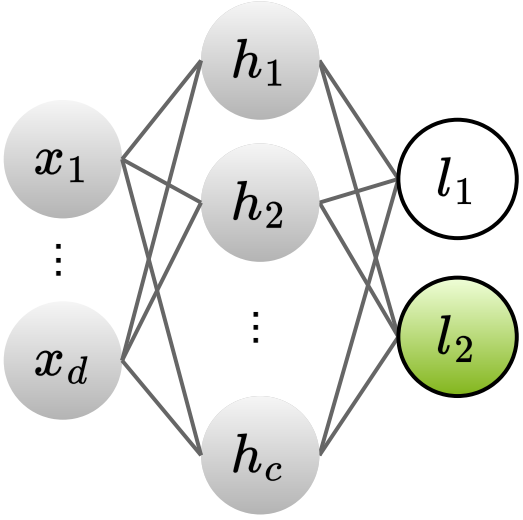
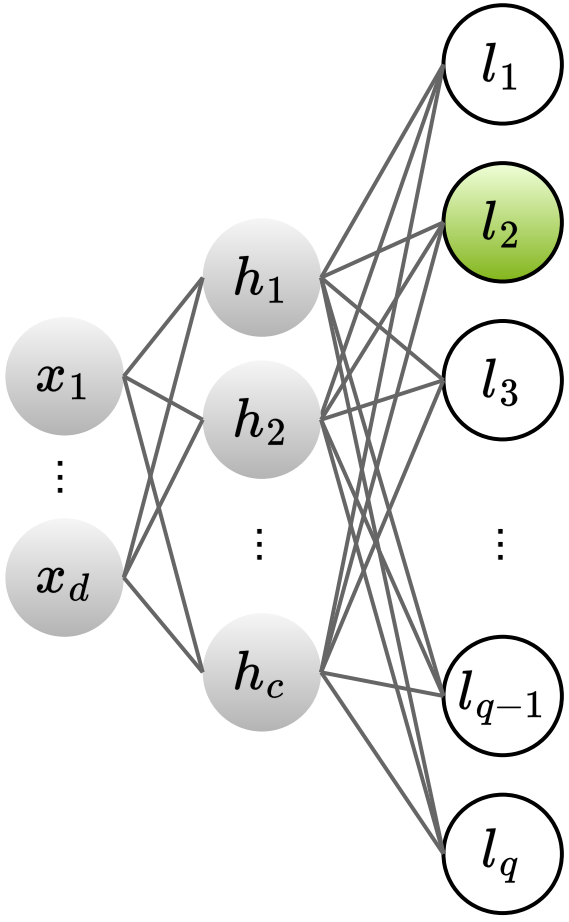
Contemporary Issues in Multi-label Representation Learning: Consistency and Adversarial Training

Kaan Demir

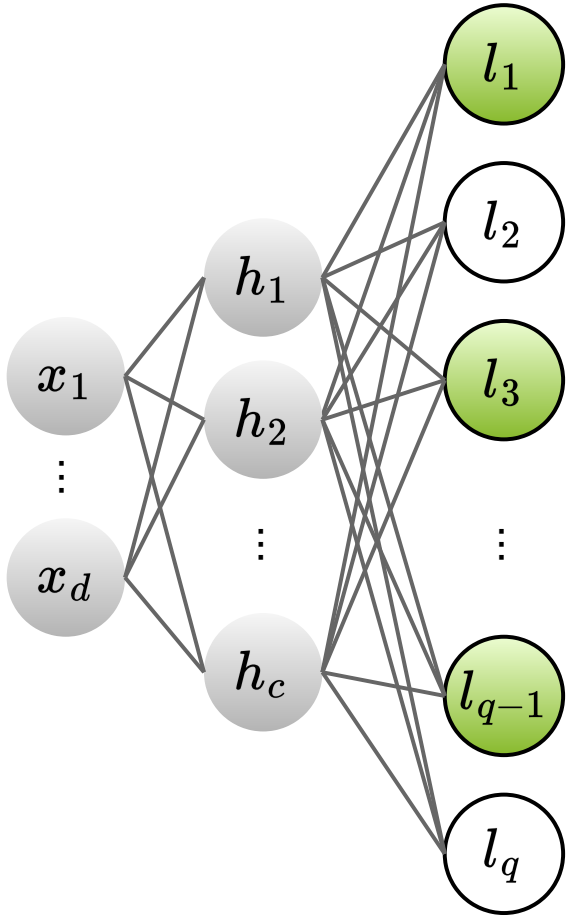
Multi-label Classification



Binary

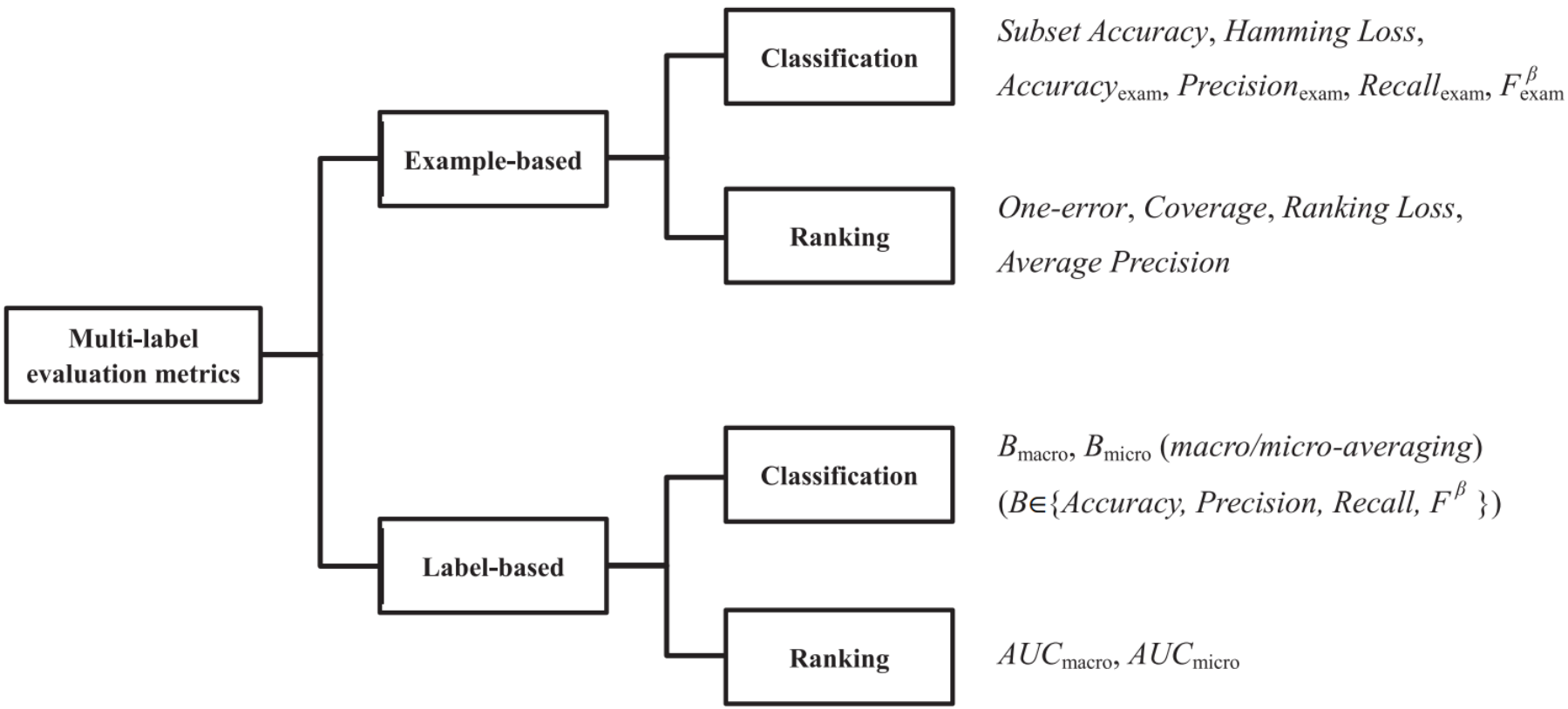


Multi-class



Multi-label

Multi-label Performance Metrics



1

[1] Zhang, M.L. and Zhou, Z.H. A review on multi-label learning algorithms. In IEEE Transactions on Knowledge and Data Engineering, 26(8), 2013.

Multi-label Performance Metrics (Conflict)

- Hamming-loss and Micro-F_1 are known to be conflicting [2]
- Label ranking average precision and Hamming-loss also conflict[3]
- More importantly, Micro-F_1 and label ranking average precision are also notorious to optimise directly, as they are highly discontinuous and non-differentiable [4] [5]
- ...

---

[2] Shi, C., Kong, X., Yu, P.S. and Wang, B. Multi-objective multi-label classification. In Proceedings of the SIAM International Conference on Data Mining, 2012.

[3] Yin, J., Tao, T. and Xu, J. A multi-label feature selection algorithm based on multi-objective optimization. In International Joint Conference on Neural Networks, 2015.

[4] N. Ghamrawi and A. McCallum. Collective multilabel classification. In The Conference on Information and Knowledge Management, 2005.

[5] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In Data Mining and Knowledge Discovery Handbook, 2010.

Multi-label Performance Metrics (Conflict)

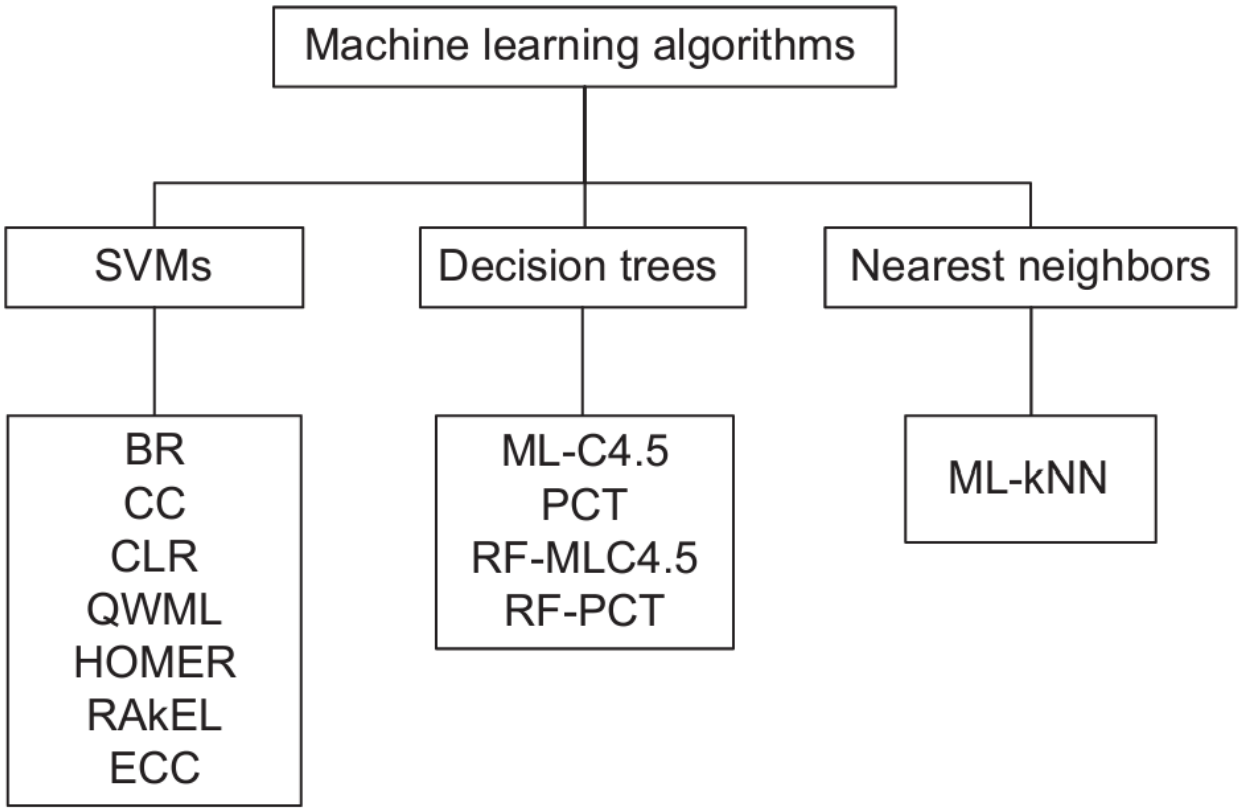


Fig. 2. The multi-label learning methods used in this study divided into groups based on the base machine learning algorithm they use.

6

[6] Madjarov, Gjorgji, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. An extensive experimental comparison of methods for multi-label learning. In Pattern Recognition, 2012.

Multi-label Performance Metrics (Conflict)

Table B12
The performance of the multi-label learning approaches in terms of the *micro-*F_1 measure. DNF (*Did Not Finish*) stands for the algorithms that did not construct a predictive model within one week under the available resources.

Dataset	BR	CC	CLR	QWML	HOMER	ML-C4.5	PCT	ML-*k*NN	RAkEL	ECC	RFML-C4.5	RF-PCT
emotions	0.509	0.503	0.512	0.528	0.588	0.655	0.571	0.457	0.533	0.554	0.647	0.672
scene	0.761	0.757	0.758	0.756	0.764	0.593	0.516	0.661	**0.772**	0.762	0.717	0.669
yeast	0.652	0.650	0.655	0.654	**0.673**	0.610	0.577	0.625	0.656	0.658	0.593	0.617
medical	0.343	0.350	0.721	0.722	**0.773**	0.756	0.356	0.634	0.714	0.714	0.374	0.693
enron	0.564	0.482	0.585	0.535	**0.591**	0.512	0.349	0.466	0.548	0.582	0.496	0.537
corel5k	0.059	0.059	**0.293**	**0.293**	0.275	0.004	0.000	0.030	0.000	0.002	0.010	0.018
tmc2007	0.932	0.936	0.930	0.930	0.927	0.135	0.547	0.682	0.890	0.869	0.777	**0.945**
mediamill	0.533	0.509	0.118	0.119	0.553	0.007	0.477	0.545	0.440	0.453	0.546	**0.563**
bibtex	0.457	**0.462**	0.448	0.454	0.429	0.093	0.108	0.206	DNF	0.247	0.123	0.230
delicious	0.234	0.236	DNF	DNF	**0.339**	0.000	0.000	0.175	DNF	DNF	0.269	0.248
bookmarks	DNF	DNF	DNF	DNF	DNF	**0.268**	0.236	0.232	DNF	DNF	0.199	0.236

Multi-label Performance Metrics (Conflict)

Table B18
The performance of the multi-label learning approaches in terms of the *average precision* measure. DNF (*Did Not Finish*) stands for the algorithms that did not construct a predictive model within one week under the available resources.

