Efficient Failure Pattern Identification of Predictive Algorithms (Supplementary material)

Bao Nguyen[1,2] **Viet Anh Nguyen**[3]

[1]School of Information and Communication Technology, Hanoi University of Science and Technology, Vietnam
[2]College of Engineering & Computer Science, VinUni-Illinois Smart Health Center, VinUniversity, Vietnam
[3]The Chinese University of Hong Kong

A ADDITIONAL EXPERIMENTS

A.1 MOTIVATION FOR THE SPECIFIC USER'S DEFINED FAILURE MODE DEFINITION

In this section, we provide the motivation, theoretical justification, and practical effectiveness of the failure mode definition based on mutual nearest neighbor graph[1] on embedding space.

- We make a similar assumption with d'Eon et al. [2022] and Sohoni et al. [2020] that the classifier's activations layer contains essential information about the semantic features used for classification. The proximity between two points in this embedding space could indicate their semantic similarity. Hence, issuing an edge between two points as in the mutual nearest neighbor graph likely guarantees that two connected points have much more semantic similarity than other pairs. This would ensure semantic cohesion for the points within a failure mode according to our definition.
- Regarding the theoretical aspect, we use the mutual nearest neighbor graph, which is effective in clustering and outliers detection (see Song et al. [2022b], Song et al. [2022a] and Brito et al. [1997]). Moreover, Brito et al. [1997, Theorem 2.2] stated that with the reasonable choice of k_{nn}, connected components (a.k.a. maximally connected subgraphs) in a mutual k_{nn}-graph are consistent for the identification of its clustering structure.
- In terms of more visual representation, we show images of four failure patterns of dataset id_1 in Figure A.1 to show the effectiveness of this definition on detecting semantic-cohesion clusters. We can observe that each failure pattern has a common concept recognizable by humans and includes images that are all misclassified. The top-left mode includes images of blonde-haired girls with tanned skin. The top-right mode includes images of girls wearing earrings. The bottom-left mode contains photos with tilted angles, and the bottom-right mode contains images with dark backgrounds.

A.2 DATASETS AND IMPLEMENTATION DETAILS

We describe fifteen datasets used in our work in Table 1.

Preprocessing: We single out 15 datasets from Eyuboglu et al. [2022], each includes three features: Activation, True Label, and Pseudo Label. After that, we preprocess the data using a standard scaler for the Activation feature.

Ground truth generation: It is necessary to assign values of k_{nn} and M to each preprocessed dataset. The value of M expresses the level of evidence required for confirming the failure patterns. A higher value of M indicates a greater emphasis on the patterns that exist most frequently in the dataset. As M decreases to 1, the problem transforms into identifying misclassified data points, where each failure data point constitutes a pattern. Moreover, the users choose M so that they can perceive the shared concept of M samples. If M is too small, then the concept may not be distinctive enough between clusters, while if M is too large, the users may have a bottleneck in identifying the shared concept. The value of k_{nn} signifies

[1] A mutual k_{nn}-nearest neighbor graph is a graph where there is an edge between x_i and x_j if x_i is one of the k_{nn} nearest neighbors of x_j and x_j is one of the k_{nn} nearest neighbors of x_i.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

