in more realistic settings suggests that utilizing instance-specific explanations provides a powerful and adaptive mechanism for personalized and effective feature acquisition.

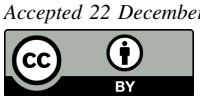
Keywords: active feature acquisition, Shapley additive explanations, missing data, machine learning

Aktivno pridobivanje značilk z razlago napovedi

V mnogih praktičnih aplikacijah strojnega učenja, npr. v medicinski diagnostiki, je pridobivanje dodatnih značilk lahko drag in zamuden postopek. Aktivno pridobivanje značilk (AFA) naslavlja ta problem, a enostavne strategije pogosto uporabljajo statične, globalne metrike, ki so neodvisne od konteksta posameznega primera in zato suboptimalne. V članku predlagamo novo strategijo, ki jo vodijo razlage SHAP, personalizirane za posamezen primer. Izvedli smo sistematično primerjavo strategij za aktivno pridobivanje značilk (naključna, statična in na razlagah osnovana) na desetih podatkovnih zbirkah, z mehanizmoma manjkajočih podatkov MCAR in MNAR. Empirični rezultati kažejo, da je strategija, osnovana na razlagah, značilno boljša od statičnih metod, zlasti v kompleksnih in realističnih pogojih MNAR. Personalizirane razlage predstavljajo močan in prilagodljiv način za učinkovito pridobivanje dodatnih značilk.

Ključne besede: aktivno pridobivanje značilk, Shapleyeve razlage, manjajoči podatki, strojno učenje.

Received 14 November 2025
Accepted 22 December 2026



1 INTRODUCTION

Machine learning models are often developed under the assumption that complete data is readily and inexpensively available for both training and inference. However, this assumption does not hold true in many real-world domains, where obtaining feature values can incur significant costs in terms of time, financial resources, or even physical risk [24]. A typical example is medical diagnosis, where physicians must make decisions based on incomplete data, as no hospital can afford to conduct all possible diagnostic tests. Faced with a patient's symptoms and basic test results, a physician has to determine which additional diagnostic tests – such as blood work, imaging scans, or biopsies – should be requested. Each test comes with its own utility, which consists of diagnostic value, cost, and potential risks to the patient, necessitating careful consideration [14, 38].

The primary motivation for our work is a specific, yet common, AFA scenario that mirrors the clinical workflow: a two-stage acquisition process [38]. An instance (patient) is first observed with a subset of all possible features (e.g., history, signs and symptoms, initial

diagnostic tests). Based on this partial information, a decision is made regarding which additional features to acquire in order to improve the predictive performance of the classification (diagnostic) task. This reflects the situation in which a physician, after reviewing initial test results, orders a follow-up set of more specialized tests from a stored biological sample [38].

Our central claim is that an AFA strategy guided by *instance-specific* feature importance is more effective and robust, particularly in complex scenarios with significant feature interactions. We propose a novel approach, *Shapley-based active feature acquisition (AFA-SHAP)*, that leverages Shapley additive explanations [16] to drive acquisition decisions. For each instance we generate SHAP values from a predictive model's prediction on partial data. These values quantify the contribution of each missing feature to the current prediction, conditional on the observed features. By selecting the missing features with the largest absolute SHAP values, we create a dynamic, context-aware acquisition strategy personalized to each instance. We evaluate the proposed hypothesis with XGBoost-trained models [6] which conveniently provide native missing data handling.

2 Related Work

The sequential, utility-based decision-making process is the central challenge addressed by Active Feature Acquisition (AFA). In many domains, the number of sequential steps may be limited; e.g., in medical diagnostics, physicians typically order blood work in one or two batches. AFA is a dynamic, instance-wise approach [8] that contrasts with traditional static feature selection, which identifies a single, fixed set of features to be used for all instances in a dataset. The dynamic nature of AFA allows for personalized data acquisition, which is essential in fields like medicine, where the most relevant tests for one patient may be irrelevant for another with different symptoms [34, 38].

From a machine learning (ML) perspective, AFA is closely related to handling missing data (unacquired feature values), which may be missing in some systematic way. Given an instance $\mathbf{x} = (x_1, \ldots, x_d)$, the AFA task is to acquire *best* missing features from $\mathbf{x}$ to maximize some performance metric, given the acquisition budget [8, 24]. A ML model $\mathcal{P}$ then uses both original and newly acquired features to predict the outcome y.

2.1 Active feature acquisition (AFA)

AFA can be broadly categorized into greedy, embedded, and reinforcement learning approaches [11, 14, 24].

- *Greedy* methods represent the most direct approach, where at each step of a sequential acquisition process, the feature that maximizes a specific objective function is chosen [8, 34]. Recently, explainability [11] has also been considered to guide the acquisition process. However, acquiring one feature at a time is often impractical in real-life applications.
- *Embedded* methods integrate the feature acquisition process directly into the learning algorithm's training or inference procedure, e.g., cost-sensitive decision trees, which modify the splitting criteria to incorporate the acquisition cost of features [24].
- In *reinforcement learning (RL)* methods, AFA is formulated as a Markov Decision Process (MDP) [24, 34]. However, RL-based AFA methods often suffer from sparse rewards (a reward is only received after the final prediction), high-dimensional action space (one action per missing feature), and general training instability, which can severely inhibit performance in practice [1, 34]. Recent results [8, 11] show that RL-based AFA methods often perform worse than much simpler greedy methods.

In practice, greedy approaches are often the most useful because they are simple to implement and non-intrusive. Common implementations involve precomputing static feature importance and then acquiring the subset of most important missing features for a given instance. The proposed AFA-SHAP method can be considered partly a sophisticated greedy approach, where the utility function is derived from local, non-myopic model explanations, and partly an embedded approach, as the feature acquisition process is guided by the model itself.

2.2 Static feature acquisition (SFA)

As opposed to the dynamic, instance-wise nature of AFA, SFA seeks to identify a single optimal subset of features for the entire dataset using feature importance metrics, traditionally used for *feature selection*. Many feature importance metrics exist; we included – as reference for our study – some of the most popular: Information Gain Ratio [22], Minimum Description Length (MDL) [12], and ReliefF [27]. We also use the XGBoost model-specific feature importance (SFA-XGBoost), built into the XGBoost algorithm [6], to quantify feature contributions after model training. While convenient and computationally efficient, these metrics can be inconsistent, producing markedly different feature rankings for the same dataset. Averaged Shapley values ($\overline{\text{SHAP}}$, Sec. 4.4), precomputed from the training data, can also serve as a static feature importance measure [17].

2.3 Missing data mechanisms and AFA

AFA is closely related to the study of missing data. An unacquired feature is, by itself, simply a missing value. However, in AFA, missingness is not a passive problem to be handled by imputation, but an active state to be resolved by acquisition [38]. Understanding its nature is crucial for evaluation of an AFA strategy. Three canonical missingness mechanisms are [29]:

- Missing Completely at Random (MCAR): The probability that a feature value is missing is independent of both the observed and unobserved values in the dataset. This represents a scenario where data loss is a purely random process*.
- Missing at Random (MAR): The probability of a value being missing depends only on other *observed* feature values. E.g., in a survey, a person might be less likely to report the income if their education level (an observed feature) is low.
- Missing Not at Random (MNAR): The probability of a feature value being missing depends on the value itself (e.g., high-income individuals not disclosing it) or on an unobserved variable, such as a physician's clinical judgment.

AFA operates within this context, where the initial state is a dataset with missing values. The decision to not acquire a feature is a form of MNAR, as the choice is often based on an implicit estimation of its potential (unobserved) value. The connection between MNAR and instance-specific AFA makes a lot of sense: MNAR implies that the missingness pattern contains implicit information about which unobserved features are most likely to be useful. A static, global importance metric is blind to this instance-level context. On the other hand, an instance-specific method such as AFA-SHAP is well-suited to exploit this context because it conditions its importance calculations on the specific pattern of observed features.

2.4 Synthetic generation of missingness

Performance evaluation of AFA methods, much like for imputation algorithms, involves simulating missingness on fully or partially complete data. Missingness can be introduced in a single feature (univariate configuration) or across several features (multivariate configuration), at different rates, and according to MCAR, MAR, or MNAR mechanisms. As the simulation process defines the basis for the experimental evaluation, it is essential that it is applied appropriately [31].

Two common strategies for simulating MNAR missingness are Missingness Based on Own Values (MBOV) [21, 37] and Missingness Based on Unobserved Values (MBUV) [9, 21]. In MBOV, missingness probability depends on the value itself (e.g., low-range values are missing more often); in MBUV, it relates to an unobserved feature (e.g., the target variable).

3 Materials

Experimental evaluation is conducted on a collection of ten datasets from the UCI Machine Learning Repository, Kaggle, and a proprietary source (*medic* dataset). The selection covers a range of characteristics, including binary and multi-class classification tasks, varying numbers of instances and features, and a mix of numerical and categorical data types. This diversity ensures that the conclusions drawn from the study are generalizable and not specific to a narrow problem domain. A brief description of each dataset is provided below, with summary statistics presented in Table 1.

- *breast cancer* (Wisconsin): A binary classification task to predict whether a breast mass is malignant or benign. Features are computed from a digitized image of a fine needle aspirate (FNA) of the mass, describing characteristics of the cell nuclei [36].
- *adult_income*: A binary classification task to predict whether an individual's annual income exceeds $50,000 based on U.S. census data [3].
- *credit_card*: A binary classification task to predict whether a customer will default on his payment in the next month. The dataset consists of a mix of continuous and nominal features [40].
- *medic*: A proprietary multi-class medical dataset for differential diagnosis of over 200 diseases with high number of features and a very high missing rate (a superset of [13]).
- *online_shoppers*: This dataset is used for a binary classification task to predict whether a visitor to an e-commerce website will complete a purchase (generate revenue) based on their clickstream data and session information [30].
- *mushroom*: A binary classification task to determine whether a given mushroom is edible or poisonous. The dataset consists entirely of categorical features describing physical characteristics such as cap shape, color, and odor [32].
- *bank_marketing*: This dataset is related to direct marketing campaigns of a Portuguese banking institution. The binary classification goal is to predict if a client will subscribe to a term deposit based on client data and campaign contact information [20].
- *forest_cover*: A multi-class classification problem with seven distinct classes. The task is to predict the forest cover type based on cartographic variables such as elevation, slope, and soil type [4].
- *human_activity*: This is a multi-class classification task to recognize one of six human activities (e.g., walking, sitting, laying) based on 3-axial sensor data from a waist-mounted smartphone [26].
- *steel_plate*: A multi-class classification problem with seven fault types. The goal is to identify the type of surface defect in steel plates using 27 geometric and outline-based indicators [5].

