On the Comparison of Learned Classifiers

Soaibuzzaman, Jenny Döring, Srinivasulu Kasi, and Jan Oliver Ringert

14 November, 2025, Toledo, Spain

The most similar real face of ID 8 (no access)

MLDiff Witness:
MLP: ID 8 (no access)
LogReg: access

The most similar real face with access

Introduction & Example

- Numerous complex, real-world applications rely on Machine Learning (ML) classifiers

- Example classification problem
  - A Support Vector Machine (SVM) & Decision Tree (DT) are both trained on the Iris dataset
  - Accuracy: SVM=96% and DT=96%

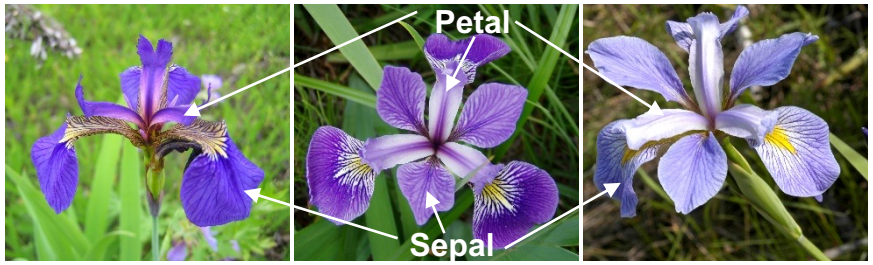


Figure: Iris flower data set (https://en.wikipedia.org/wiki/Iris_flower_data_set)

- Should we use the SVM or the DT?
- How are the two classifiers different?
- Is it relevant which one to use if they agree on known data (the train/test dataset)?

DT classifies some instances as *Virginica* (**medical use**), while SVM classifies them as *Versicolor* (**poisonous**)

Motivation of MLDiff

- State-of-the-art formal analysis of ML models lacks systematic methods to compare multiple classifiers
- We constantly need to update the models and compare multiple classifiers.
  - Old vs New
- Understanding classifier variants during software design and evolution is crucial for improving model quality and trust
- MLDiff aims to uncover and present differences (witnesses), i.e., disagreements, of classifiers (even those not observable in the dataset)

Example: Different Class Labels

- MLP: Classifies faces into 10 employee IDs.
  - 0-5 have access and 6-9 have no access
- LogReg: Classifies 16 faces as simple "access" or "no access"
- Different datasets (10 vs 16 faces) and different class labels
- Will the LogReg classifier ever grant "access" to a person that the MLP classifies within 6-9 (no access)?

The LogReg classifier grants access to employee ID 8 when the MLP classifier denies it.

Figure: A conflict of a Multi-Layer Perceptron and a Logistic Regression classifier detected by MLDiff
(Source: The Olivetti faces dataset from scikit-learn)