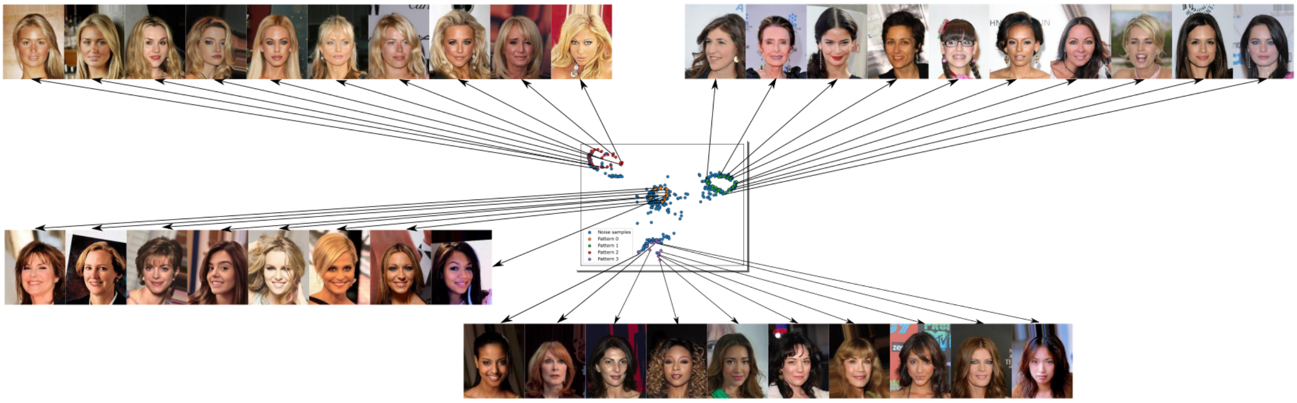


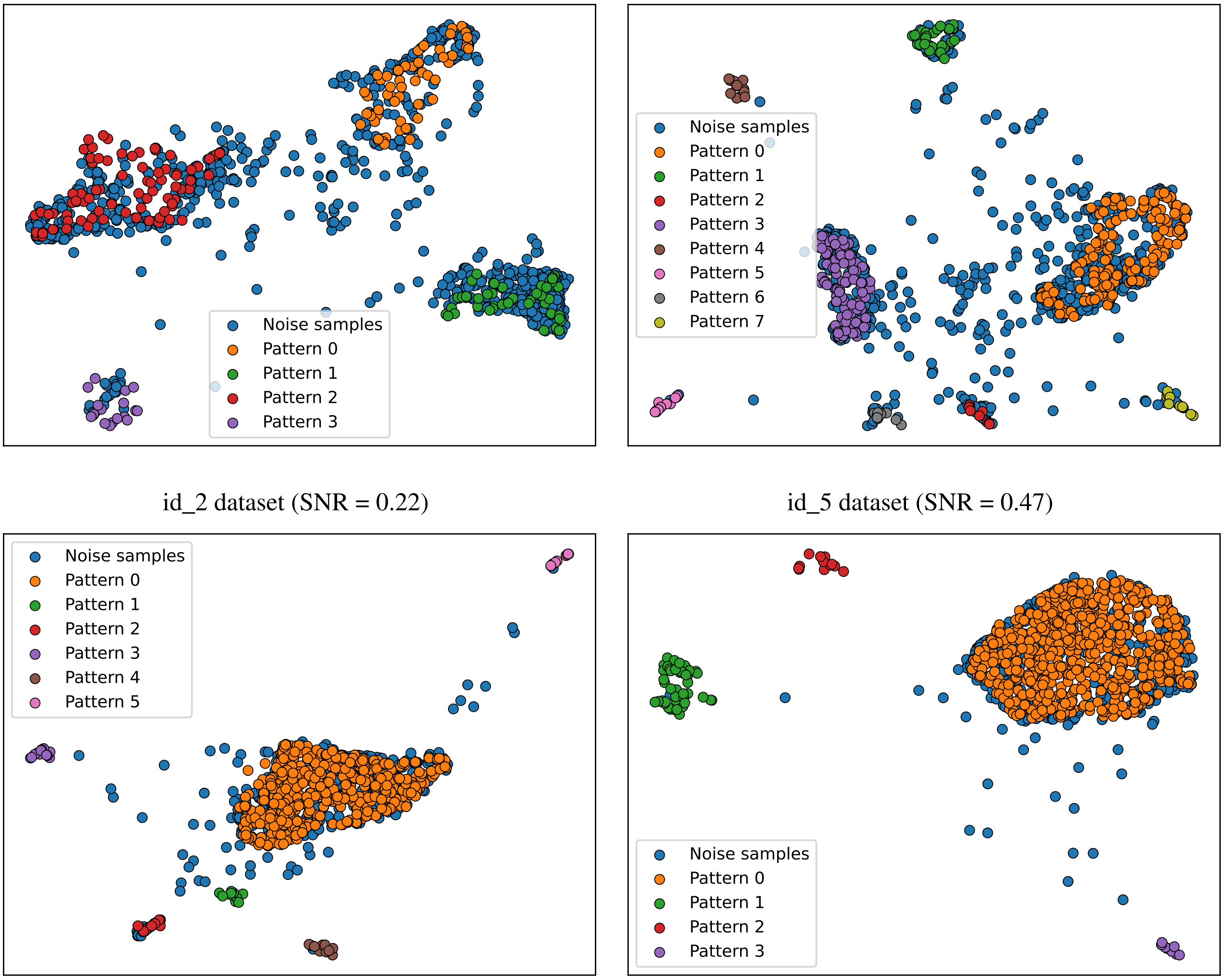
Figure A.1: Failure patterns existing in dataset id_1. One can observe four distinct failure patterns in this dataset.

Table 1: The description of 15 datasets that are used in the numerical experiments.

