UNIVERSITY OF HULL

SafeML: An Approach for Safety Monitoring of Machine Learning Classifiers through Statistical Difference Measure

Koorosh Aslansefat, Ioannis Sorokos, Declan Whiting, Ramin Tavakoli Kolagari, and Yiannis Papadopoulos

Email: k.aslansefat-2018@hull.ac.uk

Fraunhofer IESE

TECHNISCHE HOCHSCHULE NÜRNBERG GEORG SIMON OHM

APD An NEC Company

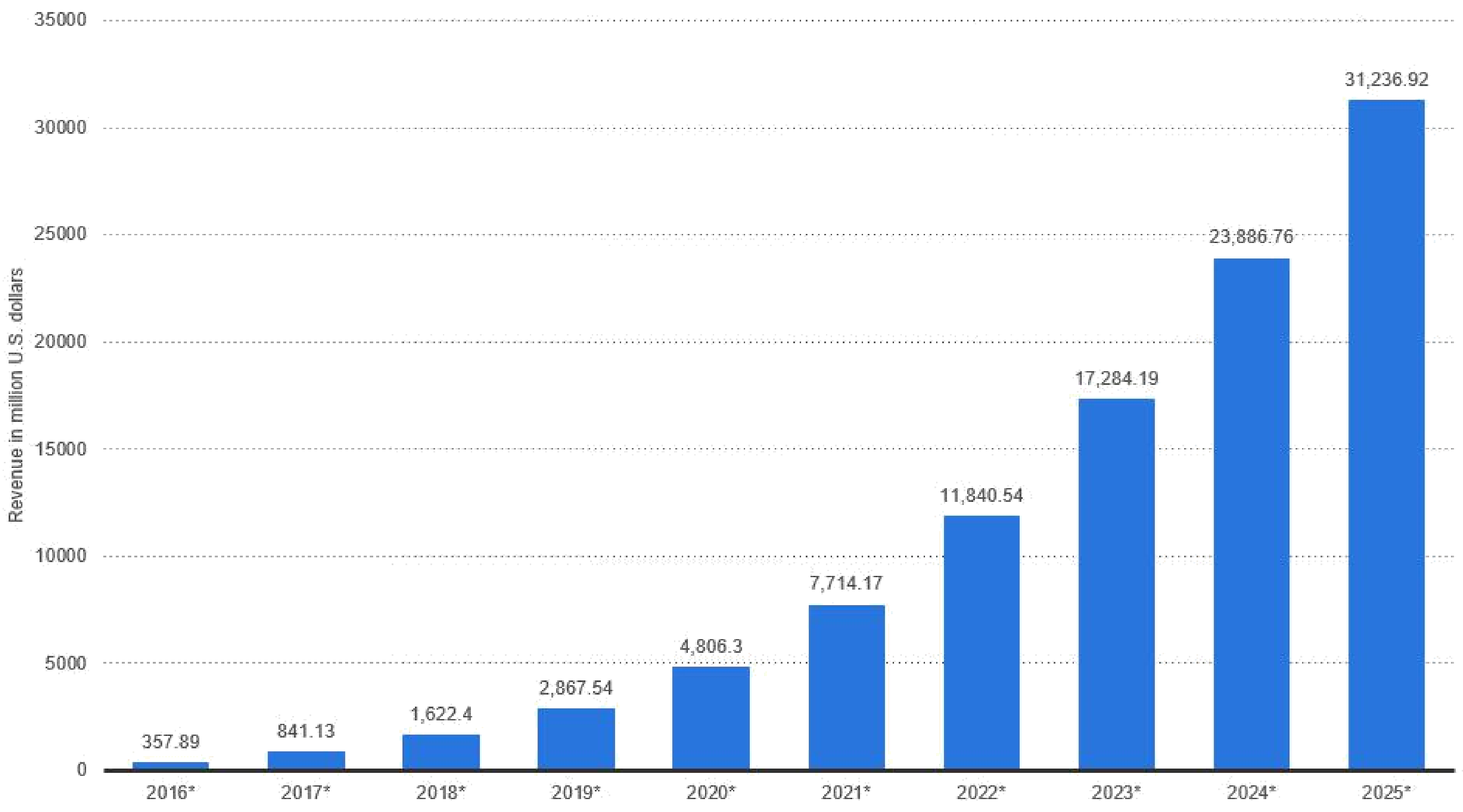
UNIVERSITY OF HULL

Table of Content

What I am going to discuss

Enterprise artificial intelligence market revenue worldwide 2016-2025

Revenues from the artificial intelligence for enterprise applications market worldwide, from 2016 to 2025 (in million U.S. dollars)



https://www.statista.com

Uber self-driving car kills a pedestrian

2018 in Review: 10 AI Failures, https://medium.com/syncedreview/2018-in-review-10-ai-failures-c18faadf5983

