AI

MDPI

*Article*

# Addressing Data Scarcity and Class Imbalance Using Active Learning and Synthetic Data Generation: Experimentation on Clinical Datasets

**Lorella Bottino [1,2,†,‡], Chiara Zucco [2,†], Allan Tucker [3] and Mario Cannataro [2,*]**

[1] Department of Medicine and Surgery, Kore University of Enna, 94100 Enna, Italy; lorella.bottino@unikore.it
[2] Data Analytics Research Center, Department of Medical and Surgical Sciences, University Magna Graecia of Catanzaro, 88100 Catanzaro, Italy; chiara.zucco@unicz.it
[3] Department of Computer Science, Brunel University of London, London UB8 3PH, UK; allan.tucker@brunel.ac.uk
* Correspondence: cannataro@unicz.it
† These authors contributed equally to this work.
‡ This work was made when Lorella Bottino was PhD Student at the University Magna Graecia of Catanzaro, during her visiting research period at the Brunel University of London.

## Abstract

The rapid advancement of Machine Learning (ML) technologies has revolutionized healthcare, particularly in the domains of disease prediction and diagnosis. Nevertheless, some data-related issues, such as annotated data scarcity, low-quality labels, and class imbalance in the dataset, can compromise the effectiveness of the learning process. This paper first gives an overview of Active Learning (AL), which is a strategy in machine learning that helps models to achieve better performance using fewer labeled examples. It then reviews various strategies for addressing dataset imbalance, in particular focusing on data-level rebalancing approaches. The paper provides a review of existing works on active learning in the presence of class imbalance and on the integration of active learning with synthetic oversampling techniques. The central core of the paper is the implementation of four different strategies for data selection from publicly available clinical datasets: Random, Random + SMOTE, Uncertainty, and Uncertainty + SMOTE. We calculated and plotted balanced accuracy and average precision by iterating over 20 different seed values, using cumulative batches of labeled data, for each combination of dataset and classifier. Moreover we also carried out a statistical test to assess the extent to which the difference between the Random and Uncertainty distributions, and that between the Uncertainty and Uncertainty + SMOTE distributions, was statistically significant. Experiments show that uncertainty-based strategy can improve predictive performance compared with random acquisition, particularly with Random Forest, although the observed benefits vary across datasets, classifiers, performance metrics, and annotation budgets. The additional contribution of SMOTE is also context dependent and does not yield statistically significant improvements over uncertainty-based acquisition alone. For the WBCD dataset, it can be seen that Uncertainty + SMOTE is comparable to, or in some cases below, the uncertainty-only strategy. Metrics such as precision, recall, specificity, F1-score, AUROC and AUPRC are evaluated and provided.

**Keywords:** machine learning; active learning; synthetic data generation; data scarcity; class imbalance; uncertainty sampling; data sampling; data labeling

Check for updates

Academic Editor: Giovanni Diraco

Received: 19 August 2026
Revised: 3 November 2026
Accepted: 6 November 2026
Published: 13 November 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Modern supervised machine learning (ML) models have achieved remarkable success in healthcare; however, their performance remains fundamentally constrained by the availability of labeled training data [1].

In real-world applications, obtaining high-quality labels is often prohibitively expensive, time-consuming, and sometimes infeasible, especially in specialized domains where expert annotations are required. For example, in medical imaging and remote sensing, labeled samples often constitute less than 1% of the available data [2]. Consequently, practitioners are often faced with massive pools of data, only a small fraction of which is labeled.

To mitigate this challenge, Active Learning (AL) trains a baseline model using only the small pool of data that is already labeled and, rather than requesting labels for the rest of the data, actively identifies the most informative instances for labeling, by ranking them according to an acquisition score (i.e. uncertainty score) [3–5]. Among various active learning strategies, uncertainty sampling is widely adopted due to its intuitive nature and computational efficiency.

This study investigates how to find the best-performing classifier while labeling as few data as possible. In particular, we simulate the active learning process using publicly available clinical tabular datasets from the UCI Machine Learning Repository and evaluate the labeling budget required to achieve a desired level of predictive performance.

Although active learning reduces the number of instances that require manual annotation, experts still need to label the selected samples. In contrast, synthetic data generation can increase the size of the training set without requiring additional manual labeling once the generative model has been trained [6]. For this reason, this study investigates the use of synthetic data generation and evaluates the effectiveness of combining synthetically generated samples with those selected through active learning (i.e., uncertainty sampling) by assessing their impact on predictive performance across the same clinical datasets.

In this work, we formulate the following key research questions (RQs) and seek to answer them.

(1) RQ1: Can a predictive model utilizing an active learning approach achieve comparable or superior performance to a model trained on a larger, randomly labeled dataset while minimizing the need for a large amount of expertly labeled training data?
More specifically, the first research question addressed in this work is whether iteratively adding the most uncertain samples selected from the unlabeled data pool leads to faster improvements in model performance than adding randomly selected samples, while requiring substantially less labeled training data.

(2) RQ2: Can the addition of synthetic data improve the performance of a predictive model trained using uncertainty sampling compared to uncertainty sampling alone?
More specifically, this second research question investigates whether augmenting the training set with synthetic data can further enhance the performance of a model trained using uncertainty sampling, one of the most widely adopted active learning strategies.
In other words, this question aims to determine whether the generation of synthetic data could be a complement to active learning and further reduce the need to obtain new labels from an expert.

The novelty of this study lies in quantifying the benefit of active learning over random sample acquisition in a clinical machine-learning setting, while further investigating whether SMOTE provides an additional benefit beyond intelligent sample selection and how this benefit changes as the available labeling budget increases.

In particular, the contribution of the present manuscript is twofold:

- A first research contribution of our work is an experimental pipeline that allows us to quantify the benefits of active learning over random sample acquisition on typical biomedical datasets, with potential follow-up in a clinical setting.
- A second research contribution regards investigating whether SMOTE provides an additional benefit beyond intelligent sample selection and how this benefit changes as the available labeling budget increases.

The rest of the paper is organized as follows: the theoretical background relevant to this work is presented in Section 2. Section 3 details the analysis workflow and methodologies employed. Section 4 presents the results of the analysis and Section 5 concludes the paper underlying potential directions for future research.

## 2. Related Works

Several studies have investigated the application of AL techniques to the fine-tuning of models. Relatively less attention has been devoted to examining how uncertainty-based sample selection interacts with synthetic oversampling when only a very small number of labeled clinical examples are initially available. Our study addresses this gap by evaluating the interaction between uncertainty-based sample acquisition and SMOTE-based data augmentation under limited-label conditions.

### *2.1. Active Learning*

The development of robust and reliable supervised machine learning models typically requires sufficiently large labeled datasets. However, obtaining reliable labels can be costly and time-consuming, particularly in clinical settings, where annotation often requires specialized expertise. Active Learning (AL) addresses this constraint by iteratively selecting unlabeled observations for which annotation is expected to be more informative, querying an oracle for their labels, and incorporating the newly annotated instances into the training set [3,7].

AL paradigms are commonly distinguished according to the mechanism through which unlabeled instances are made available for selection. In pool-based AL, which is the setting considered in this study, a predefined pool of unlabeled observations is available, from which a subset is selected for annotation. In stream-based settings, observations are processed sequentially as they become available, and a querying decision is made for each instance [3,7]. A further distinction concerns the criterion adopted to quantify the usefulness of candidate observations. Informativeness-based approaches assign priority to samples associated with high predictive uncertainty or with regions close to the current decision boundary. Representativeness-based approaches emphasize coverage of the underlying data distribution, while hybrid criteria combine different aspects of sample informativeness and representativeness [3].

Among informativeness-based approaches, uncertainty sampling represents one of the most widely adopted strategies because of its simple formulation and relatively limited computational requirements [7,8]. Under uncertainty sampling, higher acquisition priority is assigned to observations for which predictive confidence is lower. In binary classification, such observations are frequently located close to the current decision boundary, where similar posterior probabilities may be assigned to the two classes. Under least-confidence sampling, adopted in this study, uncertainty is quantified from the highest predicted class probability, and observations associated with lower maximum probability receive higher acquisition priority [8,9].

The relative effectiveness of uncertainty-based strategies may vary across datasets, classifiers, and acquisition settings. Nevertheless, uncertainty sampling provides a well-

established reference strategy for investigating whether informative acquisition can attain a given level of predictive performance using fewer labeled observations [7–9].

The effectiveness of an AL strategy may also depend on the class distribution within the labeled set and the unlabeled pool. When one class is underrepresented, the composition of the observations selected for annotation contributes to determining the information subsequently available for model estimation. At the same time, the existing class distribution may affect predictive uncertainty and, consequently, the acquisition scores assigned to unlabeled observations [10–14]. These aspects become particularly relevant in classification problems characterized by substantial class imbalance.

### 2.2. Class Imbalance and Data-Level Rebalancing

Class imbalance occurs when observations are unevenly distributed across the classes of a classification problem. Under these conditions, model estimation may be influenced predominantly by observations belonging to the majority class, with a consequent reduction in sensitivity and predictive performance for minority-class observations [14,15]. This issue assumes particular relevance when the minority class represents an outcome of specific clinical interest.

Several methodological families have been proposed for class-imbalance mitigation. Data-level strategies modify the empirical training distribution through undersampling, oversampling, or combinations of these procedures. Algorithm-level strategies modify the learning procedure through class-dependent costs, sample weighting, or class-specific loss functions. In the context of AL, additional categories include probability-based, weight-based, loss-based, and optimization-based strategies [14]. Class imbalance can therefore be addressed at different stages of the learning process, including data construction, model estimation, and sample acquisition.

Among data-level approaches, the Synthetic Minority Over-sampling Technique (SMOTE) is one of the most established methods [15]. In SMOTE, synthetic observations are obtained through interpolation between minority-class instances and their nearest minority-class neighbors. An increase in minority-class representation can therefore be obtained without direct replication of existing observations. The characteristics of the resulting synthetic data depend on the minority observations available for interpolation and on their distribution within the feature space. Theoretical analyses have shown that the distribution induced by SMOTE may differ from the underlying minority-class distribution [16]. Particular caution is therefore warranted when only a limited number of minority-class observations is available.

Class imbalance and limited annotation are closely related in an AL setting. The class composition of the labeled set is progressively determined by the observations selected at successive acquisition rounds. The class distribution available at a given stage may in turn influence predictive uncertainty and subsequent acquisition scores. This interaction has motivated the development of AL approaches in which information about class representation is explicitly incorporated into the acquisition or learning procedure [10,11,13,14].

### 2.3. Active Learning Under Class Imbalance

The relationship between AL and class imbalance was explicitly investigated by Ertekin et al. [10]. Their study examined active selection under imbalanced classification conditions and reported that the acquisition of observations close to the decision boundary can contribute to reducing labeling requirements while improving the representation of informative minority-class observations. The study provided early evidence that the behavior of AL under imbalance requires dedicated evaluation.

The same problem has also been considered in biomedical applications. Doyle et al. [17] proposed an active-learning classification strategy for the minority-class problem in histopathology annotation. Informative sample acquisition was combined with explicit consideration of minority-class representation, providing an example of the joint treatment of annotation scarcity and class imbalance in a biomedical setting.

More recent studies have incorporated class-distribution information directly into the acquisition procedure. Bengar et al. [11] proposed a class-balanced active-learning framework for image classification in which batch acquisition accounts for both informativeness and expected class composition. Jin et al. [12] introduced Balanced Active Learning (BAL), in which estimated probabilities of minority-class membership are incorporated into acquisition decisions. In both cases, class representation is considered during sample selection, thereby affecting the composition of the labeled set as annotation progresses.

A weighting-based formulation was proposed by Fairstein et al. [13]. In their approach, conventional acquisition scores are adjusted through inverse-propensity weighting based on estimated class-membership probabilities. This formulation is designed to improve the representation of minority observations within a fixed annotation budget and can be combined with different score-based AL criteria.

Further developments have considered decision-boundary structure explicitly. DIRECT [18], for example, is based on the identification of class-separation regions and the acquisition of observations located close to those regions under conditions of class imbalance and label noise.

A recent review by Kwon and Kang [14] organizes imbalance-aware AL methods into data-level, probability-based, weight-based, loss-based, and optimization-based categories. This classification reflects the diversity of mechanisms through which class information can be incorporated into AL, including modifications of acquisition probabilities, weighting schemes, objective functions, and training-set composition.

Clinical applications have also been considered. Gabriel et al. [19] evaluated AL for lung-disease severity classification from chest radiographs under class imbalance. Several acquisition criteria, including least confidence, entropy, and margin sampling, were evaluated together with a weighted loss. Repeated random initializations and imbalance-sensitive performance measures, including macro F1, AUROC, and AUPRC, were included in the experimental design. These results further support the relevance of jointly considering annotation efficiency and class imbalance in clinical classification.

Overall, existing approaches address class imbalance at different stages of the AL workflow. Class information may be incorporated into acquisition probabilities, weighting mechanisms, training objectives, or the composition of the labeled set. Data-level rebalancing performed during iterative acquisition provides an additional formulation in which class-imbalance mitigation is integrated directly into the evolution of the training data.

### *2.4. Combining Active Learning and Synthetic Oversampling*

Synthetic oversampling has also been incorporated into AL-based classification frameworks. Mi [20] considered the combination of Active Learning and SMOTE for imbalanced classification and evaluated the resulting performance through measures including F-measure, G-mean, and AUC.

Deng et al. [21] proposed AL-SVMSMOTE-DBN for imbalanced multiclass classification. The framework combines an SVM-based SMOTE procedure with an AL criterion based on uncertainty and representativeness. Synthetic oversampling is included within the construction of a more balanced training representation, while informative observations are subsequently identified through the AL component.

In the medical domain, Sena and Ben Hamida [22] proposed ACTIVE SMOTE for imbalanced medical data classification. Uncertainty and diversity measures are used to identify minority-class observations suitable for guiding synthetic sample generation. The method was evaluated on datasets related to diabetes and breast-cancer diagnosis, with AL-related information incorporated directly into the oversampling procedure.

Islam et al. [23] investigated the automated classification of patient-safety event reports using a pool-based AL setting. Several acquisition criteria were considered, including least confidence, margin, entropy, and a combined informativeness–representativeness criterion, together with random sampling. At each acquisition round, newly labeled observations were incorporated into the training set, followed by SMOTE-based rebalancing and classifier retraining. The evaluation therefore included iterative acquisition of real observations and subsequent synthetic rebalancing within the same clinical classification workflow.