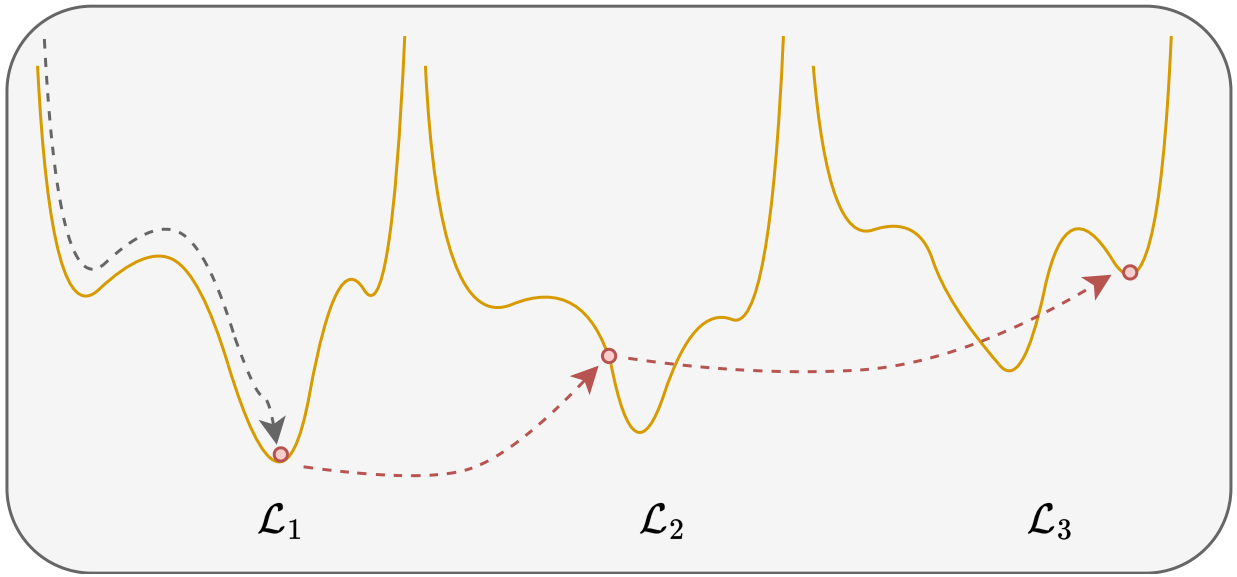
Dataset	BR	CC	CLR	QWML	HOMER	ML-C4.5	PCT	ML-*k*NN	RAkEL	ECC	RFML-C4.5	RF-PCT
emotions	0.721	0.724	0.718	0.679	0.698	0.759	0.713	0.694	0.713	0.687	**0.812**	**0.812**
scene	**0.893**	0.881	0.886	0.864	0.848	0.751	0.745	0.851	0.862	0.856	0.862	0.874
yeast	0.768	0.755	**0.768**	0.698	0.740	0.706	0.724	0.758	0.715	0.734	0.749	0.757
medical	0.896	**0.901**	0.864	0.862	0.786	0.823	0.522	0.784	0.676	0.684	0.817	0.868
enron	0.693	0.695	**0.699**	0.604	0.604	0.629	0.546	0.635	0.522	0.576	0.680	0.698
corel5k	0.303	0.293	**0.352**	0.311	0.222	0.196	0.208	0.266	0.088	0.014	0.314	0.334
tmc2007	0.978	0.981	0.972	0.938	0.945	0.842	0.700	0.844	0.939	0.935	0.945	**0.996**
mediamill	0.686	0.672	0.450	0.492	0.583	0.669	0.654	0.703	0.492	0.453	0.728	**0.737**
bibtex	0.597	**0.599**	0.579	0.498	0.407	0.392	0.212	0.349	DNF	0.228	0.418	0.525
delicious	0.351	0.343	DNF	DNF	0.231	0.321	0.206	0.326	DNF	DNF	0.359	**0.395**
bookmarks	DNF	DNF	DNF	DNF	DNF	0.378	0.213	0.381	DNF	DNF	0.423	**0.480**

Multi-label Performance Metrics (Conflict)

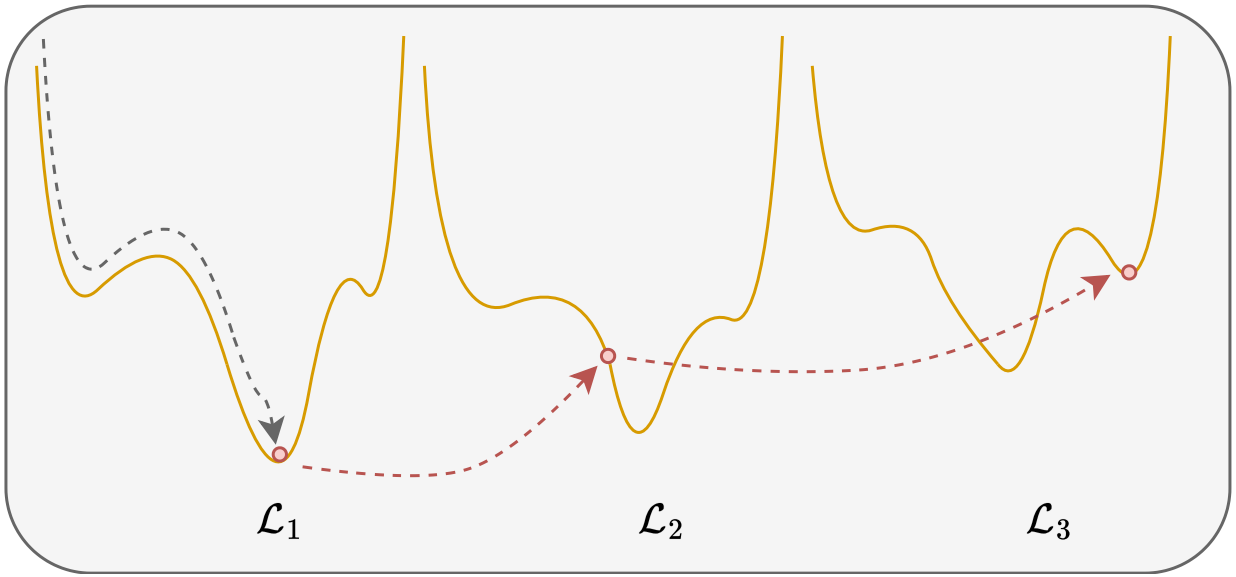
Table B1
The performance of the multi-label learning approaches in terms of the *Hamming loss* measure. DNF (*Did Not Finish*) stands for the algorithms that did not construct a predictive model within one week under the available resources.

Dataset	BR	CC	CLR	QWML	HOMER	ML-C4.5	PCT	ML-*k*NN	RAkEL	ECC	RFML-C4.5	RF-PCT
emotions	0.257	0.256	0.257	0.254	0.361	0.247	0.267	0.294	0.282	0.281	0.198	**0.189**
scene	0.079	0.082	0.080	0.081	0.082	0.141	0.129	0.099	**0.077**	0.085	0.116	0.094
yeast	**0.190**	0.193	**0.190**	0.191	0.207	0.234	0.219	0.198	0.192	0.207	0.205	0.197
medical	0.077	0.077	0.017	**0.012**	**0.012**	0.013	0.023	0.017	**0.012**	0.014	0.022	0.014
enron	**0.045**	0.064	0.048	0.048	0.051	0.053	0.058	0.051	**0.045**	0.049	0.047	0.046
corel5k	0.017	0.017	0.012	0.012	0.012	0.010	**0.009**	**0.009**	**0.009**	**0.009**	**0.009**	**0.009**
tmc2007	0.013	0.013	0.014	0.014	0.015	0.093	0.075	0.058	0.021	0.026	0.037	**0.011**
mediamill	0.032	0.032	0.043	0.043	0.038	0.044	0.034	0.031	0.035	0.035	0.030	**0.029**
bibtex	**0.012**	**0.012**	**0.012**	**0.012**	0.014	0.016	0.014	0.014	DNF	0.013	0.014	0.013
delicious	**0.018**	**0.018**	DNF	DNF	0.022	0.019	0.019	**0.018**	DNF	DNF	0.018	0.018
bookmarks	DNF	DNF	DNF	DNF	DNF	**0.009**	**0.009**	**0.009**	DNF	DNF	**0.009**	**0.009**

Multi-label Performance Metrics (Consistency)

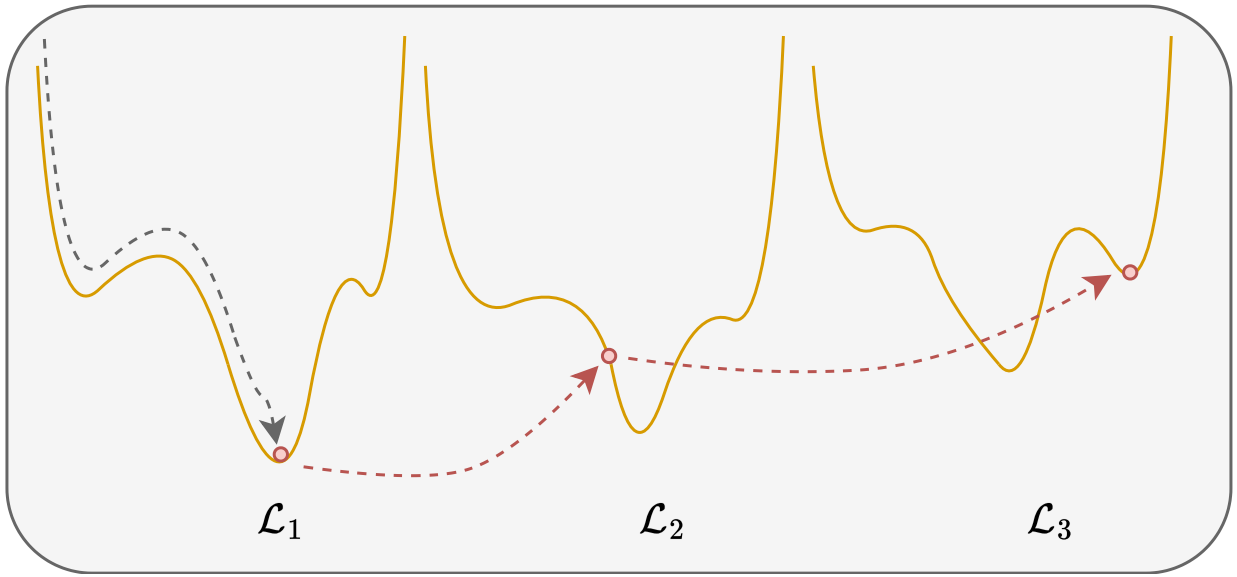


Multi-label Performance Metrics (Consistency)



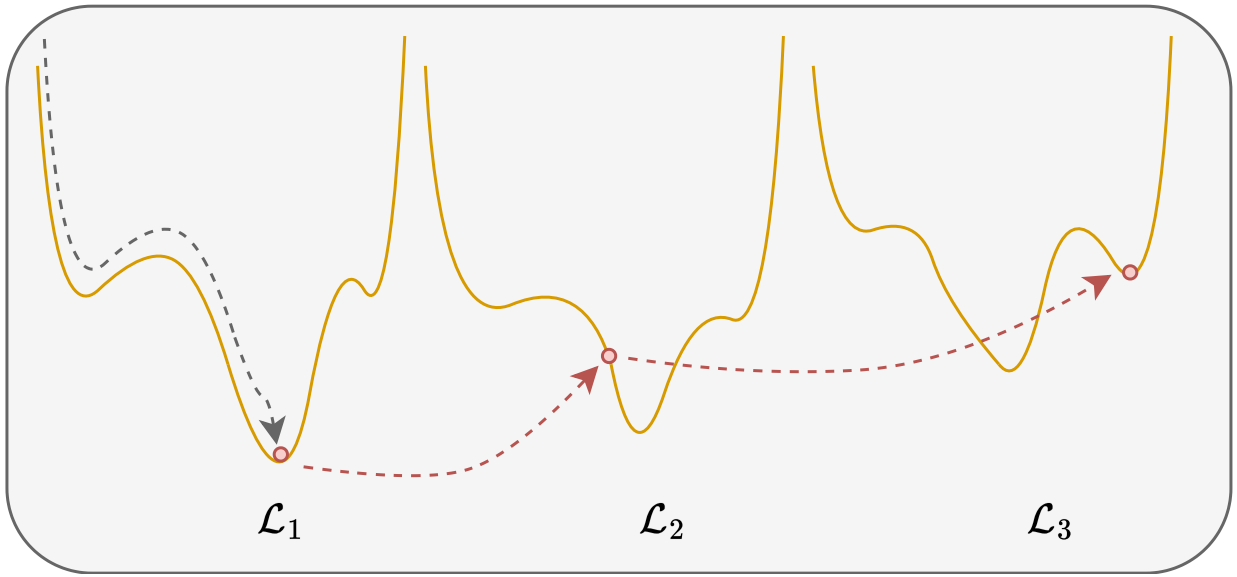
- Non-differentiable?

Multi-label Performance Metrics (Consistency)



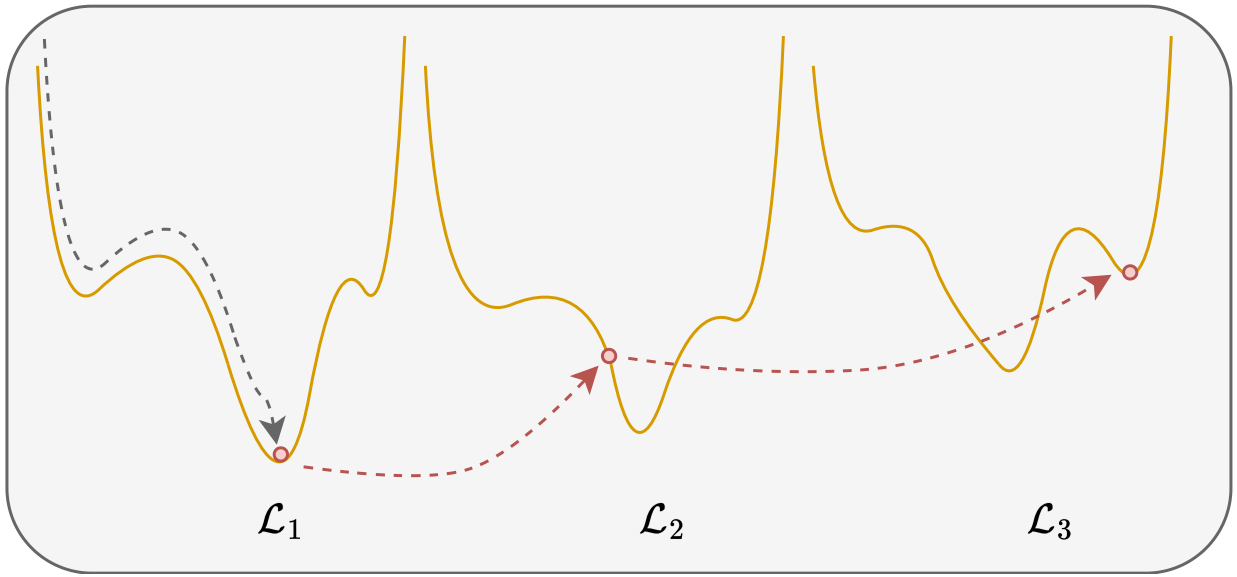
- Non-differentiable?
- Approximating proxy objective functions?