3.1 Separate evaluation on the medic dataset

The *medic* dataset is evaluated separately from the other nine benchmark datasets due to its unique and

*It is reasonable to expect XGBoost feature importance (SFA-XGBoost) to work well in this setting, as it is closely related to the underlying model and will thus always acquire features that are expected to be the most important (useful) for the particular model.

Table 1. Descriptive statistics of the datasets used in evaluation.

Dataset name	Number of instances	Number of features	Missing rate	Number of classes	Majority class prevalence	Entropy of class distribution
breast_cancer	569	30	0.00	2	0.63	0.95
adult_income	32561	14	0.01	2	0.76	0.80
medic	122093	275	0.82	220	0.06	6.51
credit_card	30000	23	0.00	2	0.78	0.76
online_shoppers	12330	17	0.00	2	0.86	0.64
mushroom	8124	22	0.00	2	0.52	1.00
bank_marketing	41188	19	0.00	2	0.89	0.51
forest_cover	581012	54	0.00	7	0.49	1.74
human_activity	10299	562	0.00	6	0.19	2.58
steel_plate	1941	27	0.00	7	0.35	2.41

somewhat extreme characteristics. Averaging its performance with the others would distort the aggregate results and obscure specific insights. Primary reasons for the separate analysis are:

- Very high natural missingness. The *medic* dataset has inherent MNAR missingness rate of 82%. This contrasts sharply with the other datasets, which are either complete or nearly complete ($\leq 1\%$ missing values). The dataset already embodies the complex, real-world challenge this paper addresses.
- Massive multi-class complexity. The dataset represents a differential diagnosis task with 220 distinct classes. This is an order of magnitude more complex than the next-largest dataset, which has only 7 classes. This results in a very high class entropy (6.51), a complex classification problem.
- High dimensionality: with 275 features, *medic* is a high-dimensional dataset.

Given these properties, the *medic* dataset is not a typical benchmark dataset, but rather the prime motivating case for this study. By analyzing it in isolation (Figures 7–11, we can perform a focused, in-depth evaluation on a high-complexity, high-missingness dataset. This prevents its extreme characteristics from skewing the aggregate results of the other, more standard benchmarks (Figures 2–6).

4 METHODS

The particular problem we are dealing with is modeled on a common two-stage medical scenario [7, 38]. A physician receives initial diagnostic results and must select a single, final batch of follow-up tests. Unlike in many AFA studies, we assume that real-world logistical and biological constraints preclude a sequential, iterative testing process. The AFA task is therefore to leverage the initial results to select the batch of tests that assures the highest likelihood of yielding a definitive diagnosis.

4.1 Problem formulation

Let a complete dataset $\mathcal{D}$ consist of N instances

$$\mathcal{D} = \bigcup_{i=1}^{N} \{(\mathbf{x}_i, y_i)\} \tag{1}$$

where each instance $\mathbf{x}_i$ is a vector of d features from the feature space $X = \{X_1, \ldots, X_d\}$, and y_i is the corresponding class label. In the two-stage AFA scenario, for a given instance $\mathbf{x}$, we observe an initial feature subset $O \subseteq X$, leaving a complementary missing set $M = X - O$. We define the *missingness rate* as $r = |M|/|X|$. The *acquisition set* $A \subseteq M$ has a size $|A| = \lceil \rho \cdot |M| \rceil$, where $\rho \in [0, 1]$ is the acquisition rate. After acquiring A, the final observed set is $O' = O \cup A$. A predictive model $\mathcal{P} : X \to Y$ then predicts the label $\hat{y} = \mathcal{P}(\mathbf{x}_{O'})$ based on this updated observation. The goal of an AFA strategy is to choose the set A that, for given ρ, maximizes the expected performance of the model $\mathcal{P}$.

4.2 Predictive model: XGBoost

The predictive model $\mathcal{P}$, used throughout this study, is an XGBoost (Extreme Gradient Boosting)-trained model [6]. XGBoost is a state-of-the-art tree-ensemble algorithm, widely recognized for high predictive accuracy on tabular data across a broad range of tasks [10, 23, 28, 35] and its ability to outperform even modern deep learning approaches with minimal hyperparameter tuning [35]. Crucially, it has a robust, built-in mechanism for handling missing feature values, as described in Section 4.3. This allows us to directly pass partial feature vectors to the model without requiring an external imputation step, which could otherwise act as an additional, possibly confounding, variable.

4.3 XGBoost and missing values

Both the predictive model and the explanation method must be compatible with the presence of missing data.

XGBoost combines strong generalization, efficient handling of missing data [6], and parallelized tree construction with flexible objectives and evaluation metrics, making it a popular tool for real-world applications and imperfect data. During the construction of each decision tree, when considering a split on a feature, the algorithm evaluates two potential gain scores: one for sending all instances with a missing value for that feature to the left child node, and one for sending them to the right. It then learns a *default direction* for missing values at that split by choosing the direction that yields a higher gain. This allows the model to learn the optimal path for incomplete data based on patterns in the training set. No prior imputation step is required.

4.4 Shapley explanations – SHAP

Explaining a prediction for an instance with missing feature values requires a systematic approach. TreeSHAP [17], an efficient algorithm for calculating SHAP values for tree-based models, accomplishes this by approximating the conditional expectation $E(\mathbf{x}_O)$, where O is the set of observed features. This is achieved through a process of interventional feature perturbation [17], where the effect of missing features is estimated by averaging model predictions on a background dataset (a small random subset of the training data). SHAP value for a feature, whether observed or missing, represents its expected contribution to changing the model output from the base value (the average prediction over the background data) to the prediction for the specific instance. This mechanism provides a strong motivation for using SHAP in AFA. The core task in AFA is to estimate the utility of a missing feature, which can be defined as the expected change in the model's prediction if that feature's true value were revealed. The SHAP value for a missing feature is a direct quantification of this utility.

A loosely related approach was recently proposed in [11], where SHAP values are used to establish the feature ranking for each training instance, which in turn is used to train a decision transformer to predict which feature to acquire next. This is a complex, sequential RL approach, which can be impractical in many medical settings. Our AFA-SHAP method, on the other hand, applies a simpler, part greedy and part embedded strategy that computes SHAP values directly at inference time to support batch acquisition.

4.5 Comparison of rankings for static feature importance metrics across datasets

To compare the static feature importance measures, we analyzed their relative rank differences (Table 2) and Spearman's correlation coefficients (Table 3) across all benchmark datasets (excluding *medic*). We define relative rank difference with respect to the total number of features (Eq. 2); for 10 features, the relative rank difference 0.2 reflect the absolute rank difference 2.

$$\mathrm{rrd}(\mathbf{x}_i, X_j) = \frac{|\mathit{rank}_1(\mathbf{x}_i, X_j) - \mathit{rank}_2(\mathbf{x}_i, X_j)|}{|X|} \quad (2)$$

The comparison also includes a static SHAP value, $\overline{\mathrm{SHAP}}(X_j)$, calculated as the mean SHAP value for feature X_j over internal cross-validation results. There is a significant lack of consensus among the methods. The rankings proved highly dissimilar, with an average relative rank difference of 0.27 (Table 2). The correlation analysis confirmed this, showing only moderate agreement between methods (Table 3). The few statistically significant ($p < 0.05$) correlations – between GainRatio and MDL (0.67) and XGBoost and MDL (0.56) – are too low to imply that the metrics are interchangeable.

This analysis indicates that there is no single, authoritative ground truth for static feature importance ranking. The fact that these methods disagree so strongly on a global feature hierarchy is a strong argument against using a static list for active feature acquisition. Feature importance results (Figure 1) confirmed our hypothesis: the model-specific SFA-XGBoost considerably outperformed all other static methods. It is intrinsically aligned with the predictive model and represents the best-case scenario for a static approach.

4.6 Acquisition Strategies

We evaluate three distinct strategies for selecting the acquisition set A.

4.6.1 Random acquisition: This strategy serves as a naive baseline to establish a lower bound on performance. For an instance with a set of missing features M, the acquisition set $A \subseteq M$ is formed by selecting $k = \lceil \rho \cdot |M| \rceil$ features uniformly at random from M.

$$A_{\mathit{Random}} = \mathrm{random\_sample}(M, k) \quad (3)$$

4.6.2 Static feature acquisition: SFA represents the traditional, instance-agnostic approach. Global feature importance ranking, R, is pre-computed on the training subset. For a given instance with missing features M, the acquisition set A consists of the $k = \lceil \rho \cdot |M| \rceil$ features in M that have the highest rank in R.

$$A_{\mathit{SFA}} = \underset{A \subset M, |A|=k}{\mathrm{argmin}} \sum_{j \in A} \mathrm{rank}_R(X_j) \quad (4)$$

We used each of four static methods (Information Gain Ratio, MDL, ReliefF and XGBoost feature importance) to compute the global feature ranking R. However, due to the clear superiority of SFA-XGBoost (Figure 1) only this method was used in further experiments.

4.6.3 Shapley-based active feature acquisition: In the proposed AFA-SHAP instance-specific strategy, for each partial test instance $\mathbf{x}_O$, with observed features O and missing features M, we use the pre-trained XGBoost

Table 2. Relative mean rank differences across datasets, ranging between 0.16 and 0.34 (mean 0.27 ± 0.05).