Dataset	DcBench	Noise Magnitude	SNR	M	k_{nn}	Sample size	Number of misclassified samples
id_1	p_72799	Low	0.15	10	7	6088	572
id_2	p_122144	Low	0.22	10	7	6103	1076
id_3	p_121880	Low	0.38	10	7	5969	1259
id_4	p_122653	Low	0.47	10	7	6019	1088
id_5	p_118660	Low	0.47	10	8	5994	1019
id_6	p_122145	Medium	0.69	10	11	6135	1141
id_7	p_121753	Medium	0.96	10	10	6138	1612
id_8	p_122406	Medium	1.17	10	16	6072	937
id_9	p_118049	Medium	1.38	10	12	5979	1051
id_10	p_122150	Medium	1.39	10	10	6107	1304
id_11	p_121948	High	1.75	10	15	6027	1438
id_12	p_122417	High	1.85	10	19	6035	1096
id_13	p_122313	High	1.91	10	15	6048	1011
id_14	p_121977	High	2.19	10	17	6117	1153
id_15	p_121854	High	3.78	10	24	6017	1554

the coherence required for data points within a pattern. The users choose a smaller k_{nn} if they need strong tightness between samples in a failure mode. Brito et al. [1997] recommended choosing k_{nn} of order $\log(N)$ for consistent identification of the clustering structure. A smaller value of k_{nn} imposes a more stringent condition to create an edge in the k_{nn} graph. When k_{nn} = 0, each data point is only connected with itself. If k_{nn} is sufficiently high, all misclassified data points merge to form a single failure pattern. From Figure A.2, we notice that as the increase of k_{nn} and SNR, there is a tendency to appear big patterns with a large number of data points. We could explain it as follows. When increasing k_{nn}, more edges are additionally created, which could initially connect separate patterns or augment more data points into the patterns. In practical applications of this problem, it is important to note that the two parameters k_{nn} and M rely heavily on the users, the machine learning tasks, and the nature of the dataset. In this study, we have established a fixed value of M equal to 10 for all datasets, and we have varied the value of k_{nn} to generate diverse scenarios of Signal-to-Noise Ratio (SNR). With the defined value of k_{nn}, we have constructed the k_{nn} graph of the re-scaled Activation feature. Subsequently, we have employed a simple Depth First Search algorithm on the sub-graph of only misclassified data points to collect all maximally connected components with cardinality greater than M. These components represent patterns that are the focus of the recommending algorithms. We add one additional feature named Pattern to each data point which indicates the pattern of it or -1 if it does not belong to any patterns.

Finally, the complete dataset for our problem consists of four information: Activation, True Label, Pseudo Label, and Pattern.



id_2 dataset (SNR = 0.22)

id_5 dataset (SNR = 0.47)

id_8 dataset (SNR = 1.17)

id_12 dataset (SNR = 1.85)

Figure A.2: The 2-D visualization of the Activation feature in four datasets. To downsample from a 512-dimension vector to a 2-dimension vector, we utilize the Supervised Dimension Reduction technique introduced by McInnes et al. [2018].

Table 2: Benchmark of Effectiveness (at 10% of sample size) on different noise magnitudes. Larger values are better. Bolds indicate the best methods for each dataset.

Dataset	US	DS_0.0	DS_0.25	DS_0.5	DS_0.75	DS_1.0	Coreset	BADGE
id_1	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
id_2	0.00±0.00	**0.25±0.00**	0.00±0.00	0.25±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
id_3	0.00±0.00	**0.14±0.00**	**0.14±0.00**	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
id_4	0.00±0.00	0.00±0.00	**0.33±0.00**	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
id_5	0.00±0.00	**0.12±0.00**	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
id_6	0.00±0.00	**0.33±0.00**	0.17±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
id_7	0.00±0.00	**0.25±0.00**	**0.25±0.00**	**0.25±0.00**	**0.25±0.00**	0.00±0.00	0.00±0.00	0.00±0.00
id_8	0.01±0.03	**0.17±0.00**	0.00±0.00	**0.17±0.00**	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
id_9	0.00±0.00	**0.20±0.00**	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
id_10	0.00±0.00	0.25±0.00	**0.50±0.00**	**0.50±0.00**	0.25±0.00	0.00±0.00	0.00±0.00	0.00±0.00
id_11	0.00±0.00	**0.33±0.00**	0.00±0.00	0.00±0.00	**0.33±0.00**	0.00±0.00	0.00±0.00	0.00±0.00
id_12	0.01±0.04	**0.25±0.00**	**0.25±0.00**	**0.25±0.00**	**0.25±0.00**	0.00±0.00	0.00±0.00	0.00±0.00
id_13	0.00±0.00	**0.33±0.00**	**0.33±0.00**	**0.33±0.00**	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
id_14	0.01±0.04	**0.50±0.00**	0.00±0.00	0.25±0.00	**0.50±0.00**	0.00±0.00	0.00±0.00	0.00±0.00
id_15	0.07±0.17	**0.50±0.00**	**0.50±0.00**	**0.50±0.00**	**0.50±0.00**	0.00±0.00	0.00±0.00	0.00±0.00
Overall	0.01±0.05	**0.24±0.14**	0.17±0.18	0.17±0.18	0.14±0.18	0.00±0.00	0.00±0.00	0.00±0.00

A.3 ADDITIONAL NUMERICAL RESULTS

In the main paper, we present the numerical results for groups categorized into three levels of Signal-to-Noise Ratio (SNR). In this section, we offer a comprehensive breakdown of the results for each individual dataset in Tables 2, 3, and 4, respectively.