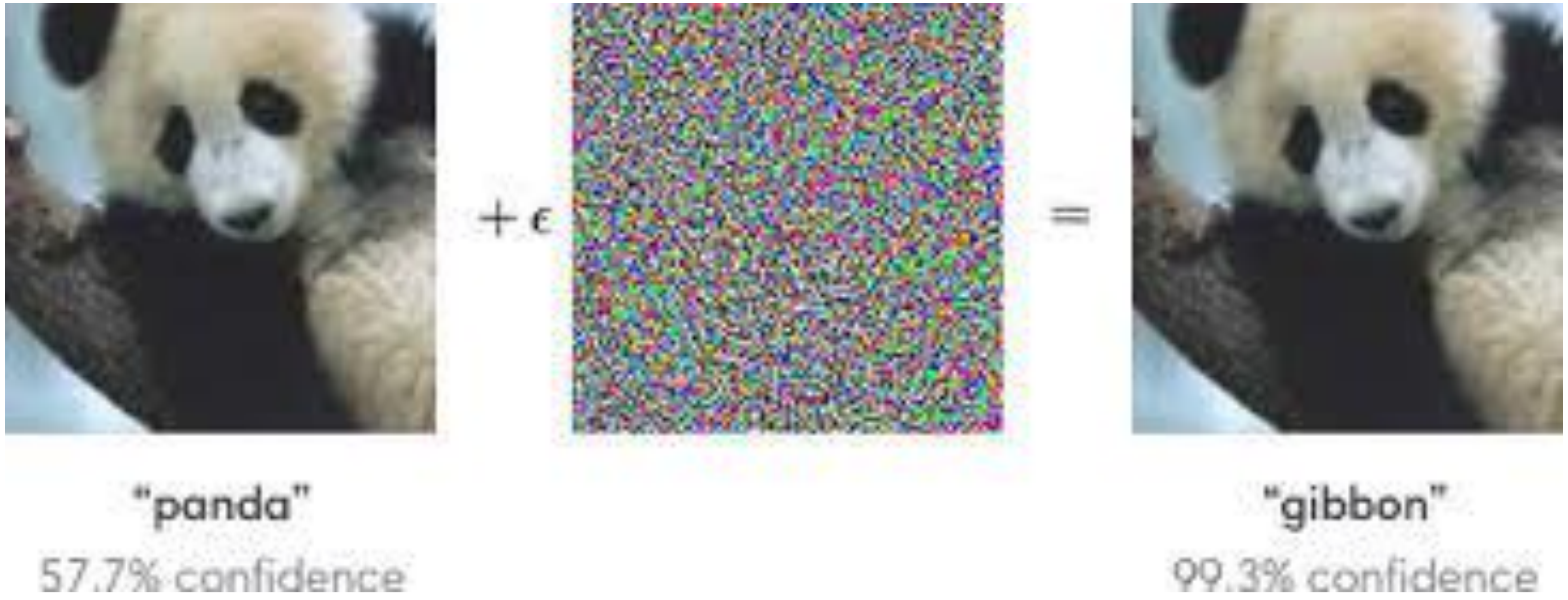
SafeML Problem Statement

(a) Input 1

(b) Input 2 (darker version of 1)

K. Pei, et al. K., Cao, Y., Yang, J., & Jana, S. (2017). Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles* (pp. 1-18).

SafeML Problem Statement



https://openai.com/blog/adversarial-example-research/

SafeML Problem Statement

https://www.reddit.com/r/ProgrammerHumor/comments/cl2rve/so_a_friend_of_mine_was_working_on_an_opencvml/

AI Safety Issues

Amodei et al. (2016). ***Concrete Problems in AI Safety***.