More recently, Uncertainty Guided Oversampling (UGO) was proposed by Sağlam and Cengiz [24]. In this formulation, classifier-specific epistemic and aleatoric uncertainty estimates are incorporated into the synthetic-generation procedure. Uncertainty information is therefore used directly to determine the regions in which additional synthetic observations are introduced.

These studies illustrate several possible forms of integration between active acquisition and synthetic data generation. Differences concern the role assigned to uncertainty, the stage at which synthetic observations are introduced, and the contribution of synthetic data to subsequent model estimation and sample acquisition.

A related investigation was conducted by Moles et al. [25] on imbalanced tabular classification. Random oversampling, G-SMOTE, and CTGAN-based augmentation were evaluated together with entropy-based AL across three multiclass datasets. One of the experimental scenarios considered iterative augmentation following the acquisition of real observations, with subsequent model estimation based on both real and synthetic data. The analysis included minority-class performance and the amount of real data required to attain reference predictive performance.

The present study focuses on binary clinical classification and on the interaction between a conventional uncertainty-based acquisition criterion and SMOTE-based rebalancing. Standard SMOTE is adopted as the data-level rebalancing procedure, while least-confidence sampling is used for informative acquisition. Random and uncertainty-based acquisition are evaluated both with and without SMOTE using Support Vector Machine and Random Forest classifiers. The complete experimental procedure is repeated across matched data partitions and initial labeled sets, and performance is assessed at successive annotation budgets and over the complete learning trajectory. Paired statistical comparisons are used to evaluate the consistency of the observed effects across repetitions. This experimental design allows the joint influence of acquisition strategy, synthetic rebalancing, classifier choice, and annotation budget to be examined under limited-label clinical conditions.

## 3. Materials and Methods

The experimental procedure adopted in this study consists of a series of phases, including data preprocessing, model development, active learning, synthetic data generation, and performance evaluation.

First, the predictive models were trained on a small labeled training set. Next, the training set was iteratively expanded by adding newly labeled instances selected using two different sampling strategies, namely uncertainty sampling and random sampling, and the resulting model performance was evaluated. Finally, the contribution of synthetic data

generation was investigated by assessing its impact on the performance of the predictive models in combination with the active learning framework.

The entire workflow, from data selection to model evaluation, is described in detail in the following subsections, together with the corresponding pseudocode algorithms, which provide a clear representation of the experimental procedure for each phase.

### 3.1. Experimental Setting and Data Partitioning

The experimental evaluation was conducted on two publicly available clinical datasets obtained from the University of California, Irvine (UCI) Machine Learning Repository: the Heart Failure Clinical Records (HFCR) dataset and the Breast Cancer Wisconsin (Diagnostic) (WBCD) dataset. The principal dataset characteristics and the corresponding experimental partitions are summarized jointly in Table 1.

The HFCR dataset (Available at https://doi.org/10.24432/C5Z89R) contains the medical history records of 299 patients with heart problems. The death event is the target variable that binary classification models attempt to predict; the target is encoded as 1 in the case of deceased patient and 0 in the case of alive patient. The dataset contains no null values. Regarding the dataset imbalance, the number of survivors (death event = 0) are 203, while the number of deaths (death event = 1) are 96.

The WBCD dataset (Available at https://doi.org/10.24432/C5DW2B) contains 30 tumor characteristics from 569 patients, together with an ID attribute and a diagnosis label. The features are computed by analyzing the cell nuclei characteristics of digitalized images obtained by Fine Needle Aspiration (FNA) of a breast mass. The target feature is diagnosis which can assume two types of values: benign (B) or malignant (M). The dataset contains 212 malignant and 357 benign samples. The dataset contains no null values.

For each dataset, the observations allocated to model development were divided into an initial labeled set and an unlabeled pool used for subsequent acquisition. A separate test set was retained exclusively for performance evaluation. For HFCR, 19 observations were initially labeled, 202 observations formed the unlabeled pool, and 78 observations were assigned to the test set. For WBCD, the corresponding numbers were 20, 435, and 114 observations, respectively.

**Table 1.** Dataset characteristics and experimental data partitioning.

| Characteristic | HFCR | WBCD |
|---|---|---|
| Source | UCI ML Repository | UCI ML Repository |
| Number of instances | 299 | 569 |
| Number of predictors | 12 | 30 |
| Positive class | Death (96) | Malignant (212) |
| Negative class | Survival (203) | Benign (357) |
| Feature type | Continuous/Binary | Numerical |
| Missing values | No | No |
| Initial labeled set ($D_L$) | 19 | 20 |
| Unlabeled pool ($D_U$) | 202 | 435 |
| Independent test set | 78 | 114 |

All ML models were implemented using scikit-learn library in Python 3.15.

#### 3.1.1. Classification Models and Preprocessing

To assess whether the effects of the acquisition and rebalancing strategies depended on the underlying predictive model, both classifiers considered in the original experimental design were evaluated on both datasets. Specifically, a Random Forest (RF) classifier and

a Support Vector Machine (SVM) classifier were trained and evaluated on both HFCR and WBCD.

The RF classifier consisted of 100 decision trees (*n_estimators* = 100), with no maximum-depth constraint in the main analysis and with the remaining hyperparameters set to the corresponding scikit-learn default values unless otherwise specified. The SVM classifier employed a linear kernel with regularization parameter $C = 1$. Probability estimation was enabled (*probability* = *True*) so that the same probability-based least-confidence criterion could be used for both classifiers.

To maintain a consistent preprocessing pipeline across classifiers, the predictors were standardized using a `StandardScaler` fitted exclusively on the currently available labeled training data. The fitted transformation was subsequently applied to the unlabeled pool and to the independent test set. In the experimental conditions including SMOTE, standardization preceded synthetic oversampling and classifier fitting.

#### 3.1.2. Repeated Experimental Design

To assess the robustness of the experimental findings with respect to data partitioning and initialization, the complete experimental procedure was repeated over 20 predefined random seeds. For each seed, a new stratified development/test partition was generated. Within each repetition, the same partition and the same initial labeled subset were shared across all acquisition strategies, rebalancing conditions, and classifiers, thereby enabling paired comparisons under identical experimental conditions.

For WBCD, the initial labeled subset contained 20 observations with a fixed composition of 14 benign and 6 malignant cases. For HFCR, the initial labeled subset contained 19 observations, including 17 survival and 2 death cases. The identities of the observations assigned to the initial labeled subset varied across repetitions, whereas its size and class composition were kept fixed. The remaining development observations constituted the unlabeled pool.

The following seeds were used throughout the experiments:

29, 179, 209, 40, 71, 220, 163, 20, 227, 203, 138, 46, 249, 17, 201, 199, 187, 151, 139, 42.

The initial labeled set, unlabeled pool, and predictive model are denoted by $D_L$, $D_U$, and $m$, respectively, throughout the following sections.

### *3.2. Acquisition Strategies and Annotation Budget*

Two acquisition conditions were considered. Under random sampling, observations were selected uniformly from the unlabeled pool without replacement. Under uncertainty sampling, acquisition priority was determined from least-confidence scores computed from the class probabilities estimated by the current classifier. The least-confidence score for an unlabeled observation $x$ was defined as

$$\mathcal{U}_{\mathrm{LC}}(x) = 1 - \max_{y \in \mathcal{Y}} P_m(y \mid x), \tag{1}$$

where $P_m(y|x)$ is the conditional class probability of the class $y$ given the unlabeled instance $x$, $\max_{y \in Y} P_m(y \mid x)$ is the class label of $x$ with the highest probability using the model $m$ and $Y$ represents the set of possible class labels.

The goal is to select the instance $x$ from the unlabeled data pool $D_U$ that maximizes this score,i.e. that instance with the highest value of $U_{LC}(x)$. A high value of $U_{LC}(x)$ indicates that the model is uncertain about its prediction, whereas a low value indicates high confidence. Thus, instances for which the model assigns relatively similar probabilities to multiple classes are prioritized for annotation.

Uncertainty sampling is easy to implement and fairly intuitive. It often provides an improvement over random sampling by focusing annotation efforts on instances about which the current model is most uncertain.

For the uncertainty-based condition, observations associated with the largest least-confidence scores were therefore assigned the highest acquisition priority. The same acquisition definitions were applied to both classifiers and datasets.

Batch Acquisition and Annotation Budget

All experimental results were obtained using true batch acquisition with a fixed batch size $B = 10$. At each Active Learning round, the classifier was first fitted on the currently available labeled set. The acquisition score was then computed for every observation remaining in the unlabeled pool, and the 10 observations with the highest least-confidence scores were selected simultaneously from the same fitted model. Their labels were subsequently revealed, the complete batch was incorporated into the labeled set, and the classifier was retrained only after the entire batch had been acquired. Random sampling followed the same batch procedure, with 10 observations sampled uniformly without replacement.

The annotation budget $b$ was defined as the number of additional real observations labeled after initialization. Model performance was evaluated at the predefined budgets

$$b \in \{0, 10, 20, 30\}.$$

Accordingly,

$$|D_L^{(b)}| = |D_L^{(0)}| + b,$$

where $D_L^{(0)}$ denotes the initial labeled set. The four reporting points therefore correspond to total labeled-set sizes of 20, 30, 40, and 50 observations for WBCD, and 19, 29, 39, and 49 observations for HFCR, respectively.

The independent test set was evaluated at these predefined reporting budgets to characterize the learning trajectory. Test-set results were never used for query selection, stopping decisions, model selection, or hyperparameter tuning.

The maximum annotation budget is denoted by $A_{\max}$ in the pseudocodes; in the main experiments, $A_{\max} = 30$ additional real labels.

The two acquisition procedures are summarized in Algorithms 1 and 2. In the experimental protocol, $B = 10$ was selected as the number of observations acquired at each iteration.

**Algorithm 1** Random Sampling Strategy

**Require:** $D_L$ (initial labeled dataset), $D_U$ (unlabeled data pool), $m$ (classifier)
**Require:** $B$ (batch size), $A_{\max}$ (maximum annotation budget)
**Ensure:** $D_L$, $D_U$, $m$ (updated labeled dataset, pool, and classifier)
1: Train classifier $m$ using $D_L$
2: **while** the number of additional real labels acquired is lower than $A_{\max}$ **do**
3: Select a batch $Q \subseteq D_U$ containing $B$ instances uniformly at random without replacement
4: Query the oracle for the labels of all instances in $Q$
5: $D_L \leftarrow D_L \cup Q$
6: $D_U \leftarrow D_U \setminus Q$
7: Retrain classifier $m$ using updated $D_L$
8: **end while**
9: **return** $D_L, D_U, m$

https://doi.org/10.3390/ai7110485

**Algorithm 2** Uncertainty Sampling Strategy

**Require:** $D_L$ (initial labeled dataset), $D_U$ (unlabeled data pool), $m$ (classifier)
**Require:** $B$ (batch size), $A_{\max}$ (maximum annotation budget)
**Ensure:** $D_L$, $D_U$, $m$ (updated labeled dataset, pool, and classifier)
Train classifier $m$ using $D_L$
**while** the number of additional real labels acquired is lower than $A_{\max}$ **do**
  **for all** $x_i \in D_U$ **do**
    Compute the uncertainty score $u(x_i, m)$
  **end for**
  Select the batch $Q \subseteq D_U$ containing the $B$ instances with the highest uncertainty scores
  Query the oracle for the labels of all instances in $Q$
  $D_L \leftarrow D_L \cup Q$
  $D_U \leftarrow D_U \setminus Q$
  Retrain classifier $m$ using the updated $D_L$
**end while**
**return** $D_L, D_U, m$

### *3.3. Synthetic Rebalancing and Combined AL–SMOTE Procedure*

SMOTE was used as the data-level rebalancing procedure. Synthetic minority-class observations were obtained by interpolation between an available minority-class observation and one of its nearest minority-class neighbors.

The interpolation process adopted by SMOTE can be mathematically expressed as :

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda(\mathbf{x}_{\text{nn}} - \mathbf{x}_i), \qquad \lambda \in [0, 1]. \tag{2}$$

where $\mathbf{x}_i$ denotes a minority class sample, $\mathbf{x}_{\text{nn}}$ is one of its $k$ nearest minority class neighbors, $\lambda$ is a random value uniformly sampled from the interval $[0, 1]$, and $\mathbf{x}_{\text{new}}$ is the generated synthetic sample.

In the combined acquisition–rebalancing conditions, SMOTE was applied exclusively to the currently available labeled training observations; neither the unlabeled pool nor the independent test set contributed to synthetic sample generation. In the combined acquisition–rebalancing conditions, oversampling was applied after standardization and before classifier fitting at each annotation budget. Model estimation was performed on the resulting temporary rebalanced training set. The default number of nearest neighbors ($k = 5$) was used whenever the number of currently available minority-class observations allowed it. For smaller minority-class subsets, the number of neighbors was adaptively defined as

$$k_{\text{SMOTE}} = \min(5, n_{\text{minority}} - 1),$$

where $n_{\text{minority}}$ denotes the number of minority-class observations currently available in the labeled set. The stochastic components of SMOTE were deterministically controlled by the seed associated with the corresponding experimental repetition and annotation budget. For uncertainty-based acquisition, the classifier fitted on this rebalanced set was subsequently used to compute acquisition scores over the real unlabeled pool. After each batch of real observations had been acquired, the real labeled set was updated, a new temporary SMOTE-rebalanced training set was constructed, and the classifier was fitted again.

Synthetic observations were used only to construct the rebalanced training data at the current iteration. They were not treated as newly annotated real observations, were not transferred to the unlabeled pool, and did not contribute to the annotation budget.

The random- and uncertainty-based rebalancing procedures are summarized in Algorithms 3 and 4, respectively.

We report the pipeline that reflects the overall architecture (see Figure 1).