We also provide charts that illustrate the progress of algorithms over iterations in dataset id_10, as depicted in Figure A.3. The blue line represents the percentage of queried samples, which appears linear due to the fixed size of the queried batch at each iteration. The orange line indicates the percentage of detected misclassified samples out of the total misclassified ones in the dataset. The green line represents the percentage of detected failure modes out of the total number of failure modes in the dataset. It is evident that the orange line, corresponding to methods that incorporate our exploiting component (Gaussian process component) such as DS_0.0, DS_0.25, DS_0.5, and DS_0.75, consistently outperforms the blue lines significantly. This trend clearly demonstrates the effectiveness of our exploiting term in identifying misclassified samples.

However, DS_0.0 shows inferior performance, as evidenced by the green line consistently falling below the blue line throughout the iterations, despite its effectiveness in identifying misclassified samples. In contrast, DS_0.25, DS_0.5, and DS_0.75 exhibit superb performance in detecting all failure patterns within approximately 100 iterations (40% of the dataset samples). This difference can be attributed to the absence of the exploration term in DS_0.0 when dealing with a high SNR level in dataset id_10.

A.4 ANALYSIS OF SAMPLING COMPLEXITY

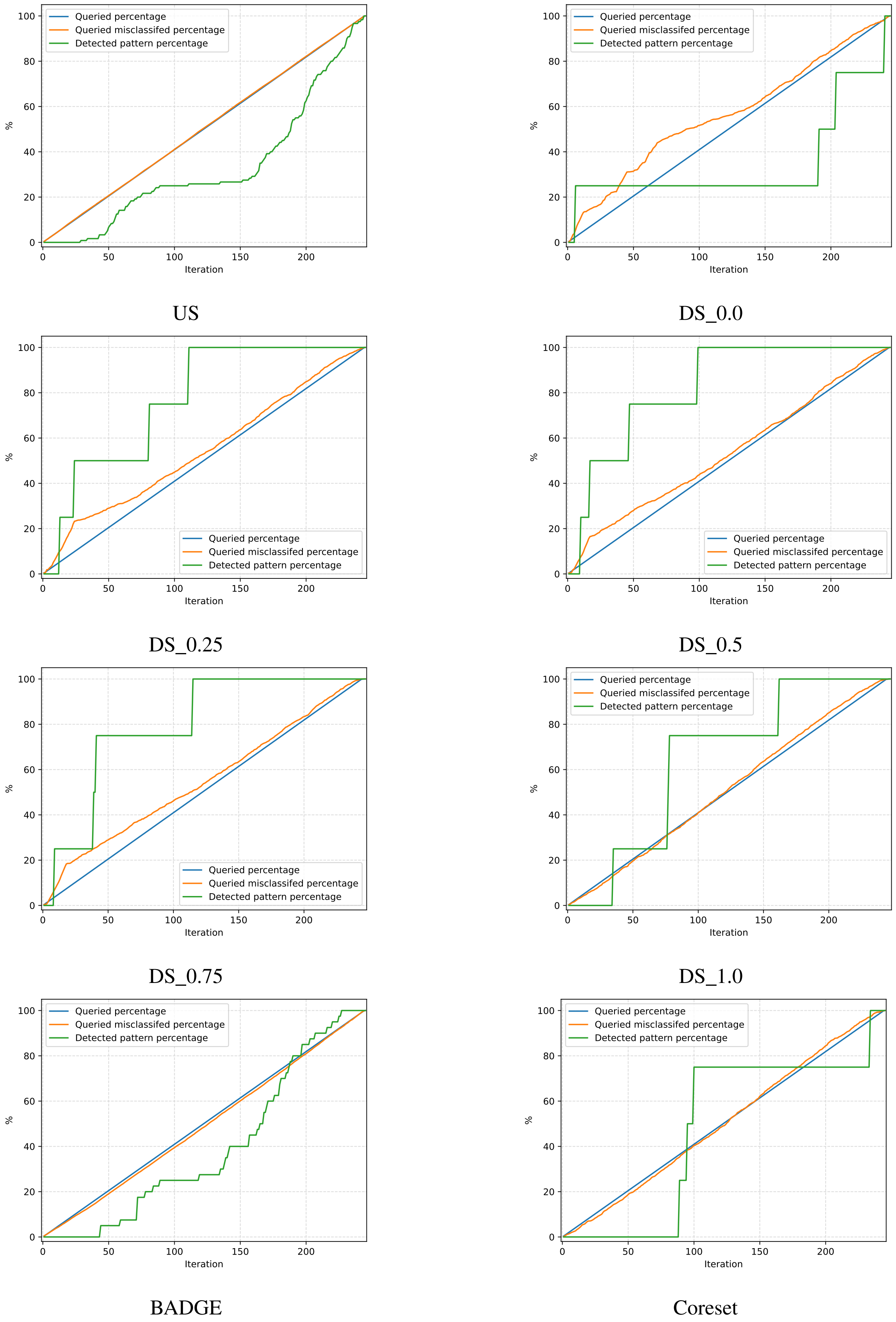
Each iteration in our framework consists of two main phases. The first phase determines which samples to be labeled next, the most costly computation in this phase is the matrix inversion and computing matrix determinant. The maximum size of the matrix is N, so the time complexity is $O(N^3)$. If we use the optimized CW-like algorithm for matrix inversion, then the complexity can be as low as $O(N^{2.373})$. The second phase includes updating information and confirming detected failure modes. Updating information involves matrix inversions and multiplications, with cost $O(N^{2.373})$. A low-cost Depth First Search is implemented to check detected failure modes, which costs $O(N)$. In conclusion, the cost of an iteration is $O(N^{2.373})$.