	XGBoost	Gain Ratio	MDL	ReliefF	$\overline{\text{SHAP}}$	Random
XGBoost		0.26 ± 0.07	0.19 ± 0.06	0.26 ± 0.07	0.21 ± 0.05	0.33 ± 0.04
Gain Ratio	0.26 ± 0.07		0.16 ± 0.09	0.28 ± 0.10	0.30 ± 0.05	0.34 ± 0.04
MDL	0.19 ± 0.06	0.16 ± 0.09		0.24 ± 0.12	0.24 ± 0.08	0.34 ± 0.04
ReliefF	0.26 ± 0.07	0.28 ± 0.10	0.24 ± 0.12		0.26 ± 0.09	0.32 ± 0.05
$\overline{\text{SHAP}}$	0.21 ± 0.05	0.30 ± 0.05	0.24 ± 0.08	0.26 ± 0.09		0.32 ± 0.03
Random	0.33 ± 0.04	0.34 ± 0.04	0.34 ± 0.04	0.32 ± 0.05	0.32 ± 0.03	

Table 3. Spearman's rank-order correlation coefficients across datasets are predominantly small to moderate (between 0.1 and 0.5), with a mean of 0.25 ± 0.23. Statistically significant correlations ($p < 0.05$) are emphasized.

	XGBoost	Gain Ratio	MDL	ReliefF	$\overline{\text{SHAP}}$	Random
XGBoost		0.31 ± 0.29	**0.56 ± 0.25**	0.32 ± 0.32	0.52 ± 0.18	-0.02 ± 0.16
Gain Ratio	0.31 ± 0.29		**0.67 ± 0.28**	0.17 ± 0.44	0.13 ± 0.23	-0.03 ± 0.24
MDL	**0.56 ± 0.25**	**0.67 ± 0.28**		0.39 ± 0.51	0.40 ± 0.35	-0.02 ± 0.14
ReliefF	0.32 ± 0.32	0.17 ± 0.44	0.39 ± 0.51		0.30 ± 0.38	0.04 ± 0.22
$\overline{\text{SHAP}}$	0.52 ± 0.18	0.13 ± 0.23	0.40 ± 0.35	0.30 ± 0.38		0.01 ± 0.15
Random	-0.02 ± 0.16	-0.03 ± 0.24	-0.02 ± 0.14	0.04 ± 0.22	0.01 ± 0.15	

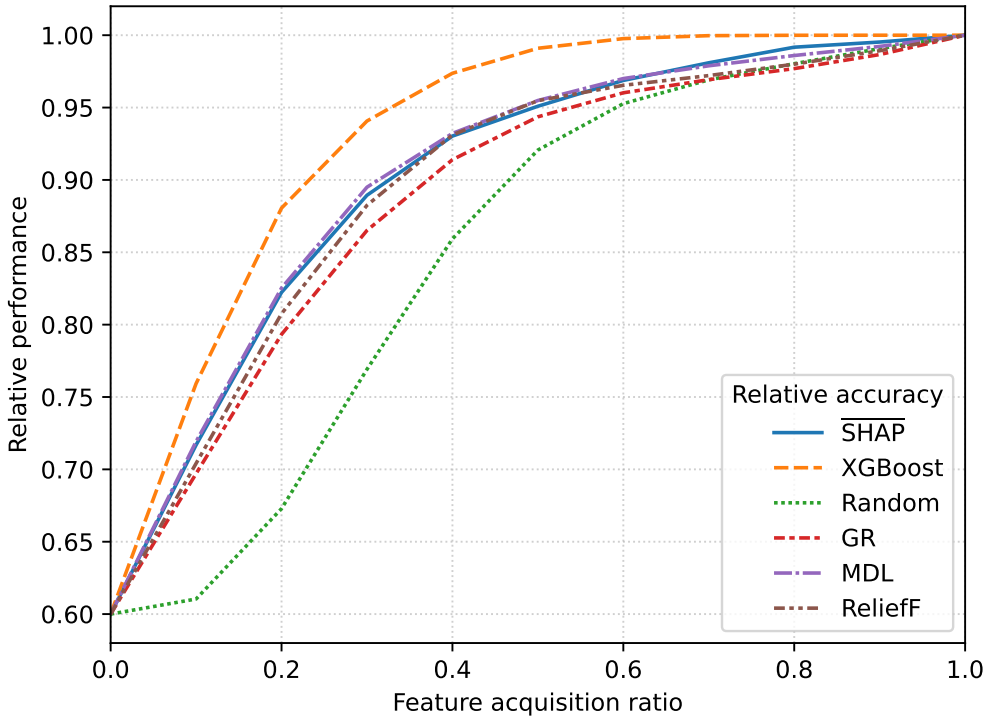


Figure 1. Relative accuracy (relative to the best method using all available features), averaged across all datasets and missingness rates. Unsurprisingly, XGBoost feature importance is superior at all acquisition rates, there is not much difference between other static importance methods, while random selection is consistently (and expectedly) worst.

model $\mathcal{P}$ to predict the outcome y and generate an explanation.

1) The TreeSHAP algorithm is employed to compute the SHAP value, $\phi_j(\mathbf{x}_O)$, for every feature X_j in the full feature space X [17]. This computation is conditional on the observed values $\mathbf{x}_O$ and marginalizes over the missing features in M using a background distribution derived from the training data. Thus we get $\phi_j(\mathbf{x}_O)$ for all features X_j, even for missing ones.

2) The acquisition score for each missing feature $X_j \in M$ is defined as the magnitude of its SHAP value, $|\phi_j(\mathbf{x}_O)|$. A higher magnitude indicates a greater expected influence of feature X_j on the model's prediction for this specific instance, and thus higher rank.
3) The acquisition set A is formed by selecting the $k = \lceil \rho \cdot |M| \rceil$ missing features with the highest absolute SHAP values $|\phi_j(\mathbf{x}_O)|$:

$$A_{\textit{AFA-SHAP}} = \underset{A \subseteq M, |A|=k}{\operatorname{argmax}} \sum_{X_j \in A} |\phi_j(\mathbf{x}_O)| \qquad (5)$$

This strategy is dynamic, as the ranking of missing features is re-computed for every partial observation $\mathbf{x}_O$.

5 Experimental Protocol

In order to ensure a fair and comprehensive comparison of acquisition strategies and prevent any data leakage we established a three-phase experimental protocol.

Phase 1: Model training and parameter tuning

10-fold stratified cross-validation is used to split data in each step into a *current* training set (90%) and a *current* testing set (10%) of instances. The current training set is used exclusively for model preparation: the model $\mathcal{P}$ (used for both predictions and SHAP explanations) is trained, its hyperparameters are optimized via internal tuning, and all static importance rankings are pre-computed.

Phase 2: Simulating missingness

For each instance in the current testing set, an initial state of partial observations is simulated. The mechanism for creating this initial missingness depends upon experimental conditions:

- MCAR: For each test instance, the required fraction r of known feature values is marked as missing, selected uniformly at random.
- MNAR: For each test instance, half of input features are marked as MBOV, the other half as MBUV. The required fraction r of feature values is marked as missing according to each principle.

We simulate missingness rates between 0.10 and 0.90 in 0.10 increments using the *mdatagen* library [18], based on existing feature values, and not including the *natural* dataset missingness (e.g., if a dataset is naturally missing 10% values, for a missingness rate 0.20 we mark additional 20% of known values as missing).

Phase 3: Evaluation loop

The evaluation proceeds instance-by-instance on the current testing set. For each instance, each acquisition rate ρ and $k = \lceil \rho \cdot |M| \rceil$:

1) For an instance $\mathbf{x}$ with true class y, the initial partial observation $\mathbf{x}_O$ is generated according to the specified missingness rate and mechanism (MCAR or MNAR), as described in Phase 2.
2) Each AFA strategy (Random, SFA-XGBoost and AFA-SHAP) is independently applied to create a ranking of all missing features and select an acquisition set A of size k.
3) The true values of the features in A are revealed, creating the updated observed feature vector $\mathbf{x}_{O'}$, where $O' = O \cup A$.
4) The updated feature vector $\mathbf{x}_{O'}$ is passed to the trained XGBoost model $\mathcal{P}$, which generates a new prediction. Any features still missing are handled by the model's internal mechanism.
5) The prediction is compared against the true label to calculate performance.

5.1 Performance Metrics

In order to aggregate and visualize results over diverse datasets, we use *relative* performance metrics (relative to the reference performance with no simulated missing values). This approach standardizes the results by setting a consistent reference for each dataset.

- Establish the reference performance score: for each dataset we first measure the performance *perf*$(\mathcal{P})$ of the model on untouched testing set with no simulated missing values. This value serves as the reference performance score (1.0 or 100%).
- Calculate relative performance scores: for each missingness rate r and each acquisition rate ρ we measure the performance $\textit{perf}_{r,\rho}(\mathcal{P})$ of the model on testing set with simulated missing values and express it as a fraction of the reference performance:

$$\textit{relative\_perf}_{r,\rho}(\mathcal{P}) = \frac{\textit{perf}_{r,\rho}(\mathcal{P})}{\textit{perf}(\mathcal{P})} \qquad (6)$$

E.g., let the untouched testing set accuracy be 0.90 (90%). For missingness rate $r = 0.20$ (20%) and acquisition rate $\rho = 0.10$ (10%) let the accuracy be 0.72 (72%). The relative performance is therefore $0.72/0.90 = 0.80$ (80%). This allows us to plot the relative performance on a standardized chart, and fairly aggregate results with other datasets (which might have different reference performances). All experiments are performed in 10-fold stratified cross-validation setting with internal hyperparameter tuning. The mean and standard deviation of the relative performance metrics (either accuracy or F1-score) are reported or depicted.

5.1.1 Key performance metrics: For comprehensive assessment, we evaluate the model performance *perf*$(\mathcal{P})$ using two established classification metrics: accuracy and F1-score (macro averaged). They were chosen to measure both overall correctness and robustness to class imbalance, which is present in most datasets:

- Accuracy: the proportion of correctly classified instances.

- Macro-averaged F1-score: the unweighted mean of the F1-scores for each class, robust to class imbalance.

5.1.2 Derived lift metric: To quantify effectiveness of the proposed AFA-SHAP strategy against the static benchmark (SFA-XGBoost), we utilize the *lift* metric. Lift is defined as the ratio of the performance of the AFA-SHAP strategy to the performance of the SFA-XGBoost strategy at the same feature acquisition rate (Eq. 7). Let $perf_{r,\rho}$ be the performance metric (accuracy or F1-score) for missingness rate r and acquisition rate ρ of either AFA-SHAP or SFA-XGBoost strategy . The lift is then calculated as:

$$lift_{r,\rho}(\text{AFA-SHAP}) = \frac{perf_{r,\rho}(\text{AFA-SHAP})}{perf_{r,\rho}(\text{SFA-XGBoost})} \quad (7)$$

This metric normalizes performance and provides a clear, interpretable measure of the magnitude of the improvement. A lift value greater than 1.0 indicates that the dynamic AFA-SHAP strategy is more effective than the static SFA-XGBoost benchmark. It is particularly useful for visualizing the relative efficiency gain of AFA-SHAP especially at low acquisition rates where prioritizing the most impactful features is critical.