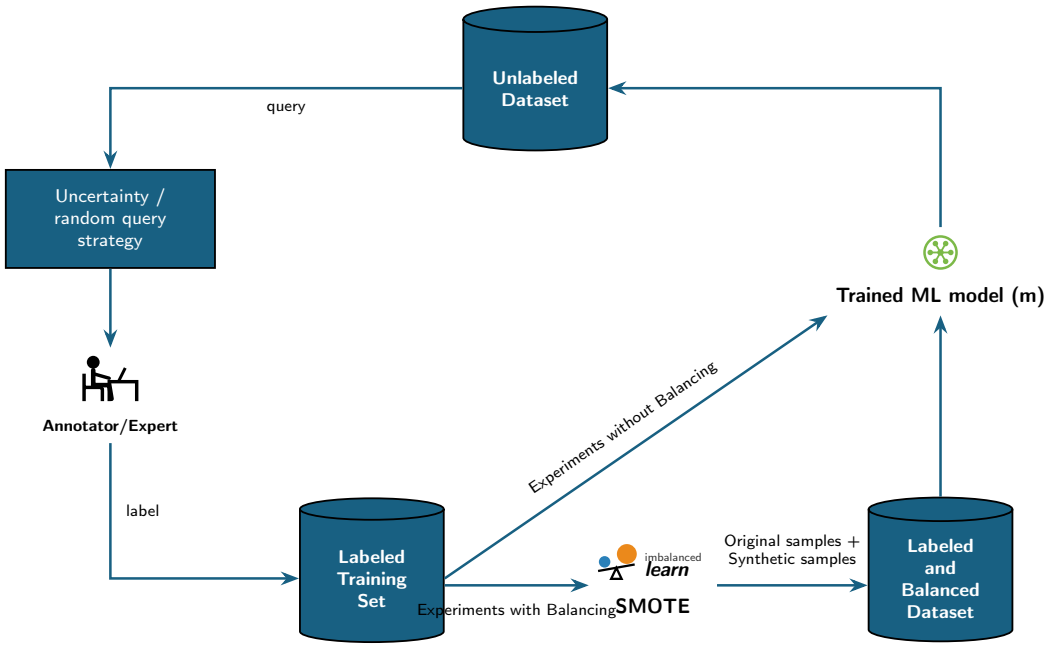



**Figure 1.** Overview of the proposed Active Learning and SMOTE-based integration workflow. The process starts with the initial labeled dataset. The model is trained using the small labeled dataset and subsequently used to identify the most informative samples from the unlabeled pool according to the adopted sample-selection strategy. The selected samples are labeled and incorporated into the training set, after which SMOTE (Synthetic Minority Over-sampling Technique) is applied to rebalance the updated training data. The model is then retrained using the SMOTE-balanced dataset and evaluated. This procedure is repeated iteratively, with each Active Learning iteration followed by SMOTE-based rebalancing, until the predefined stopping criterion is satisfied.

**Algorithm 3** Random Sampling with Synthetic Data Generation

**Require:** $D_L$ (initial real labeled dataset), $D_U$ (unlabeled data pool), $m$ (classifier)
**Require:** $B$ (batch size), $A_{\max}$ (maximum annotation budget)
**Ensure:** $D_L$, $D_U$, $m$ (updated real labeled dataset, pool, and classifier)

Generate a set of synthetic samples $S$ from the current real labeled set $D_L$ using SMOTE
Construct the temporary rebalanced training set $\widetilde{D}_L \leftarrow D_L \cup S$
Train classifier $m$ using $\widetilde{D}_L$
**while** the number of additional real labels acquired is lower than $A_{\max}$ **do**
    Select a batch $Q \subseteq D_U$ containing $B$ instances uniformly at random without replacement
    Query the oracle for the labels of all instances in $Q$
    $D_L \leftarrow D_L \cup Q$
    $D_U \leftarrow D_U \setminus Q$
    Generate a new set of synthetic samples $S$ from the updated real labeled set $D_L$ using SMOTE
    Construct the new temporary rebalanced training set $\widetilde{D}_L \leftarrow D_L \cup S$
    Retrain classifier $m$ using $\widetilde{D}_L$
**end while**
**return** $D_L, D_U, m$

**Algorithm 4** Uncertainty Sampling with Synthetic Data Generation

**Require:** $D_L$ (initial labeled dataset), $D_U$ (unlabeled data pool), $m$ (classifier)
**Require:** $B$ (batch size), $A_{\max}$ (maximum annotation budget)
**Ensure:** $D_L$, $D_U$, $m$ (updated labeled dataset, pool, and classifier)

```
Generate a set of synthetic samples S from the current real labeled set D_L using SMOTE
Construct the temporary rebalanced training set D̃_L ← D_L ∪ S
Train classifier m using D̃_L
while the number of additional real labels acquired is lower than A_max do
    for all x_i ∈ D_U do
        Compute the uncertainty score u(x_i, m)
    end for
    Select the batch Q ⊆ D_U containing the B instances with the highest uncertainty scores
    Query the oracle for the labels of all instances in Q
    D_L ← D_L ∪ Q
    D_U ← D_U \ Q
    Generate a new set of synthetic samples S from the updated real labeled set D_L using SMOTE
    Construct the temporary rebalanced training set D̃_L ← D_L ∪ S
    Retrain classifier m using D̃_L
end while
return D_L, D_U, m
```

### *3.4. Measures Used for Performance Evaluation*

Because both datasets exhibit unequal class distributions, predictive performance was characterized using complementary threshold-dependent and threshold-independent measures. Let $TP$, $TN$, $FP$, and $FN$ denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. Malignant tumors were considered the positive class for WBCD, whereas death events represented the positive class for HFCR.

Accuracy, precision, recall (sensitivity), specificity, and F1-score were defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{4}$$

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN}, \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{6}$$

$$F_1 = 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{7}$$

Balanced Accuracy was computed as

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}, \tag{8}$$

thereby assigning equal weight to performance on the positive and negative classes.

The Area Under the Receiver Operating Characteristic Curve (AUROC) summarized discrimination over the complete range of classification thresholds:

$$\text{AUROC} = \int_0^1 TPR(FPR)\, dFPR, \tag{9}$$

where $TPR = TP/(TP + FN)$ and $FPR = FP/(FP + TN)$.

The Area Under the Precision–Recall Curve (AUPRC) was calculated by trapezoidal integration of precision as a function of recall:

$$\text{AUPRC} = \int_0^1 P(R)\, dR. \tag{10}$$

Average Precision (AP) was additionally computed as the non-interpolated weighted mean of precision values over successive increases in recall:

$$\text{AP} = \sum_n (R_n - R_{n-1}) P_n, \tag{11}$$

where $P_n$ and $R_n$ denote precision and recall at the $n$-th decision threshold.

Unlike AUPRC, which was obtained by trapezoidal integration of the precision–recall curve, AP was computed as the weighted mean of precision values associated with successive increments in recall. The two quantities were therefore retained as distinct, complementary summaries of precision–recall performance.

Learning curves were constructed for Balanced Accuracy and AP at cumulative annotation budgets of 0, 10, 20, and 30 additional real labels beyond the initial labeled set. Balanced Accuracy was considered useful for accounting for sensitivity and specificity under class imbalance, whereas AP was used to summarize precision–recall performance across decision thresholds, with particular attention to the underrepresented positive class. Accuracy, precision, recall, specificity, F1-score, AUROC, and AUPRC were retained as complementary performance measures. In particular, AUPRC was reported as an additional summary of precision–recall behavior, while AP was used for the principal learning-curve and inferential analyses.

### 3.5. Repeated-Run and Statistical Analysis

For each dataset, classifier, acquisition strategy, annotation budget, and performance measure, results were summarized over the $S = 20$ predefined repetitions. Let $M_s(b)$ denote the value of a given performance measure obtained for seed $s$ at annotation budget $b$. Mean performance was computed as

$$\bar{M}(b) = \frac{1}{S} \sum_{s=1}^{S} M_s(b). \tag{12}$$

Variability across repetitions was summarized using the sample standard deviation and a two-sided 95% confidence interval for the mean:

$$\bar{M} \pm t_{1-\alpha/2,S-1} \frac{s_M}{\sqrt{S}}, \qquad \alpha = 0.05, \tag{13}$$

where $s_M$ denotes the sample standard deviation across repetitions. Learning curves were constructed separately for Balanced Accuracy and AP by reporting the mean and corresponding 95% confidence interval across the 20 repetitions at annotation budgets 0, 10, 20, and 30. The boxplots report the distribution of the 20 seed-level values obtained at the same annotation budget.

Because all strategies and both classifiers were evaluated using the same split and the same initial labeled subset within each seed, pairwise comparisons were performed using a paired design. The two-sided Wilcoxon signed-rank test was applied to Balanced Accuracy and AP at annotation budgets 10, 20, and 30. The principal strategy comparisons corresponded to the two research questions, namely Uncertainty sampling versus Random sampling and Uncertainty sampling + SMOTE versus Uncertainty sampling alone. All six

pairwise comparisons among the four experimental strategies were additionally retained for complementary analysis.

To control the family-wise error rate arising from multiple pairwise strategy comparisons, $p$-values were adjusted using Holm's step-down procedure within each classifier–metric–budget family. For comparisons based on the complete learning trajectory, the correction was applied within each classifier–metric family. Direct SVM-versus-RF comparisons were also performed in a paired manner within each strategy and annotation budget; the corresponding $p$-values were Holm-adjusted within each metric–budget family. Statistical significance was assessed at $\alpha = 0.05$ using the adjusted $p$-values.

Performance over the complete annotation trajectory was additionally summarized using a normalized Area Under the Learning Curve (AULC). For annotation budgets $b_0, \ldots, b_K$, the normalized AULC for seed $s$ was computed by trapezoidal integration:

$$\mathrm{AULC}_s = \frac{1}{b_K - b_0} \sum_{i=0}^{K-1} \frac{M_s(b_i) + M_s(b_{i+1})}{2} (b_{i+1} - b_i), \tag{14}$$

with $(b_0, b_1, b_2, b_3) = (0, 10, 20, 30)$. Normalization by the total annotation-budget range retains the scale of the underlying performance measure.

### 3.6. Sensitivity Analyses

Additional sensitivity analyses were conducted to assess whether the main findings depended materially on selected experimental choices.

For the linear SVM, the regularization parameter was evaluated at $C \in \{0.1, 1, 10\}$, whereas for Random Forest the maximum tree depth was evaluated at `max_depth` $\in \{3, 5, 10, \mathrm{None}\}$. These analyses were performed on both datasets and under all four acquisition/rebalancing strategies.

The sensitivity to acquisition granularity was additionally examined by repeating the experimental procedure with sequential acquisition ($B = 1$). Under this setting, the classifier was retrained and acquisition scores were recomputed after each newly acquired real observation. The same 20 data partitions and initial labeled subsets were retained, and the maximum annotation budget remained fixed at 30 additional labels.

As an additional imbalance-aware control, uncertainty sampling with class-weighted learning (`class_weight='balanced'`) was evaluated for both classifiers as an example of algorithm-level class-imbalance comparator within the Active Learning framework. This condition retained the same least-confidence acquisition rule while modifying classifier training through class-dependent weights, without synthetic oversampling. The results of these sensitivity and control analyses are reported in Appendix A.

## 4. Results

This section presents the results in accordance with the study objectives detailed in Section 3. Specifically, we outline the findings of the performance analysis of active learning (Research Question 1) and the performance analysis of active learning combined with synthetic data generation (Research Question 2).

Results are reported across the 20 matched experimental repetitions for the four experimental conditions: Random, Random + SMOTE, Uncertainty, and Uncertainty + SMOTE. Performance was evaluated at cumulative acquisition levels corresponding to 0, 10, 20, and 30 additional real labels beyond the initial labeled set. Balanced Accuracy (BA) and Average Precision (AP) are used as the principal measures in the learning-curve and inferential analyses, while the remaining performance measures are reported as complementary results.

Experiment 0 was considered as the baseline condition and corresponds to the initial training dataset before any sample acquisition through active learning. Experiment 1

includes 10 newly acquired samples, Experiment 2 includes a further 10 samples, and so on, resulting in a cumulative increase of the labeled dataset size.

The active learning procedure was stopped after a predefined number of iterations defined a priori. The test set was not used to determine the stopping point.

The boxplots reported below summarize the distribution of the 20 repetition-level values at each fixed acquisition level; they therefore represent between-repetition variability and not variation within an acquisition batch.

### *4.1. Research Question 1: Effect of Uncertainty-Based Acquisition*

RQ1 was examined by comparing Uncertainty with Random acquisition separately for each dataset and classifier. The comparison was assessed at 10, 20, and 30 additional real labels and over the complete acquisition trajectory using normalized AULC.

#### 4.1.1. HFCR

Table 2 reports the complete BA and AP trajectories for the four experimental conditions on HFCR.

**Table 2.** Mean Balanced Accuracy (BA) and Average Precision (AP) across 20 repetitions for HFCR. Values correspond to cumulative numbers of additional real labels beyond the initial labeled set.

| | | 0 | | 10 | | 20 | | 30 | |
|---|---|---|---|---|---|---|---|---|---|
| **Classifier** | **Strategy** | **BA** | **AP** | **BA** | **AP** | **BA** | **AP** | **BA** | **AP** |
| RF | Random | 0.522 | 0.741 | 0.648 | 0.754 | 0.691 | 0.799 | 0.727 | 0.800 |
| RF | Random + SMOTE | 0.566 | 0.680 | 0.675 | 0.691 | 0.708 | 0.737 | 0.726 | 0.756 |
| RF | Uncertainty | 0.522 | 0.741 | 0.704 | 0.797 | 0.770 | 0.803 | 0.781 | 0.805 |
| RF | Uncertainty + SMOTE | 0.566 | 0.680 | 0.726 | 0.789 | 0.782 | 0.797 | 0.793 | 0.798 |
| SVM | Random | 0.620 | 0.609 | 0.657 | 0.625 | 0.709 | 0.672 | 0.713 | 0.703 |
| SVM | Random + SMOTE | 0.619 | 0.609 | 0.675 | 0.640 | 0.730 | 0.692 | 0.729 | 0.693 |
| SVM | Uncertainty | 0.620 | 0.609 | 0.671 | 0.648 | 0.693 | 0.691 | 0.706 | 0.712 |
| SVM | Uncertainty + SMOTE | 0.619 | 0.609 | 0.673 | 0.645 | 0.716 | 0.696 | 0.724 | 0.717 |

As shown in Figure 2, with RF uncertainty-based acquisition was associated with higher mean BA than Random at the three post-initialization acquisition levels: 0.704 versus 0.648 at 10 additional labels, 0.770 versus 0.691 at 20, and 0.781 versus 0.727 at 30. The paired difference was statistically supported after Holm correction at 20 additional labels ($\Delta = 0.078$, $p_{\text{Holm}} = 0.003$) and at 30 additional labels ($\Delta = 0.054$, $p_{\text{Holm}} = 0.025$). The same comparison over the complete trajectory yielded a BA AULC difference of +0.054 ($p_{\text{Holm}} = 0.005$) (see Figure 3). For AP, the numerical advantage was smaller and the paired comparisons were not statistically supported after correction (see Figure 4).