UNIVERSITY OF HULL

SafeML Project Goal

Accuracy Estimation

Estimating the ML Classifier Accuracy through Statistical Differences

Safety Monitoring

Safety Monitoring through a proposed human-in-loop procedure

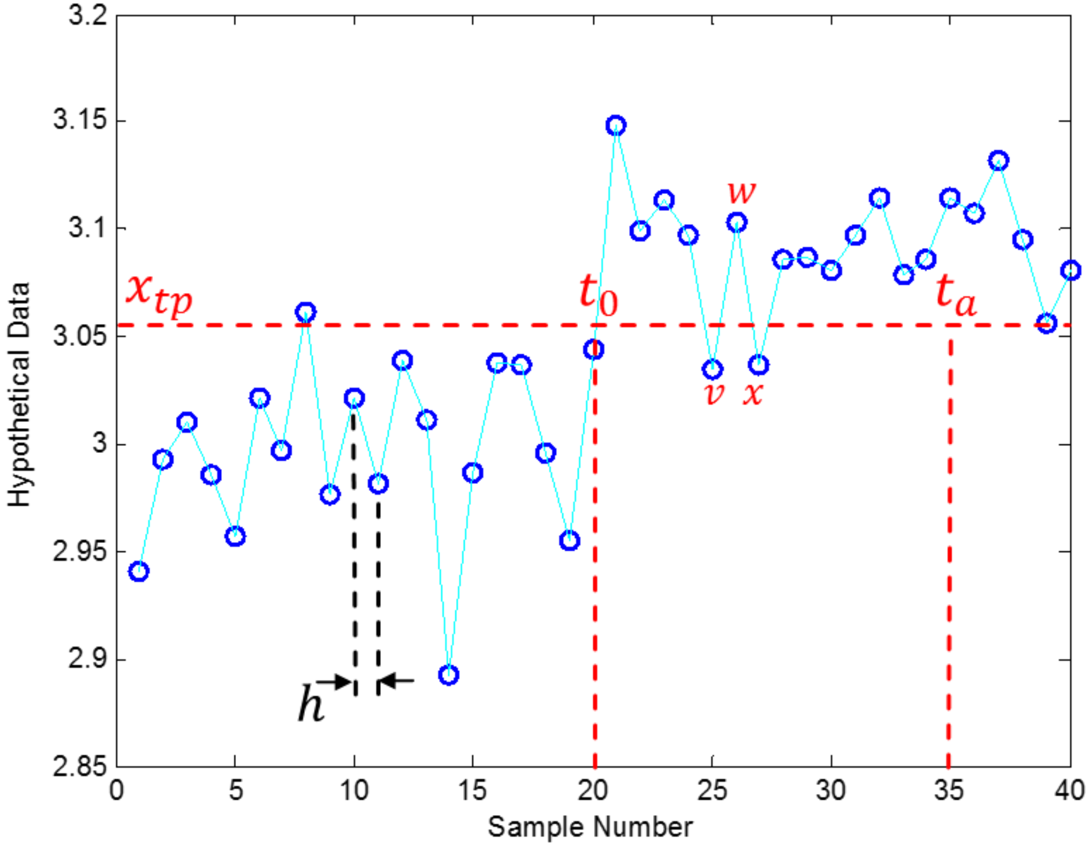
XAI: Explainable Artificial Intelligence