5.1.3 Derived efficiency metric: An important property of any AFA method is its performance at low acquisition rates, as it represents a test of its efficiency. Important features are often sparse: a small subset of features contributes the vast majority of information, while the remaining are either redundant or irrelevant. An effective AFA strategy must therefore identify such crucial subset quickly. Measuring performance after acquiring only 10% of missing features (e.g., just 5 features out of 50 missing) thus provides a realistic test of ability to prioritize the most impactful features.

To quantify the practical efficiency gain of the AFA-SHAP strategy, we measure the *equivalence acquisition rate*. It is calculated by first recording the performance (accuracy or F1-score) of AFA-SHAP at a fixed low acquisition rate $\rho = 0.10$ and then identifying the acquisition rate (ρ_{equiv}) that the benchmark SFA-XGBoost method requires to achieve the same performance level. The resulting efficiency gain is then expressed as the ratio $\rho_{equiv}/0.10$. For example, an equivalence rate of $\rho = 0.38$ for the SFA-XGBoost, as seen in the *medic* MNAR results (Figure 9 and Table 4), corresponds to a $3.8\times$ efficiency gain for AFA-SHAP, meaning that this strategy achieves the same performance using only $1/3.8$ (26.3%) of the features required by the static SFA-XGBoost benchmark strategy.

5.2 Statistical evaluation of results

Experimental results are statistically compared using the two-sided Wilcoxon signed-rank test and the Common Language Effect Size. The Wilcoxon signed-rank test [39] is a non-parametric method chosen because resulting performance distributions cannot be assumed to be normal. Its p-values allow us to assess statistical significance of results. However, a p-value alone does not quantify the magnitude of their differences. For this purpose we use the Common Language Effect Size (CLES) statistic [19]. CLES is an intuitive, probabilistic measure, closely related to AUC (area under the ROC curve), that reports the probability that a score randomly selected from the one set results will be higher than a score randomly selected from another set of results. This allows us to gauge the practical superiority of a method, rather than just its statistical likelihood.

6 Results

We performed comprehensive experimental evaluation to compare three acquisition strategies: Random acquisition (Random), Static feature acquisition using the XGBoost model-based importance (SFA-XGBoost), and the proposed Shapley-based active feature acquisition (AFA-SHAP). We first compare them on the aggregate benchmark dataset pool, and then perform additional analysis on the high-complexity *medic* dataset, as described in Section 3.1.

6.1 Aggregate performance (all datasets but medic)

We first analyze the mean relative performance for both accuracy and F1-score across the nine benchmark datasets, distinguishing by the missingness mechanism (MCAR or MNAR). Figures 2-5 depict their mean relative performance. Dashed blue lines mark AFA-SHAP performance at $\rho = 0.10$ while dotted orange lines mark SFA-XGBoost performance at $\rho = 0.10$ and its equivalence acquisition rate ρ_{equiv}. Results indicate that the AFA-SHAP and SFA-XGBoost methods perform similarly in the MCAR scenario. AFA-SHAP's slight advantage becomes significantly more pronounced under the MNAR condition. This highlights the robustness of the AFA-SHAP strategy: it clearly outperforms the static benchmark in complex, realistic MNAR settings while maintaining a small performance edge in simpler MCAR scenarios.

6.1.1 MCAR Missingness: Under the MCAR scenario (Figures 2 and 3), the AFA-SHAP (orange) and SFA-XGBoost (blue) strategies exhibit closely matched performance. As expected, both methods considerably outperform the Random baseline (green). At low missingness rates (up to 0.3), AFA-SHAP shows a marginal advantage at low acquisition rates ($\rho \leq 0.10$). This small advantage is confirmed by the lift curves (Figures 6a and 6c), which show a slight peak near $\rho = 0.10$ but otherwise hover near 1.0, indicating parity between methods.

As summarized in Table 4, at an acquisition rate of $\rho = 0.10$, AFA-SHAP achieves a mean relative accuracy of 0.88, which is statistically significantly better ($p < 0.01$) than SFA-XGBoost's 0.85, though the

Table 4. Comparative performance (accuracy and F1-score) of SFA-XGBoost as the best-performing static method, and AFA-SHAP across all missingness rates. We compare performances at acquisition rate 0.10 (Rate AFA-SHAP), and at the *equivalence acquisition rate* (eq. rate) where SFA-XGBoost reaches the same performance. Results are statistically evaluated using the two-sided Wilcoxon signed-rank test (p-value) and the CLES effect size.

missing-ness	data-set(s)	Performance metric		Rate AFA-SHAP	Metric AFA-SHAP	Metric SFA-XGBoost	Eq. rate SFA-XGBoost	p-value (metric)	p-value (eq. rate)	CLES (metric)	CLES (eq. rate)
MCAR	all (no medic)	accuracy	mean	0.10	0.88	0.85	0.14	$< \mathbf{0.01}$	$< \mathbf{0.01}$	small	**medium**
			st. dev.		0.09	0.10	0.02				
		F1-score	mean	0.10	0.88	0.85	0.14	$< \mathbf{0.01}$	$< \mathbf{0.01}$	small	**medium**
			st. dev.		0.09	0.10	0.02				
MNAR	all (no medic)	accuracy	mean	0.10	0.87	0.78	0.18	$< \mathbf{0.01}$	$< \mathbf{0.01}$	**medium**	**large**
			st. dev.		0.11	0.16	0.02				
		F1-score	mean	0.10	0.86	0.77	0.17	$< \mathbf{0.01}$	$< \mathbf{0.01}$	**medium**	**large**
			st. dev.		0.12	0.18	0.02				
MCAR	medic	accuracy	mean	0.10	0.95	0.82	0.34	$< \mathbf{0.01}$	$< \mathbf{0.01}$	**large**	**large**
			st. dev.		0.05	0.12	0.05				
		F1-score	mean	0.10	0.94	0.78	0.36	$< \mathbf{0.01}$	$< \mathbf{0.01}$	**large**	**large**
			st. dev.		0.06	0.15	0.05				
MNAR	medic	accuracy	mean	0.10	0.96	0.75	0.38	$< \mathbf{0.01}$	$< \mathbf{0.01}$	**large**	**large**
			st. dev.		0.04	0.09	0.01				
		F1-score	mean	0.10	0.96	0.78	0.37	$< \mathbf{0.01}$	$< \mathbf{0.01}$	**large**	**large**
			st. dev.		0.04	0.09	0.01				

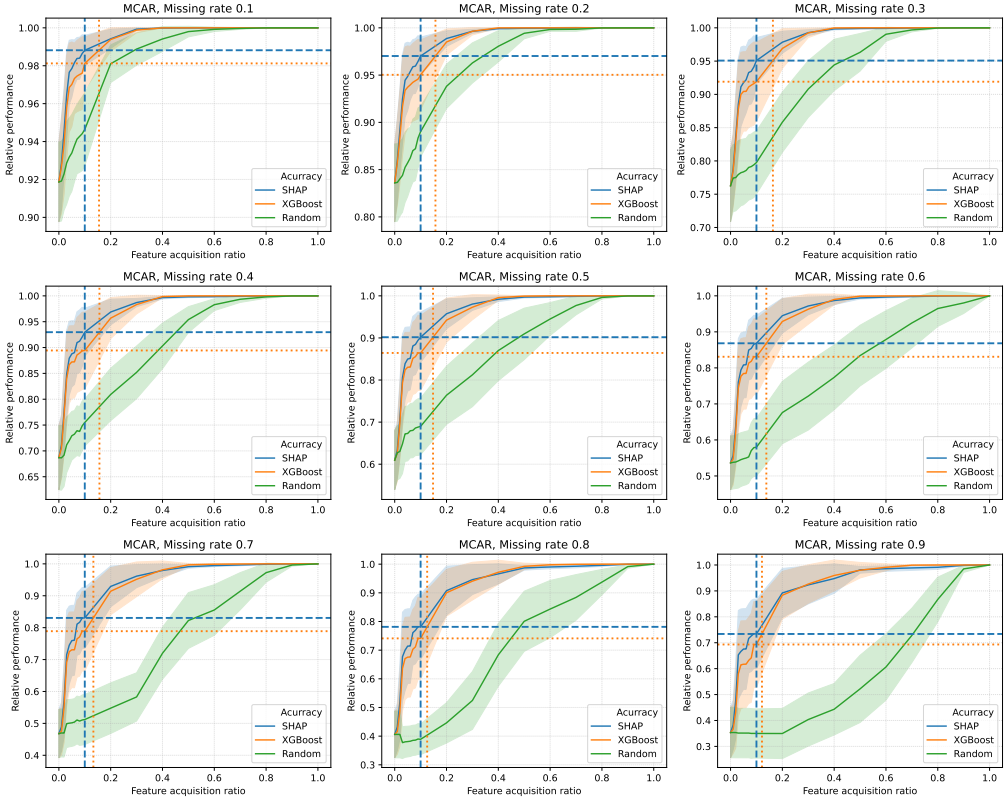


Figure 2. Accuracy results (means with standard deviation bands) on all datasets (excluding medic), MCAR.