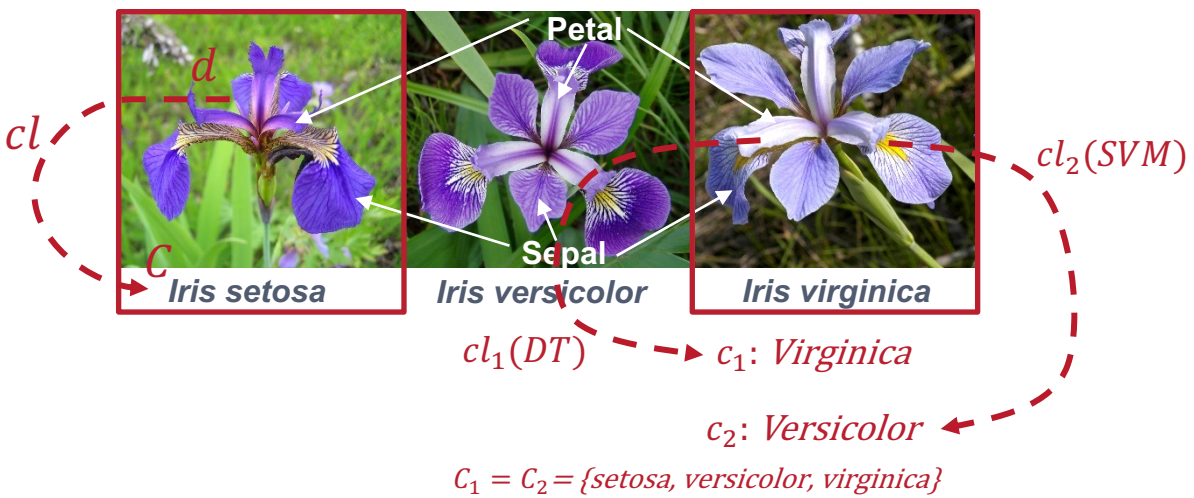
MLDiff Concept

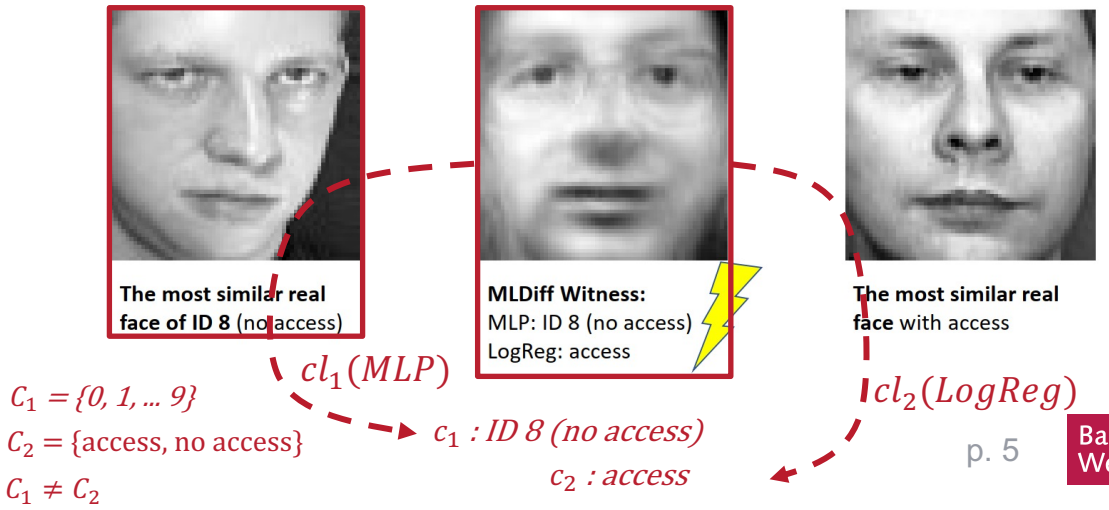
- **Instance** (d): A valuation of features $(\mathrm{x} \in X)$, typically $d \in D = \mathbb{R}^{|x|}$
- **Classifier** (cl) : A classifier is a total function that maps an instance to a class $(c \in C)$

$$cl\colon\ \mathbb{R}^{|x|} \to C$$

- **Classifier Combination:** For two classifiers cl_1 on features $X_1$and cl_2 on features X_2, we **encode SMT assertions** for

$$\forall d \in \mathbb{R}^{|X_1 \cup X_2|}\colon cl_1 \oplus cl_2\ (d) = cl_1(d|_{X_1}) \times cl_2(d|_{X_2})$$