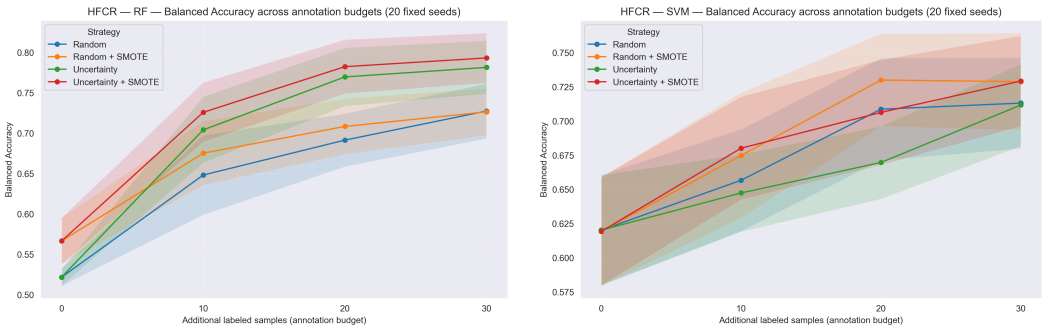



**Figure 2.** HFCR Balanced Accuracy across cumulative acquisition levels. Mean performance across 20 matched repetitions is shown with 95% confidence intervals. RF is shown on the left and SVM on the right.

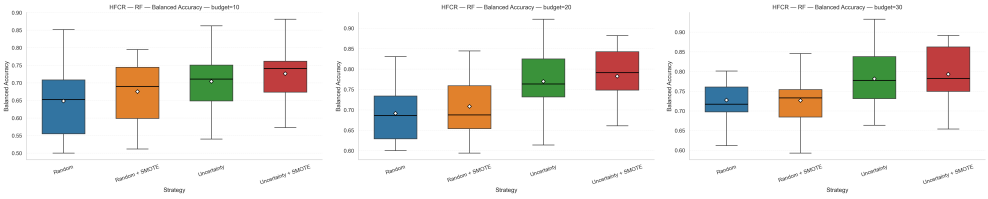


**Figure 3.** HFCR RF Balanced Accuracy distributions across the 20 matched repetitions at 10, 20, and 30 additional real labels (left to right).

With SVM, the BA pattern was less consistent: Uncertainty was slightly higher than Random at 10 additional labels, but slightly lower at 20 and 30 labels, and the trajectory-level AULCs were nearly identical (see Figure 5). AP was numerically higher under Uncertainty at each post-initialization level, although these differences were not statistically supported. In this setting, the benefit associated with uncertainty-based acquisition therefore depended on the classifier and on the performance measure considered. This classifier-dependent behavior is compatible with previous work emphasizing that uncertainty-based acquisition can be sensitive to the learner and to the class distribution [8,14].

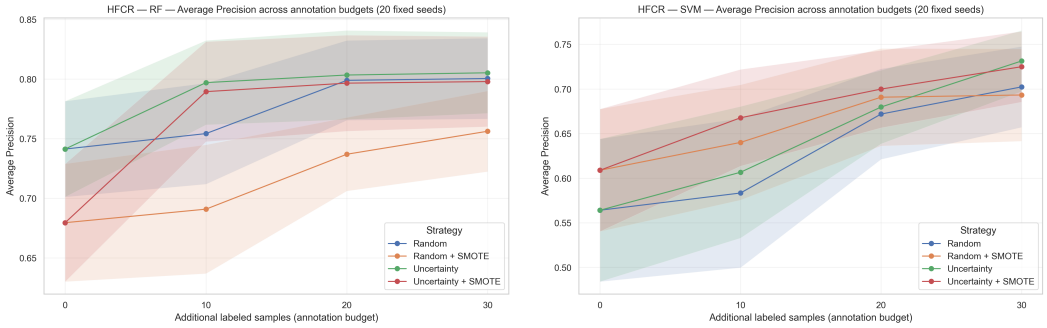



**Figure 4.** HFCR Average Precision across cumulative acquisition levels. Mean performance across 20 matched repetitions is shown with 95% confidence intervals. RF is shown on the left and SVM on the right.

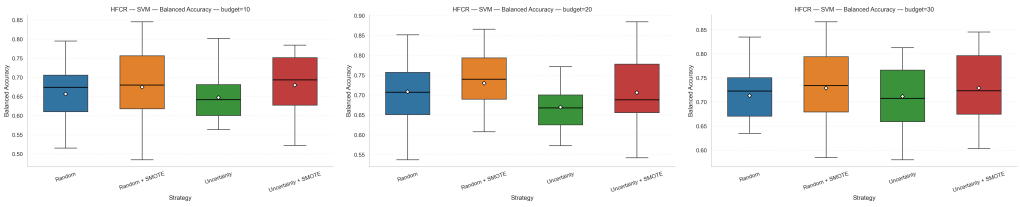


**Figure 5.** HFCR SVM Balanced Accuracy distributions across the 20 matched repetitions at 10, 20, and 30 additional real labels (left to right).

Experiment 0 is not included in the boxplots because it represents the initial labeled training set, which serves as the starting point of the active learning process.

At the maximum acquisition level, the complementary metrics in Table 3 help clarify the BA results. With RF, the higher BA observed for Uncertainty than for Random was accompanied by higher recall (0.662 versus 0.502), while specificity remained high (0.901 versus 0.953). With SVM, Uncertainty produced higher specificity but lower recall than Random, yielding similar overall BA. These results suggest that, for HFCR, the acquisition strategy affected the balance between minority-class detection and majority-class discrimination differently across the two learners.

**Table 3.** Performance at the maximum acquisition level (30 additional real labels) for HFCR. Values are mean ± standard deviation across 20 repetitions.

| Classifier | Strategy | Acc. | BA | Prec. | Rec. | Spec. | F1 | AUROC | AUPRC | AP |
|---|---|---|---|---|---|---|---|---|---|---|
| RF | Random | 0.808 ± 0.046 | 0.727 ± 0.073 | 0.848 ± 0.100 | 0.502 ± 0.156 | 0.953 ± 0.034 | 0.613 ± 0.137 | 0.885 ± 0.039 | 0.799 ± 0.075 | 0.800 ± 0.072 |
| RF | Random + SMOTE | 0.784 ± 0.042 | 0.726 ± 0.063 | 0.705 ± 0.082 | 0.566 ± 0.136 | 0.887 ± 0.045 | 0.620 ± 0.097 | 0.867 ± 0.038 | 0.751 ± 0.077 | 0.756 ± 0.072 |
| RF | Uncertainty | 0.824 ± 0.050 | 0.781 ± 0.071 | 0.770 ± 0.092 | 0.662 ± 0.149 | 0.901 ± 0.052 | 0.700 ± 0.100 | 0.875 ± 0.044 | 0.803 ± 0.077 | 0.805 ± 0.073 |
| RF | Uncertainty + SMOTE | 0.835 ± 0.049 | 0.793 ± 0.066 | 0.784 ± 0.087 | 0.676 ± 0.125 | 0.910 ± 0.041 | 0.720 ± 0.093 | 0.880 ± 0.048 | 0.797 ± 0.083 | 0.798 ± 0.081 |
| SVM | Random | 0.767 ± 0.055 | 0.713 ± 0.071 | 0.678 ± 0.117 | 0.562 ± 0.157 | 0.864 ± 0.076 | 0.600 ± 0.106 | 0.821 ± 0.066 | 0.693 ± 0.102 | 0.703 ± 0.096 |
| SVM | Random + SMOTE | 0.757 ± 0.060 | 0.729 ± 0.076 | 0.619 ± 0.090 | 0.650 ± 0.156 | 0.808 ± 0.075 | 0.626 ± 0.104 | 0.808 ± 0.074 | 0.684 ± 0.116 | 0.693 ± 0.110 |
| SVM | Uncertainty | 0.778 ± 0.042 | 0.706 ± 0.064 | 0.726 ± 0.081 | 0.504 ± 0.138 | 0.908 ± 0.045 | 0.584 ± 0.107 | 0.821 ± 0.057 | 0.703 ± 0.098 | 0.712 ± 0.094 |
| SVM | Uncertainty + SMOTE | 0.779 ± 0.052 | 0.724 ± 0.074 | 0.686 ± 0.083 | 0.570 ± 0.148 | 0.878 ± 0.040 | 0.616 ± 0.115 | 0.814 ± 0.069 | 0.709 ± 0.101 | 0.717 ± 0.097 |

#### 4.1.2. WBCD

Table 4 reports the complete BA and AP trajectories for the four experimental conditions on WBCD.

**Table 4.** Mean Balanced Accuracy (BA) and Average Precision (AP) across 20 repetitions for WBCD. Values correspond to cumulative numbers of additional real labels beyond the initial labeled set.

| | | 0 | | 10 | | 20 | | 30 | |
|---|---|---|---|---|---|---|---|---|---|
| **Classifier** | **Strategy** | **BA** | **AP** | **BA** | **AP** | **BA** | **AP** | **BA** | **AP** |
| RF | Random | 0.893 | 0.971 | 0.905 | 0.977 | 0.909 | 0.979 | 0.917 | 0.980 |
| RF | Random + SMOTE | 0.897 | 0.970 | 0.909 | 0.973 | 0.921 | 0.978 | 0.923 | 0.979 |
| RF | Uncertainty | 0.893 | 0.971 | 0.920 | 0.982 | 0.951 | 0.990 | 0.957 | 0.989 |
| RF | Uncertainty + SMOTE | 0.897 | 0.970 | 0.916 | 0.980 | 0.942 | 0.988 | 0.946 | 0.989 |
| SVM | Random | 0.904 | 0.962 | 0.922 | 0.976 | 0.929 | 0.981 | 0.932 | 0.980 |
| SVM | Random + SMOTE | 0.903 | 0.962 | 0.923 | 0.976 | 0.929 | 0.981 | 0.930 | 0.979 |
| SVM | Uncertainty | 0.904 | 0.962 | 0.928 | 0.976 | 0.941 | 0.984 | 0.948 | 0.982 |
| SVM | Uncertainty + SMOTE | 0.903 | 0.962 | 0.927 | 0.976 | 0.942 | 0.983 | 0.949 | 0.983 |

With RF, the separation between Uncertainty and Random was more evident at the larger acquisition levels. Mean BA increased from 0.909 under Random to 0.951 under Uncertainty at 20 additional labels and from 0.917 to 0.957 at 30 labels (see Figure 6). Both differences were statistically supported after Holm correction ($p_{\text{Holm}} = 0.002$ and $p_{\text{Holm}} < 0.001$, respectively) (see Figure 7). A similar pattern was observed for AP at 20 and 30 additional labels, and both BA and AP showed statistically supported positive AULC differences over the complete trajectory (see Figure 8).

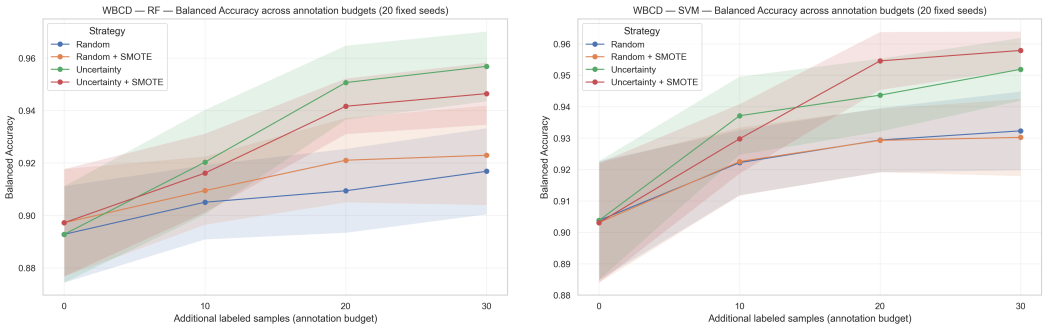



**Figure 6.** WBCD Balanced Accuracy across cumulative acquisition levels. Mean performance across 20 matched repetitions is shown with 95% confidence intervals. RF is shown on the left and SVM on the right.

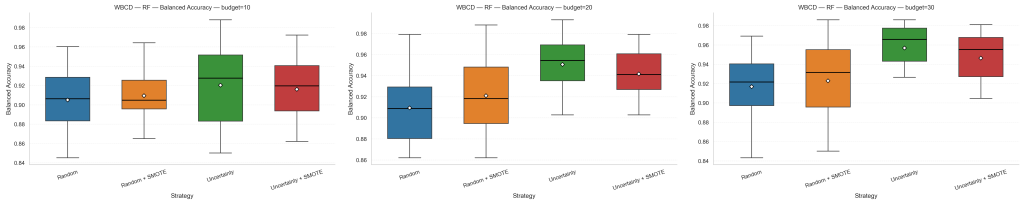


**Figure 7.** WBCD RF Balanced Accuracy distributions across the 20 matched repetitions at 10, 20, and 30 additional real labels (left to right).

With SVM, mean BA was also higher under Uncertainty at the three post-initialization levels, with the largest numerical difference at 30 additional labels (+0.016) (see Figure 9). These differences, however, were not statistically supported after multiplicity correction. AP was high across all SVM conditions and showed only small differences among strategies. Within WBCD, the evidence in favor of uncertainty-based acquisition was therefore stronger for RF, whereas the SVM results suggested a smaller effect. This finding should be interpreted as specific to the learner–dataset configuration considered here and does not establish a general ranking between the two classifiers.

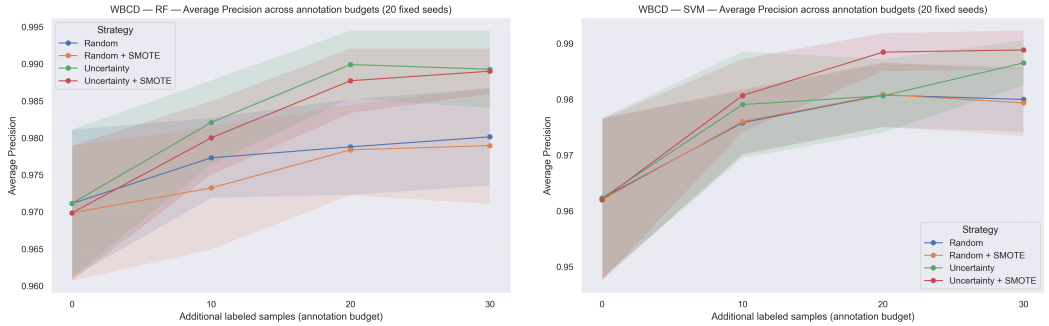



**Figure 8.** WBCD Average Precision across cumulative acquisition levels. Mean performance across 20 matched repetitions is shown with 95% confidence intervals. RF is shown on the left and SVM on the right.

