Explainability-Guided Active Data Acquisition for Classification

Cristian Pessatti dos Anjos[1]

[1]UFPR, Curitiba, Paraná, Brazil

Abstract

This proposed project investigates how active learning and active feature acquisition can be redesigned when the main scientific objective is to improve the quality of explanations rather than only predictive performance. The central motivation is that, in many real applications, labels and feature values are costly, progressively acquired, and sometimes affected by distributional change, while the explanations derived from trained models are later used to justify consequential decisions. The proposed research develops a methodology in which explanations guide the acquisition process itself. Instead of asking only which instance should be labeled next or which feature should be acquired next, the proposed methodology asks which action is most useful for improving the interpretability, stability, and reliability of the classifier and its explanations under a limited budget. The focus is on classification models in general, while models such as logistic regression and support vector machines are treated as particularly attractive examples because they combine desirable statistical properties, including convex training objectives, with comparatively transparent explanatory structures. The central hypothesis is that explanation-aware acquisition policies can produce models that are more interpretable and more stable than policies based only on uncertainty, entropy, or margin, while maintaining competitive predictive performance under the same acquisition budget. The expected contribution is a statistically grounded methodology for explanation-guided active data acquisition, supported by methodological development, theoretical analysis, and empirical evaluation.

Keywords

Explainable AI, Active Learning, Active Feature Acquisition, Classification, Explanation Stability

1. Context and motivation

Explainable artificial intelligence has become a central research topic because predictive models are increasingly used in high-stakes domains, such as medicine, finance, public administration, and scientific decision support. In these contexts, predictive accuracy alone is not sufficient. Models are also expected to provide explanations that are intelligible, stable, and useful for human oversight. This expectation is particularly important when model outputs influence decisions that require justification, auditing, or contestability.

At the same time, many practical learning systems operate under constrained information. Labels are often expensive because they require expert annotation or delayed verification. Feature values may also be costly because they depend on laboratory tests, imaging procedures, manual inspection, or administrative collection. As a result, the final explanatory behavior of a model is shaped not only by the learning algorithm but also by the way information is selectively acquired during training and deployment.

Despite this, most work in explainable AI assumes that the available data are fixed, while most work in active learning and active feature acquisition optimizes predictive performance and acquisition efficiency without treating explanation quality as a primary objective [5, 7]. This separation is unsatisfactory. If explanations are later used to justify predictions, then the process by which labels and variables are acquired should itself be designed with explanatory quality in mind.

This project is motivated by the idea that explainability should enter the learning pipeline earlier, at the level of data acquisition. Rather than viewing explanation as a post hoc add-on, the project

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil
cristian.anjos@ufpr.br (C. P. d. Anjos)

treats explanation quality as a quantity that can be improved through acquisition decisions. In this setting, the learner should not only ask which example is most uncertain or which feature is most informative for prediction, but also which acquisition action is most useful for producing a classifier whose explanations are more stable, more parsimonious, and more trustworthy.

This perspective is especially relevant when explainability is understood as a statistical problem. Bordt, Raidl, and von Luxburg [1] argue that explainable machine learning should be approached as applied statistics for high-dimensional functions, where explanation methods are interpreted as statistics of the learned predictor rather than as purely visual artifacts. Under this view, explanation quality has to be discussed in terms of interpretation, uncertainty, and sensitivity to the sampled data. This motivates a new question: if explanations are statistics of a fitted model, which data should be acquired to improve those statistics?

The project studies classification models in general, while using logistic regression and support vector machines as reference examples of models with desirable statistical and optimization properties. Logistic regression provides a probabilistic model with interpretable coefficients, odds ratios, and well-established links to statistical inference and calibration. Support vector machines provide a complementary geometric perspective through maximum-margin classification, strong generalization theory, and the possibility of moving from linear to kernel-based non-linear decision boundaries [3]. Both are also attractive because their training problems are convex in standard formulations, making them useful anchor models for a more general methodology.

The main goal of the proposed project is therefore to develop a methodology for *explanation-guided active data acquisition*, in which active learning and active feature acquisition are used not as final goals, but as methodological tools for constructing classifiers with better explanatory properties under constrained budgets.

2. Key related work framing the research

This project lies at the intersection of explainable AI, active learning, active feature acquisition, and statistically grounded interpretable classification.

A first relevant line of work concerns the conceptual foundations of explainable machine learning. Bordt, Raidl, and von Luxburg [1] argue that explainable machine learning should be rethought as applied statistics, emphasizing that explanations are statistics of learned functions and therefore require careful attention to interpretation, intended use, and statistical uncertainty. This perspective is important for the present project because it supports the idea that explanation quality is not merely cosmetic and can be treated as a formal object of optimization.