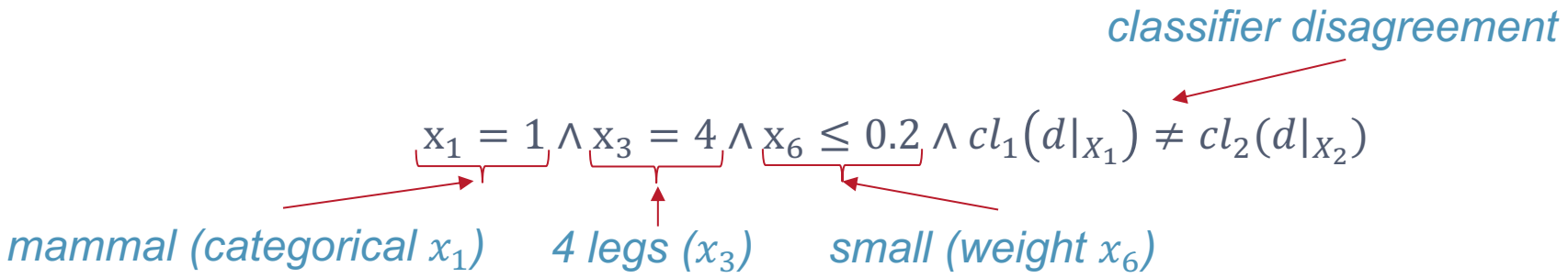




Use Cases and Queries

- MLDiff enables generation and inspection of witnesses in safety- or interpretability-critical contexts where ML models require rapid evaluation and iteration, e.g., in certification, debugging, or auditing scenarios

- MLDiff supported Queries
  - Differences: $cl_1(d|_{X_1}) \neq cl_2(d|_{X_2})$
  - Extension with **custom/domain constraints**:



Feature Constraints (Meaningful Witnesses)

- MLDiff finds instance that satisfy a given query
- Without constraints, MLDiff may produce "undesired witnesses" that are mathematically correct but **nonsensical in the real world**
- Queries can be **extended with constraints on feature values**

- Supported Constraints in the Prototype
  - Upper and lower bounds (e.g., $x \geq 0$).
  - Restriction of feature values to Int or Real values.
  - Support for categorial features

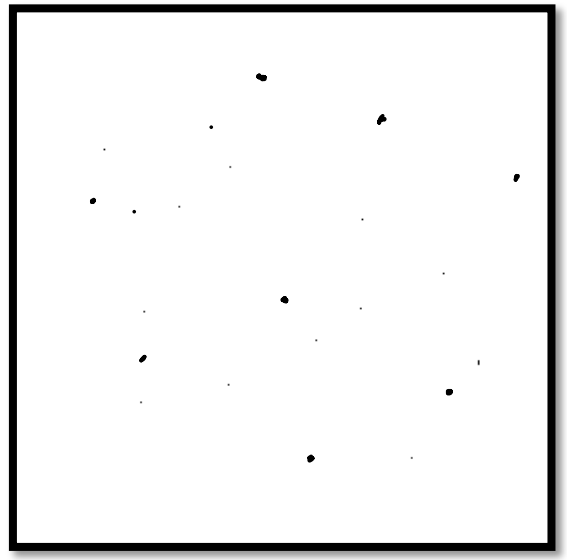


Figure: A Diff Witness in Digits Dataset

MLDiff SMT-based Prototype

- First, declare the **input features** (e.g., petal length, pixel values)
- Also declare two variables to hold the final class predictions, cl_1 and cl_2
- Next, encode the ***entire logic* of each classifier** as a set of **assertions**
- Finally, we add our **query**
- Our prototype can do this for **any combination** of these models