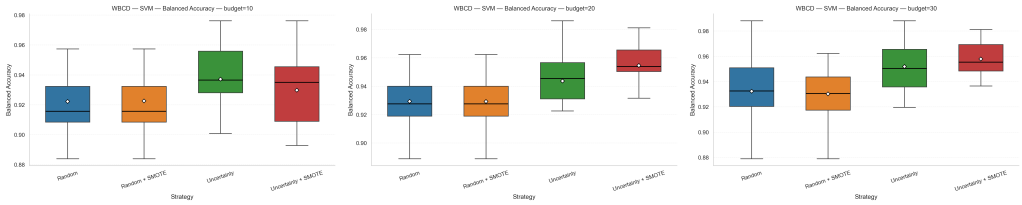


**Figure 9.** WBCD SVM Balanced Accuracy distributions across the 20 matched repetitions at 10, 20, and 30 additional real labels (left to right).

The complementary metrics at 30 additional labels are reported in Table 5. With RF, the increase in BA under Uncertainty was accompanied by higher recall, specificity, F1, AUROC, and AP than under Random acquisition. With SVM, the same metrics showed smaller numerical differences. The high AP values observed across WBCD conditions indicate limited remaining headroom for further improvement on this measure, particularly for SVM.

**Table 5.** Performance at the maximum acquisition level (30 additional real labels) for WBCD. Values are mean ± standard deviation across 20 repetitions.

| Classifier | Strategy | Acc. | BA | Prec. | Rec. | Spec. | F1 | AUROC | AUPRC | AP |
|---|---|---|---|---|---|---|---|---|---|---|
| RF | Random | 0.929 ± 0.029 | 0.917 ± 0.035 | 0.939 ± 0.054 | 0.869 ± 0.073 | 0.965 ± 0.037 | 0.900 ± 0.043 | 0.986 ± 0.012 | 0.981 ± 0.014 | 0.980 ± 0.014 |
| RF | Random + SMOTE | 0.931 ± 0.036 | 0.923 ± 0.041 | 0.926 ± 0.067 | 0.892 ± 0.083 | 0.954 ± 0.050 | 0.905 ± 0.050 | 0.986 ± 0.013 | 0.980 ± 0.016 | 0.979 ± 0.017 |
| RF | Uncertainty | 0.963 ± 0.022 | 0.957 ± 0.028 | 0.964 ± 0.031 | 0.935 ± 0.058 | 0.979 ± 0.019 | 0.948 ± 0.033 | 0.992 ± 0.008 | 0.990 ± 0.011 | 0.989 ± 0.011 |
| RF | Uncertainty + SMOTE | 0.954 ± 0.019 | 0.946 ± 0.025 | 0.957 ± 0.028 | 0.918 ± 0.055 | 0.975 ± 0.017 | 0.936 ± 0.028 | 0.993 ± 0.005 | 0.989 ± 0.006 | 0.989 ± 0.007 |
| SVM | Random | 0.941 ± 0.024 | 0.932 ± 0.027 | 0.942 ± 0.054 | 0.900 ± 0.060 | 0.965 ± 0.039 | 0.918 ± 0.033 | 0.985 ± 0.009 | 0.980 ± 0.012 | 0.980 ± 0.012 |
| SVM | Random + SMOTE | 0.938 ± 0.023 | 0.930 ± 0.026 | 0.935 ± 0.055 | 0.900 ± 0.060 | 0.960 ± 0.039 | 0.915 ± 0.032 | 0.985 ± 0.009 | 0.979 ± 0.013 | 0.979 ± 0.013 |
| SVM | Uncertainty | 0.955 ± 0.018 | 0.948 ± 0.019 | 0.955 ± 0.042 | 0.924 ± 0.038 | 0.973 ± 0.027 | 0.938 ± 0.024 | 0.988 ± 0.008 | 0.982 ± 0.012 | 0.982 ± 0.012 |
| SVM | Uncertainty + SMOTE | 0.954 ± 0.018 | 0.949 ± 0.020 | 0.951 ± 0.040 | 0.926 ± 0.039 | 0.971 ± 0.026 | 0.937 ± 0.025 | 0.989 ± 0.009 | 0.983 ± 0.013 | 0.983 ± 0.013 |

The paired inferential results addressing RQ1 are summarized in Table 6. Taken together, the present experiments support a benefit of uncertainty-based acquisition most clearly for RF, particularly for BA on both datasets and for AP on WBCD. The effect was less consistent for SVM and for AP on HFCR.

**Table 6.** RQ1 paired comparisons: Uncertainty versus Random acquisition.

| Dataset | Classifier | Metric | 10 Labels | 20 Labels | 30 Labels | AULC |
|---|---|---|---|---|---|---|
| HFCR | RF | BA | +0.056 (0.116) | +0.078 (0.003) | +0.054 (0.025) | +0.054 (0.005) |
| HFCR | RF | AP | +0.043 (0.072) | +0.004 (1.000) | +0.005 (1.000) | +0.017 (0.248) |
| HFCR | SVM | BA | +0.014 (1.000) | −0.015 (1.000) | −0.007 (1.000) | −0.001 (1.000) |
| HFCR | SVM | AP | +0.023 (1.000) | +0.019 (1.000) | +0.009 (1.000) | +0.015 (1.000) |
| WBCD | RF | BA | +0.015 (0.124) | +0.041 (0.002) | +0.040 (<0.001) | +0.026 (<0.001) |
| WBCD | RF | AP | +0.005 (0.200) | +0.011 (<0.001) | +0.009 (<0.001) | +0.007 (<0.001) |
| WBCD | SVM | BA | +0.006 (1.000) | +0.012 (0.650) | +0.016 (0.218) | +0.008 (0.417) |
| WBCD | SVM | AP | +0.000 (1.000) | +0.003 (1.000) | +0.002 (1.000) | +0.002 (1.000) |

Values are mean paired differences (first condition minus second condition); Holm-adjusted Wilcoxon *p*-values are reported in parentheses.

### *4.2. Research Question 2: Incremental Contribution of SMOTE-Based Rebalancing*

The same datasets used in the active learning experiments were also employed to evaluate the proposed combination of active learning and synthetic data generation.

RQ2 was assessed primarily through the paired comparison between Uncertainty + SMOTE and Uncertainty. The objective was therefore to determine whether, within the present experimental setting, SMOTE-based rebalancing provided an additional contribution once uncertainty-based acquisition was already in use.

#### 4.2.1. HFCR

For HFCR, adding SMOTE to Uncertainty produced modest numerical increases in BA for both classifiers. With RF, the mean difference was +0.022 at 10 additional labels, +0.013 at 20, and +0.012 at 30; with SVM, the corresponding differences were +0.002, +0.022, and +0.018. None of the fixed-budget comparisons remained statistically supported after Holm correction. Over the complete trajectory, the RF BA AULC difference was +0.021, with an unadjusted Wilcoxon *p*-value of 0.030 but a Holm-adjusted value of 0.059. AP did not show a corresponding incremental advantage.

The complementary metrics suggest that the numerical BA changes reflected different trade-offs across classifiers. At 30 additional labels, RF Uncertainty + SMOTE showed slightly higher recall and specificity than Uncertainty alone, whereas for SVM the increase in recall was accompanied by lower specificity and precision. Thus, within HFCR, the effect associated with SMOTE was metric- and learner-dependent rather than uniformly favorable.

#### 4.2.2. WBCD

For WBCD, Uncertainty and Uncertainty + SMOTE were nearly overlapping with SVM across both BA and AP. With RF, Uncertainty + SMOTE showed slightly lower mean BA

than Uncertainty at 20 and 30 additional labels (−0.009 and −0.010, respectively), while AP differences were small. None of these incremental comparisons was statistically supported after Holm correction, and the corresponding AULC differences were also small.

In this setting, the addition of SMOTE therefore did not appear to provide a consistent incremental benefit over uncertainty-based acquisition alone. This finding is restricted to the datasets, learners, imbalance levels, and acquisition budgets considered here. It does not imply that SMOTE is generally ineffective within Active Learning. Previous studies have reported benefits from integrating oversampling and Active Learning under different tasks and experimental configurations [22,23,25], which further supports interpreting the present result as context-dependent.

The complete paired comparisons addressing RQ2 are reported in Table 7.

**Table 7.** RQ2 paired comparisons: Uncertainty + SMOTE versus Uncertainty.

| Dataset | Classifier | Metric | 10 Labels | 20 Labels | 30 Labels | AULC |
|---|---|---|---|---|---|---|
| HFCR | RF | BA | +0.022 (0.344) | +0.013 (0.490) | +0.012 (0.251) | +0.021 (0.059) |
| HFCR | RF | AP | −0.008 (0.596) | −0.007 (1.000) | −0.007 (1.000) | −0.016 (0.248) |
| HFCR | SVM | BA | +0.002 (1.000) | +0.022 (0.333) | +0.018 (1.000) | +0.011 (0.251) |
| HFCR | SVM | AP | −0.003 (1.000) | +0.005 (1.000) | +0.006 (1.000) | +0.001 (1.000) |
| WBCD | RF | BA | −0.004 (0.812) | −0.009 (0.140) | −0.010 (0.104) | −0.005 (0.232) |
| WBCD | RF | AP | −0.002 (0.985) | −0.002 (0.109) | −0.000 (1.000) | −0.002 (0.409) |
| WBCD | SVM | BA | −0.000 (1.000) | +0.000 (0.650) | +0.000 (0.774) | −0.000 (0.774) |
| WBCD | SVM | AP | +0.000 (1.000) | −0.001 (1.000) | +0.001 (1.000) | −0.000 (1.000) |

Values are mean paired differences (first condition minus second condition); Holm-adjusted Wilcoxon $p$-values are reported in parentheses.

Overall, RQ2 did not show a statistically supported incremental advantage of SMOTE when added to uncertainty-based acquisition across the four dataset–classifier combinations considered here. Numerical differences were observed in individual settings, particularly for BA on HFCR, but their magnitude and direction varied across classifiers and performance measures.

### *4.3. Discussion on the Trade-Off Between SMOTE and Uncertainty Sampling*

The combination of SMOTE with uncertainty-based acquisition was motivated by the possibility of obtaining additional predictive benefits without increasing the annotation budget. Uncertainty-based acquisition identifies informative unlabeled instances for which obtaining ground-truth labels entails an additional annotation cost, whereas SMOTE generates synthetic minority-class samples from the existing labeled dataset without requiring further manual labeling. Combining these approaches therefore provides a potentially cost-effective strategy: uncertainty-based acquisition improves the informativeness of the labeled dataset, while SMOTE may improve minority-class representation at no additional annotation cost. However,the absence of a statistically supported incremental benefit from SMOTE when combined with uncertainty-based acquisition can be understood through the complementary roles of data selection and data augmentation. Uncertainty-based acquisition preferentially selects informative samples, typically those located near the current decision boundary or in regions where the model exhibits limited confidence. Consequently, the acquired samples may already provide substantial information about minority-class decision boundaries, reducing the marginal contribution of synthetic oversampling. Although SMOTE increases the representation of minority-class instances through interpolation between neighboring observations, it does not introduce genuinely new information and may instead generate redundant samples in regions that are already adequately represented. Moreover, when minority-class samples are sparse, noisy, or distributed across heterogeneous subregions, interpolation may produce synthetic

observations that add limited discriminative value or insufficiently capture the underlying class distribution. These effects suggest that the utility of SMOTE depends not only on class imbalance but also on the informational content and local distribution of the samples selected by the acquisition strategy. Thus, the lack of a statistically significant improvement should not be interpreted as evidence that SMOTE is universally ineffective; rather, it suggests that, under the evaluated experimental conditions, its additional contribution beyond uncertainty-driven sample selection is limited relative to the information already provided by the actively acquired observations. This finding highlights the importance of evaluating resampling methods in conjunction with acquisition strategies rather than assuming that their individual benefits necessarily translate into additive gains.

Another plausible interpretation of the limited incremental contribution of SMOTE concerns the highly data-scarce and imbalanced regime examined in this study. In HFCR, the initial labeled set contains only two minority-class observations. Under the rule adopted in the study, this corresponds to $k_{\mathrm{SMOTE}} = 1$ at initialization. The initial synthetic samples were therefore generated from a very limited representation of the minority class, in the specific dataset through interpolation between the two available cases. Although the labeled set increases during Active Learning, the maximum annotation budget considered in the present experiments still results in a relatively small real training set, with uncertainty sampling not explicitly controlling the class composition of the acquired observations. Under these conditions, the additional contribution of SMOTE may remain modest or variable across repetitions and classifiers, which is compatible with the small differences observed in our experiments, none of which were statistically supported after correction. The present results are consequently intended to characterize the interaction between uncertainty-based acquisition and SMOTE within a particularly challenging limited-data regime. They do not establish whether the same behavior would be observed with larger labeled datasets or under different degrees of class imbalance. Further studies should therefore examine the AL–SMOTE interaction across controlled variations in sample size, minority-class prevalence, and classifier choice, in order to determine whether the incremental contribution of synthetic rebalancing changes as the amount and composition of the available real training data increase.

## 5. Conclusions

This study investigated the interaction between active learning and SMOTE-based class rebalancing under limited-label conditions in two clinical tabular classification tasks. The analysis addressed two complementary questions: whether uncertainty-based acquisition improves predictive performance compared with random acquisition, and whether SMOTE provides an additional contribution when uncertainty-based acquisition is already applied.

The main findings of this study can be summarized in the following key points:

- The results of the comparison between two acquisition strategies showed a clear dependence on the dataset, classifier, performance measure, and annotation budget, demonstrating that our study can be weakly generalized. With Random Forest, uncertainty-based acquisition yielded higher Balanced Accuracy than random acquisition on both HFCR and WBCD, with statistically supported differences at selected annotation budgets and over the complete acquisition trajectory; on WBCD, statistically supported improvements were also observed for Average Precision. With SVM, uncertainty-based acquisition showed a less consistent pattern on HFCR, where no statistically supported advantage over random acquisition was observed, whereas on WBCD higher Balanced Accuracy was obtained at the larger annotation budgets and over the complete acquisition trajectory.

- The additional contribution associated with SMOTE was also context dependent and limited under highly data-scarce and imbalanced conditions, as already discussed in Section 4.3. On HFCR, combining SMOTE with uncertainty-based acquisition produced numerical increases in Balanced Accuracy for both classifiers, although these differences were not statistically supported after correction for multiple comparisons. On WBCD, the differences between Uncertainty and Uncertainty + SMOTE were smaller and varied across classifiers, metrics, and annotation budgets.
- The inherent limitations of this study, arising from the use of a real-world clinical scenario characterized by limited data availability, resulted in performance improvements that, although consistently observed in some settings, did not always reach statistical significance. The choice of very small and imbalanced datasets is motivated by the fact that such situations occur very frequently in medical scenarios.
- Another important contribution of the article is the development of a reproducible data analytics pipeline to support the systematic evaluation of acquisition strategies and the contribution of SMOTE, providing a framework that can be readily replicated and extended to other datasets in future studies.