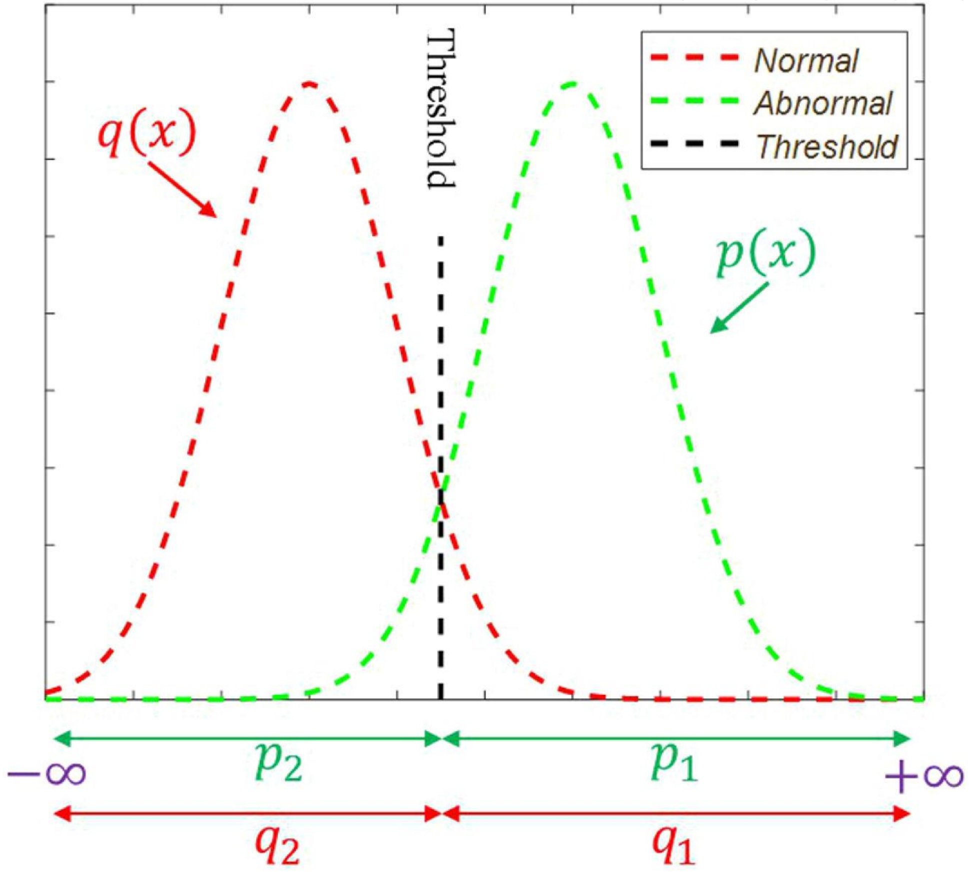
Providing Explainable Artificial Intelligence using Statistical Differences

SafeML

An Example

$$\begin{cases} Class\ 1:\ x(t) \sim N(3,1) & t_0 < 20h \\ Class\ 2:\ x(t) \sim N(5,1) & t_0 > 40h \end{cases}$$





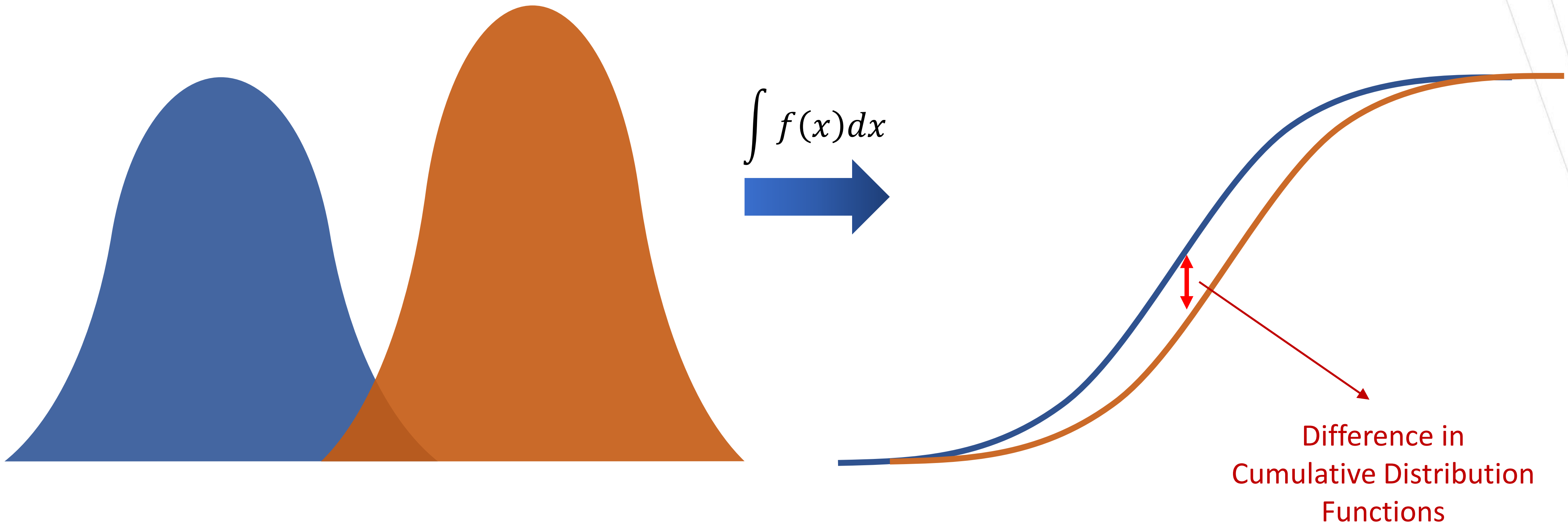
$$FDR(T_d) = q_1(T_d) = \int_{T_d}^{+\infty} q(x)\,\mathrm{dx}$$