Multi-label Performance Metrics (Consistency)



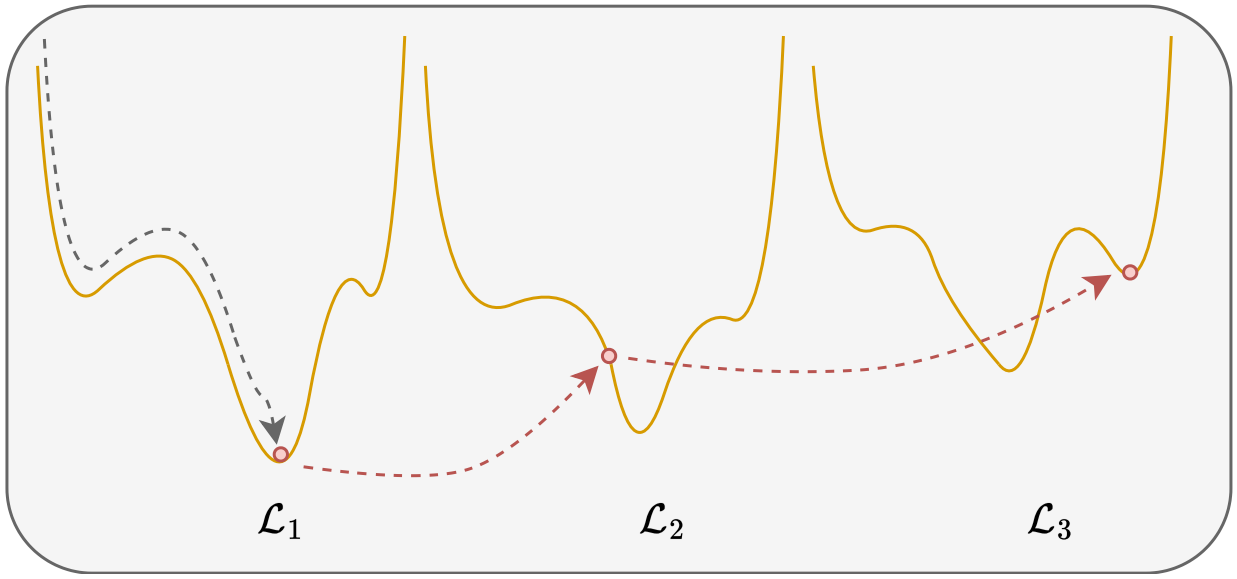
- Non-differentiable?
- Approximating proxy objective functions?
- Conflict? Consistency?

Multi-label Performance Metrics (Consistency)



- Non-differentiable?
- Approximating proxy objective functions?
- Conflict? Consistency?

Multi-label Performance Metrics (Consistency)



- Non-differentiable?
- Approximating proxy objective functions?
- Conflict? Consistency?

$$loss\,(x,y) = -\frac{1}{N}\frac{1}{C}\sum_{i=1}^{N}\sum_{j=1}^{C}\left(y_{ij}\log\frac{1}{1+e^{-x_{ij}}} + (1-y_{ij})\log\frac{e^{-x_{ij}}}{1+e^{-x_{ij}}}\right)$$

Multi-label Consistency

- Gao, W., Zhou, Z. H. On the consistency of multi-label learning. In Proceedings of the 24th annual conference on learning theory. In JMLR Workshop and Conference Proceedings, 2011.
- Koyejo, O., Natarajan, N., Ravikumar, P., and Dhillon, I. S. Consistent multilabel classification. In NeurIPS, 2015.
- Menon, A. K., Rawat, A. S., Reddi, S., and Kumar, S. Multilabel reductions: what is my loss optimizing? In NeurIPS, 2019.
- Zhang, M., Ramaswamy, H. G., and Agarwal, S. Convex calibrated surrogates for the multi-label f-measure. In ICML, 2020.
- Wu, G., Li, C., Xu, K., and Zhu, J. Rethinking and reweighting the univariate losses for multi-label ranking: Consistency and generalization. In NeurIPS, 2021.

Defining Multi-label Consistency

- Let $\mathcal{X} \in \mathbb{R}^d$, $\mathcal{Y} \in \{0,1\}^q$, and $\Omega \in \mathbb{R}^P$ respectively denote the input, output, and learnable parameter space.
- Let $\mathcal{P}$ be a joint p.d. over $\mathcal{X} \times \mathcal{Y}$. Let $f : \mathbb{R}^d \to \mathbb{R}^q$ represent a DNN drawn from $\Omega \in \mathbb{R}^P$, and trained on n samples drawn from $\mathcal{P}$.
- Given $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$, we denote $p(\mathbf{y}|\mathbf{x})_{\mathbf{y}\in\mathcal{Y}}$ as the conditional probability of $\mathbf{y}$.
-

$$\kappa = \{p(\mathbf{y}|\mathbf{x}) : \sum_{\mathbf{y}\in\mathcal{Y}} p(\mathbf{y}|\mathbf{x}) = 1 \wedge p(\mathbf{y}|\mathbf{x}) \geq 0\}. \tag{1}$$

Defining Multi-label Consistency (continued)

and the conditional risk of f given surrogate loss (ψ), loss ($\mathcal{L}$), the conditional probability of sample $\mathbf{x}$ and the label set $\mathbf{y}$:

$$\begin{aligned} \mathcal{L}^c(p(\mathbf{y}|\mathbf{x}), f) &= \sum_{\mathbf{y}\in\mathcal{Y}} p(\mathbf{y}|\mathbf{x})\mathcal{L}(f(\mathbf{x}), \mathbf{y}) \\ \psi^c(p(\mathbf{y}|\mathbf{x}), f) &= \sum_{\mathbf{y}\in\mathcal{Y}} p(\mathbf{y}|\mathbf{x})\psi(f(\mathbf{x}), \mathbf{y}). \end{aligned} \tag{2}$$

Definition (Conditional Risk)

The expected conditional risk R, and the Bayesian risk R^B, of a model representation f given $\mathcal{L}$ is defined as:

$$\begin{aligned} R(f) &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}[\mathcal{L}^c(p(\mathbf{y}|\mathbf{x}), f)] \\ R^B(f) &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}[\inf_{f'}[\mathcal{L}^c(p(\mathbf{y}|\mathbf{x}), f')]]. \end{aligned} \tag{3}$$

Defining Multi-label Consistency (continued)

Definition (Bayes Predictors)

The set of Bayes predictors:

$$B(p(\mathbf{y}|\mathbf{x})) = \{f : \mathcal{L}^c(p(\mathbf{y}|\mathbf{x}), f) = \inf_{f'}[\mathcal{L}^c(p(\mathbf{y}|\mathbf{x}), f')]\}. \tag{4}$$

determine that ψ *can* be multi-label consistent w.r.t. $\mathcal{L}$ if the following holds for every $p(\mathbf{y}|\mathbf{x}) \in \kappa$:

$$R_\psi^B(f) < \inf_{f'}\{R_\psi(f') : \forall f' \in \Omega, f' \notin B\}. \tag{5}$$

i.e., A model ψ *can* be considered multi-label consistent for a loss function $\mathcal{L}$ if it performs optimally not just for specific cases but consistently across the entire distribution $\mathcal{P}$, *e.g.*, the optimal $R_\psi^B(f)$ is strictly better than the best sub-optimal $R_\psi(f')$

Defining Multi-label Consistency (continued)

Theorem (Multi-label Consistency)

ψ can only be multi-label consistent w.r.t. $\mathcal{L}$ iff it holds for any sequence of $f^{(n)}$ that:

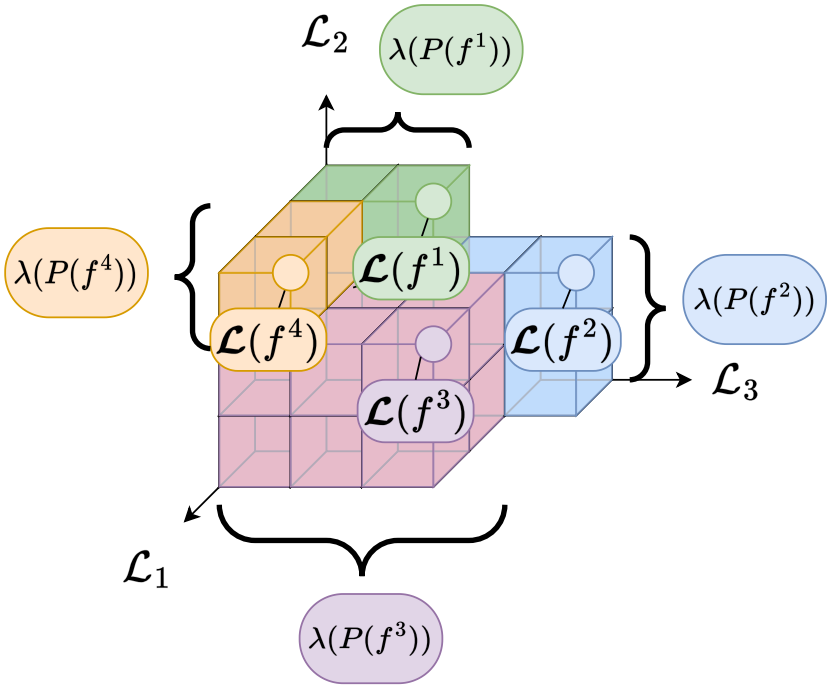
$$R_\psi(f^{(N)}) \to R_\psi^B(f) \quad \textit{then} \quad R_\mathcal{L}(f^{(N)}) \to R_\mathcal{L}^B(f). \tag{6}$$

Contemporary Issues in Consistency

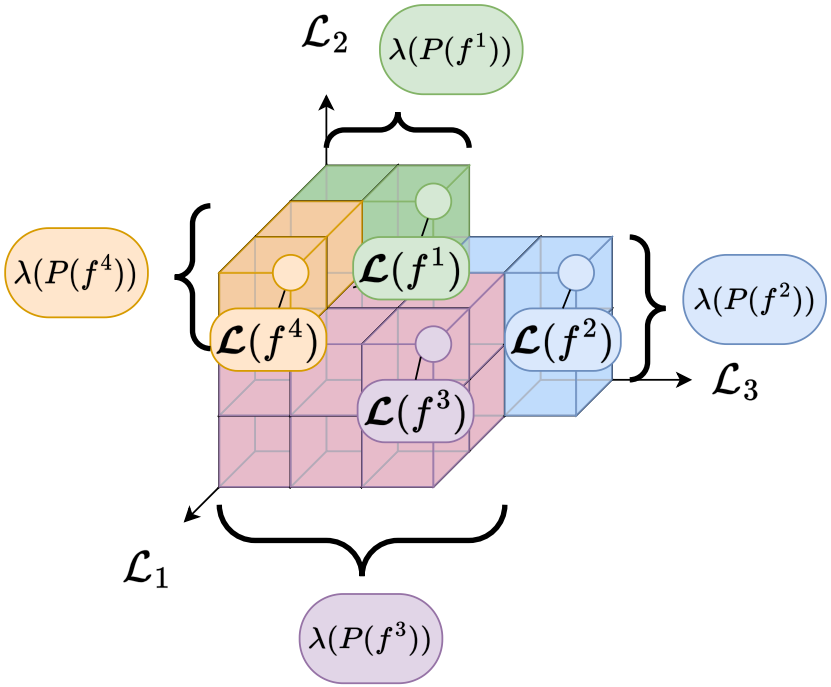
We know that multi-label performance metrics are potentially conflicting, and that they are highly discontinuous, non-smooth, or practically impossible to optimise directly. Yet...