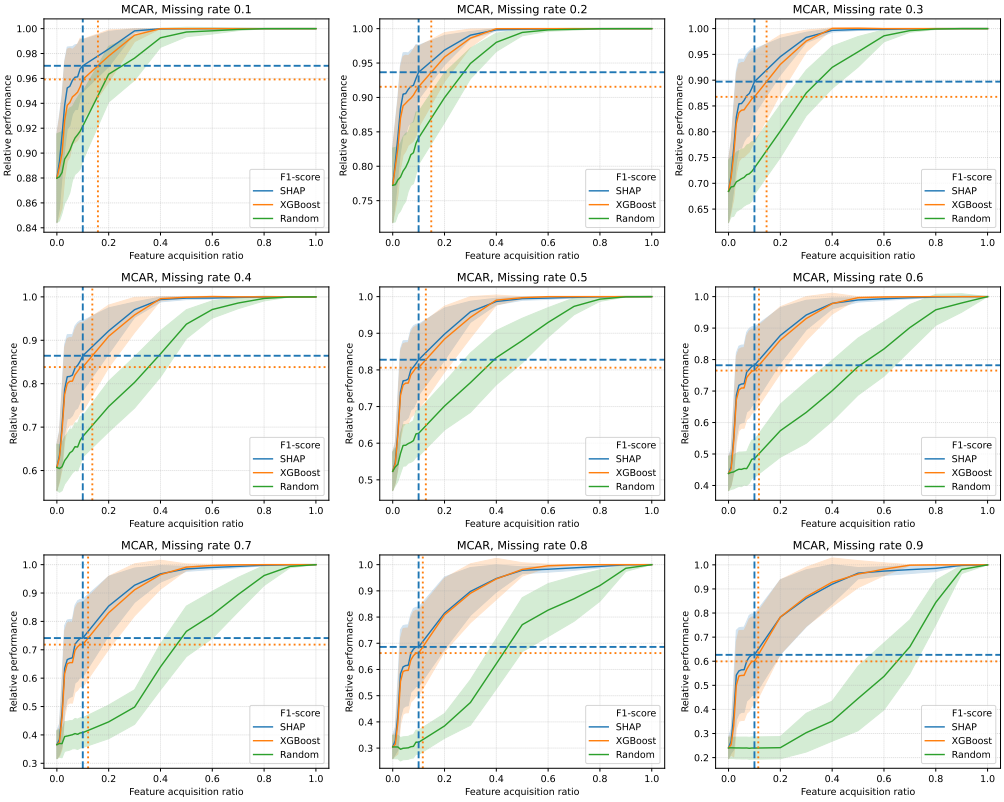


Figure 3. F1-score results (means with standard deviation bands) on all datasets (excluding medic), MCAR.

corresponding effect size is small. The efficiency gain is statistically significant, and with a medium effect size: AFA-SHAP achieves at $\rho = 0.10$ the performance that SFA-XGBoost reaches at $\rho = 0.14$, meaning that SFA-XGBoost needs $1.4\times$ more features for the same accuracy. For F1-scores the conclusions are similar.

6.1.2 MNAR Missingness: The MNAR scenario (Figures 4 and 5) reveals a distinct and consistent advantage for the AFA-SHAP strategy. Across all missingness rates, the AFA-SHAP curve (orange) is clearly superior to the SFA-XGBoost curve (blue). AFA-SHAP's performance curve is also significantly steeper, indicating that it achieves high accuracy with a considerably smaller number of acquired features.

In Table 4 at $\rho = 0.10$ AFA-SHAP achieves mean relative accuracy of 0.87, compared to 0.78 for SFA-XGBoost (the numbers for F1-score are similar). This difference is statistically significant ($p < 0.01$) with a medium effect size. The efficiency gain is noticeable: AFA-SHAP at $\rho = 0.10$ provides the same accuracy as SFA-XGBoost at $\rho = 0.18$, a $1.8\times$ efficiency improvement. The lift curves (Figures 6b and 6d) are similar, showing a sustained lift for AFA-SHAP, peaking at approximately 1.10 (accuracy) and 1.15 (F1-score), a 10% to 15% improvement for $\rho \leq 0.2$. Besides improving accuracy, AFA-SHAP also slightly improves the F1-score (meaning that it is also better in handling imbalanced class distributions).

6.2 Aggregate performance (medic dataset)

The medic dataset, with its high dimensionality (275 features), massive multi-class complexity (220 classes), and high natural missingness (82%), is a realistic test for AFA. Figures 7-10 depict the mean relative performance. Dashed blue lines mark AFA-SHAP performance at $\rho = 0.10$ while dotted orange lines mark SFA-XGBoost performance at $\rho = 0.10$ and its equivalence acquisition rate ρ_{equiv}. On all figures, the gap between AFA-SHAP and SFA-XGBoost is much more pronounced.

6.2.1 MCAR Missingness: In the MCAR scenario (Figures 7 and 8), AFA-SHAP demonstrates a clear advantage over SFA-XGBoost, particularly as the missingness rate increases. While visibly better at 10% missingness, AFA-SHAP is dominantly superior from 40% missingness onward.

In Table 4 at $\rho = 0.10$, AFA-SHAP relative accuracy is 0.95 versus SFA-XGBoost's 0.82. This difference is statistically significant ($p < 0.01$) and exhibits a large effect size. The efficiency gain is substantial: AFA-SHAP at the acquisition rate of $\rho = 0.10$ achieves the same performance as SFA-XGBoost at $\rho = 0.34$ (SFA-XGBoost needs $3.4\times$ more features). The lift plots (Figure 11a and 11c) show a peak lift of 1.1-1.2, indicating a 10-

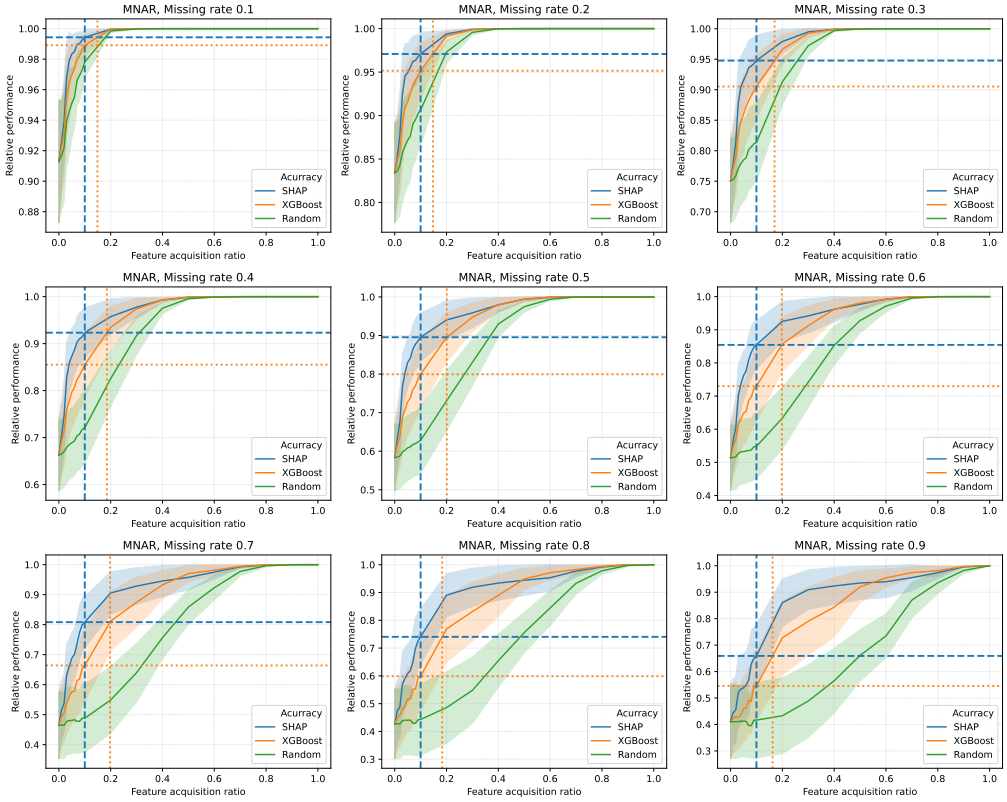


Figure 4. Accuracy results (means with standard deviation bands) on all datasets (excluding medic), MNAR.

20% performance gain at low acquisition rates (for both accuracy and F1-score).

6.2.2 MNAR Missingness: The MNAR scenario on the medic dataset (Figures 9 and 10) provides the most convincing results. At all missingness rates, AFA-SHAP (orange) achieves near-maximal accuracy (0.95-0.98) within the first 10-20% of features acquired. SFA-XGBoost (blue) lags significantly, requiring 3-4$\times$ as many features for the same performance level. E.g., at the $r = 0.50$ missingness rate (Figure 9, middle panel), AFA-SHAP at $\rho = 0.10$ achieves relative accuracy 0.96, while SFA-XGBoost is only at 0.75, and reaches the same performance at $\rho_{equiv} = 0.40$ acquisition rate.

Table 4 shows that AFA-SHAP performs better than SFA-XGBoost with significant improvements and large effect sizes. The equivalence acquisition rate for SFA-XGBoost is $\rho_{equiv} = 0.38$, meaning AFA-SHAP is, on average, $3.8\times$ more efficient. The lift curves in Figures 11b and 11d highlight this improvement, with lift values reaching between 1.6 (accuracy) and 1.9 (F1-score). This indicates a 60-90% performance boost for AFA-SHAP compared to SFA-XGBoost at important low feature acquisition rates. A high macro-averaged F1-score lift suggests that AFA-SHAP not only excels in classification accuracy but also better supports less frequent classes during feature acquisition, particularly at low acquisition rates.

7 Discussion

The experimental results confirm our hypothesis that explanation-guided feature acquisition (AFA-SHAP) is more effective than static, global importance-based acquisition (SFA-XGBoost). This advantage varies based on the context and the mechanism of missing data.

The difference in performance between MCAR and MNAR scenarios sheds light on the effectiveness of the AFA-SHAP. In MCAR settings, missingness is random, independent of feature values meaning there is no hidden context to exploit. Thus, a feature's average global utility (as estimated by SFA-XGBoost) is a very good proxy for its average local utility. The AFA-SHAP strategy, by calculating instance-specific local utility, often arrives at a similar conclusion and closely matched performance (Figures 2 and 3). The slight advantage of AFA-SHAP, especially in the more complex *medic* dataset (Figures 7 and 8), likely stems from its ability to account for feature interactions relative to the observed values ($\mathbf{x}_O$).