Table 3: Benchmark of Effectiveness (at 20% of sample size) on different noise magnitudes. Larger values are better. Bolds indicate the best methods for each dataset.

Dataset	US	DS_0.0	DS_0.25	DS_0.5	DS_0.75	DS_1.0	Coreset	BADGE
id_1	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.000±0.000
id_2	0.01±0.04	**0.25±0.00**	0.00±0.00	**0.25±0.00**	**0.25±0.00**	0.00±0.00	0.00±0.00	0.000±0.000
id_3	0.00±0.00	**0.29±0.00**	0.14±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.000±0.000
id_4	0.01±0.06	**0.67±0.00**	0.33±0.00	0.33±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.000±0.000
id_5	0.00±0.00	**0.25±0.00**	0.00±0.00	0.12±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.000±0.000
id_6	0.00±0.00	**0.67±0.00**	0.17±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.000±0.000
id_7	0.02±0.08	**0.25±0.00**	**0.25±0.00**	**0.25±0.00**	**0.25±0.00**	0.00±0.00	0.00±0.00	0.000±0.000
id_8	0.01±0.03	**0.17±0.00**	**0.17±0.00**	**0.17±0.00**	0.00±0.00	0.00±0.00	0.00±0.00	0.000±0.000
id_9	0.02±0.06	**0.20±0.00**	**0.20±0.00**	**0.20±0.00**	**0.20±0.00**	0.00±0.00	0.00±0.00	0.000±0.000
id_10	0.05±0.10	0.25±0.00	0.50±0.00	**0.75±0.00**	**0.75±0.00**	0.25±0.00	0.00±0.00	0.005±0.100
id_11	0.06±0.12	0.33±0.00	0.33±0.00	**0.67±0.00**	0.33±0.00	0.00±0.00	0.00±0.00	0.003±0.100
id_12	0.12±0.12	0.25±0.00	**0.50±0.00**	0.25±0.00	**0.50±0.00**	0.00±0.00	0.00±0.00	0.000±0.000
id_13	0.04±0.11	0.33±0.00	**0.67±0.00**	**0.67±0.00**	0.33±0.00	0.00±0.00	0.00±0.00	0.000±0.000
id_14	0.11±0.12	**0.50±0.00**	**0.50±0.00**	**0.50±0.00**	**0.50±0.00**	0.25±0.00	0.25±0.00	0.100±0.120
id_15	0.48±0.09	**0.50±0.00**	**0.50±0.00**	**0.50±0.00**	**0.50±0.00**	0.50±0.00	0.00±0.00	0.150±0.230
Overall	0.06±0.14	**0.33±0.18**	0.28±0.21	0.31±0.24	0.24±0.23	0.07±0.14	0.02±0.06	0.020±0.090

Table 4: Benchmark of Sensitivity on different noise magnitudes. Smaller values are better. Bolds indicate the best methods in each dataset.