- To date, performance gain (compared to traditional methods) of deep-learning using approximate loss has outweighed inconsistency risk
- Existing consistency research is highly focused:
  - Koyejo et al., NeurIPS 2015, characterised Bayes optimal classifiers for a given single metric
  - Menon et al., NeurIPS 2019, focuses on the consistency of reduction techniques (problem transformation) rather than the natural multi-label problem
  - Zhang et al., ICML 2020, focuses on optimising the F-measure after problem transformation

Characterising the Problem - Hypervolumes

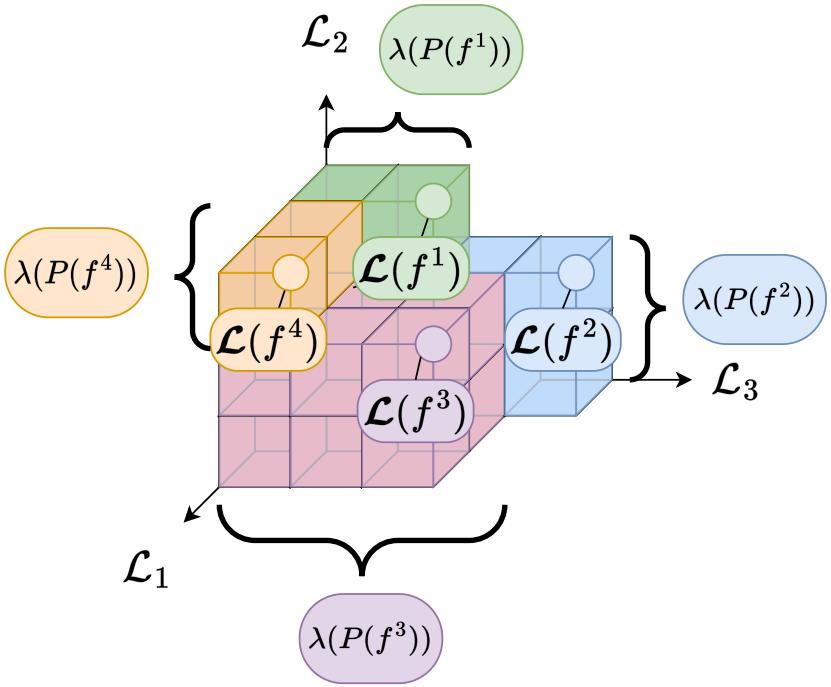


Characterising the Problem - Hypervolumes



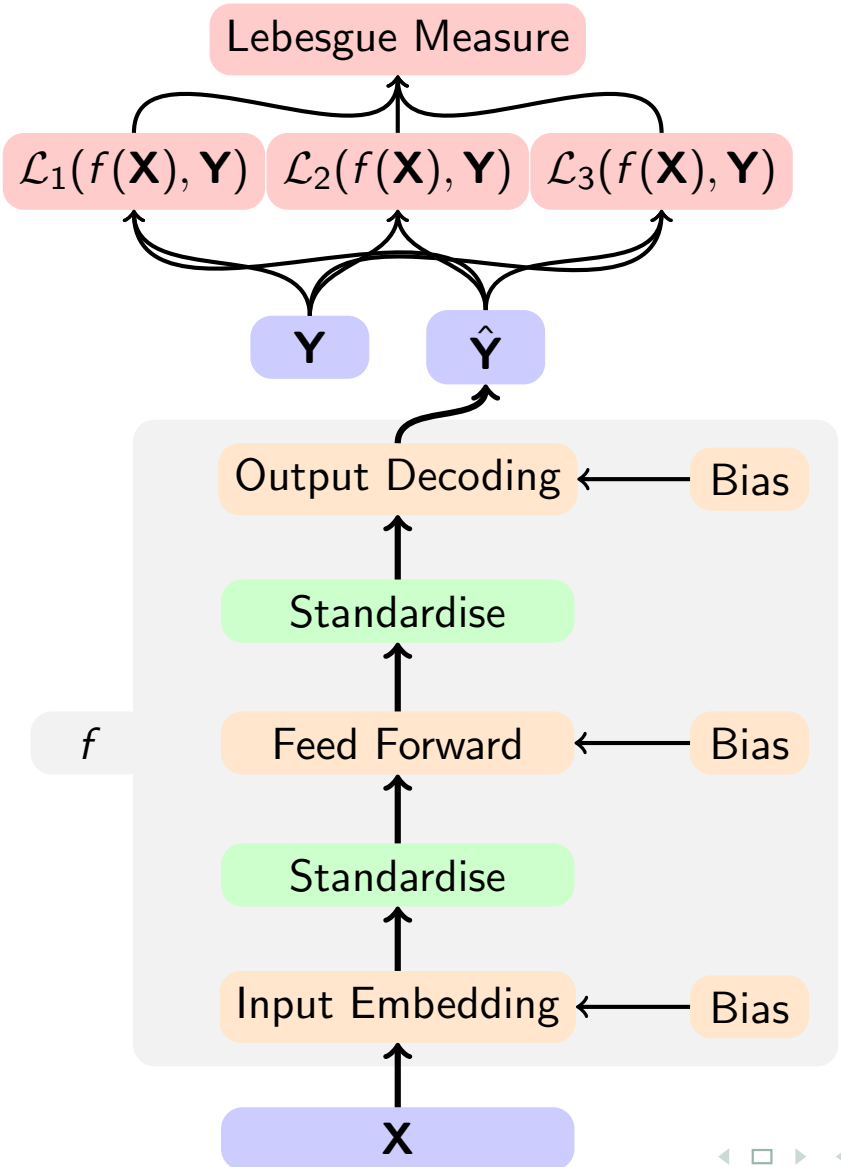
- We *care* about how well f^i performs in terms of $\mathcal{L}_1$ compared to $\mathcal{L}_2$ and $\mathcal{L}_3$

Characterising the Problem - Hypervolumes

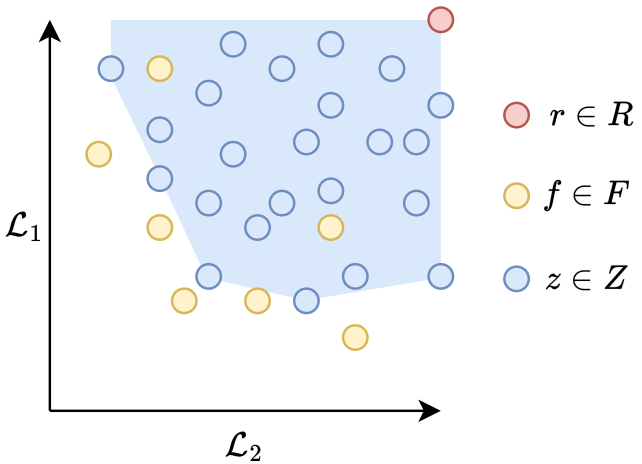


- We *care* about how well f^i performs in terms of $\mathcal{L}_1$ compared to $\mathcal{L}_2$ and $\mathcal{L}_3$
- We *care* about the raw quantities, *i.e.*, the original loss values s.t. the multi-label consistency holds (more on this later)

A Consistent Lebesgue Measure-based Multi-label Learner (CLML)



The Lebesgue Measure



$$H(F,R) := \{\mathbf{z} \in Z \quad | \quad \exists f \in F, \quad \exists \mathbf{r} \in R : \mathcal{L}(f(\mathbf{X}), \mathbf{Y}) \preceq \mathbf{z} \preceq \mathbf{r}\}.$$

$$\lambda(H(F,R)) = \int_{\mathbb{R}^o} \mathbf{1}_{H(F,R)}(\mathbf{z}) d\mathbf{z} \tag{7}$$

$$P(f) = H(\{f\}, R) \backslash H(F \backslash \{f\}, R). \tag{8}$$

Hence, the Lebesgue contribution of f, $\lambda(P(f)) = \int_{\mathbb{R}^o} \mathbf{1}_{P(f)}(\mathbf{z}) d\mathbf{z}$.[a]

[a] Auger, A., Bader, J., Brockhoff, D., Zitzler, E. Hypervolume-based multiobjective optimization: Theoretical foundations and practical implications. In Theoretical Computer Science, 2012.

The Consistency of the Lebesgue Measure

Recall that a surrogate loss function ψ can only be multi-label consistent w.r.t. $\mathcal{L}$ iff it holds for any sequence of $f^{(n)}$ that:

$$R_\psi(f^{(N)}) \to R_\psi^B(f) \quad \text{then} \quad R_\mathcal{L}(f^{(N)}) \to R_\mathcal{L}^B(f). \tag{9}$$

Theorem (A Consistent Lebesgue Measure)

Given a sequence $F^{(N)}$, the maximisation of the Lebesgue measure $\lambda(H(F^{(N)}, R))$ is consistent with the minimisation of $\mathcal{L}_1, \mathcal{L}_2$, and $\mathcal{L}_3$:

$$\begin{aligned} \lim_{N\to\infty} \lambda(H(F^{(N)}, R)) \to \lambda(H(\mathbb{P}^B, R)) \quad \text{then} \\ R_{\mathcal{L}_1}(f^{(N)}) \to R_{\mathcal{L}_1}^B(f) \wedge \\ R_{\mathcal{L}_2}(f'^{(N)}) \to R_{\mathcal{L}_2}^B(f') \wedge \\ R_{\mathcal{L}_3}(f''^{(N)}) \to R_{\mathcal{L}_3}^B(f''). \end{aligned} \tag{10}$$

The Consistency of the Lebesgue Measure (continued)

In other words, the maximisation of $\lambda(H(F^{(N)}, R))$ tends to the convergence toward the Bayes risk for each loss function $\mathcal{L}_i$ $\forall i : 1 \leq i \leq 3$, $f^{(N)}, f'^{(N)}, f''^{(N)} \in F^{(N)}$ and that $f, f', f'' \in \mathbb{P}^B$.

Proof of Theorem 4.

We proceed by contradiction. Suppose the following function exists: $f^\gamma \notin \mathbb{P}^B$, $f^\gamma \in \Omega$ s.t. $\exists v : R_{\mathcal{L}_v}(f^\gamma) = R^B_{\mathcal{L}_v}(f^\gamma)$, *i.e.*, f^γ is a Bayes predictor for the v^{th} loss $\mathcal{L}_v$. Now suppose another function $f^\beta \in \Omega$ exists s.t. $f^\beta \in \mathbb{P}^B$. By this condition, $f^\beta \prec f^\gamma$ as $f^\gamma \notin \mathbb{P}^B$, hence $\forall i : 1 \leq i \leq o : \mathcal{L}_i(f^\beta) \leq \mathcal{L}_i(f^\gamma)$, and $\exists \mathcal{L}_k : \mathcal{L}_k(f^\beta) < \mathcal{L}_k(f^\gamma)$. This result has two implications:

1. If $k = v$ then $\mathcal{L}_k(f^\beta) < \mathcal{L}_k(f^\gamma)$ would contradict f^γ being a Bayes predictor. This would imply a Bayes predictor *cannot exist outside* $\mathbb{P}^B$.
2. If $k \neq v$, then $\forall i : 1 \leq i \leq o : \mathcal{L}_i(f^\beta) \leq \mathcal{L}_i(f^\gamma)$. For this condition to hold, when $i = v$, $\mathcal{L}_i(f^\beta) \leq \mathcal{L}_i(f^\gamma)$ would imply that f^β is *also* a Bayes predictor of $\mathcal{L}_i$, when there is strict equality, and implication 1 when there is inequality. Therefore, the Bayes predictor of $\mathcal{L}_i$ already exists within $\mathbb{P}^B$.

□

CLML Representation

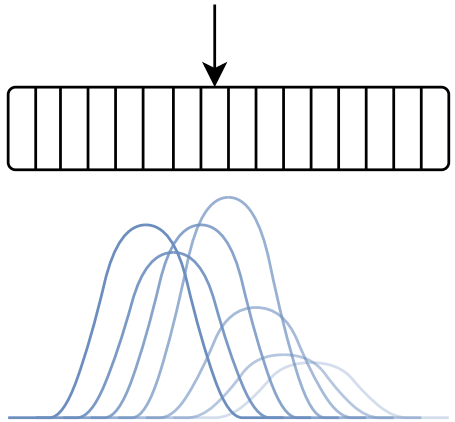
Encoder $\mathbf{E} : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times C}$ with bias $\mathbf{W}_b^{\mathbf{E}}$, where $C << d$. Feedforward layer $\mathbf{W}^{\mathbf{L}}$ and bias $\mathbf{W}_b^{\mathbf{L}}$. Decoder $\mathbf{D} : \mathbb{R}^{n \times C} \to \mathbb{R}^{n \times q}$ with bias $\mathbf{W}_b^{\mathbf{D}}$.

$$\hat{\mathbf{Y}} = \sigma(\sigma(\gamma(\sigma(\gamma(\mathbf{XE} + \mathbf{W}_b^{\mathbf{E}}))\mathbf{W}^{\mathbf{L}} + \mathbf{W}_b^{\mathbf{L}}))\mathbf{D} + \mathbf{W}_b^{\mathbf{D}}). \tag{11}$$

After each layer, we apply a sigmoid activation function (σ) and row-standardisation (γ). Activation functions such as ReLU and GELU are more tailored to deeper architectures and address specific issues such as vanishing gradients.
Tight-bound complexities: $\Theta(ndC)$, $\Theta(nqC)$, $\Theta(nC^2)$.

CLML Optimisation

$$\theta_i^f = [\mathbf{E}, \mathbf{W}_b^{\mathbf{E}}, \mathbf{W}^{\mathbf{L}}, \mathbf{W}_b^{\mathbf{L}}, \mathbf{D}, \mathbf{W}_b^{\mathbf{D}}]$$



$$\theta_i^f \sim \mathbf{m} + \sigma\mathcal{N}_i(0, \mathbf{C}) \quad \forall i, \quad 1 \leq i \leq \lambda \tag{12}$$
$$\lambda(P(f))$$

Use CMA-ES!

Methodology

Table: Datasets with number of instances (n), features (d), and labels (q).

Dataset	n	d	q	Domain	Cardinality
flags	194	19	7	image	3.392
CAL500	502	68	174	music	26.044
emotions	593	72	6	music	1.869
genbase	662	1186	27	biology	1.252
enron	1702	1001	53	text	3.378
yeast	2417	103	14	biology	4.237
tmc2007-500	28,596	500	22	text	2.158
mediamill	43907	120	101	video	4.376
IMDB-F	120,900	1001	28	text	2.000

Methodology (continued)

- Dual perspective of label-specific feature learning for multi-label classification (DELA). [7]
- Collaborative learning of label semantics and deep label-specific features (CLIF). [8]
- Learning deep latent spaces for multi-label classification (C2AE). [9]
- MLkNN [10]
- Gaussian Naive Bayes with binary relevance (GNB-BR) and classifier chain (GNB-CC) transformations.

---

[7] Hang, Jun-Yi, and Min-Ling Zhang. Dual perspective of label-specific feature learning for multi-label classification. In ICML, 2022.

[8] Hang, Jun-Yi, and Min-Ling Zhang. Collaborative learning of label semantics and deep label-specific features for multi-label classification. TPAMI, 2021.

[9] Yeh, Chih-Kuan, et al. Learning deep latent space for multi-label classification. In AAAI, 2017.

[10] Zhang, Min-Ling, and Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. In Pattern recognition, 2007.