In MNAR scenarios, missingness is not random and depends on unobserved values; the pattern of missing data contains information. AFA-SHAP leverages this by conditioning its SHAP explanations on the observed features $\mathbf{x}_O$. This allows AFA-SHAP to infer, e.g, that if features A and B are present, the unobserved feature C is locally significant. In contrast, SFA-XGBoost is static

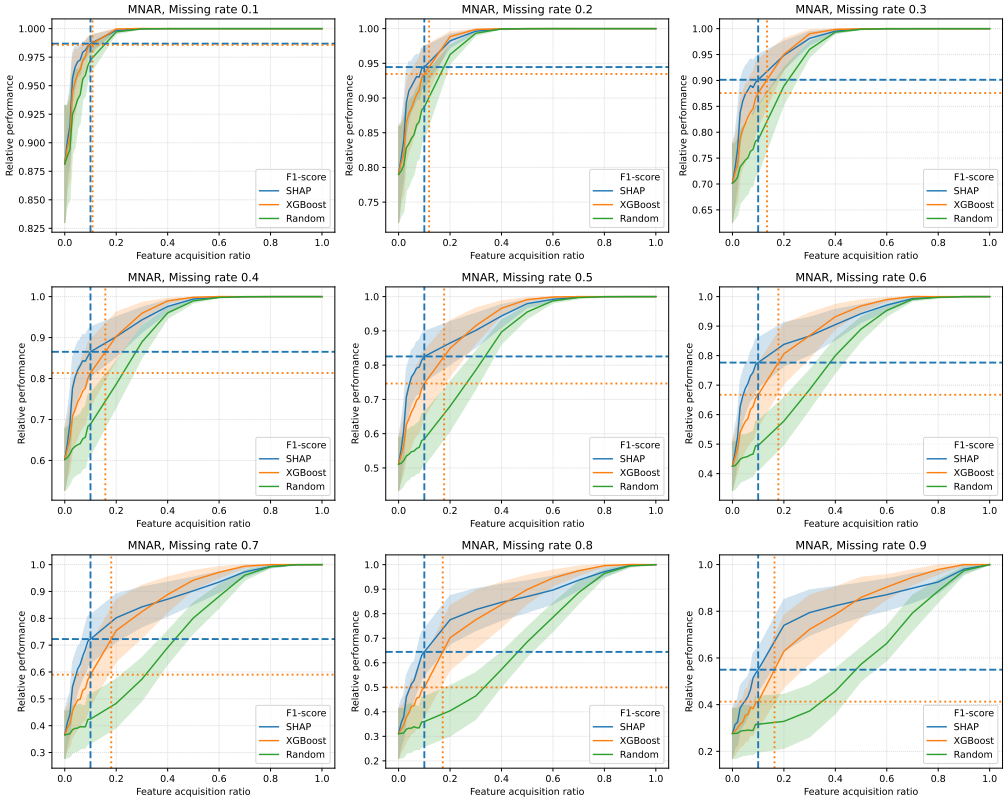


Figure 5. F1-score macro results (means with standard deviation bands) on all datasets (excluding medic), MNAR.

and does not consider this instance-specific context. As a result, it may waste resources acquiring data on globally important features that could be redundant or irrelevant given the specific observed data for that instance.

The *medic* dataset demonstrates this. The 82% natural missingness is MNAR, as it reflects physicians' unobserved judgments – not requesting a test is therefore an informative action. AFA-SHAP leverages the results of the requested tests (the observed features) to model the context and determine the additional set of tests with the highest diagnostic utility. This is why its performance is superior in the MNAR case (Figures 9-11), delivering $3.8\times$ efficiency gains on average.

7.1 Implications

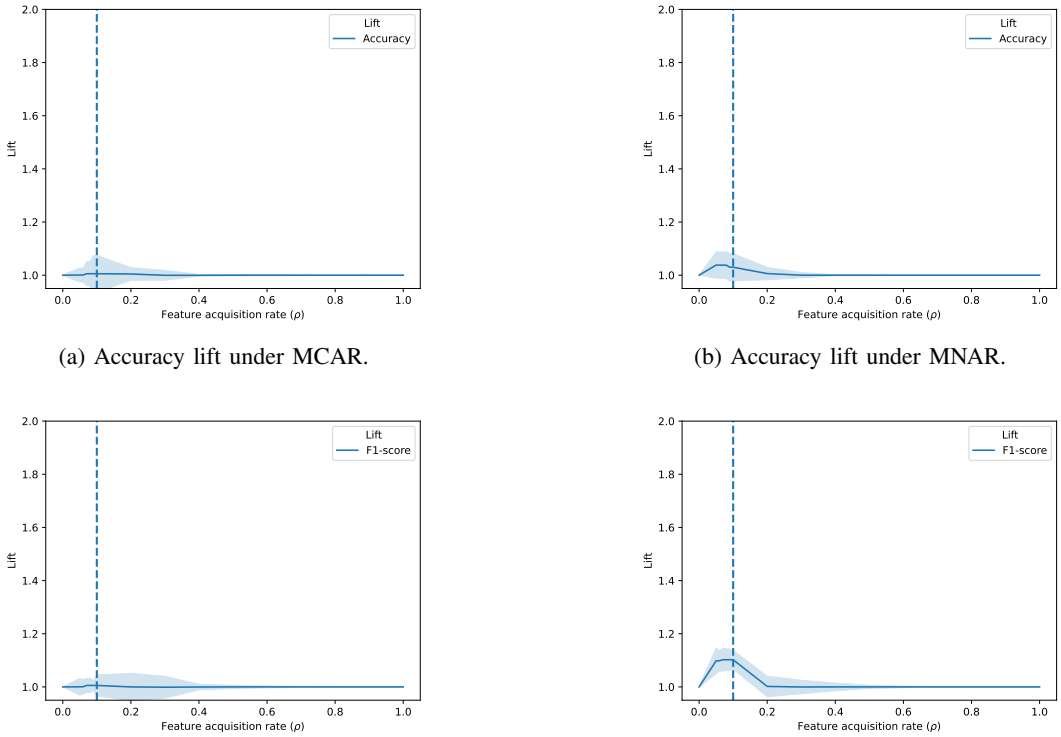
For resource-constrained domains such as medical diagnostics, finance, or industrial monitoring, our findings provide a clear recommendation: AFA systems should not rely on static feature importance lists. The use of an instance-specific strategy like AFA-SHAP offers superior efficiency, especially in complex environments with non-random missingness. This translates to direct savings in cost, time, or risk, by enabling decision-making with fewer, more intelligently acquired features. The AFA-SHAP strategy is a practical method to personalize and optimize data acquisition.

This work also contributes to the growing field of explainable AI (XAI) [2] by showing that explanations can be more than post-hoc interpretations [25]. The SHAP explanation is not the only goal, but an important input to decision-making process (feature acquisition). Explanations are actionable components of a larger, adaptive system, where the model's internal reasoning (via SHAP explanations) is used to guide its future actions (via AFA).

7.2 Cost of the AFA-SHAP strategy

A potential critique of the proposed AFA-SHAP strategy is the computational overhead of generating instance-specific explanations at inference time. We argue this represents a highly favorable cost-benefit trade-off. In resource-constrained domains, this minor computational cost is vastly outweighed by the efficiency gains in data acquisition (e.g., expensive lab tests).

Within the emerging XAI landscape, SHAP explanations [17] are now considered a state-of-the-art, widely adopted method. Most production-level models are already computing SHAP explanations for every prediction as part of their standard operational pipelines and providing them to end-users. The computational cost of ranking the already computed SHAP values to create a per-instance feature acquisition list is marginal. Thus, the AFA-SHAP is almost free in any system where



(a) Accuracy lift under MCAR.

(b) Accuracy lift under MNAR.

(c) F1-score lift under MCAR.

(d) F1-score lift under MNAR.

Figure 6. Accuracy and F1-score lift curves depicting the ratio between AFA-SHAP and SFA-XGBoost (means with standard deviation bands) under MCAR and MNAR missingness mechanisms on all datasets (excluding medic).

model explainability is already a requirement, and in others, it is not difficult to implement.

7.3 Limitations and future work

While our results are promising, this study has several limitations that open avenues for future research:

- Feature acquisition costs: we assumed uniform acquisition cost for all features. In reality, acquiring some features is more expensive than others (e.g., MRI imaging vs. blood pressure measurement). It would be easy to integrate heterogeneous costs by ranking features based on a cost-normalized utility [15], such as $|\phi_j(\mathbf{x}_i)|/cost_j$. Also, some features may need to be acquired in fixed batches (like blood work panels), and this should be considered as well. Determining and optimizing costs depends heavily on the specific application and cannot be readily generalized.
- Generalization to other model types: the AFA-SHAP strategy is based on explanations generated by the TreeSHAP algorithm. TreeSHAP is an efficient implementation of the generalized Shapley explanations [16] for decision trees. AFA-SHAP strategy can be extended to any model, but the AFA process may be less computationally efficient.
- The performance of our method depends on the underlying SHAP explanations. This is an active area of concurrent research, with recent concerns about the instability of SHAP explanations. Future work should investigate how it affects our strategy.
- Batch vs. sequential feature acquisition. Our protocol reflects a two-stage batch acquisition process. A fully sequential AFA, where a single feature is acquired and then the whole process is repeated, might be more optimal but would be significantly more computationally intensive. Also, it is questionable whether in practice we would really be able to acquire features incrementally, one by one.
- Our work aligns with emerging efforts to standardize AFA evaluation [33]. We will validate our AFA-SHAP strategy with the proposed frameworks as these standards reach maturity.