Dataset	US	DS_0.0	DS_0.25	DS_0.5	DS_0.75	DS_1.0	Coreset	BADGE
id_1	0.55±0.00	0.76±0.00	**0.23±0.00**	0.57±0.00	0.44±0.00	0.62±0.00	0.76±0.00	0.660±0.009
id_2	0.35±0.00	0.09±0.00	0.23±0.00	**0.05±0.00**	0.14±0.00	0.51±0.00	0.60±0.00	0.600±0.100
id_3	0.59±0.00	**0.05±0.00**	0.07±0.00	0.42±0.00	0.31±0.00	0.49±0.00	0.55±0.00	0.550±0.007
id_4	0.44±0.00	0.13±0.00	**0.06±0.00**	0.11±0.00	0.22±0.00	0.57±0.00	0.60±0.00	0.530±0.006
id_5	0.53±0.00	**0.08±0.00**	0.21±0.00	0.19±0.00	0.23±0.00	0.33±0.00	0.57±0.00	0.470±0.006
id_6	0.48±0.00	**0.03±0.00**	0.04±0.00	0.47±0.00	0.29±0.00	0.32±0.00	0.55±0.00	0.400±0.007
id_7	0.37±0.00	**0.03±0.00**	0.10±0.00	0.07±0.00	0.07±0.00	0.35±0.00	0.37±0.00	0.340±0.004
id_8	0.30±0.00	0.08±0.00	0.12±0.00	**0.03±0.00**	0.30±0.00	0.46±0.00	0.45±0.00	0.400±0.006
id_9	0.28±0.00	**0.03±0.00**	0.14±0.00	0.13±0.00	0.15±0.00	0.45±0.00	0.48±0.00	0.330±0.005
id_10	0.20±0.00	**0.02±0.00**	0.05±0.00	0.04±0.00	0.04±0.00	0.14±0.00	0.36±0.00	0.280±0.006
id_11	0.26±0.00	**0.01±0.00**	0.13±0.00	0.15±0.00	0.06±0.00	0.31±0.00	0.32±0.00	0.300±0.005
id_12	0.13±0.00	**0.02±0.00**	0.04±0.00	0.04±0.00	0.07±0.00	0.21±0.00	0.22±0.00	0.290±0.003
id_13	0.22±0.00	0.05±0.00	0.05±0.00	**0.04±0.00**	0.19±0.00	0.29±0.00	0.40±0.00	0.340±0.006
id_14	0.18±0.00	**0.03±0.00**	0.11±0.00	0.10±0.00	**0.03±0.00**	0.11±0.00	0.18±0.00	0.230±0.004
id_15	0.23±0.00	**0.02±0.00**	0.05±0.00	0.05±0.00	0.04±0.00	0.19±0.00	0.26±0.00	0.210±0.003
Overall	0.34±0.14	**0.10±0.18**	0.11±0.07	0.16±0.17	0.17±0.12	0.36±0.15	0.44±0.16	0.400±0.150



US

DS_0.0

DS_0.25

DS_0.5

DS_0.75

DS_1.0

BADGE

Coreset

Figure A.3: The percentage of misclassified detected samples, the percentage of detected patterns, and the percentage of queried samples along with queried iterations in dataset id_10

A.5 PRINCIPAL HYPER-PARAMETERS AND USER-DEFINED HYPER-PARAMETERS

Our proposed framework is applied to human-machine cooperation systems. Therefore, some terms depend on the user such as the failure mode definition which is defined by two factors: (i) how to determine whether two samples have a common concept; (ii) what the structure of a failure pattern is. In our experiments, we consider the case that the user defines an edge (common concept) by using the mutual k_{nn}-graph under the Euclidean distance on the embedding space. The connectivity criterion is maximally connected subgraphs (a.k.a. connected components). With this indication, the user also provides two hyper-parameters k_{nn} and M. The meaning of k_{nn} and M are mentioned in Appendix A.2. From the algorithmic viewpoint, our approach depends mainly on one main hyper-parameter ϑ. The parameter ϑ regulates the exploration-exploitation trade-off in the sampling procedure ($\vartheta = 0$ means pure exploitation, $\vartheta = 1$ means pure exploration). We experimented with five values of ϑ throughout the paper.

