Appendix A: Analysis of Uncertainty Measurement Behaviours in Active Learning for Binary Classification

There are three main approaches for uncertainty sampling in active learning. However, in a binary classification setting (which is what we use) these approaches perform identically to each other. We explain the different approaches here. Figure 1 shows the behaviour of these uncertainty sampling methods graphically.

We implement 'Least Confidence' in our code.

- *Least Confidence*: for a given input x and an output label $\hat{y}$, we can measure the posterior probability $P(\hat{y}|x;\theta)$ of observing $\hat{y}$ given x via the current model (parameterised by θ). The Least Confidence method selects data points x^* with the smallest maximum posterior probability across all labels:
$$x^* = \operatorname*{argmin}_x \max_{\hat{y}} P(\hat{y}|x;\theta) \tag{1}$$
- *Margin-based*: this approach takes the two highest posterior probability values for each input data point x and calculates their difference. The smaller the difference, the less certain the model is about its prediction and vice versa. More formally, let $\hat{y_1}$ and $\hat{y_2}$ the output labels with the highest and second-highest posterior probabilities for a given input x, respectively, the queried points x^* are chosen as:
$$x^* = \operatorname*{argmin}_x P(\hat{y_1}|x;\theta) - P(\hat{y_2}|x;\theta) \tag{2}$$
- *Entropy-based*: this approach takes into account the posterior probability values across *all* output classes. The idea is to select the data points x^* where there is a high entropy among the predicted output labels:
$$x^* = \operatorname*{argmax}_x - \sum_i P(\hat{y}|x;\theta) \log P(\hat{y}|x;\theta) \tag{3}$$

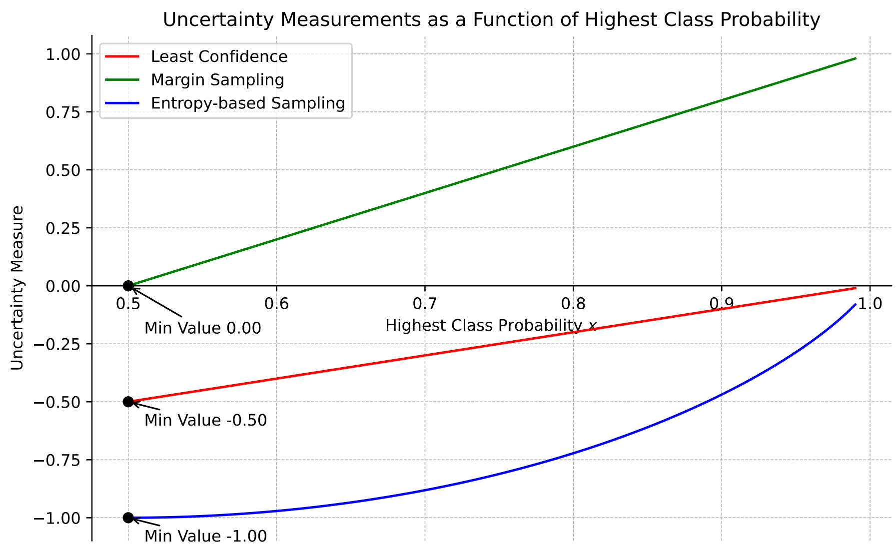

Figure 1: Uncertainty measurements as a function of the highest class probability. The red curve represents the Least Confidence uncertainty (LC) calculated as $LC = x - 1$, the green curve denotes Margin Sampling (MS) using the formula $MS = x - (1 - x)$, and the blue curve illustrates the Entropy-based method ($H(x) = -[x \log_2(x) + (1 - x)\log_2(1 - x)]$). Critical minimum values for each method are marked with black circles and annotated to emphasise the points where the uncertainty function is minimised.