$$MDR(T_d) = p_2(T_d) = \int_{-\infty}^{T_d} p(x)\,\mathrm{dx}$$

Probability Density Functions

Consider a hypothetical one dimensional data with nine classes

Probability Density Functions and Cumulative Functions



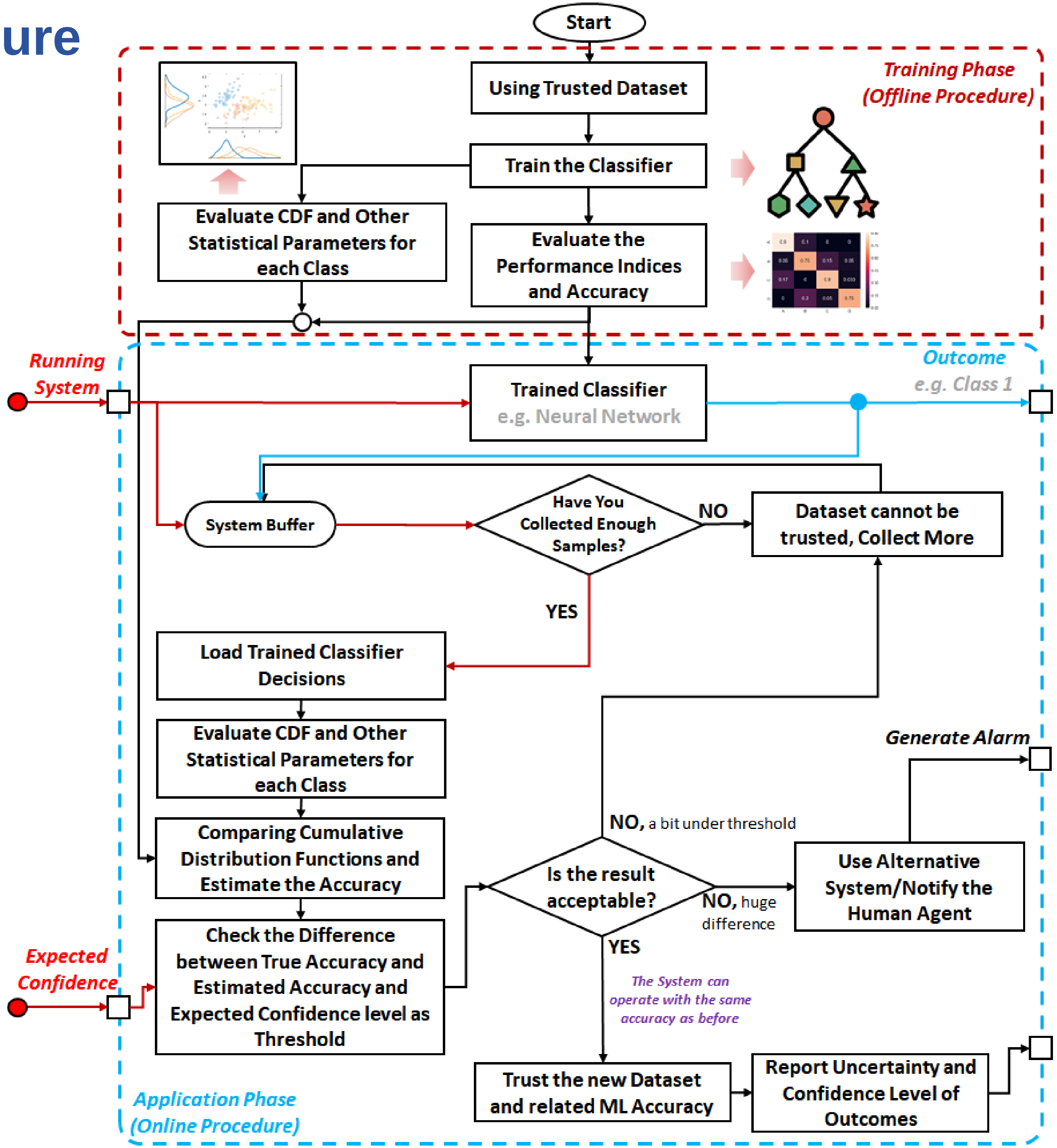
Cumulative Distribution Function (CDF) Distance Measures