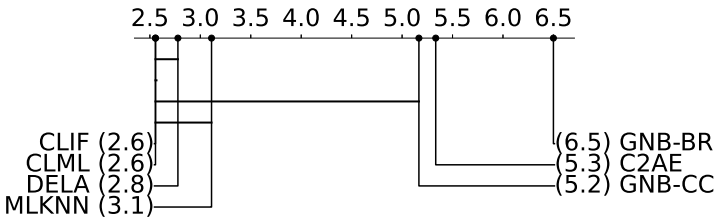
Results

$$\mu_g(\boldsymbol{\mathcal{L}}(f(\mathbf{X}), \mathbf{Y})) = \left(\prod_{i=1}^{3} \mathcal{L}_i(f(\mathbf{X}), \mathbf{Y})\right)^{\frac{1}{3}}$$

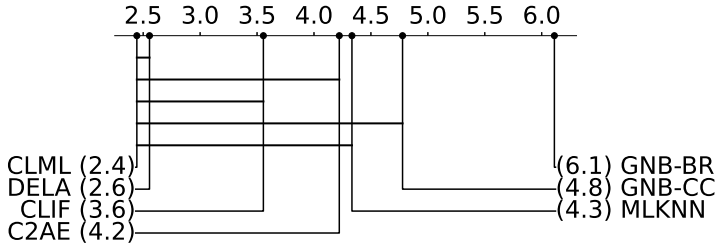
f	Median $\mu_g(\boldsymbol{\mathcal{L}}(f))$
GNB-BR	0.481
GNB-CC	0.415
MLkNN	0.249
C2AE	0.394
CLIF	0.269
DELA	0.254
CLML	**0.240**

Metric	F	C
$\lambda(P(f))$	2.88	
$\mu_g(\boldsymbol{\mathcal{L}}(f))$	21.37	
$\mathcal{L}_1$	35.62	15.51
$\mathcal{L}_2$	25.64	
$\mathcal{L}_3$	17.75	

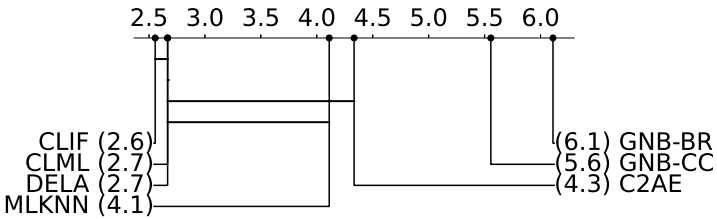
Results (continued)



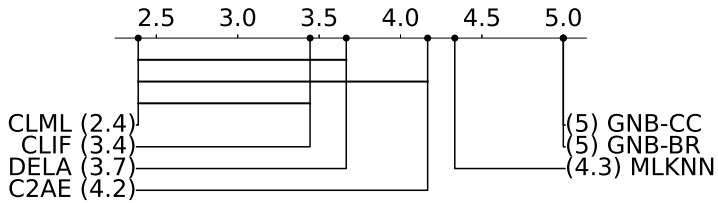
(a) $\mathcal{L}_1(f)$



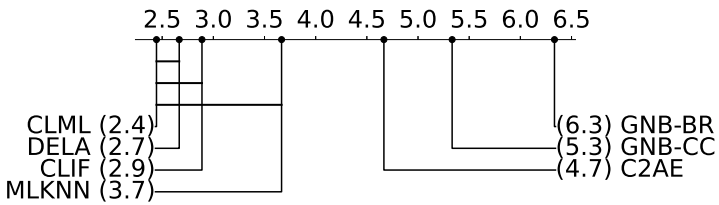
(b) $\mathcal{L}_2(f)$



(c) $\mathcal{L}_3(f)$



(d) $\lambda(P(f))$



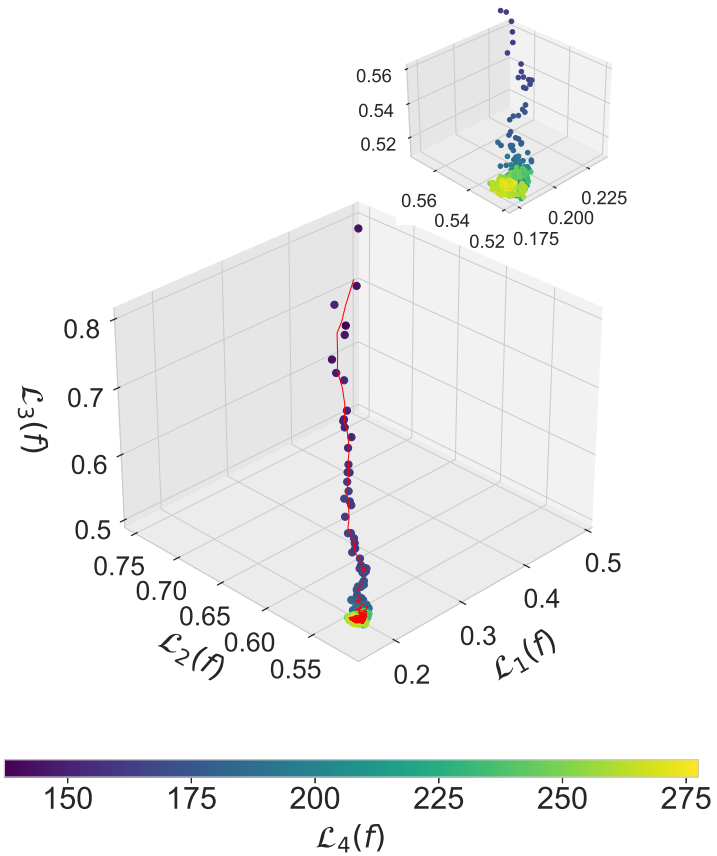
(e) $\mu_g(\mathcal{L}(f))$

Figure: Bonferroni-Dunn test critical difference plots. ($CD = 2.686$ with $K = 7$ methods and $T = 9$ datasets obtained from a studentised range table).

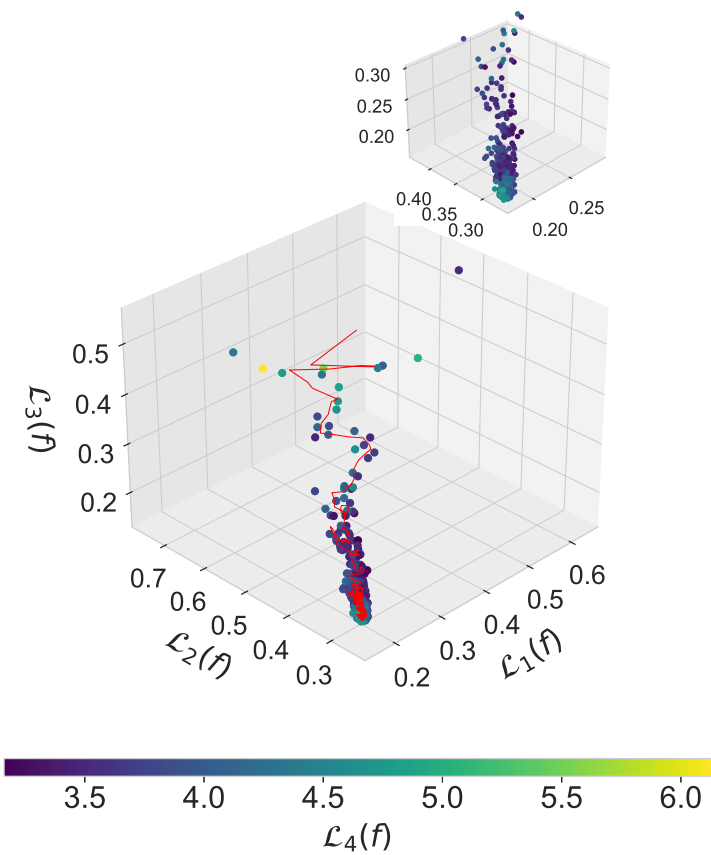
Results (continued)

- In summary, CLML achieves the best aggregate rank of 2.50 (aggregated among all measures), compared to:
- 2.90 of DELA (+13.79%),
- 3.02 of CLIF (+17.22%),
- 3.9 of MLkNN (+35.89%),
- 4.54 of C2AE (+44.93%),
- 5.18 of GNB-CC (+51.74%),
- and 6.0 of GNB-BR (+58.33%).

Results (continued)

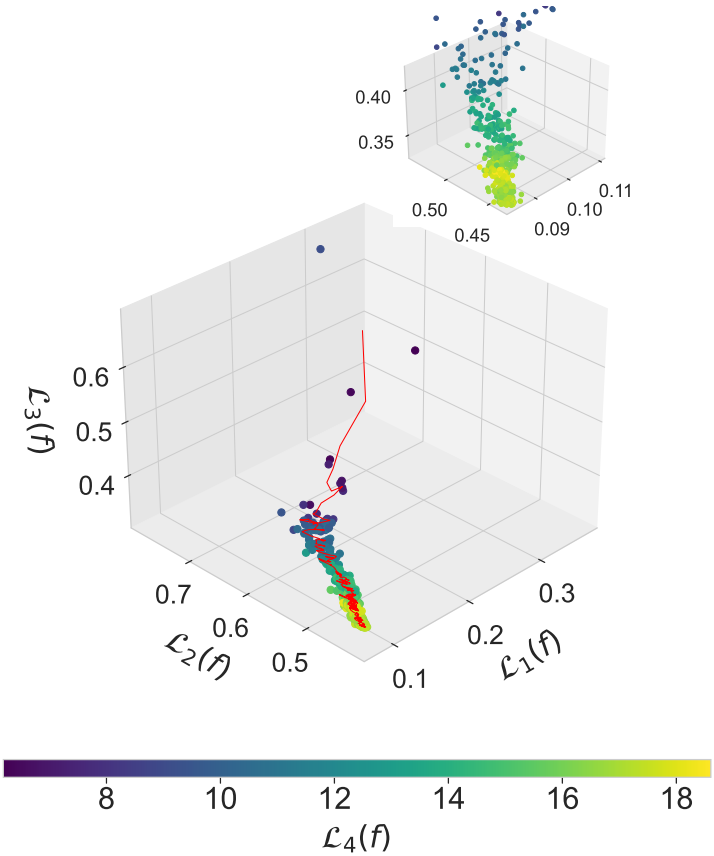


(a) CAL500

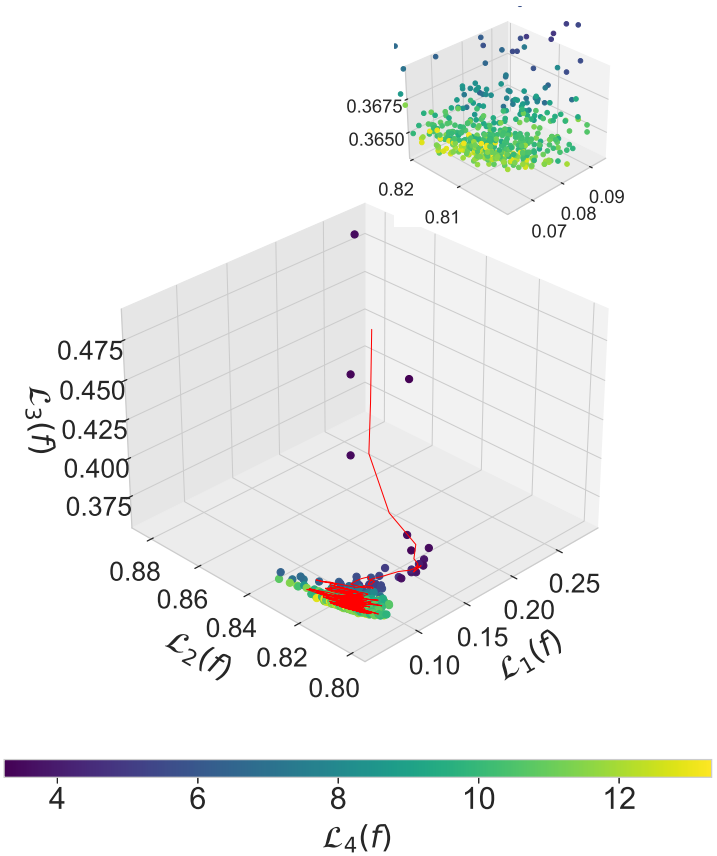


(b) emotions

Results (continued)

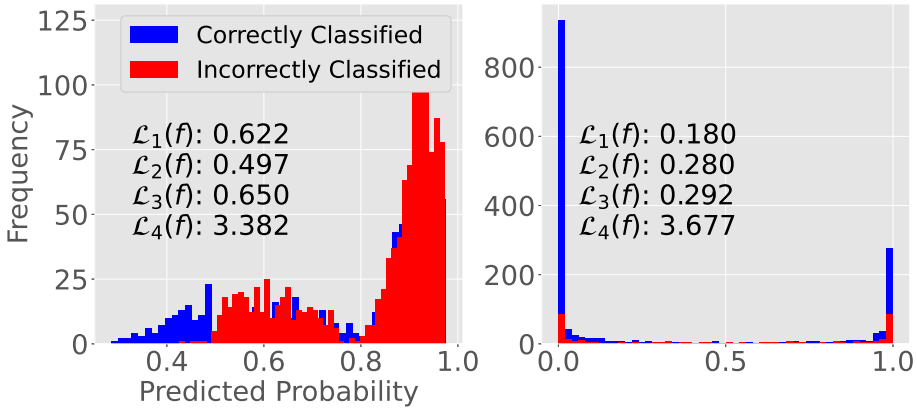


(c) tmc2007-500

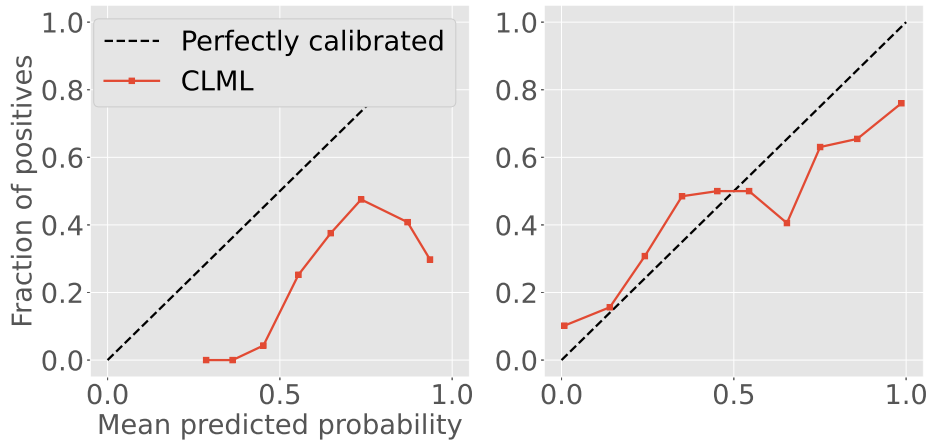


(d) IMDB-F

Results (continued)



(a) Label distributions (before and after training).



(b) Calibration curves (before and after training).

Figure: Emotions dataset before (after 1 epoch) and after training (all epoch).

Summary and Contributions

- CLML learns from multiple related, yet potentially conflicting, loss functions using a single model.
- Naturally navigates multi-dimensional loss landscape by understanding and accounting for their trade-offs.
- CLML learns to solve the problem *without* the use of a surrogate loss function, as proven consistent.
- Third, our experimental findings demonstrate that CLML consistently achieves a 13.79% to 58.33% *better* critical distance ranking against competitive state-of-the-art methods.
- These results are accentuated by the *simplicity* of CLML, which achieves SOTA results with a simple FNN.