A second line concerns statistical inference for explanations. Whitehouse, Sawarni, and Syrgkanis [2] recently proposed a semiparametric methodology for inference on global SHAP functionals, showing that commonly used global importance summaries based on SHAP can be studied with asymptotic tools and debiasing techniques. This literature reinforces the point that explanations have sampling variability and motivates data acquisition strategies that explicitly target explanatory stability or uncertainty reduction.

A third line concerns active learning. Tong and Koller [4] established support vector machine active learning through margin-based query selection, showing how pool-based active learning can exploit the geometry of the version space. Settles [5] later systematized the field, organizing active learning strategies around uncertainty, expected error reduction, query-by-committee, and variance reduction. For logistic regression specifically, Yang and Loog [6] provided an extensive benchmark showing that uncertainty-based methods remain surprisingly competitive, while also highlighting the difficulty of consistently outperforming simple baselines. These works are foundational, but their main target is predictive improvement rather than explanation quality.

A fourth line concerns active feature acquisition. In AFA, the learner sequentially decides which variables to observe under a feature-cost budget. Rahbar, Aronsson, and Haghir Chehreghani [7] provide a recent survey of this area, covering sequential decision-making approaches, cost-sensitive formulations, and open methodological challenges. Shim, Hwang, and Yang [8] proposed a joint approach for feature acquisition and classification with variable-size set encoding, showing how prediction and acquisition can be learned together in a cost-sensitive setting. This literature is crucial because it formalizes selective feature observation, but explanation quality is not usually central in the design of acquisition policies.

A fifth line relates explanations directly to acquisition. Guney, Saichandran, Elzokm, Zhang, and Kolachalama [9] proposed an explanation-driven active feature acquisition method that uses instance-specific feature importance rankings to determine the next variable to observe. This work is highly relevant because it demonstrates that explanations can actively guide acquisition instead of being used only after training. However, the current state of the art remains mostly restricted to feature acquisition. It does not yet provide a unified framework in which explanations determine the joint choice between requesting labels and requesting feature values.

A sixth line concerns explainable active learning. Ghai, Liao, Zhang, Bellamy, and Mueller [10] introduced explainable active learning as a human-centered setting in which local explanations serve as interfaces between the learner and the annotator. More recently, Uddin, Wang, Qin, Zhou, and Santosh [11] proposed an explainability-guided active learning method for medical imaging that combines uncertainty with attention misalignment to prioritize informative and clinically meaningful samples. These works show that explanations can influence sample querying, but the broader methodological question of using explanations to guide a joint label-feature acquisition process remains largely open.

A final line concerns concept drift and explanation of drift. Hinder, Vaquet, and Hammer [12] discuss the problem of locating and explaining concept drift, emphasizing that changes in the explanatory structure of a model can be as relevant as changes in predictive performance. This is especially important for the present project because explanation-aware acquisition policies may also help detect or respond to explanatory change over time.

Taken together, the literature suggests three main conclusions. First, explanation quality should be treated as a genuine scientific target. Second, active learning and active feature acquisition already provide mature tools for constrained data collection. Third, the integration of these two perspectives remains incomplete, especially when the goal is to optimize explanatory behavior rather than prediction alone. The proposed project aims to fill this gap.

3. Specific research questions, hypothesis and objectives

3.1. Main research question

How can explanations be used to guide active acquisition of labels and features so that the resulting classifier is not only accurate, but also more stable, interpretable, and reliable under constrained data budgets?

3.2. Specific research questions

RQ1. How should explanation quality be formally defined for classification models in a way that is operationally useful for data acquisition?

RQ2. Can explanations be used to decide whether the next acquisition action should be a label query or a feature query?

RQ3. Under a fixed acquisition budget, do explanation-guided policies yield more stable and interpretable models than uncertainty-based or cost-based baselines?

RQ4. How does the role of explanations differ across classifiers with different statistical and optimization properties, such as probabilistic linear models, margin-based classifiers, and more flexible non-linear models?

RQ5. Can explanation-guided acquisition remain effective under heterogeneous feature costs, missingness, and distribution shift?

3.3. Hypothesis

The central hypothesis of this project is:

> *Acquisition policies that explicitly optimize explanation quality, in addition to predictive utility and acquisition cost, produce classifiers that are more stable, more interpretable, and comparably or more accurate than standard active learning and active feature acquisition methods under the same data budget.*

A secondary hypothesis is that the benefits of explanation-guided acquisition will be especially visible in models whose statistical and optimization structure supports clearer definition of explanatory functionals, such as convex probabilistic or margin-based classifiers.

3.4. Objectives

The general objective is to develop a unified framework for explanation-guided active data acquisition in classification.

The specific objectives are:

O1. Define operational measures of explanation quality for classification models, including stability, consistency, sparsity, and agreement across re-samples or perturbations.

O2. Develop acquisition criteria that combine predictive utility, explanatory utility, and acquisition cost.