B PROOFS

B.1 PROOFS OF PROPOSITION 6.1

Proof of Proposition 6.1. We first show that the value of δ should be upper-bounded by $\sqrt{N-1}$. To see this, note that $K(h_{\mathcal{X}}, h_{\mathcal{Y}})$ is a Gram matrix, so its diagonal elements are all ones, and the off-diagonal elements are in the range $(0, 1]$. We have an upper bound that:

$$\|K(h_{\mathcal{X}}, h_{\mathcal{Y}}) - I_N\|_F \leq \sqrt{N(N-1)}.$$

To ensure the existence of $h_{\mathcal{X}}, h_{\mathcal{Y}}$, the value of δ must fulfill:

$$\delta \|I_N\|_F < \sqrt{N(N-1)} \implies \delta < \sqrt{N-1}.$$

Next, we show that condition for $h_{\mathcal{X}}$ and $h_{\mathcal{Y}}$. Squaring both sides of (8) gives

$$\|K(h_{\mathcal{X}}, h_{\mathcal{Y}}) - I_N\|_F^2 \geq \delta^2 \|I_N\|_F^2 = \delta^2 N.$$

Because the diagonal elements of $K(h_{\mathcal{X}}, h_{\mathcal{Y}})$ are all ones, the above condition is equivalent to

$$\sum_{i>j} \exp\big(- \frac{\|x_i - x_j\|_2^2}{h_{\mathcal{X}}^2} - \frac{\|\widehat{\mu}_{\hat{y}_i} - \widehat{\mu}_{\hat{y}_j}\|_2^2 + \|\widehat{\Sigma}_{\hat{y}_i} - \widehat{\Sigma}_{\hat{y}_j}\|_F^2}{h_{\mathcal{Y}}^2} \big) \geq \frac{\delta^2 N}{2}. \tag{1}$$

Using Jensen inequality for the exponential function, which is convex, we have the following lower bound:

$$\begin{aligned} &\frac{1}{\binom{N}{2}} \sum_{i>j} \exp\big(- \frac{\|x_i - x_j\|_2^2}{h_{\mathcal{X}}^2} - \frac{\|\widehat{\mu}_{\hat{y}_i} - \widehat{\mu}_{\hat{y}_j}\|_2^2 + \|\widehat{\Sigma}_{\hat{y}_i} - \widehat{\Sigma}_{\hat{y}_j}\|_F^2}{h_{\mathcal{Y}}^2} \big) \\ &\qquad \geq \exp\big(- \frac{\sum_{i>j} \|x_i - x_j\|_2^2}{h_{\mathcal{X}}^2 \binom{N}{2}} - \frac{\sum_{i>j} \|\widehat{\mu}_{\hat{y}_i} - \widehat{\mu}_{\hat{y}_j}\|_2^2 + \|\widehat{\Sigma}_{\hat{y}_i} - \widehat{\Sigma}_{\hat{y}_j}\|_F^2}{h_{\mathcal{Y}}^2 \binom{N}{2}} \big). \end{aligned}$$

Therefore, if $h_{\mathcal{X}}$ and $h_{\mathcal{Y}}$ satisfy

$$\exp\big(- \frac{\sum i > j \|x_i - x_j\|_2^2}{h_{\mathcal{X}}^2 \binom{N}{2}} - \frac{\sum_{i>j} \|\widehat{\mu}_{\hat{y}_i} - \widehat{\mu}_{\hat{y}_j}\|_2^2 + \|\widehat{\Sigma}_{\hat{y}_i} \widehat{\Sigma}_{\hat{y}_j}\|_F^2}{h_{\mathcal{Y}}^2 \binom{N}{2}} \big) \geq \frac{\delta^2}{N-1},$$

then they also satisfy the condition (1). Defining the quantities $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$ as in statement of the proposition, we find that $h_{\mathcal{X}}$ and $h_{\mathcal{Y}}$ should satisfy

$$\Leftrightarrow \frac{D_{\mathcal{X}}}{h_{\mathcal{X}}^2} + \frac{D_{\mathcal{Y}}}{h_{\mathcal{Y}}^2} \leq \ln \frac{N-1}{\delta^2}.$$

This completes the proof. □

B.2 TAYLOR EXPANSION FOR VALUE-OF-INTEREST VOI

We first use a second-order Taylor expansion to approximate $f(X) = \mathrm{VoI}(X) = (1 + \exp(-g(X))^{-1}$ around the point $X = \mu$:

$$\begin{aligned} f(X) &= f(\mu) + (X-\mu)^\top \nabla f(\mu) + \frac{1}{2}(X-\mu)^\top \nabla^2 f(\mu)(X-\mu) + \mathcal{O}(\|\Delta_X\|^3) \\ &= f(\mu) + (X-\mu)^\top \nabla f(\mu) + \frac{1}{2}\mathrm{Tr}[\nabla^2 f(\mu)(X-\mu)(X-\mu)^\top] + \mathcal{O}(\|\Delta_X\|^3). \end{aligned}$$