Break Time

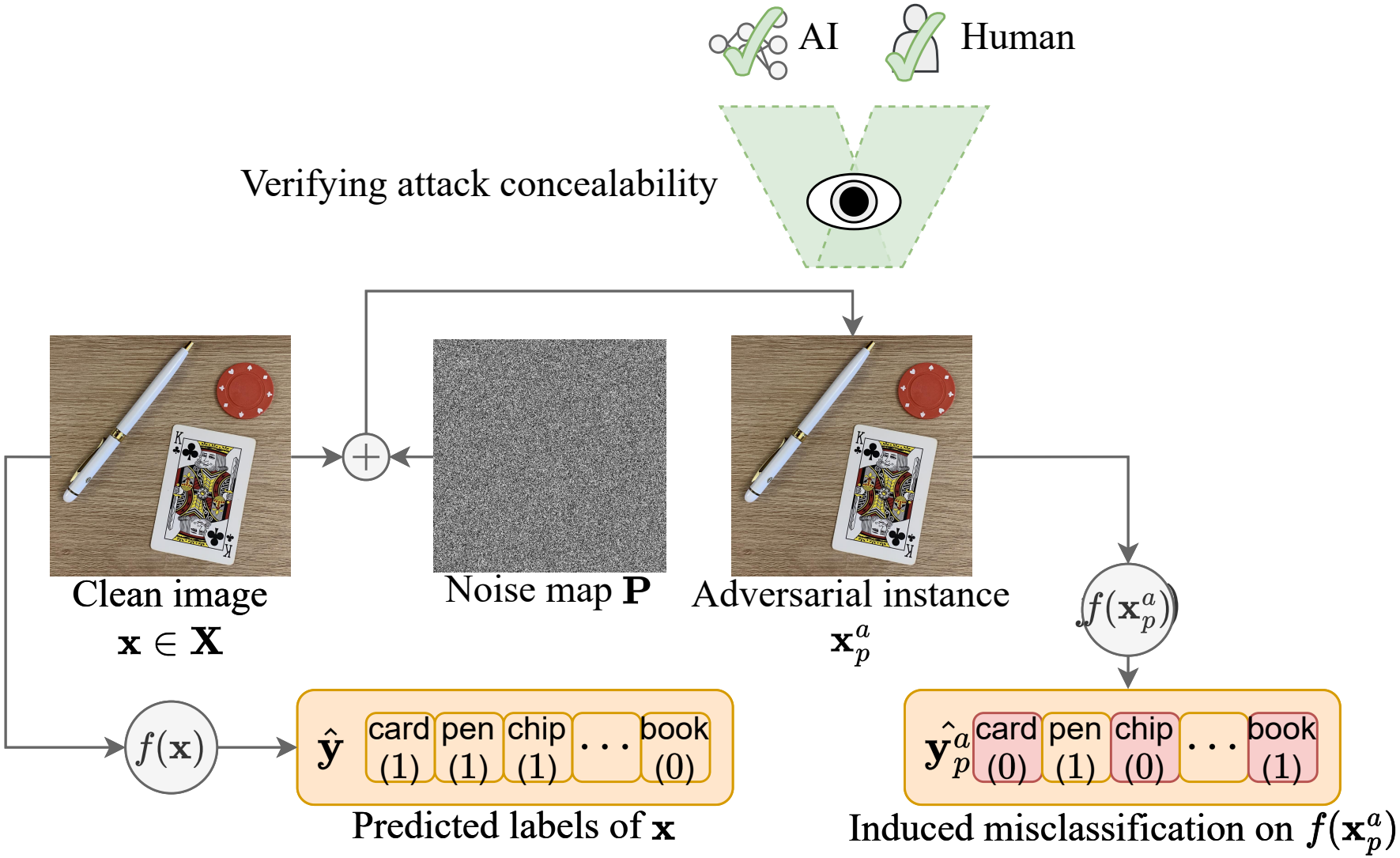
Any questions thus far?

Issues with DNN's

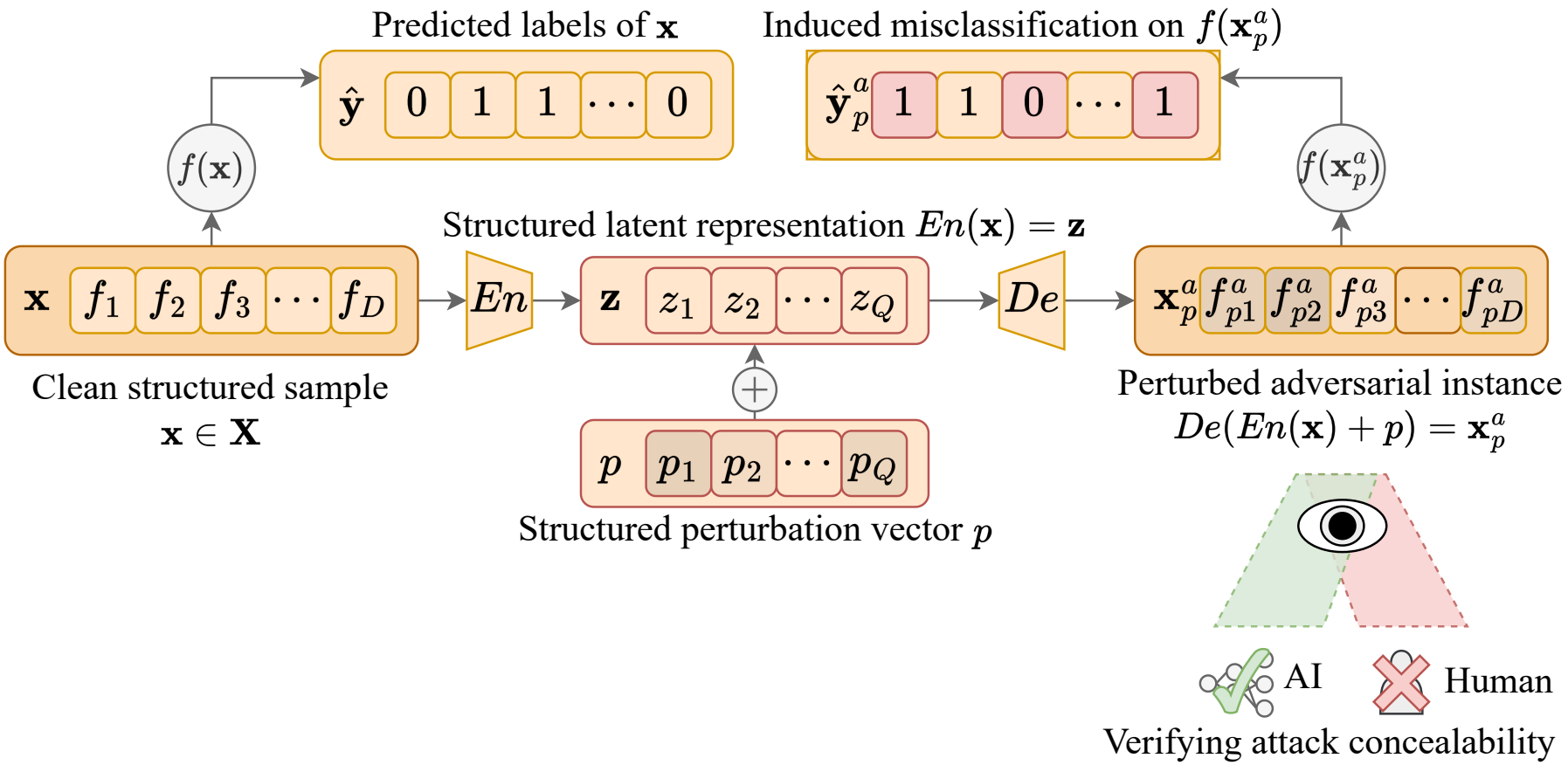
"In real-world applications such as credit checking, medical diagnosis, censoring hateful content, financial fraud, and financing of terrorism, the robustness of deep-learning models against adversarial attacks is crucial. Adversarial training addresses this by conditioning models through training on perturbed instances. Most research focuses on computer vision, leaving tabulated multi-label problems (which represents a large quantity actual real-world data) largely unexplored. This poses significant risks as adversarial methods for images do not translate well to structured tabulated data."

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In arXiv preprint arXiv:1412.6572, 2014.

Adversarial Attacks



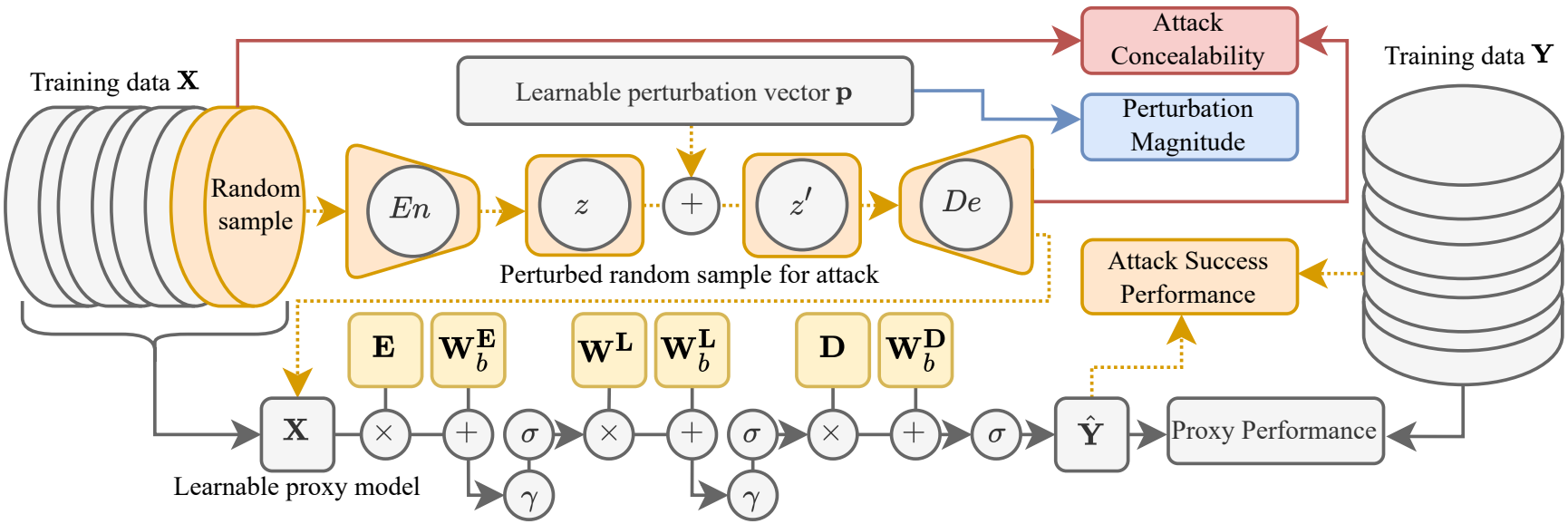
Adversarial Attacks (continued)



Challenges

- Adversarial training requires balancing attack success, attack concealability, and perturbation magnitude (multi-objective),
- attackers often do not have access to the model weights (black-box attack),
- the attack strategy is often unknown (can swap, add, or remove any quantity of labels, hence untargeted in principle),
- the concealability of structured representations is harder to define than image-based representations, and,
- generating convincing adversarial perturbations for structured (tabulated) data requires careful curation, where image-based pixel-noise methods cannot be directly applied.

Multi-label Many-objective Adversarial Perturbations (ML-MAP)



Fitness Function

$$\mathcal{L}^U = \min(1 - (\mathcal{L}^{F_1}(f(\mathbf{X}^a), \mathbf{Y}^a) - \mathcal{L}^{F_1}(f(\mathbf{X}_p^a), \mathbf{Y}^a)), 1) \tag{13}$$

$$\mathcal{L}^S = \mathcal{L}^{F_1}(f(\mathbf{X}), \mathbf{Y}) \tag{14}$$

$$\mathcal{L}^C = \frac{1}{1 + (e^{-\mu_{\mathbf{X}^a}^{MSE}})} \tag{15}$$

$$\mathcal{L}^N = \frac{1}{1 + e^{-||p||}} \tag{16}$$

To optimise this fitness landscape, we use the CLML framework by representing both proxy model weights (θ^f) and perturbation vectors ($\mathbf{p}$) as a flattened vector $[\theta^f, \mathbf{p}]$.

Methodology

- Same **nine** datasets as CLML.
- Attack **three** of the **SOTA** methods: DELA, CLIF, and End-to-end probabilistic label-specific feature learning for multi-label classification (PACA). [11]
- SOTA are trained on *clean* instances, then select a random 10% of the test set ($\mathcal{X}^{pre}$), and attack using perturbed samples ($\mathcal{X}^{pos}$).
- We assume the distributions of each dataset (test) as unknown. Non-parametric Kolmogorov-Smirnov coupled with a permutation test with a significance level of 5% to assess the earth movers distances between $\mathcal{X}^{pre}$ and $\mathcal{X}^{pos}$.
  - Repeated comparisons between a random pair of adversarial examples sampled from $\mathcal{X}^{pre} \times \mathcal{X}^{pos}$
  - Generally robust to violations of assumptions such as normality,
  - Null distribution via randomisation of the data essentially enables distribution-free testing...

[11] Hang, Jun-Yi, et al. End-to-end probabilistic label-specific feature learning for multi-label classification. In AAAI, 2022.

Attack Success

Table: Adversarial attack results. The untargeted attack results are presented in terms of micro-F_1 ($\mathcal{L}^{F_1}$), label ranking average precision ($\mathcal{L}^{AP}$), and Hamming-loss ($\mathcal{L}^{HL}$) of each clean model tested on $\mathbf{X}^a$ and $\mathbf{X}^a_p$ sampled from the test set.