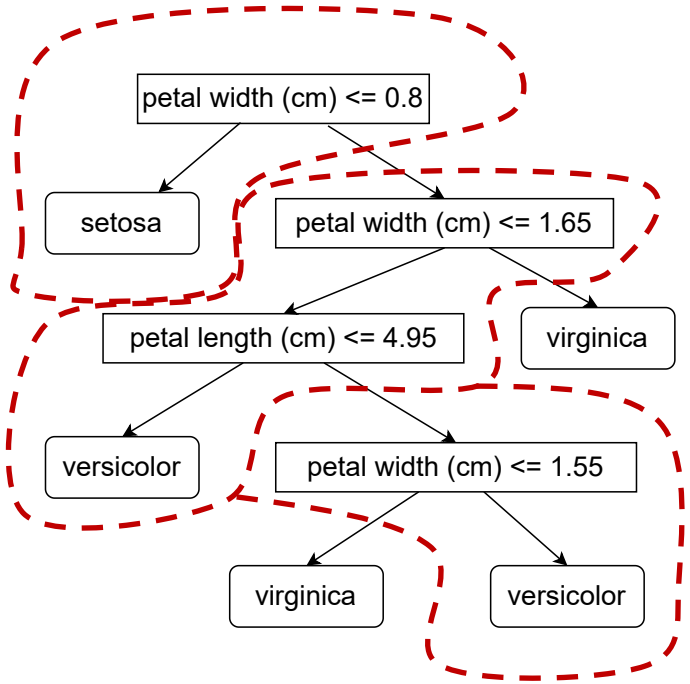
```
(declare-const x0 Real) ; one constant for each feature
; ...
(declare-const xn Real)
(declare-const cls1 Int) ; predicted class of first classifier
(declare-const cls2 Int) ; predicted class of second classifier
; assertion for classifier 1 relating x1..xn to cls1
; assertion for classifier 2 relating x1..xn to cls2
(assert (not (= cls1 cls2))) ; example query for disagreement
```

Currently supports

- *Decision Trees*
- *Logistic Regression*
- *Multi Layer Perceptron (ReLU, identity)*
- *Support Vector Machine (linear kernels)*

SMT Translation of Classifiers (Decision Tree)

- A single path is a **conjunction** of all the decisions made along the way.
- We assert that this path being true is **equivalent** to the classifier outputting that class (e.g., class 0).
- If a class has multiple paths, we combine them all with a **disjunction**.
- The complete SMT formula is one large **And** of all the rules for all the classes.




```
And( ;listing classes with all their paths through the DT
  Or(And(x3 <= 0.8)) == (cls == 0), ;setosa (only one path of length one)
  Or(And(x3 > 0.8, x3 <= 1.65, x2 > 4.95, x3 > 1.55),
     And(x3 > 0.8, x3 <= 1.65, x2 <= 4.95)) == (cls == 1), ;versicolor
  Or(And(x3 > 0.8, x3 <= 1.65, x2 > 4.95, x3 <= 1.55),
     And(x3 > 0.8, x3 > 1.65)) == (cls == 2)) ;virginica
```

Prototype Evaluation Research Questions

- **RQ1:** How many disagreements does our MLDiff comparison find?
- **RQ2:** What is the analysis cost for different classifiers?
- **RQ3:** What is the overhead of adding feature constraints?

Evaluation Setup

- Prototypical implementation can compare any combination of:
  - Decision Trees (DT)
  - Linear SVM
  - Logistic Regression (LogReg)
  - Multi-Layer Perceptrons (MLP) (with ReLU)
- Datasets: Iris, Digits, Olivetti Faces, and Breast Cancer.
  - Diverse features
- Classifiers: Trained all 4 classifiers on all 4 datasets
- Backend: Z3 SMT-solver

Model	Accuracy				F1-Score			
	i	d	o	c	i	d	o	c
DT	100	86.94	65	94.74	1	0.97	0.93	0.99
SVM	100	96.39	100	93.86	0.98	0.99	1	0.92
LOGREG	100	96.94	100	95.61	0.98	0.99	1	0.95
MLP	83.33	92.22	75	96.49	0.88	0.95	0.74	0.92

Table: Accuracy and F1-score of classifiers on datasets:
Iris (i), Digits (d), Olivetti Faces (o), and Breast Cancer (c)

RQ1: How many disagreements does our MLDiff comparison find? (Effectiveness)