These findings indicate that the effect of class rebalancing is closely related to the acquisition strategy and the learning algorithm and cannot be characterized independently of the experimental configuration. The robustness of these observations was further examined across 20 matched experimental repetitions, using both Random Forest and SVM on both datasets, paired statistical testing with Holm correction, trajectory-level AULC analysis, classifier hyperparameter sensitivity, and a comparison between the main batch-mode acquisition protocol ($B = 10$) and sequential acquisition ($B = 1$). An additional class-weighted learning condition was also evaluated to provide an imbalance-aware control without synthetic oversampling.

We underline that the main goal of this paper is to investigate the interaction between active learning and synthetic data generation through SMOTE. Accordingly, we compared random sampling and uncertainty-based acquisition, both with and without oversampling, and included an additional class-weighted learning condition as an imbalance-aware control. However, dedicated imbalance-aware acquisition algorithms were not included in the evaluation. Comparison with such methods therefore remains a limitation of the present study and represents a relevant direction for future work.

The present analysis is also limited to two small public clinical datasets and two classifiers, and it evaluates the predictive effect of SMOTE-based rebalancing without directly characterizing the fidelity of the induced synthetic-data distribution. Future studies should extend this investigation to larger and more heterogeneous clinical datasets, incorporate dedicated imbalance-aware acquisition methods, and examine alternative acquisition, rebalancing, and synthetic-data generation strategies under different annotation regimes.

**Author Contributions:** Conceptualization, L.B. and A.T.; methodology, L.B. and A.T.; software, L.B.; validation, L.B. and M.C.; formal analysis, L.B.; investigation, L.B.; resources, L.B.; data curation, L.B.; writing—original draft preparation, L.B. ; writing—review and editing, L.B. and M.C.; visualization, L.B.; supervision, A.T. and M.C.; project administration, M.C.; funding acquisition, M.C. Co-first authorship is shared between L.B. and C.Z., who contributed equally to this research. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research has been partially supported by the "HeartTrack (Driving the Patient Journey in Heart Failure)" project funded by PR CALABRIA FESR FSE 2021–2027. This research has been partially supported by the OFIDIAPlus (Operational Fire Danger preventIon plAtform Plus) project funded by the Interreg VI-A Greece-Italy 2021-2027 programme.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The datasets used in this article are publicly available in the UCI repository and they can be downloaded here: HFCR dataset, available at https://doi.org/10.24432/C5Z89R; WBCD dataset, available at https://doi.org/10.24432/C5DW2B. Moreover, the synthetic datasets used in this article have been generated using Synthetic Minority Over-sampling Technique (SMOTE), implemented through the imbalanced-learn Python library, available at https://imbalanced-learn.org/stable/ (accessed on 5 November 2026). The synthetic dataset presented in this study is available upon request from the corresponding author.

**Acknowledgments:** This work was made when L.B. was visiting PhD Student at the Brunel University of London, as part of her research activities within her PhD program at University Magna Graecia of Catanzaro. L.B. wishes to thank Brunel University for the guidance and resources provided to complete this research, and the University Magna Graecia of Catanzaro for the support during the preparation and writing of this manuscript. During the preparation of this manuscript/study, the author(s) used Google Gemini 2.5. For the purposes of spellchecking and grammar. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AI | Artificial Intelligence |
| ML | Machine Learning |
| AL | Active Learning |
| HFCR | Heart Failure Clinical Records |
| WBCD | Breast Cancer Wisconsin (Diagnostic) |

## Appendix A. Classifier Hyperparameter Sensitivity

### *Appendix A.1. SVM Regularization*

The results in Table A1 show that, on HFCR, the mean BA difference associated with SMOTE was larger at $C = 0.1$ than at $C = 1$ or $C = 10$. Under uncertainty-based acquisition, final BA was 0.761 with SMOTE and 0.696 without SMOTE at $C = 0.1$; at $C = 10$, the corresponding values were 0.705 and 0.699. BA AULC showed the same tendency. AP followed a less uniform pattern: the final-budget difference favored SMOTE at $C = 0.1$, while its direction reversed at the other two settings.

**Table A1.** Sensitivity to SVM regularization on HFCR and WBCD. Values are mean ± sample standard deviation across 20 seeds. BA: balanced accuracy; AP: average precision. Normalized AULC is computed at $b = \{0, 10, 20, 30\}$, with acquisition batch size $B = 10$.

| $C$ | Strategy | At $b = 30$ | | Normalized AULC | |
|---|---|---|---|---|---|
| | | BA | AP | BA | AP |
| HFCR | | | | | |
| 0.1 | Random | $0.673 \pm 0.072$ | $0.712 \pm 0.086$ | $0.626 \pm 0.067$ | $0.653 \pm 0.111$ |
| | Random + SMOTE | $0.758 \pm 0.062$ | $0.716 \pm 0.094$ | $0.708 \pm 0.061$ | $0.681 \pm 0.100$ |
| | Uncertainty | $0.696 \pm 0.071$ | $0.716 \pm 0.084$ | $0.632 \pm 0.058$ | $0.638 \pm 0.103$ |
| | Uncertainty + SMOTE | $0.761 \pm 0.045$ | $0.762 \pm 0.077$ | $0.712 \pm 0.057$ | $0.696 \pm 0.091$ |
| 1 | Random | $0.713 \pm 0.071$ | $0.702 \pm 0.097$ | $0.677 \pm 0.067$ | $0.629 \pm 0.109$ |
| | Random + SMOTE | $0.729 \pm 0.076$ | $0.693 \pm 0.110$ | $0.693 \pm 0.070$ | $0.661 \pm 0.112$ |
| | Uncertainty | $0.712 \pm 0.064$ | $0.731 \pm 0.073$ | $0.661 \pm 0.050$ | $0.645 \pm 0.086$ |
| | Uncertainty + SMOTE | $0.729 \pm 0.071$ | $0.725 \pm 0.085$ | $0.687 \pm 0.072$ | $0.678 \pm 0.099$ |
| 10 | Random | $0.712 \pm 0.068$ | $0.672 \pm 0.102$ | $0.674 \pm 0.063$ | $0.626 \pm 0.109$ |
| | Random + SMOTE | $0.716 \pm 0.070$ | $0.676 \pm 0.111$ | $0.675 \pm 0.065$ | $0.641 \pm 0.110$ |
| | Uncertainty | $0.699 \pm 0.073$ | $0.707 \pm 0.100$ | $0.657 \pm 0.048$ | $0.642 \pm 0.085$ |
| | Uncertainty + SMOTE | $0.705 \pm 0.057$ | $0.697 \pm 0.084$ | $0.673 \pm 0.061$ | $0.661 \pm 0.098$ |

**Table A1.** *Cont.*

| *C* | Strategy | At $b = 30$ | | Normalized AULC | |
|---|---|---|---|---|---|
| | | BA | AP | BA | AP |
| WBCD | | | | | |
| 0.1 | Random | $0.936 \pm 0.022$ | $0.987 \pm 0.007$ | $0.927 \pm 0.022$ | $0.981 \pm 0.007$ |
| | Random + SMOTE | $0.941 \pm 0.023$ | $0.986 \pm 0.006$ | $0.932 \pm 0.019$ | $0.982 \pm 0.007$ |
| | Uncertainty | $0.961 \pm 0.019$ | $0.993 \pm 0.005$ | $0.942 \pm 0.019$ | $0.985 \pm 0.009$ |
| | Uncertainty + SMOTE | $0.962 \pm 0.024$ | $0.991 \pm 0.007$ | $0.946 \pm 0.016$ | $0.987 \pm 0.007$ |
| 1 | Random | $0.932 \pm 0.027$ | $0.980 \pm 0.012$ | $0.923 \pm 0.020$ | $0.976 \pm 0.011$ |
| | Random + SMOTE | $0.930 \pm 0.026$ | $0.979 \pm 0.013$ | $0.923 \pm 0.020$ | $0.976 \pm 0.011$ |
| | Uncertainty | $0.952 \pm 0.021$ | $0.987 \pm 0.009$ | $0.936 \pm 0.019$ | $0.978 \pm 0.014$ |
| | Uncertainty + SMOTE | $0.958 \pm 0.013$ | $0.989 \pm 0.007$ | $0.938 \pm 0.014$ | $0.982 \pm 0.010$ |
| 10 | Random | $0.926 \pm 0.024$ | $0.977 \pm 0.013$ | $0.920 \pm 0.017$ | $0.975 \pm 0.012$ |
| | Random + SMOTE | $0.926 \pm 0.024$ | $0.977 \pm 0.013$ | $0.920 \pm 0.017$ | $0.975 \pm 0.012$ |
| | Uncertainty | $0.945 \pm 0.021$ | $0.980 \pm 0.017$ | $0.934 \pm 0.018$ | $0.977 \pm 0.014$ |
| | Uncertainty + SMOTE | $0.955 \pm 0.023$ | $0.988 \pm 0.011$ | $0.937 \pm 0.013$ | $0.981 \pm 0.009$ |

On WBCD, the table shows higher mean final BA for uncertainty-based than for random acquisition across the evaluated values of *C*. The differences between Uncertainty and Uncertainty + SMOTE were smaller and varied with regularization and budget. Comparing the datasets therefore suggests that the magnitude of the rebalancing effect depends on the classifier setting and data considered. This analysis characterizes sensitivity within the evaluated grid and does not establish the optimality of the main settings.

*Appendix A.2. Random Forest Maximum Depth*

The results in Table A2 show a comparatively stable acquisition pattern across the evaluated depths. On both datasets, uncertainty-based acquisition retained higher mean final BA than its random counterpart. On HFCR, Uncertainty + SMOTE yielded final BA between 0.793 and 0.799. On WBCD, the BA AULC of Uncertainty ranged from 0.929 to 0.932. Results with maximum depth 10 and unrestricted depth were very similar under these settings.

**Table A2.** Sensitivity to Random Forest maximum tree depth on HFCR and WBCD. Values are mean ± sample standard deviation across 20 seeds. BA: balanced accuracy; AP: average precision. Normalized AULC is computed at $b = \{0, 10, 20, 30\}$, with acquisition batch size $B = 10$.

| Depth | Strategy | At $b = 30$ | | Normalized AULC | |
|---|---|---|---|---|---|
| | | BA | AP | BA | AP |
| HFCR | | | | | |
| 3 | Random | $0.703 \pm 0.082$ | $0.813 \pm 0.070$ | $0.643 \pm 0.070$ | $0.783 \pm 0.071$ |
| | Random + SMOTE | $0.744 \pm 0.065$ | $0.758 \pm 0.075$ | $0.677 \pm 0.060$ | $0.720 \pm 0.070$ |
| | Uncertainty | $0.787 \pm 0.075$ | $0.813 \pm 0.072$ | $0.710 \pm 0.063$ | $0.797 \pm 0.071$ |
| | Uncertainty + SMOTE | $0.793 \pm 0.063$ | $0.797 \pm 0.082$ | $0.720 \pm 0.060$ | $0.771 \pm 0.076$ |
| 5 | Random | $0.729 \pm 0.076$ | $0.807 \pm 0.069$ | $0.656 \pm 0.067$ | $0.778 \pm 0.069$ |
| | Random + SMOTE | $0.731 \pm 0.067$ | $0.758 \pm 0.076$ | $0.679 \pm 0.062$ | $0.719 \pm 0.075$ |
| | Uncertainty | $0.786 \pm 0.062$ | $0.815 \pm 0.064$ | $0.707 \pm 0.062$ | $0.795 \pm 0.071$ |
| | Uncertainty + SMOTE | $0.799 \pm 0.077$ | $0.806 \pm 0.077$ | $0.727 \pm 0.062$ | $0.777 \pm 0.077$ |
| 10 | Random | $0.727 \pm 0.073$ | $0.800 \pm 0.072$ | $0.655 \pm 0.066$ | $0.775 \pm 0.069$ |
| | Random + SMOTE | $0.726 \pm 0.063$ | $0.756 \pm 0.072$ | $0.677 \pm 0.062$ | $0.715 \pm 0.072$ |
| | Uncertainty | $0.781 \pm 0.071$ | $0.805 \pm 0.073$ | $0.708 \pm 0.064$ | $0.791 \pm 0.073$ |
| | Uncertainty + SMOTE | $0.793 \pm 0.066$ | $0.798 \pm 0.081$ | $0.729 \pm 0.058$ | $0.775 \pm 0.078$ |
| None | Random | $0.727 \pm 0.073$ | $0.800 \pm 0.072$ | $0.655 \pm 0.066$ | $0.775 \pm 0.069$ |
| | Random + SMOTE | $0.726 \pm 0.063$ | $0.756 \pm 0.072$ | $0.677 \pm 0.062$ | $0.715 \pm 0.072$ |
| | Uncertainty | $0.781 \pm 0.071$ | $0.805 \pm 0.073$ | $0.708 \pm 0.064$ | $0.791 \pm 0.073$ |
| | Uncertainty + SMOTE | $0.793 \pm 0.066$ | $0.798 \pm 0.081$ | $0.729 \pm 0.058$ | $0.775 \pm 0.078$ |

**Table A2.** *Cont.*

| Depth | Strategy | At $b = 30$ | | Normalized AULC | |
|---|---|---|---|---|---|
| | | BA | AP | BA | AP |
| WBCD | | | | | |
| 3 | Random | $0.914 \pm 0.035$ | $0.980 \pm 0.014$ | $0.906 \pm 0.029$ | $0.977 \pm 0.013$ |
| | Random + SMOTE | $0.924 \pm 0.040$ | $0.980 \pm 0.016$ | $0.914 \pm 0.029$ | $0.976 \pm 0.012$ |
| | Uncertainty | $0.944 \pm 0.035$ | $0.988 \pm 0.010$ | $0.929 \pm 0.028$ | $0.984 \pm 0.010$ |
| | Uncertainty + SMOTE | $0.950 \pm 0.028$ | $0.990 \pm 0.008$ | $0.927 \pm 0.026$ | $0.982 \pm 0.008$ |
| 5 | Random | $0.917 \pm 0.035$ | $0.980 \pm 0.014$ | $0.906 \pm 0.030$ | $0.977 \pm 0.013$ |
| | Random + SMOTE | $0.923 \pm 0.041$ | $0.979 \pm 0.017$ | $0.914 \pm 0.029$ | $0.975 \pm 0.013$ |
| | Uncertainty | $0.954 \pm 0.030$ | $0.990 \pm 0.010$ | $0.931 \pm 0.029$ | $0.984 \pm 0.011$ |
| | Uncertainty + SMOTE | $0.947 \pm 0.027$ | $0.989 \pm 0.007$ | $0.927 \pm 0.025$ | $0.982 \pm 0.008$ |
| 10 | Random | $0.917 \pm 0.035$ | $0.980 \pm 0.014$ | $0.906 \pm 0.030$ | $0.977 \pm 0.013$ |
| | Random + SMOTE | $0.923 \pm 0.041$ | $0.979 \pm 0.017$ | $0.914 \pm 0.029$ | $0.975 \pm 0.013$ |
| | Uncertainty | $0.957 \pm 0.028$ | $0.989 \pm 0.011$ | $0.932 \pm 0.029$ | $0.984 \pm 0.011$ |
| | Uncertainty + SMOTE | $0.946 \pm 0.025$ | $0.989 \pm 0.007$ | $0.927 \pm 0.025$ | $0.982 \pm 0.008$ |
| None | Random | $0.917 \pm 0.035$ | $0.980 \pm 0.014$ | $0.906 \pm 0.030$ | $0.977 \pm 0.013$ |
| | Random + SMOTE | $0.923 \pm 0.041$ | $0.979 \pm 0.017$ | $0.914 \pm 0.029$ | $0.975 \pm 0.013$ |
| | Uncertainty | $0.957 \pm 0.028$ | $0.989 \pm 0.011$ | $0.932 \pm 0.029$ | $0.984 \pm 0.011$ |
| | Uncertainty + SMOTE | $0.946 \pm 0.025$ | $0.989 \pm 0.007$ | $0.927 \pm 0.025$ | $0.982 \pm 0.008$ |

The incremental contribution of SMOTE remained metric dependent. Uncertainty without SMOTE yielded higher mean AP AULC than Uncertainty + SMOTE at every depth on both datasets. For WBCD, the sign of the final BA difference between these two conditions also changed across depth settings. Thus, the tables suggest greater stability of the uncertainty–random comparison than of an additional SMOTE benefit.