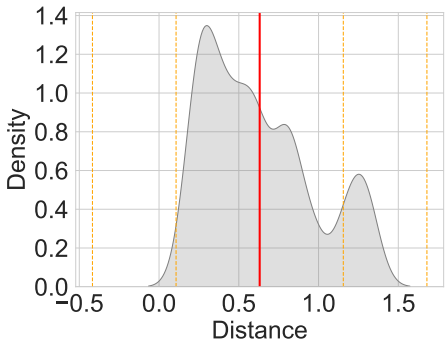
Metric	Method	emotions	flags	CAL500	enron	genbase	yeast	IMDB-F	mediamill	tmc2007-500
$\mathcal{L}^{F_1}(\downarrow)$	CLIF	0.754	0.703	0.361	0.475	0.979	0.610	0.164	0.645	0.783
	CLIF+ML-MAP	**0.429**	**0.488**	**0.330**	**0.392**	**0.000**	**0.000**	**0.070**	**0.484**	**0.280**
	DELA	0.732	0.706	0.351	0.443	1.000	0.608	0.164	0.645	0.774
	DELA+ML-MAP	**0.429**	**0.702**	**0.321**	**0.392**	**0.000**	**0.000**	**0.093**	**0.467**	**0.334**
	PACA	0.778	0.667	0.372	0.491	1.000	0.606	0.100	0.635	0.727
	PACA+ML-MAP	**0.459**	**0.293**	**0.267**	**0.061**	**0.047**	**0.431**	**0.000**	**0.403**	**0.237**
$\mathcal{L}^{AP}(\downarrow)$	CLIF	0.909	0.791	**0.380**	0.603	1.000	0.723	0.590	0.801	0.862
	CLIF+ML-MAP	**0.580**	**0.508**	0.415	**0.450**	**0.519**	**0.601**	**0.578**	**0.576**	**0.396**
	DELA	0.894	0.866	**0.404**	0.611	1.000	0.723	0.608	0.781	0.863
	DELA+ML-MAP	**0.577**	**0.697**	0.416	**0.454**	**0.487**	**0.581**	**0.587**	**0.614**	**0.421**
	PACA	0.915	0.741	0.423	0.608	1.000	0.717	0.633	0.778	0.845
	PACA+ML-MAP	**0.580**	**0.457**	**0.272**	**0.091**	**0.182**	**0.373**	**0.615**	**0.550**	**0.257**
$\mathcal{L}^{HL}(\uparrow)$	CLIF	0.157	0.262	**0.165**	0.057	0.002	0.218	0.050	0.026	0.042
	CLIF+ML-MAP	**0.370**	**0.500**	0.141	**0.063**	**0.047**	**0.301**	**0.964**	**0.037**	**0.342**
	DELA	0.176	0.238	**0.167**	0.059	0.000	0.228	0.055	0.028	0.045
	DELA+ML-MAP	**0.370**	**0.270**	0.143	**0.063**	**0.047**	**0.301**	**0.648**	**0.044**	**0.249**
	PACA	0.148	0.286	0.158	0.062	0.000	0.233	0.041	0.029	0.055
	PACA+ML-MAP	**0.364**	**0.690**	**0.225**	**0.841**	**0.080**	**0.538**	**0.036**	**0.055**	**0.325**

Attack Success (continued)

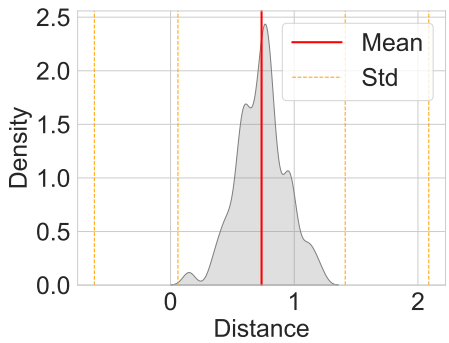
Table: Attack Success Rate (ASR) across datasets.

Method	emotions	flags	CAL500	enron	genbase	yeast	IMDB-F	mediamill	tmc2007-500
CLIF+ML-MAP	83.33%	100%	33.33%	66.0%	100%	84.7%	100%	82.1%	100%
DELA+ML-MAP	83.33%	83.33%	33.33%	66.0%	100%	73.6%	100%	81.3%	98.6%
PACA+ML-MAP	83.33%	100%	93.33%	100%	100%	98.6%	90.8%	91.2%	100%

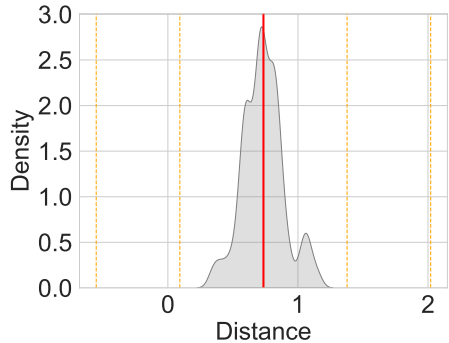
Concealability



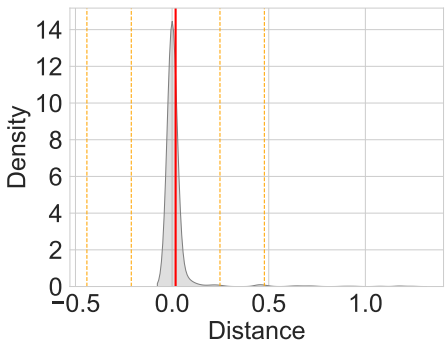
(a) flags



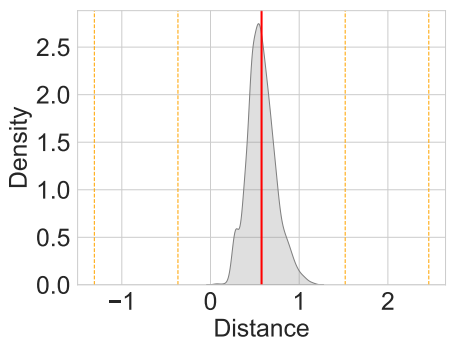
(b) CAL500



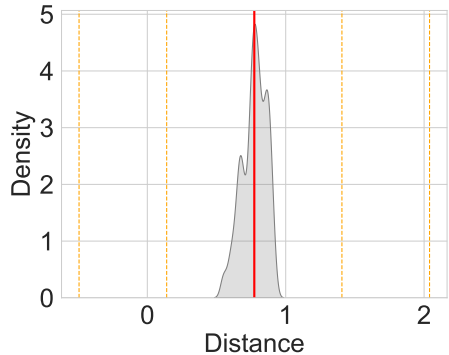
(c) emotions



(d) genbase



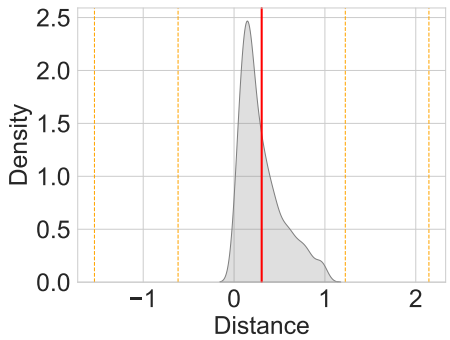
(e) enron



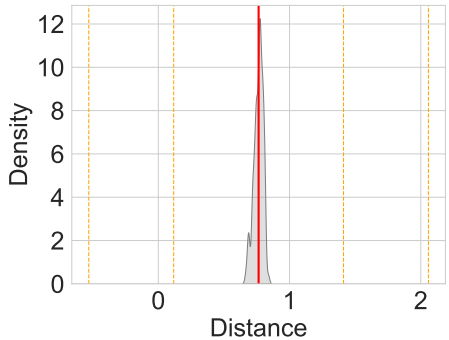
(f) yeast

Figure: Distributions of pairwise Euclidean distances between $\mathcal{X}^{pre}$ and $\mathcal{X}^{pos}$ modelled by kernel density estimation.

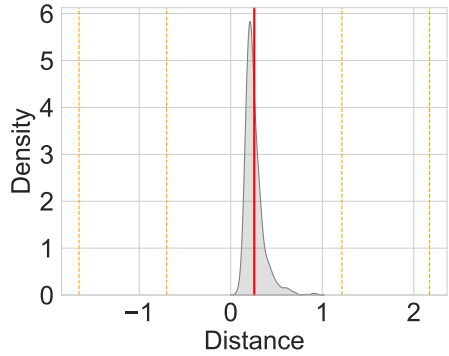
Concealability (continued)



(g) tmc2007-500



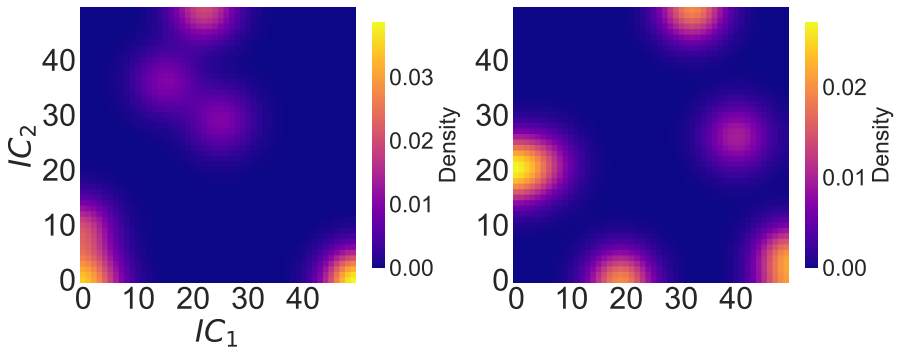
(h) mediamill



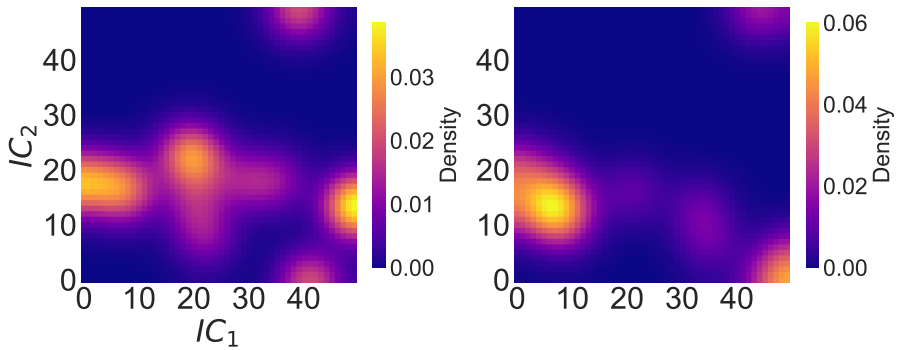
(i) IMDB-F

Figure: Distributions of pairwise Euclidean distances between $\mathcal{X}^{pre}$ and $\mathcal{X}^{pos}$ modelled by kernel density estimation.

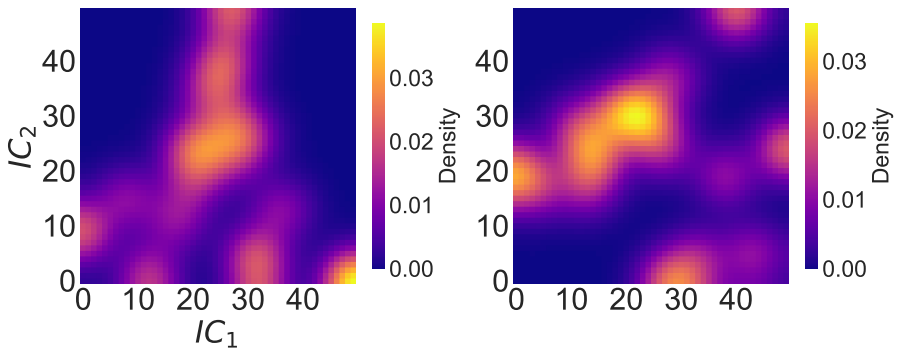
Concealability (continued)



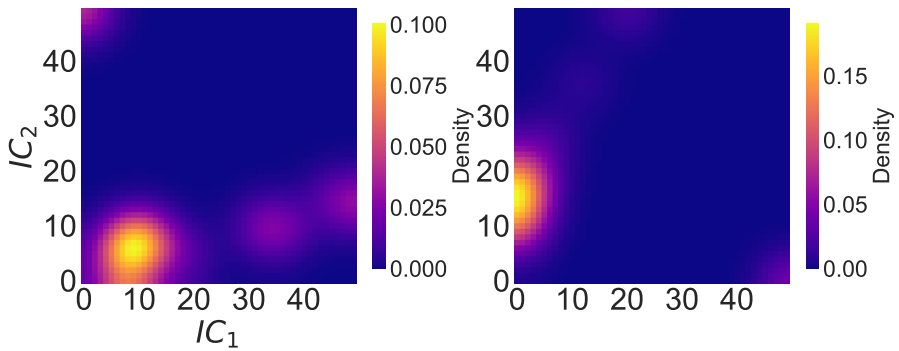
(a) flags $\mathcal{X}^{pre}$ (l) $\mathcal{X}^{pos}$ (r)



(b) CAL500 $\mathcal{X}^{pre}$ (l) $\mathcal{X}^{pos}$ (r)



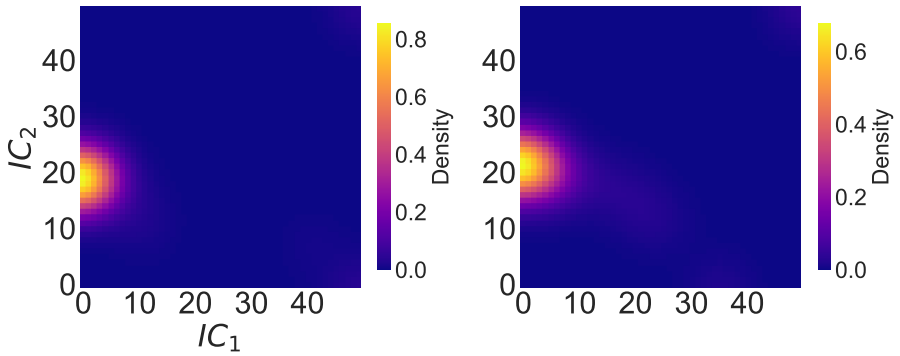
(c) emotions $\mathcal{X}^{pre}$ (l) $\mathcal{X}^{pos}$ (r)



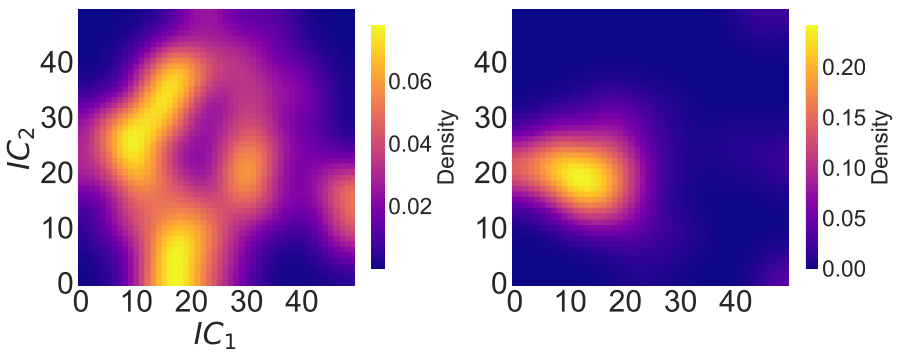
(d) genbase $\mathcal{X}^{pre}$ (l) $\mathcal{X}^{pos}$ (r)

Figure: Visualisation of Isomap projections of $\mathcal{X}^{pre}$ (left) and $\mathcal{X}^{pos}$ (right) with Guassian filtering ($\sigma = 4$) and bins ($B = 50$). Data is projected onto two Isomap components: IC_1 and IC_2.