Appendix B: Performance of 8 individual options

Figure 2 illustrates a side-by-side comparison of the following eight active learning strategies in binary classification without aggregation across configurations:

- Uncertainty Sampling (Baseline)
- Uncertainty Sampling with Timeout Predictor (TO)
- Uncertainty Sampling with Dynamic Timeout (DT)
- Uncertainty Sampling with Timeout Predictor and Dynamic Timeout (TO+DT)
- Random Sampling (Baseline)
- Random Sampling with Timeout Predictor (TO)
- Random Sampling with Dynamic Timeout (DT)
- Random Sampling with Timeout Predictor and Dynamic Timeout (TO+DT)

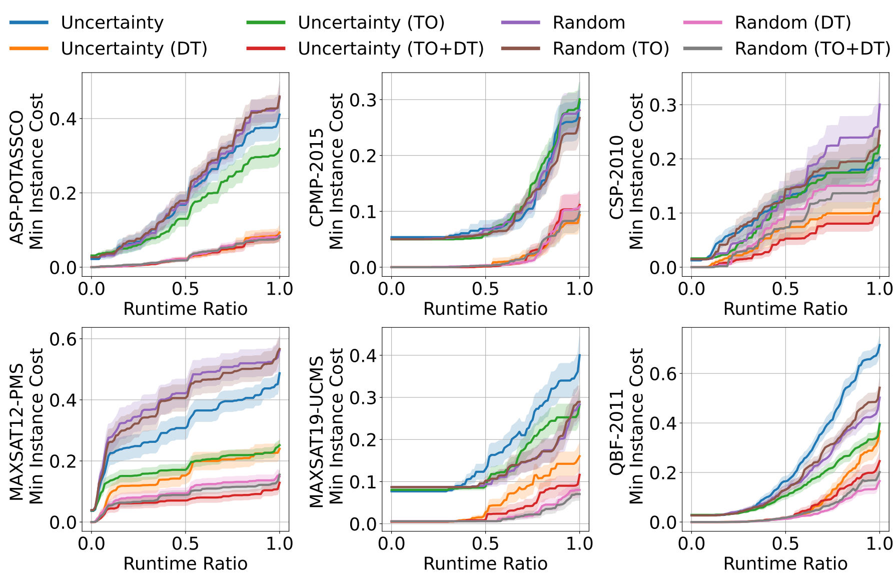

Figure 2: Comparison of performance across eight configurations as described in the paper. Each configuration was normalized according to the passive learning prediction performance ratio.

Appendix C: Description Table of Selected Datasets

Dataset	Instances	Algorithms	Features	Total Time	VBS	SBS
ASP-POTASSCO	1294	11	138	2,085h	8h	112h
CPMP-2015	527	4	22	682h	33h	134h
CSP-2010	2024	2	86	435h	49h	82h
MAXSAT12-PMS	876	6	37	1,472h	8h	85h
MAXSAT19-UCMS	572	7	54	545h	20h	52h
QBF-2011	1368	5	46	352h	28h	300h

Table 1: Descriptive statistics of selected datasets. Times rounded to the nearest whole number.

Appendix D: Timeout (TO) Configuration Impact on Passive Learning

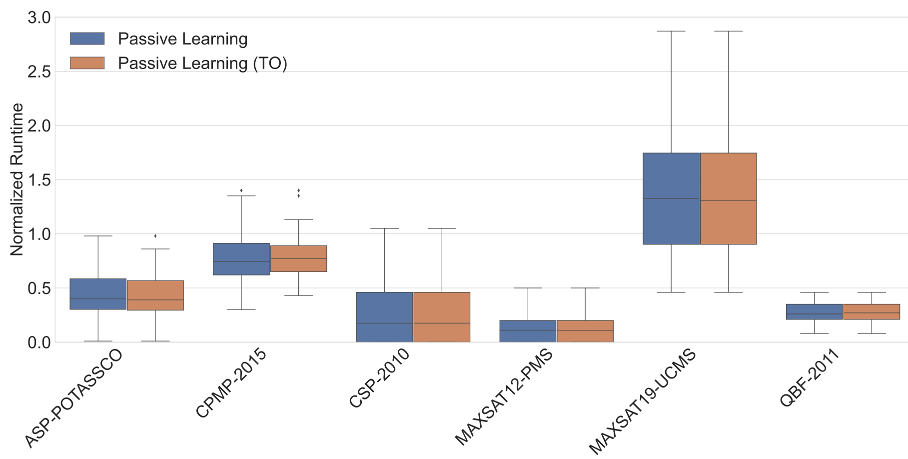

Figure 3: Comparison of Timeout (TO) Configuration Impact on Passive Learning: The graph illustrates that implementing the TO configuration in passive learning on the test set does not significantly enhance performance, yet importantly, it does not compromise prediction accuracy either.