7.4 Conclusion

The paper demonstrates that for the complex, context-dependent task of active feature acquisition, a dynamic, instance-specific strategy is not just beneficial, but essential. The proposed Shapley-based active feature acquisition (AFA-SHAP) method consistently and significantly outperforms the strongest static feature acquisition method (SFA-XGBoost) in realistic and challenging scenarios: those defined by informative missingness

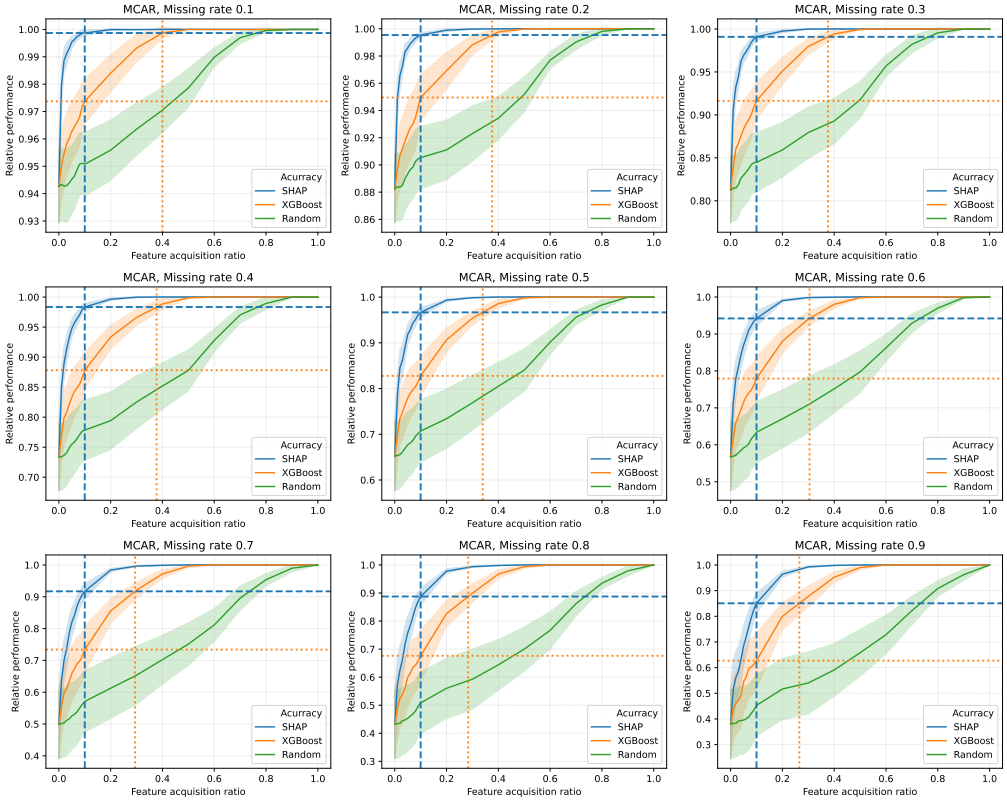


Figure 7. Accuracy results (means with standard deviation bands) on the medic dataset, MCAR missingness.

(MNAR), high problem complexity (the *medic* dataset), and high data missingness.

While static methods provide a reasonable heuristic when data is abundant and missing completely at random, their performance decreases in more challenging real-world settings. AFA-SHAP by leveraging instance-specific explanations, effectively interprets the context provided by both observed features and informative missingness patterns to create an adaptive and highly efficient acquisition strategy. The finding that AFA-SHAP can be 3-4× more effective than SFA-XGBoost on a complex diagnostic task provides a powerful economic and practical argument for its adoption in practice.

ACKNOWLEDGEMENT

The author acknowledges the financial support from the Slovenian Research Agency (research core funding No. P2-209).

DATA AVAILABILITY STATEMENT

The data (as described in Sec. 3) are available from the Zenodo repository at https://doi.org/10.5281/zenodo.18879292.

REFERENCES

[1] C. An, Q. Zhou, and S. Yang. A reinforcement learning guided adaptive cost-sensitive feature acquisition method. *Applied Soft Computing*, 117: 108437, 2022.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[3] B. Becker and R. Kohavi. Adult income dataset, 1996. UCI Machine Learning Repository, doi: 10.24432/C5XW20.

[4] J. Blackard. Forest cover type dataset, 1998. UCI Machine Learning Repository, doi: 10.24432/C50K5N.

[5] M. Buscema, S. Terzi, and W. Tastle. Steel plates faults dataset, 2010. UCI Machine Learning Repository, doi: 10.24432/C5J88N.

[6] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[7] S. Das, N. Ramanan, G. Kunapuli, P. Radivojac,

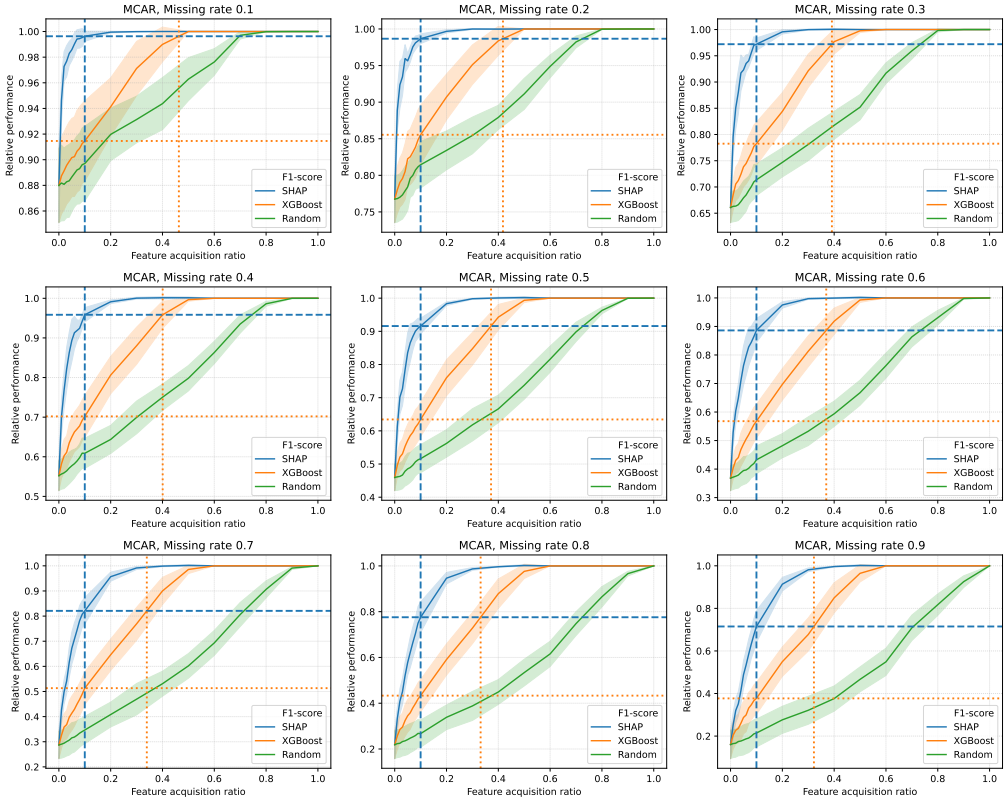


Figure 8. F1-score results (means with standard deviation bands) on the medic dataset, MCAR missingness.

and S. Natarajan. Active feature elicitation: An unified framework. *Frontiers in Artificial Intelligence*, 6, 2023. doi: 10.3389/frai.2023.1029943.

[8] S. Gadgil, I. C. Covert, and S. I. Lee. Estimating conditional mutual information for dynamic feature selection. *International Conference on Learning Representations*, 2024.

[9] U. Garciarena and R. Santana. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Syst. Appl.*, 89:52–65, 2017. ISSN 0957-4174.

[10] L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on tabular data? In *Adv. Neural Inf. Process. Syst.*, volume 35, pages 507–520, 2022.

[11] O. B. Guney, K. S. Saichandran, K. Elzokm, Z. Zhang, and V. B. Kolachalama. Active feature acquisition via explainability-driven ranking. In *Proce. 42nd Int. Conf. Mach. Learn.*, volume 267, pages 20748–20765. PMLR, 2025.

[12] I. Kononenko. On biases in estimating multi-valued attributes. In *Proc. of the 14th Int. Jt. Conf. Artif. Intell.*, volume 2, pages 1034–1040, 1995.

[13] M. Kukar and B. Bračko. Šibko nadzorovano učenje z orodjem Snorkel. In A. Žemva and A. Trost, editors, *Proceedings of the 33rd International Conference ERK'2024*, pages 448–451. Društvo Slovenska sekcija IEEE, 2024.

[14] J. Li and J. B. Oliva. Generative surrogate-guided reinforcement learning for active feature acquisition. In *Proc. International Conference on Machine Learning*, pages 6536–6546. PMLR, 2021.

[15] X. Liu, X. B. Li, and S. Sarkar. Cost-restricted feature selection for data acquisition. *Manage. Sci.*, 69(7):3976–3992, 2023.

[16] S. M. Lundberg and S. I. Lee. A unified approach to interpreting model predictions. In *Adv. Neural Inf. Process. Syst.*, volume 30, 2017.

[17] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

[18] A. D. Mangussi, M. S. Santos, F. L. Lopes, R. C. Pereira, A. C. Lorena, and P. H. Abreu. mdatagen: A python library for the artificial generation of missing data. *Neurocomputing*, 625:129478, 2025. ISSN 0925-2312.

[19] K. O. McGraw and S. P. Wong. A common language effect size statistic. *Psychol. Bull.*, 111 (2):361–365, 1992.

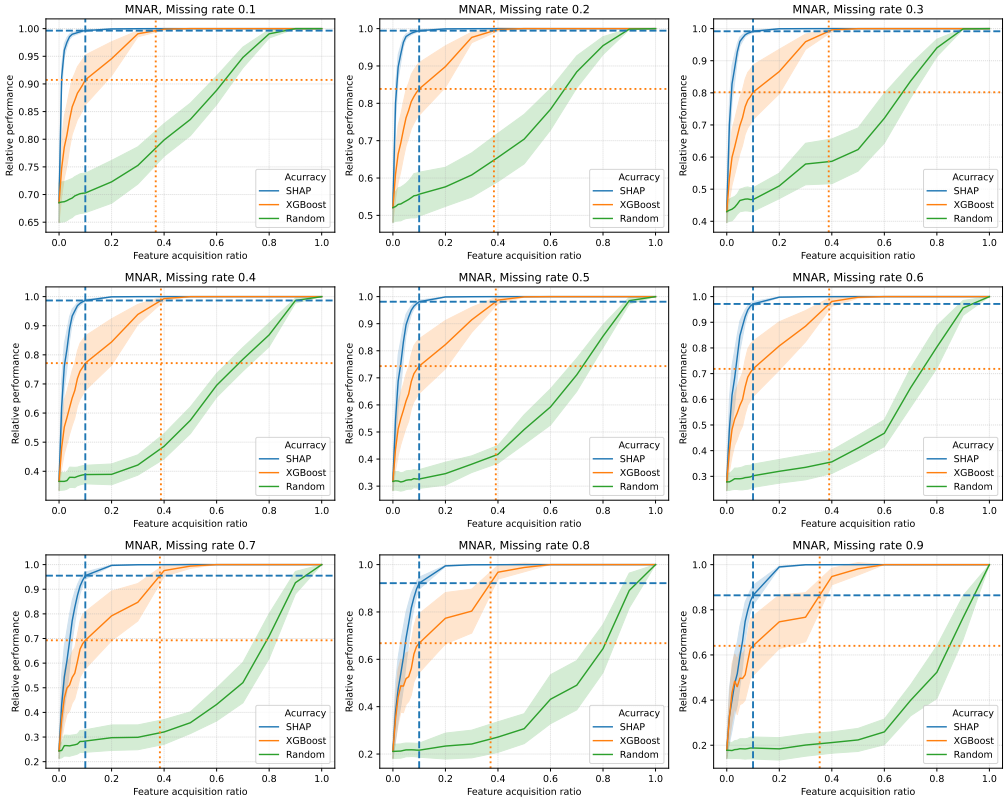


Figure 9. Accuracy results (means with standard deviation bands) on the medic dataset, MNAR missingness.