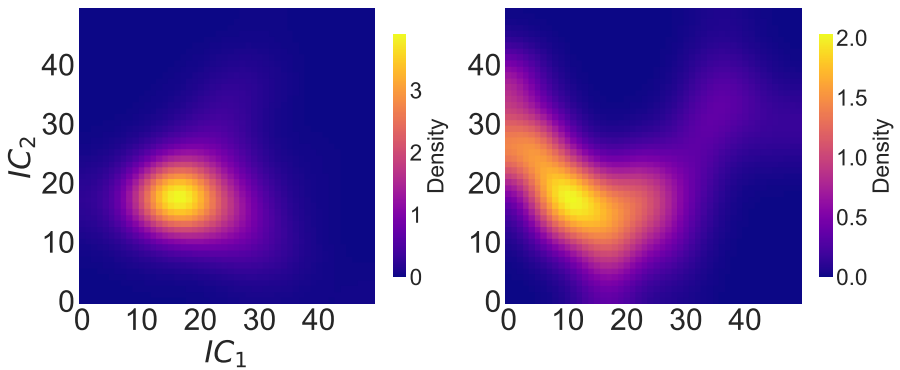
Concealability (continued)



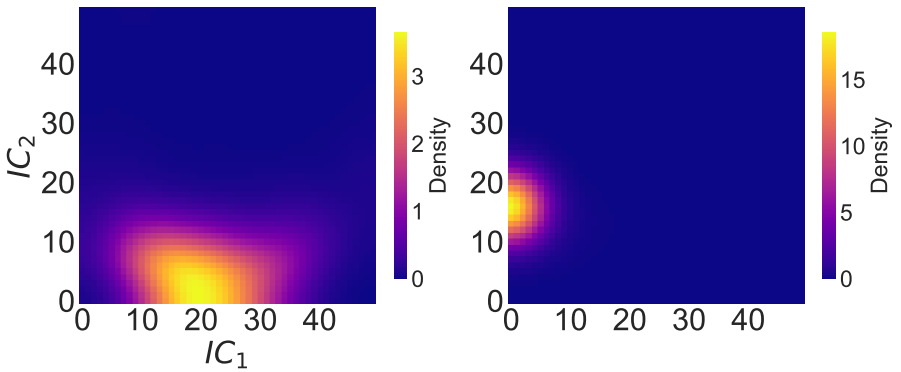
(e) enron $\mathcal{X}^{pre}$ (l) $\mathcal{X}^{pos}$ (r)



(f) yeast $\mathcal{X}^{pre}$ (l) $\mathcal{X}^{pos}$ (r)



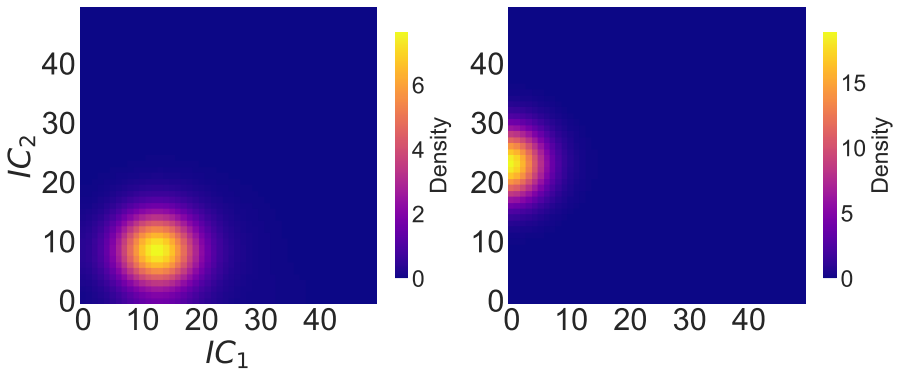
(g) tmc2007-500 $\mathcal{X}^{pre}$ (l) $\mathcal{X}^{pos}$ (r)



(h) mediamill $\mathcal{X}^{pre}$ (l) $\mathcal{X}^{pos}$ (r)

Figure: Visualisation of Isomap projections of $\mathcal{X}^{pre}$ (left) and $\mathcal{X}^{pos}$ (right) with Guassian filtering ($\sigma = 4$) and bins ($B = 50$). Data is projected onto two Isomap components: IC_1 and IC_2.

Concealability (continued)



(i) IMDB-F $\mathcal{X}^{pre}$ (l) $\mathcal{X}^{pos}$ (r)

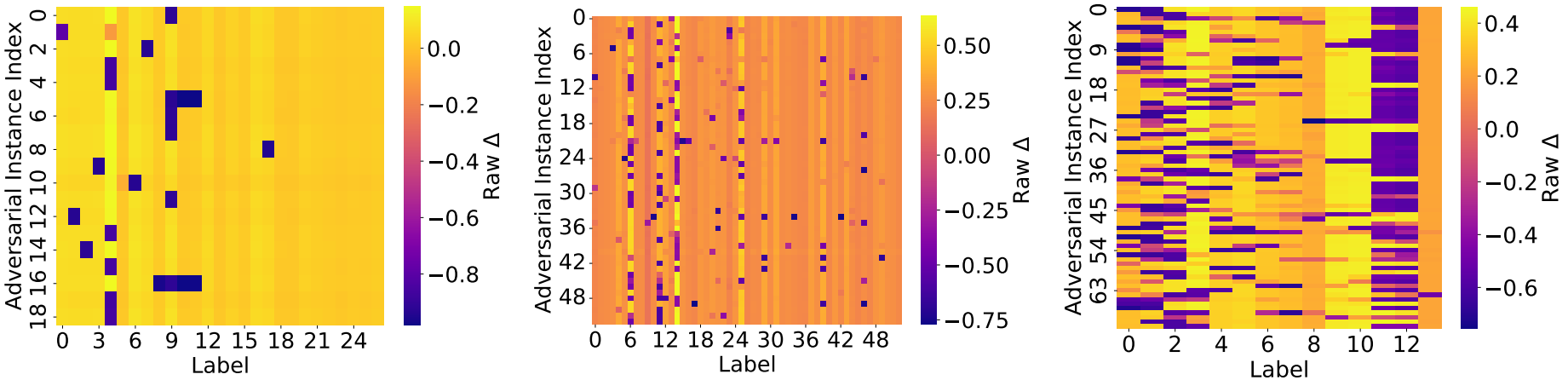
Figure: Visualisation of Isomap projections of $\mathcal{X}^{pre}$ (left) and $\mathcal{X}^{pos}$ (right) with Guassian filtering ($\sigma = 4$) and bins ($B = 50$). Data is projected onto two Isomap components: IC_1 and IC_2.

Concealability (continued)

Table: Observed earth movers distance (EMD) and Kolmogorov-Smirnov test between $\mathcal{X}^{pre}$ and $\mathcal{X}^{pos}$ via a permutation test.

	flags	CAL500	emotions	enron	genbase	yeast	tmc2007-500	mediamill	IMDB-F
EMD	$< \epsilon$	0.0004	$< \epsilon$	0.0016	0.0008	0.0036	0.0172	0.1392	0.0568
p-value	**1.0**	**1.0**	**1.0**	**0.7908**	**0.8928**	**0.4863**	**0.0983**	$< \epsilon$	$< \epsilon$

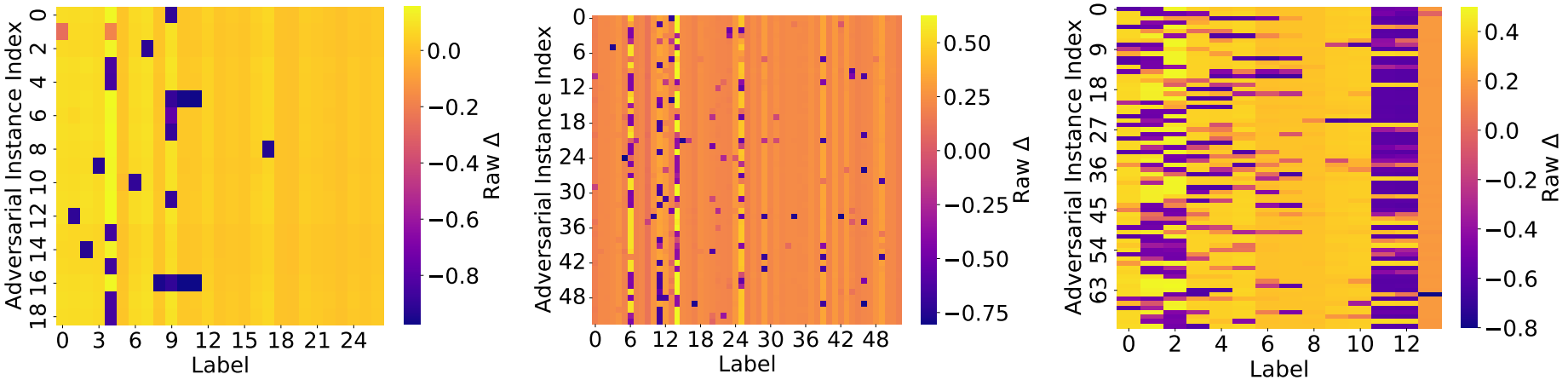
Induced Label Changes on DELA



(d) genbase (e) enron (f) yeast

Figure: Visualisation of the change in class label confidence of DELA between $\mathbf{X}^{pre}$ and X^{pos}. Results are presented on genbase through yeast (d-f).

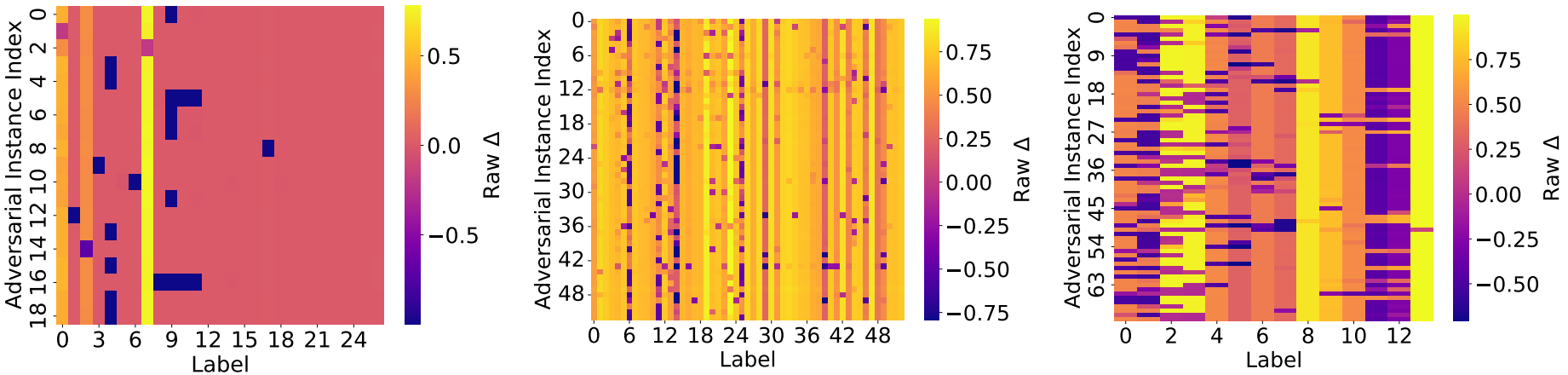
Induced Label Changes on CLIF



(d) genbase (e) enron (f) yeast

Figure: Visualisation of the change in class label confidence of CLIF between $\mathbf{X}^{pre}$ and X^{pos}. Results are presented on genbase through yeast (d-f).

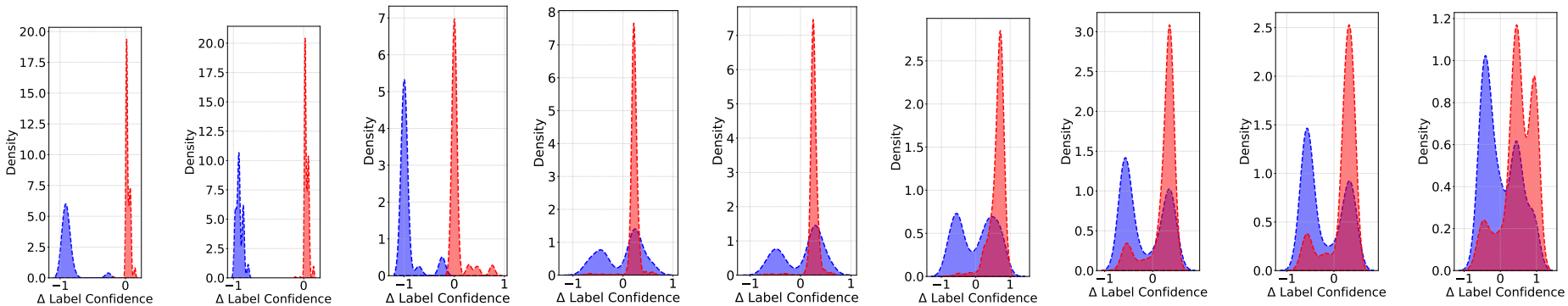
Induced Label Changes on PACA



(d) genbase (e) enron (f) yeast

Figure: Visualisation of the change (Δ) in class label confidence of PACA between $\mathbf{X}^{pre}$ and X^{pos}. Results are presented on genbase through yeast (d-f).

Emergent Attack Strategy - Exploiting Decision Boundary Weaknesses



(d) genbase (e) enron (f) yeast

Figure: Visualisation of the change (Δ) in class label confidence on each dataset genbase through yeast (d-f).

Summary and Contributions

- We proposed a novel approach to generate convincing **tabulated** adversarial examples via **perturbations** to their **latent representations**.
- The training is achieved via a novel adversarial many-objective optimisation framework, where the learned perturbation vector and proxy multi-label learner are "**adversaries**" (akin to GANs).
- ML-MAP is designed to learn convincing perturbations using a **proxy model**, simulating black-box attacks
- ML-MAP induces significant **misclassification on SOTA**, between 81.3% and 100% attack success rate on large-scale datasets.
- ML-MAP generates statistically **concealable** samples.
- Interestingly, samples generated by ML-MAP can *indirectly* learn **vulnerabilities** of a decision boundary by inducing false positives and false negatives.