- MLDiff was able to find a witness for 100% of all possible disagreement combinations across all classifier pairs and datasets.
- Compared this to using the test set alone.
- For half of the 16*4 classifier-dataset-combinations, 26% (median) or more disagreements can be found among existing dataset elements
- When excluding the binary classification Breast Cancer dataset (which only offers two possible disagreements), the median disagreements-in-dataset percentage drops to 17%, i.e., MLDiff is needed to discover the remaining 83% synthetic witnesses

RQ2: What is the analysis cost (time) to find all disagreements? (Efficiency)

- Low Cost for Simpler classifiers: DT, SVM, and LogReg (0.01 to 24.31 seconds)
- Cost increases significantly for MLPs (up to 1382.27 seconds)
  - high dimensionality and many classes (Digits and Olivetti Faces datasets)
- **The order of the assertions** encoded into the Z3 solver had a **significant impact**
  - **SMT Instability and Encoding Order**

	DT				SVM				LOGREG				MLP			
	i	d	o	c	i	d	o	c	i	d	o	c	i	d	o	c
DT	0.01	0.01	0.01	0	0.02	1.8	1.02	0.03	0.02	2.75	1.32	0.03	6.03	247.9	171.97	0.69
SVM	0.02	2.31	1	0.02	0.01	0.06	0.02	0.01	0.02	**10.89**	5.32	0.03	8.19	**1382.27**	833.39	0.65
LOGREG	0.02	2.74	0.98	0.02	0.02	**24.31**	7.89	0.03	0	0.06	0.02	0.01	7.2	1133.86	838.69	0.56
MLP	7.06	97.54	242.46	0.32	8.03	**254.06**	1009.8	0.44	8.49	324.44	1180.57	0.58	0.01	0.03	0.02	0.01

Table: Analysis time (in seconds) of finding all differences of pairs of classifiers on datasets:
Iris (i), Digits (d), Olivetti Faces (o), and Breast Cancer (c)

RQ3: What is the overhead of adding feature constraints?

- The general trend observed was that adding constraints with more features makes the SMT problem **computationally more expensive**
- The computation frequently **timed out (TO)** after the 1-hour limit, particularly when involving **complex models** (like the Multi-Layer Perceptron/MLP) and datasets with **higher dimensionality and features**
- Datasets with fewer features, specifically the **Iris and Breast Cancer** datasets, were handled **relatively quickly**, even with the imposition of feature constraints, resulting in no time outs
- Interestingly, there were a **few instances** where imposing feature constraints **reduced the computation time** instead of increasing it

	DT				SVM				LOGREG				MLP			
	i	d	o	c	i	d	o	c	i	d	o	c	i	d	o	c
DT	0.91	1.31	3.19	1.08	1.04	TO	323.72	1	0.96	TO	206.67	1	1.23	TO	TO	7.09
SVM	0.93	1441.79	188.65	0.96	0.87	1.08	2.75	1.02	1.02	TO	TO	1.02	1.54	TO	TO	2.4
LOGREG	1.02	TO	495.57	0.99	1.13	TO	TO	1.04	1.07	1.1	4.52	1.03	1.05	TO	TO	6.32
MLP	1.46	TO	TO	16.03	2.16	TO	TO	2.72	1.08	TO	TO	3.18	0.97	1.18	3.02	1.04

Table: Overhead of adding feature constraints computed as factors for times. TO indicates a timeout after 1h

Challenges and Open Problems

- Developing a domain-expert-friendly query language
- Generating relevant and interesting in-domain examples
- SMT-based prototype
  - Supporting a larger set of functions in classifiers (SMT's arithmetic limitations)
  - Scaling to complex models and queries (approximation and decomposition)
- Exploring alternatives to SMT-based prototype (see RQ1 dataset-based)

Conclusion

- Introduced MLDiff framework for witness-based comparison of classifiers
- Presented a prototypical SMT-based implementation
- Evaluation shows high effectiveness, but high computational cost for more complex classifiers and high-dimensional datasets

https://github.com/se-buw/MLDiff