[20] S. Moro, P. Rita, and P. Cortez. Bank marketing dataset, 2012. UCI Machine Learning Repository, doi: 10.24432/C5K306.

[21] R. C. Pereira, P. H. Abreu, P. P. Rodrigues, and M. A. T. Figueiredo. Imputation of data missing not at random: Artificial generation and benchmark analysis. *Expert Syst. Appl.*, 249:123654, 2024.

[22] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[23] Z. Rafie, M. S. Talab, B. E. Z. Koor, A. Garavand, C. Salehnasab, and M. Ghaderzadeh. Leveraging XGBoost and explainable AI for accurate prediction of type 2 diabetes. *BMC Medical Informatics and Decision Making*, 25(1):403, 2025.

[24] A. Rahbar, L. Aronsson, and M. H. Chehreghani. A survey on active feature acquisition strategies. *arXiv:2502.11067*, 2025.

[25] N. Ramanan, P. Odom, K. Kersting, and S. Natarajan. Active feature acquisition via human interaction in relational domains. In *Proc. 10th ACM IKDD CODS and 28th COMAD*, pages 70–78. ACM, 2023.

[26] J. L. Reyes-Ortiz, D. Anguita, A. Ghio, L. Oneto, and X. Parra. Human activity recognition using smartphones dataset, 2013. UCI Machine Learning Repository, doi: 10.24432/C54S4K.

[27] M. Robnik-Šikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(1):23–69, 2003.

[28] M. Rondinone, S. F. Dal Sasso, H. H. Aung, L. Contillo, G. Dimola, M. Schiattarella, M. Fiorentino, and V. Telesca. Assessing flood and landslide susceptibility using XGBoost: Case study of the Basento river in southern Italy. *Applied Sciences*, 15(10):5290, 2025.

[29] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[30] C. O. Sakar and Y. Kastro. Online shoppers purchasing intention dataset, 2018. UCI Machine Learning Repository, doi: 10.24432/C5F88Q.

[31] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu. Generating synthetic missing data: A review by missing mechanism. *IEEE Access*, 7:11651–11667, 2019.

[32] J. C. Schlimmer. Mushroom dataset, 1987. UCI Machine Learning Repository, doi: 10.24432/C5959T.

[33] V. Schütz, H. Wu, R. Rezvan, L. Aronsson, and M. H. Chehreghani. AFABench: A generic framework for benchmarking active feature acquisition. *arXiv:2508.14734*, 2025.

[34] H. Shim, S. J. Hwang, and E. Yang. Joint active

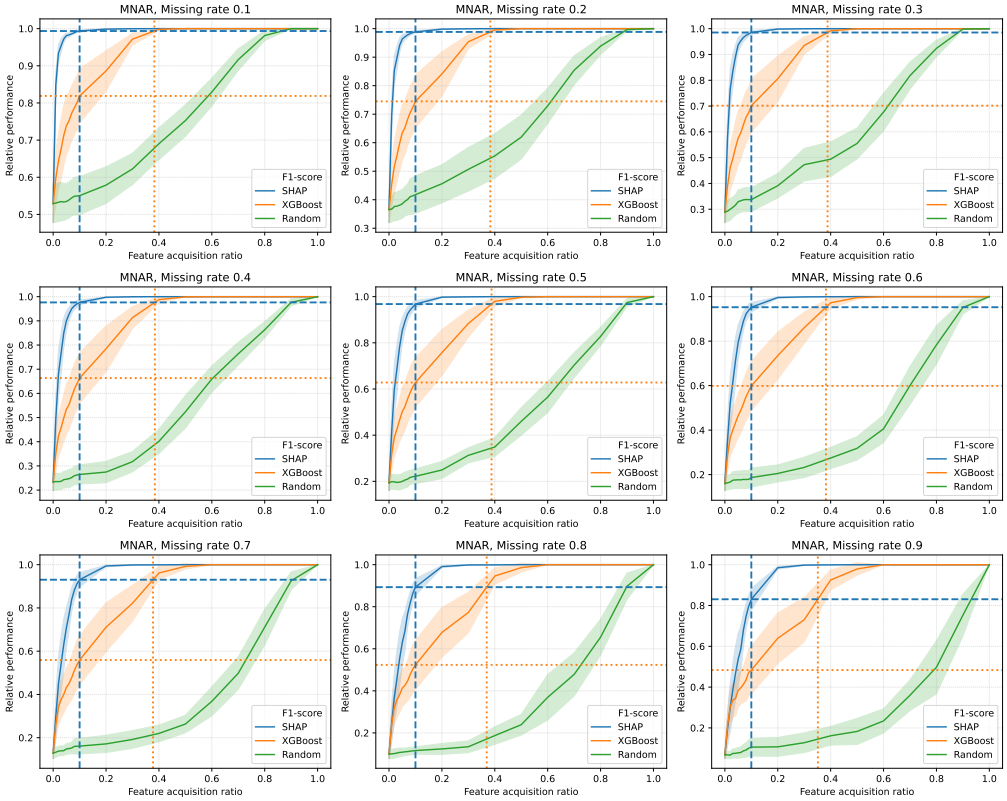


Figure 10. F1-score results (means with standard deviation bands) on the medic dataset, MNAR missingness.

feature acquisition and classification with variable-size set encoding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Adv. Neural Inf. Process. Syst.*, volume 31, page e5841df2, 2018.

[35] R. Shwartz-Ziv and A. Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[36] W. N. Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical Image Processing and Biomedical Visualization*, volume 1905, pages 861–870. SPIE, 1993.

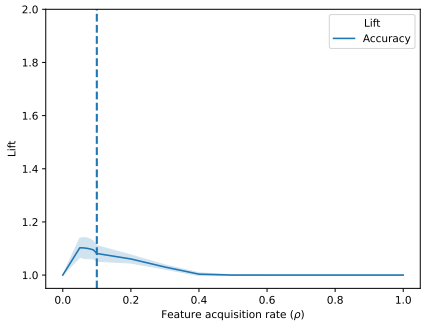
[37] B. Twala. An empirical comparison of techniques for handling incomplete data using decision trees. *Appl. Artif. Intell.*, 23(5):373–405, 2009.

[38] H. von Kleist, A. Zamanian, I. Shpitser, and N. Ahmidi. Evaluation of active feature acquisition methods for time-varying feature settings. *Journal of Machine Learning Research*, 26(60):1–84, 2025.

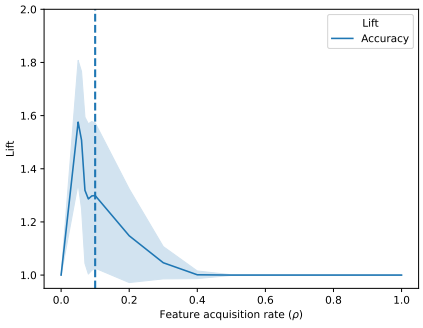
[39] F. Wilcoxon. Individual comparisons by ranking methods. *Biomet. Bull.*, 1(6):80–83, 1945.

[40] I. C. Yeh and C. H. Lien. Default of Credit Card Clients Dataset, 2009. UCI Machine Learning Repository, doi: 10.24432/C55S3H.

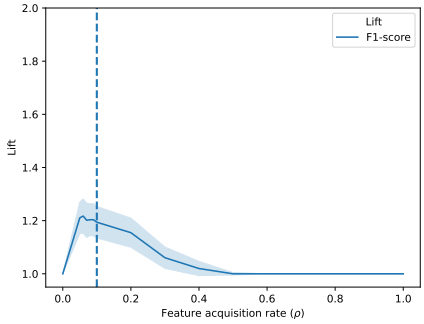
Matjaž Kukar received his Ph.D. degree in Computer and Information Science from the University of Ljubljana, Ljubljana, Slovenia, in 2001. He is an associate professor at the Faculty of Computer and Information Science, University of Ljubljana. His research interests include artificial intelligence, machine learning, medical application and dealing with missing data in machine learning. He has published his works in a range of scientific journals and international conferences and co-authored the book “Machine learning and data mining: Introduction to principles and algorithms”. He serves as an associate editor of the international journal Knowledge and Information Systems. He is a member of Slovenian Artificial Intelligence Society and the the European Association for Artificial Intelligence.



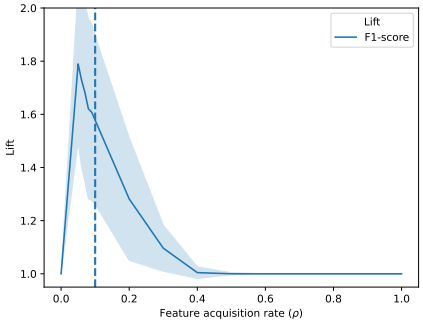
(a) Accuracy lift under MCAR



(b) Accuracy lift under MNAR



(c) F1-score lift under MCAR



(d) F1-score lift under MNAR

Figure 11. Accuracy and F1-score lift curves depicting the ratio between AFA-SHAP and SFA-XGBoost (means with standard deviation bands) under MCAR and MNAR missingness mechanisms on the medic dataset.