## Appendix B. Sensitivity to Acquisition Batch Size

Table A3 compares sequential acquisition ($B = 1$) with the main batch setting ($B = 10$), at the same maximum budget of 30 additional real labels. With $B = 1$, the model is refitted and the pool is rescored after each acquired observation. All nine initial metrics coincided between batch sizes for each matched seed and condition.

At $b = 0$, no additional real labels have been acquired, but SMOTE has already been applied in the rebalanced conditions. Random and Uncertainty therefore start together within each rebalancing condition; the initial values with and without SMOTE may differ. AULC includes these initial values and summarizes the complete acquisition–training procedure. The comparison uses the common grid $b = \{0, 10, 20, 30\}$, while the figures show all 31 sequential reporting points, averaged over the 20 seeds.

**Table A3.** Acquisition batch-size comparison. Values are mean ± sample standard deviation across 20 matched seeds. Main classifier settings are used for both batch sizes. Normalized AULC is computed on the common grid $b = \{0, 10, 20, 30\}$. BA: balanced accuracy; AP: average precision.

| Classifier | Strategy | *B* | At $b = 30$ | | Normalized AULC | |
|---|---|---|---|---|---|---|
| | | | BA | AP | BA | AP |
| HFCR | | | | | | |
| RF | Random | 10 | $0.727 \pm 0.073$ | $0.800 \pm 0.072$ | $0.655 \pm 0.066$ | $0.775 \pm 0.069$ |
| | | 1 | $0.725 \pm 0.053$ | $0.821 \pm 0.045$ | $0.635 \pm 0.061$ | $0.779 \pm 0.063$ |
| | Random + SMOTE | 10 | $0.726 \pm 0.063$ | $0.756 \pm 0.072$ | $0.677 \pm 0.062$ | $0.715 \pm 0.072$ |
| | | 1 | $0.725 \pm 0.071$ | $0.760 \pm 0.084$ | $0.662 \pm 0.058$ | $0.714 \pm 0.079$ |
| | Uncertainty | 10 | $0.781 \pm 0.071$ | $0.805 \pm 0.073$ | $0.708 \pm 0.064$ | $0.791 \pm 0.073$ |
| | | 1 | $0.795 \pm 0.062$ | $0.802 \pm 0.089$ | $0.717 \pm 0.057$ | $0.789 \pm 0.074$ |
| | Uncertainty + SMOTE | 10 | $0.793 \pm 0.066$ | $0.798 \pm 0.081$ | $0.729 \pm 0.058$ | $0.775 \pm 0.078$ |
| | | 1 | $0.806 \pm 0.055$ | $0.797 \pm 0.066$ | $0.730 \pm 0.049$ | $0.765 \pm 0.060$ |

**Table A3.** *Cont.*

| Classifier | Strategy | $B$ | At $b = 30$ | | Normalized AULC | |
|---|---|---|---|---|---|---|
| | | | BA | AP | BA | AP |
| SVM | Random | 10 | $0.713 \pm 0.071$ | $0.702 \pm 0.097$ | $0.677 \pm 0.067$ | $0.629 \pm 0.109$ |
| | | 1 | $0.725 \pm 0.058$ | $0.711 \pm 0.087$ | $0.661 \pm 0.060$ | $0.598 \pm 0.123$ |
| | Random + SMOTE | 10 | $0.729 \pm 0.076$ | $0.693 \pm 0.110$ | $0.693 \pm 0.070$ | $0.661 \pm 0.112$ |
| | | 1 | $0.721 \pm 0.052$ | $0.683 \pm 0.092$ | $0.668 \pm 0.055$ | $0.634 \pm 0.100$ |
| | Uncertainty | 10 | $0.712 \pm 0.064$ | $0.731 \pm 0.073$ | $0.661 \pm 0.050$ | $0.645 \pm 0.086$ |
| | | 1 | $0.684 \pm 0.078$ | $0.700 \pm 0.106$ | $0.655 \pm 0.062$ | $0.637 \pm 0.121$ |
| | Uncertainty + SMOTE | 10 | $0.729 \pm 0.071$ | $0.725 \pm 0.085$ | $0.687 \pm 0.072$ | $0.678 \pm 0.099$ |
| | | 1 | $0.710 \pm 0.076$ | $0.707 \pm 0.075$ | $0.674 \pm 0.061$ | $0.663 \pm 0.092$ |
| WBCD | | | | | | |
| RF | Random | 10 | $0.917 \pm 0.035$ | $0.980 \pm 0.014$ | $0.906 \pm 0.030$ | $0.977 \pm 0.013$ |
| | | 1 | $0.924 \pm 0.041$ | $0.983 \pm 0.011$ | $0.908 \pm 0.034$ | $0.977 \pm 0.012$ |
| | Random + SMOTE | 10 | $0.923 \pm 0.041$ | $0.979 \pm 0.017$ | $0.914 \pm 0.029$ | $0.975 \pm 0.013$ |
| | | 1 | $0.924 \pm 0.036$ | $0.983 \pm 0.011$ | $0.911 \pm 0.032$ | $0.976 \pm 0.012$ |
| | Uncertainty | 10 | $0.957 \pm 0.028$ | $0.989 \pm 0.011$ | $0.932 \pm 0.029$ | $0.984 \pm 0.011$ |
| | | 1 | $0.954 \pm 0.025$ | $0.989 \pm 0.009$ | $0.934 \pm 0.027$ | $0.986 \pm 0.009$ |
| | Uncertainty + SMOTE | 10 | $0.946 \pm 0.025$ | $0.989 \pm 0.007$ | $0.927 \pm 0.025$ | $0.982 \pm 0.008$ |
| | | 1 | $0.957 \pm 0.026$ | $0.991 \pm 0.008$ | $0.939 \pm 0.022$ | $0.985 \pm 0.010$ |
| SVM | Random | 10 | $0.932 \pm 0.027$ | $0.980 \pm 0.012$ | $0.923 \pm 0.020$ | $0.976 \pm 0.011$ |
| | | 1 | $0.924 \pm 0.033$ | $0.978 \pm 0.018$ | $0.920 \pm 0.027$ | $0.971 \pm 0.020$ |
| | Random + SMOTE | 10 | $0.930 \pm 0.026$ | $0.979 \pm 0.013$ | $0.923 \pm 0.020$ | $0.976 \pm 0.011$ |
| | | 1 | $0.924 \pm 0.033$ | $0.978 \pm 0.018$ | $0.920 \pm 0.027$ | $0.971 \pm 0.020$ |
| | Uncertainty | 10 | $0.952 \pm 0.021$ | $0.987 \pm 0.009$ | $0.936 \pm 0.019$ | $0.978 \pm 0.014$ |
| | | 1 | $0.958 \pm 0.020$ | $0.991 \pm 0.007$ | $0.941 \pm 0.017$ | $0.979 \pm 0.013$ |
| | Uncertainty + SMOTE | 10 | $0.958 \pm 0.013$ | $0.989 \pm 0.007$ | $0.938 \pm 0.014$ | $0.982 \pm 0.010$ |
| | | 1 | $0.960 \pm 0.015$ | $0.992 \pm 0.006$ | $0.944 \pm 0.019$ | $0.982 \pm 0.013$ |

The original Random implementation was retained, so shared seeds do not impose identical acquired subsets across batch sizes. Differences in the Random conditions can also reflect the particular observations sampled.

*Appendix B.1. HFCR: Random Forest*

Comparing the learning curves in Figure A1, both uncertainty-based conditions show a more rapid early increase in mean BA than the random conditions, followed by a smaller separation between Uncertainty and Uncertainty + SMOTE at later budgets. SMOTE is associated with higher BA during part of the early trajectory, but the two uncertainty-based curves cross or approach each other as acquisition progresses. AP shows a different pattern: Uncertainty generally remains above Uncertainty + SMOTE, while Random + SMOTE stays below Random over most of the trajectory. The final values and AULCs in Table A3 therefore suggest a persistent BA advantage of uncertainty-based acquisition under both batch sizes, without a corresponding uniform AP benefit from SMOTE.

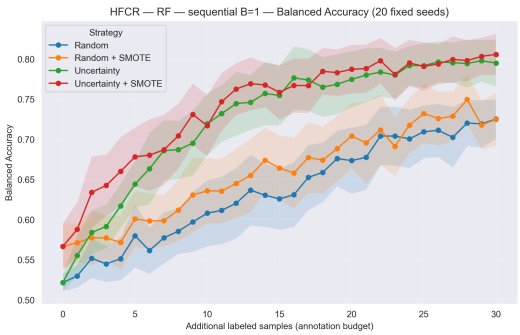



**Figure A1.** *Cont.*

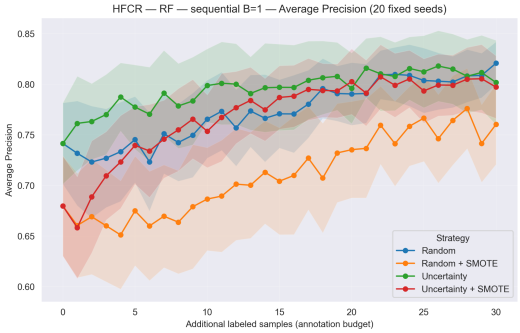



**Figure A1.** HFCR: sequential acquisition ($B = 1$) with Random Forest. BA is shown above and AP below. Each point is the mean metric across 20 seeds at the indicated cumulative budget, including initialization at $b = 0$; shaded bands show pointwise 95% $t$-intervals for the mean. All budgets from 0 to 30 are shown. SMOTE is applied before the initial fit in the conditions that include it.

### *Appendix B.2. HFCR: SVM*

The curves in Figure A2 show a less consistent uncertainty–random separation than for RF. For BA, Uncertainty + SMOTE lies above Uncertainty for most post-initialization budgets. The random conditions, however, approach or exceed the uncertainty-based conditions at later budgets, and Random has the highest mean final BA. For AP, Uncertainty + SMOTE remains above the other conditions for much of the trajectory, but the curves become closer near the final budget. Its higher initial AP is already present before acquisition. Taken together with Table A3, these trends suggest metric- and budget-dependent differences and do not support a general claim that sequential uncertainty sampling improves on random acquisition on HFCR.

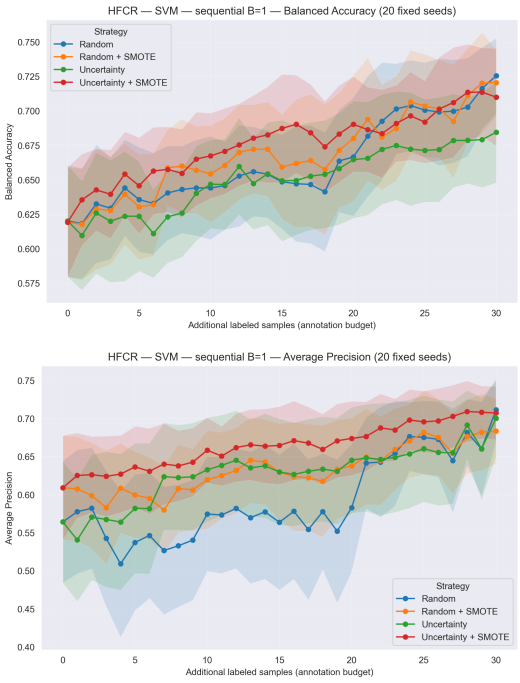



**Figure A2.** HFCR: sequential acquisition ($B = 1$) with SVM. BA is shown above and AP below. Each point is the mean metric across 20 seeds at the indicated cumulative budget, including initialization at $b = 0$; shaded bands show pointwise 95% $t$-intervals for the mean. All budgets from 0 to 30 are shown. SMOTE is applied before the initial fit in the conditions that include it.

*Appendix B.3. WBCD: Random Forest*

Figure A3 shows that the uncertainty-based curves separate from the random curves as the sequential budget increases, for both BA and AP. Random and Random + SMOTE remain close, whereas Uncertainty and Uncertainty + SMOTE alternate in their relative position. At larger budgets, the uncertainty-based curves are near one another and show smaller successive changes. Comparing these trends with Table A3 suggests that the qualitative uncertainty–random pattern persists under both batch sizes. The effect of SMOTE is smaller than the acquisition-strategy separation and does not show to have a consistent direction across budgets and measures.