O3. Build algorithms for joint active learning and active feature acquisition in which the next action is selected from a unified action space.

O4. Theoretically analyze the properties of the proposed criteria and their relationship with explanation stability and predictive risk.

O5. Empirically compare the proposed methods with strong baselines on synthetic and real tabular datasets.

O6. Characterize the scenarios in which explanation-guided acquisition brings the largest benefits.

4. Research approach, methods, and rationale for testing the hypothesis

4.1. General framework

The proposed framework treats data acquisition as a sequential decision problem. At each iteration, the learner has access to a partially observed labeled set, a partially observed unlabeled pool, a current classifier f, and an explanation operator E. The action space contains at least two families of actions, namely label-query actions and feature-acquisition actions.

The key methodological idea is that action selection should not be based only on predictive uncertainty. Each candidate action a will instead be scored by a composite acquisition criterion:

$$S(a) = \frac{\alpha U_{\text{pred}}(a) + \beta U_{\text{expl}}(a) + \gamma U_{\text{stab}}(a)}{C(a)}, \tag{1}$$

where $U_{\text{pred}}(a)$ is the expected predictive gain, $U_{\text{expl}}(a)$ is the expected explanatory gain, $U_{\text{stab}}(a)$ is the expected gain in explanation stability, $C(a)$ is the acquisition cost, and $\alpha, \beta, \gamma \geq 0$ are tunable weights.

The main research challenge is to define these utilities in a way that is both statistically meaningful and computationally feasible.

4.2. Model classes

The project targets classification models in general. However, particular attention will be given to families whose statistical and optimization structure facilitates methodological analysis. Logistic regression and support vector machines are especially relevant examples because they offer desirable statistical properties and convex optimization problems in standard formulations [3, 6]. These families provide useful starting points for methodological development, while the long-term aim is to build acquisition principles that can transfer more broadly across classifiers.

4.3. Explanation operators

A central methodological step is to define explanation operators appropriate for each model. The project will investigate four complementary families of explanation summaries.

1. **Global importance.** Examples include standardized coefficients, permutation importance, and aggregated SHAP summaries.
2. **Local contribution.** Examples include signed feature contributions for individual instances, local surrogate explanations, and SHAP-based decompositions.
3. **Stability measures.** Examples include rank correlation between repeated explanation vectors, top-k overlap, bootstrap variance of explanation values, and agreement of local explanatory rankings across re-fits.
4. **Parsimony and consistency measures.** Examples include concentration of explanatory mass on a small set of features and agreement with known relevant variables in synthetic data.

The project will initially prioritize stability-oriented measures, because explanation stability is measurable across multiple model families and is directly connected to the trustworthiness of explanations.

4.4. Acquisition criteria

The project will study several explanation-guided acquisition rules.

Criterion A: explanation instability reduction. For a candidate action a, estimate the expected reduction in variability of local or global explanations across bootstrap fits. Actions that most reduce explanation variance will be prioritized.

Criterion B: explanation disagreement resolution. When repeated fits or committees disagree about the important features of an instance, the learner will prioritize actions that are expected to reduce explanatory disagreement.

Criterion C: explanation-aware expected model change. Classical active learning often selects points that are expected to change the fitted model. Here, the analogous idea is to select actions that are expected to improve the explanatory structure of the fitted model.

Criterion D: explanation sparsification. If acquisition leads to more concentrated and interpretable explanation profiles without harming prediction, such actions may be preferred.

4.5. Algorithmic phases

The project will proceed in four phases.

Phase 1. Formalization. Define explanation-quality functionals and derive explanation-guided acquisition scores for selected benchmark classifiers.

Phase 2. Algorithm design. Develop sequential algorithms for joint label and feature acquisition, starting with binary tabular classification and then extending to multiclass settings.

Phase 3. Theoretical analysis. Study the relationship between the proposed acquisition criteria and explanation stability. Whenever feasible, derive analytical results or bounds under simplifying assumptions.

Phase 4. Empirical evaluation. Benchmark the proposed methods against standard active learning, standard active feature acquisition, and explanation-guided feature acquisition baselines.

4.6. Experimental design

Experiments will use three groups of datasets.

1. **Synthetic datasets with known relevant variables.** These will allow direct evaluation of explanatory recovery and stability.
2. **Real tabular datasets with heterogeneous feature costs.** These will test budgeted acquisition in realistic scenarios.
3. **Optional non-stationary datasets.** These will be used to explore whether explanation-guided acquisition is robust to explanatory drift.

Evaluation will include predictive metrics such as accuracy, AUC, log-loss, and calibration error, as well as explanatory metrics such as rank stability, bootstrap variance of local explanations, and top-k overlap. Acquisition efficiency will also be assessed through total label cost, total feature cost, and performance under fixed budget curves.