Moreover, we set μ as the expected value $\mathbb{E}[X]$, and taking expectations on both sides of the above equation gives

$$\begin{aligned} \mathbb{E}[f(X)] &= \mathbb{E}\big[f(\mu)\big] + \mathbb{E}\big[(X-\mu)^\top \nabla f(\mu)\big] + \frac{1}{2}\mathbb{E}\big[\mathrm{Tr}[\nabla^2 f(\mu)(X-\mu)(X-\mu)^\top]\big] + \mathcal{O}(\|\Delta\|^3) \\ &= f(\mu) + \frac{1}{2}\Sigma_{t,i}^* \nabla^2 f(\mu) + \mathcal{O}(\|\Delta\|^3), \end{aligned}$$

where the second equality follows from the relationship

$$\mathbb{E}\big[(X-\mu)^\top \nabla f(\mu)\big] = \mathbb{E}\big[(X-\mu)\big]^\top \nabla f(\mu) = (\mathbb{E}[X]-\mu)]^\top \nabla f(\mu) = 0,$$

and from the definition of the covariance matrix

$$\mathbb{E}\big[(X-\mu)(X-\mu)^\top\big] = \Sigma_{t,i}^*.$$

It now suffices to verify the expressions for α_i and β_i. Note that $\alpha_i = f(\mu) = (1 + \exp(-\mu))^{-1}$ and β_i is the second-order derivative

$$\beta_i = \nabla^2 f(\mu) = \alpha_i(1-\alpha_i)(1-2\alpha_i),$$

where the second equality follows from the property of the sigmoid function.

C SOCIAL IMPACT

One important social impact of this research lies in its potential to improve the accuracy and reliability of machine learning classifiers. By identifying misclassification patterns, the framework enables the refinement and improvement of classifiers, reducing the likelihood of wrong predictions in various domains. This can have wide-ranging implications, such as improving the performance of automated systems in critical areas where accurate classification is of utmost importance like healthcare diagnosis [Shaban-Nejad et al., 2021, Rudin and Ustun, 2018, Albahri et al., 2023], or autonomous vehicles [Glomsrud et al., 2019, Wagner and Koopman, 2015].

Another significant social impact of this research is its potential to address biases and fairness issues in machine learning systems [Caton and Haas, 2020, Mehrabi et al., 2021, Pessach and Shmueli, 2022]. By identifying misclassification patterns, the framework can shed light on potential biases in the data or algorithmic models. This knowledge is crucial for developing fairer and more equitable machine learning systems which are obligatory for bringing machine learning models to practical implementations.

Moreover, the collaborative nature of the framework promotes human-machine interaction, fostering a symbiotic relationship that combines human expertise and algorithmic capabilities. This approach not only empowers human annotators by involving them in the decision-making process but also allows them to contribute their domain knowledge and intuition [Wu et al., 2022, Xin et al., 2018].

References

AS Albahri, Ali M Duhaim, Mohammed A Fadhel, Alhamzah Alnoor, Noor S Baqer, Laith Alzubaidi, OS Albahri, AH Alamoodi, Jinshuai Bai, Asma Salhi, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. *Information Fusion*, 2023.

Maria R Brito, Edgar L Chávez, Adolfo J Quiroz, and Joseph E Yukich. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. *Statistics & Probability Letters*, 35(1):33–42, 1997.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1962–1981, 2022.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022.

Jon Arne Glomsrud, André Ødegårdstuen, Asun Lera St Clair, and Øyvind Smogeli. Trustworthy versus explainable ai in autonomous vessels. In *Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC)*, volume 37, 2019.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

Cynthia Rudin and Berk Ustun. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, 48(5):449–466, 2018.

Arash Shaban-Nejad, Martin Michalowski, John S Brownstein, and David L Buckeridge. Guest editorial explainable ai: towards fairness, accountability, transparency and trust in healthcare. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2374–2375, 2021.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 19339–19352, 2020.

Yunsheng Song, Jing Zhang, and Chao Zhang. A survey of large-scale graph-based semi-supervised classification algorithms. *International Journal of Cognitive Computing in Engineering*, 3:188–198, 2022a.

Zixing Song, Xiangli Yang, Zenglin Xu, and Irwin King. Graph-based semi-supervised learning: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems*, 2022b.

Michael Wagner and Philip Koopman. A philosophy for developing trust in self-driving cars. In *Road Vehicle Automation 2*, pages 163–171. Springer, 2015.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 2022.

Doris Xin, Litian Ma, Jialin Liu, Stephen Macke, Shuchen Song, and Aditya Parameswaran. Accelerating human-in-the-loop machine learning: Challenges and opportunities. In *Proceedings of the second workshop on data management for end-to-end machine learning*, pages 1–4, 2018.