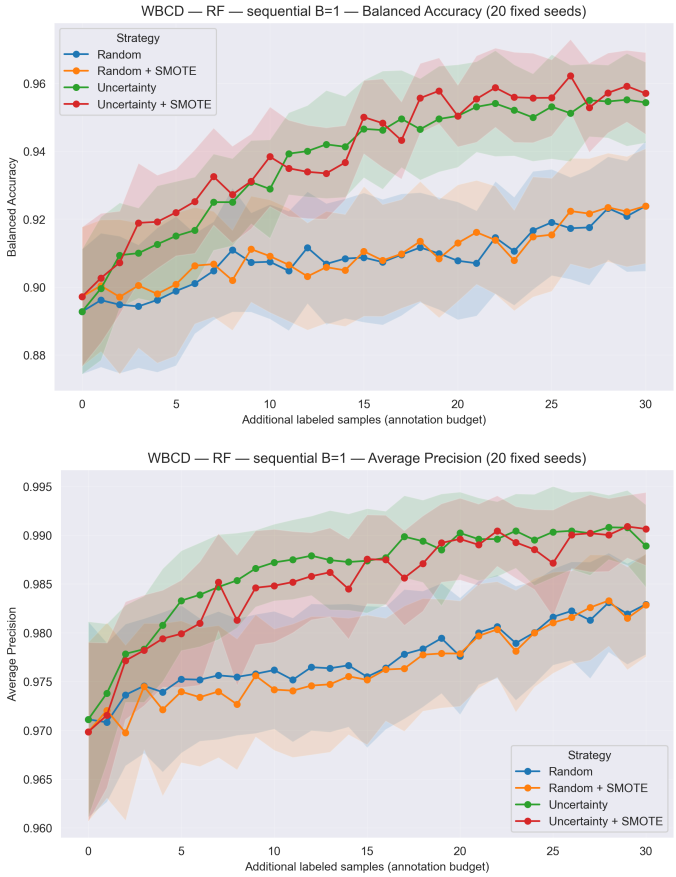



**Figure A3.** WBCD: sequential acquisition ($B = 1$) with Random Forest. BA is shown above and AP below. Each point is the mean metric across 20 seeds at the indicated cumulative budget, including initialization at $b = 0$; shaded bands show pointwise 95% $t$-intervals for the mean. All budgets from 0 to 30 are shown. SMOTE is applied before the initial fit in the conditions that include it.

*Appendix B.4. WBCD: SVM*

The learning curves in Figure A4 show a progressive separation between uncertainty-based and random acquisition. Random and Random + SMOTE are closely aligned for both measures. The two uncertainty-based conditions remain close, with Uncertainty + SMOTE showing slightly higher means over part of the later trajectory. The differences between these two curves are smaller than those separating them from the random conditions. Table A3 shows the same qualitative BA pattern under $B = 10$. The sequential results therefore suggest that the acquisition pattern is retained, while the incremental contribution of SMOTE remains comparatively small.

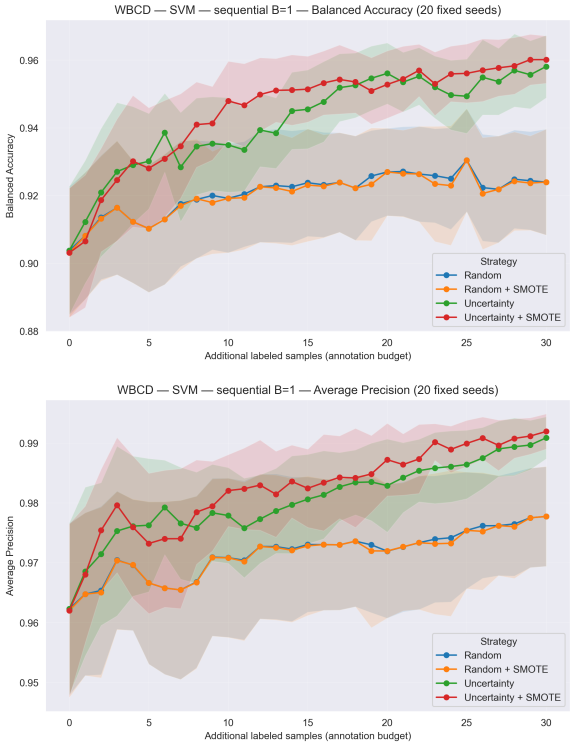



**Figure A4.** WBCD: sequential acquisition ($B = 1$) with SVM. BA is shown above and AP below. Each point is the mean metric across 20 seeds at the indicated cumulative budget, including initialization at $b = 0$; shaded bands show pointwise 95% $t$-intervals for the mean. All budgets from 0 to 30 are shown. SMOTE is applied before the initial fit in the conditions that include it.

## Appendix C. Comparison with Class-Weighted Training

Table A4 adds uncertainty sampling with class-weighted training (`class_weight='balanced'`) and without synthetic oversampling. The condition uses the same 20 initial partitions, the main classifier settings, and $B = 10$. The least-confidence acquisition rule is retained; the modification concerns classifier fitting. The comparison below is descriptive.

**Table A4.** Uncertainty sampling with no rebalancing, with SMOTE, and with class-weighted training. Values are mean ± sample standard deviation across 20 seeds, using $B = 10$, $C = 1$ for SVM, and unrestricted depth for RF. BA: balanced accuracy; AP: average precision. Normalized AULC uses $b = \{0, 10, 20, 30\}$.

| Classifier | Strategy | At $b = 30$ | | Normalized AULC | |
|---|---|---|---|---|---|
| | | BA | AP | BA | AP |
| HFCR | | | | | |
| RF | Uncertainty | $0.781 \pm 0.071$ | $0.805 \pm 0.073$ | $0.708 \pm 0.064$ | $0.791 \pm 0.073$ |
| | Uncertainty + SMOTE | $0.793 \pm 0.066$ | $0.798 \pm 0.081$ | $0.729 \pm 0.058$ | $0.775 \pm 0.078$ |
| | Uncertainty + ClassWeight | $0.770 \pm 0.086$ | $0.798 \pm 0.069$ | $0.690 \pm 0.065$ | $0.769 \pm 0.084$ |
| SVM | Uncertainty | $0.712 \pm 0.064$ | $0.731 \pm 0.073$ | $0.661 \pm 0.050$ | $0.645 \pm 0.086$ |
| | Uncertainty + SMOTE | $0.729 \pm 0.071$ | $0.725 \pm 0.085$ | $0.687 \pm 0.072$ | $0.678 \pm 0.099$ |
| | Uncertainty + ClassWeight | $0.707 \pm 0.063$ | $0.729 \pm 0.073$ | $0.655 \pm 0.054$ | $0.635 \pm 0.085$ |
| WBCD | | | | | |
| RF | Uncertainty | $0.957 \pm 0.028$ | $0.989 \pm 0.011$ | $0.932 \pm 0.029$ | $0.984 \pm 0.011$ |
| | Uncertainty + SMOTE | $0.946 \pm 0.025$ | $0.989 \pm 0.007$ | $0.927 \pm 0.025$ | $0.982 \pm 0.008$ |
| | Uncertainty + ClassWeight | $0.955 \pm 0.029$ | $0.990 \pm 0.012$ | $0.931 \pm 0.027$ | $0.984 \pm 0.012$ |
| SVM | Uncertainty | $0.952 \pm 0.021$ | $0.987 \pm 0.009$ | $0.936 \pm 0.019$ | $0.978 \pm 0.014$ |
| | Uncertainty + SMOTE | $0.958 \pm 0.013$ | $0.989 \pm 0.007$ | $0.938 \pm 0.014$ | $0.982 \pm 0.010$ |
| | Uncertainty + ClassWeight | $0.953 \pm 0.019$ | $0.986 \pm 0.009$ | $0.935 \pm 0.018$ | $0.978 \pm 0.013$ |

On HFCR, the results show higher mean BA with SMOTE than with class weighting for both classifiers, at the final budget and over the complete trajectory. For RF, BA AULC was 0.729 with SMOTE and 0.690 with class weighting; for SVM it was 0.687 and 0.655, respectively. AP shows a different pattern: the final RF means were nearly equal, and the final SVM mean was slightly higher with class weighting, although its AP AULC was lower than with SMOTE.

On WBCD, class weighting remained close to Uncertainty without rebalancing. For RF, both conditions yielded slightly higher mean BA than Uncertainty + SMOTE; with SVM, SMOTE yielded slightly higher mean BA and AP than the other two conditions. Comparing the datasets suggests that the relative behavior of class weighting and synthetic rebalancing seems to be learner, dataset, and measure dependent.

Overall, the comparison suggests a numerical advantage of SMOTE over class-weighted training within the evaluated settings, particularly for balanced accuracy on HFCR. SMOTE yielded higher final balanced accuracy and normalized balanced-accuracy AULC in three of the four dataset–classifier combinations. On HFCR, this pattern was observed with both classifiers, whereas the direction of the differences on WBCD depended on the classifier. Average precision showed a less uniform pattern, particularly at the final annotation budget.

## References


1. Naeem, S.; Ali, A.; Anam, S.; Ahmed, M.M. An unsupervised machine learning algorithms: Comprehensive review. *Int. J. Comput. Digit. Syst.* **2023**, *13*, 911–921.
2. Myakala, P.K. Scaling ai with limited labeled data: A self-supervised learning approach. *ICCK Trans. Emerg. Top. Artif. Intell.* **2025**, *2*, 26–35.
3. Tharwat, A.; Schenck, W. A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics* **2023**, *11*, 820.
4. Jahan, S.; Islam, M.R.; Hasib, K.M.; Naseem, U.; Islam, M.S. Active learning with an adaptive classifier for inaccessible big data analysis. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–7.
5. Li, X.; Wang, X.; Chen, X.; Lu, Y.; Fu, H.; Wu, Y.C. Unlabeled data selection for active learning in image classification. *Sci. Rep.* **2024**, *14*, 424.
6. Lu, Y.; Chen, L.; Zhang, Y.; Shen, M.; Wang, H.; Wang, X.; van Rechem, C.; Fu, T.; Wei, W. Machine learning for synthetic data generation: a review. *arXiv* **2023**, arXiv:2302.04062.
7. Settles, B. From Theories to Queries: Active Learning in Practice. In Proceedings of the Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010, Sardinia, Italy, 16 May 2010; Proceedings of Machine Learning Research; PMLR: Brookline, MA, USA, 2011; Volume 16, pp. 1–18.
8. Nguyen, V.L.; Shaker, M.H.; Hüllermeier, E. How to Measure Uncertainty in Uncertainty Sampling for Active Learning. *Mach. Learn.* **2022**, *111*, 89–122. https://doi.org/10.1007/s10994-021-06003-9.
9. Lewis, D.D.; Gale, W.A. A Sequential Algorithm for Training Text Classifiers. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 3–6 July 1994; pp. 3–12. https://doi.org/10.1007/978-1-4471-2099-5_1.
10. Ertekin, S.; Huang, J.; Bottou, L.; Giles, C.L. Learning on the Border: Active Learning in Imbalanced Data Classification. In Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, New York, NY, USA, 6–10 November 2007; pp. 127–136. https://doi.org/10.1145/1321440.1321461.
11. Bengar, J.Z.; van de Weijer, J.; Fuentes, L.L.; Raducanu, B. Class-balanced active learning for image classification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 1536–1545.
12. Jin, Q.; Yuan, M.; Wang, H.; Wang, M.; Song, Z. Deep Active Learning Models for Imbalanced Image Classification. *Knowl.-Based Syst.* **2022**, *257*, 109817. https://doi.org/10.1016/j.knosys.2022.109817.
13. Fairstein, Y.; Kalinsky, O.; Karnin, Z.; Kushilevitz, G.; Libov, A.; Tolmach, S. Class balancing for efficient active learning in imbalanced datasets. In Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII), St. Julian's, Malta, 22 March 2024; pp. 77–86.
14. Kwon, B.A.; Kang, K. Active Learning Strategies for the Class Imbalance Problem. *IEEE Access* **2026**, *14*, 74081–74091. https://doi.org/10.1109/ACCESS.2026.3693476.

15. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. https://doi.org/10.1613/jair.953.
16. Elreedy, D.; Atiya, A.F.; Kamalov, F. A Theoretical Distribution Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Imbalanced Learning. *Mach. Learn.* **2024**, *113*, 4903–4923. https://doi.org/10.1007/s10994-022-06296-4.
17. Doyle, S.; Monaco, J.; Feldman, M.; Tomaszewski, J.; Madabhushi, A. An Active Learning Based Classification Strategy for the Minority Class Problem: Application to Histopathology Annotation. *BMC Bioinform.* **2011**, *12*, 424. https://doi.org/10.1186/1471-2105-12-424.
18. Nuggehalli, S.; Zhang, J.; Jain, L.K.; Nowak, R.D. Improved Algorithm for Deep Active Learning under Imbalance via Optimal Separation. In Proceedings of the 42nd International Conference on Machine Learning, Vancouver, QC, Canada, 13–19 July 2025; Proceedings of Machine Learning Research; PMLR: Brookline, MA, USA, 2025; Volume 267, pp. 46815–46836.
19. Gabriel, R.M.; Zandehshahvar, M.; van Assen, M.; Kittisut, N.; Peters, K.; De Cecco, C.N.; Adibi, A. Deep Active Learning for Lung Disease Severity Classification from Chest X-rays: Learning with Less Data in the Presence of Class Imbalance. *J. Imaging Inform. Med.* **2025**, 1–12. https://doi.org/10.1007/s10278-025-01752-8.
20. Mi, Y. Imbalanced Classification Based on Active Learning SMOTE. *Res. J. Appl. Sci. Eng. Technol.* **2013**, *5*, 944–949. https://doi.org/10.19026/rjaset.5.5044.
21. Deng, J.; Sun, J.; Peng, W.; Zhang, D.; Vyatkin, V. Imbalanced multiclass classification with active learning in strip rolling process. *Knowl.-Based Syst.* **2022**, *255*, 109754.
22. Sena, R.; Ben Hamida, S. ACTIVE SMOTE for Imbalanced Medical Data Classification. In *Proceedings of the Advances in Information Systems, Artificial Intelligence and Knowledge Management*; Lecture Notes in Business Information Processing; Springer: Cham, Switzerland, 2024; Volume 486, pp. 81–97. https://doi.org/10.1007/978-3-031-51664-1_6.
23. Islam, S.; Alfred, M.; Wilson, D.; Cohen, E. Evaluating Active Learning Strategies for Automated Classification of Patient Safety Event Reports in Hospitals. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2024**, *68*, 465–472. https://doi.org/10.1177/10711813241260676.
24. Sağlam, F.; Cengiz, M.A. Uncertainty Guided Oversampling: A Classifier-Specific Approach for Effective Handling of Class Imbalance in Machine Learning. *Knowl.-Based Syst.* **2026**, *342*, 115886. https://doi.org/10.1016/j.knosys.2026.115886.
25. Moles, L.; Andres, A.; Echegaray, G.; Boto, F. Exploring Data Augmentation and Active Learning Benefits in Imbalanced Datasets. *Mathematics* **2024**, *12*, 1898. https://doi.org/10.3390/math12121898.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.