4.7. Rationale for hypothesis testing

The hypothesis will be tested by comparing the proposed methods against strong baselines under equal budgets. The main baselines will include uncertainty sampling, margin sampling, entropy-based feature acquisition, and recent explanation-guided AFA methods [4, 5, 9]. If explanation-guided joint acquisition systematically improves explanation stability and interpretability while preserving competitive predictive performance, then the central hypothesis will be supported.

The rationale for initially focusing on classifiers with clearer statistical and optimization structure is that they allow a more transparent interpretation of what explanation improvement actually means. This makes them especially appropriate for a project whose main contribution is methodological and conceptual rather than scale-driven.

5. Expected next research steps and expected final contribution to knowledge

The next step of the research will be to implement a first baseline version of the methodology for binary tabular classification. This version will combine standard uncertainty-based acquisition with explanation-stability criteria and will serve as the initial experimental benchmark.

The second step will be to refine the acquisition criteria. Several candidate definitions of explanatory utility will be compared, including reduction in bootstrap variance, increase in top-k explanation agreement, and sparsification of local explanation profiles. This stage will determine which explanatory targets are most useful for acquisition.

The third step will extend the methodology to joint acquisition, where the action space contains both label-query actions and feature-acquisition actions. A major question at this point will be whether a unified acquisition score is sufficient or whether separate decision modules are needed for label and feature actions.

The fourth step will focus on more flexible model classes and multiclass settings. This stage is important because it will test whether the benefits of explanation-guided acquisition remain visible when the explanatory structure is less directly readable than in simple linear models.

The fifth step will be theoretical. The aim will be to derive analytical results connecting acquisition decisions with the variability of explanation summaries. Even partial theory under simplified assumptions would already constitute a meaningful contribution.

A later stage of the project may investigate explanatory drift. In this extension, the same acquisition methodology could be used not only to improve explanatory quality, but also to detect when the explanatory structure of a classifier changes over time [12].

The expected final contribution to knowledge is fivefold.

First, the project is expected to contribute a new conceptual framing in which data acquisition is treated as part of the explainability problem rather than external to it.

Second, it is expected to produce a unified methodology for explanation-guided active acquisition that integrates label querying and feature querying.

Third, it is expected to provide algorithms that remain grounded in statistically interpretable principles, thereby offering transparent alternatives to purely black-box acquisition systems.

Fourth, it is expected to generate empirical evidence about when optimizing explanation quality improves practical model behavior and when purely predictive acquisition remains sufficient.

Fifth, it is expected to contribute to the broader XAI literature by showing that trustworthy explanations depend not only on the explanation method itself, but also on the way information is selectively collected.

In summary, the final research output is expected to demonstrate that explainability can be embedded earlier in the learning pipeline, at the stage of data acquisition, instead of being attached only after model fitting. This may help establish a more statistically principled and operationally useful view of explainable learning under real-world constraints.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT in order to support brainstorming, language refinement, and structural revision. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] Sebastian Bordt, Eric Raidl, and Ulrike von Luxburg. Position: Rethinking explainable machine learning as applied statistics. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 81130–81142. PMLR, 2025.

[2] Justin Whitehouse, Ayush Sawarni, and Vasilis Syrgkanis. Statistical inference and learning for Shapley additive explanations (SHAP). *arXiv preprint arXiv:2602.10532*, 2026.

[3] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[4] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2(Nov):45–66, 2001.

[5] Burr Settles. *Active Learning*. Morgan & Claypool Publishers, 2012.

[6] Yazhou Yang and Marco Loog. A benchmark and comparison of active learning for logistic regression. *Pattern Recognition*, 83:401–415, 2018.

[7] Arman Rahbar, Linus Aronsson, and Morteza Haghir Chehreghani. A survey on active feature acquisition strategies. *arXiv preprint arXiv:2502.11067*, 2025.

[8] Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. In *Advances in Neural Information Processing Systems*, volume 31, pages 1368–1378, 2018.

[9] Osman Berke Guney, Ketan Suhaas Saichandran, Karim Elzokm, Ziming Zhang, and Vijaya B. Kolachalama. Active feature acquisition via explainability-driven ranking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 20748–20765. PMLR, 2025.

[10] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Klaus Mueller. Explainable active learning (XAL): Toward AI explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):235:1–235:28, 2021.

[11] Ifrat Ikhtear Uddin, Longwei Wang, Xiao Qin, Yang Zhou, and K. C. Santosh. Learning to select like humans: Explainable active learning for medical imaging. *arXiv preprint arXiv:2602.13308*, 2026.

[12] Fabian Hinder, Valerie Vaquet, and Barbara Hammer. One or two things we know about concept drift—a survey on monitoring in evolving environments. Part B: locating and explaining concept drift. *Frontiers in Artificial Intelligence*, 7:1330258, 2024.