- Wasserstein
- Kolmogorov-Smirnov
- Kuiper
- Anderson-Darling
- Cramer-Von Mises
- Wasserstein + Cramer-Von Mises (DTS)

UNIVERSITY
OF HULL

SafeML Procedure

Proposed Procedure



UNIVERSITY
OF HULL

SafeML

Numerical Examples

UNIVERSITY OF HULL

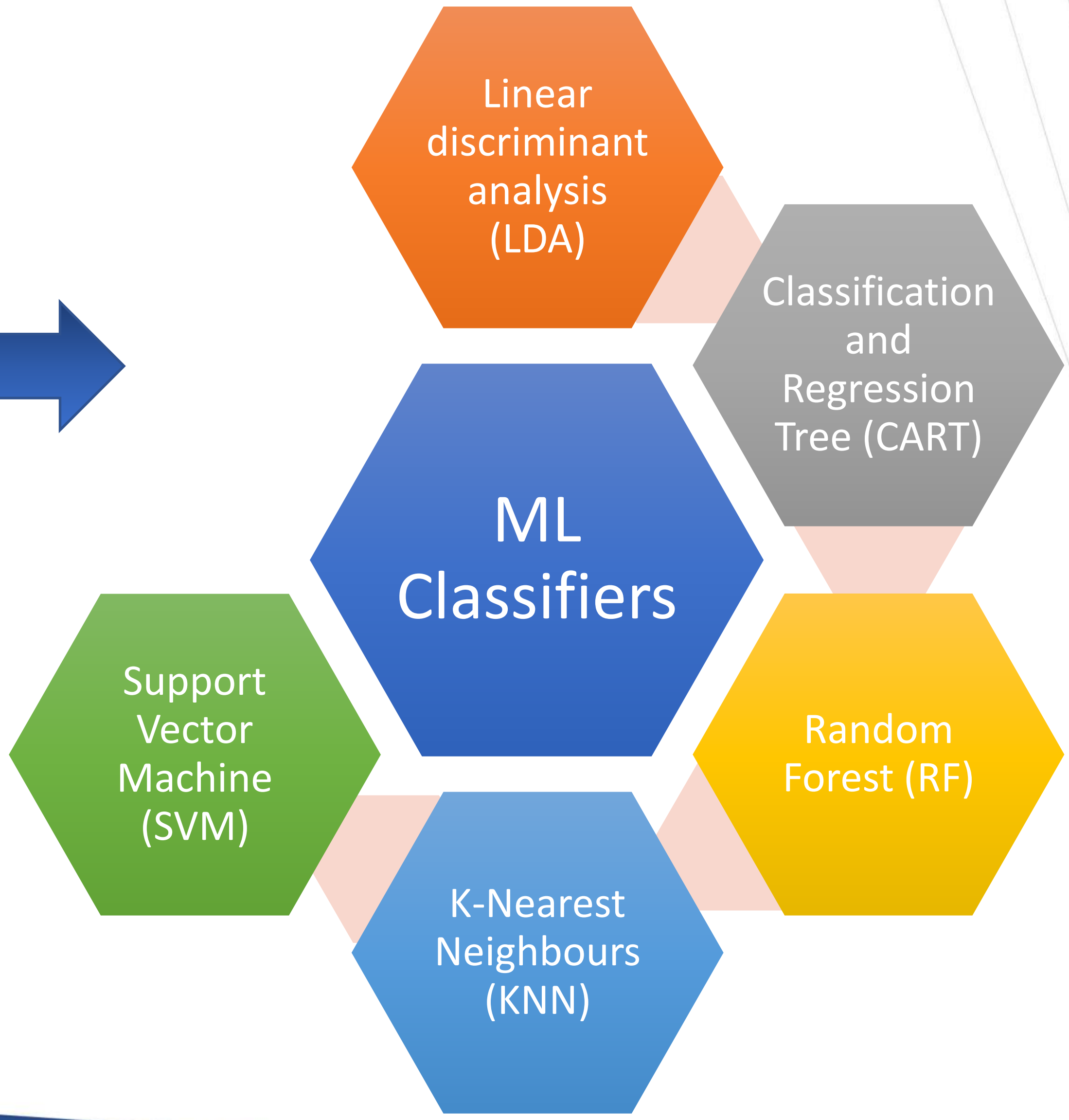
Cross-Validation and ML Classifiers

70% Train

15% Test

15% Validation

K = 10



UNIVERSITY OF HULL

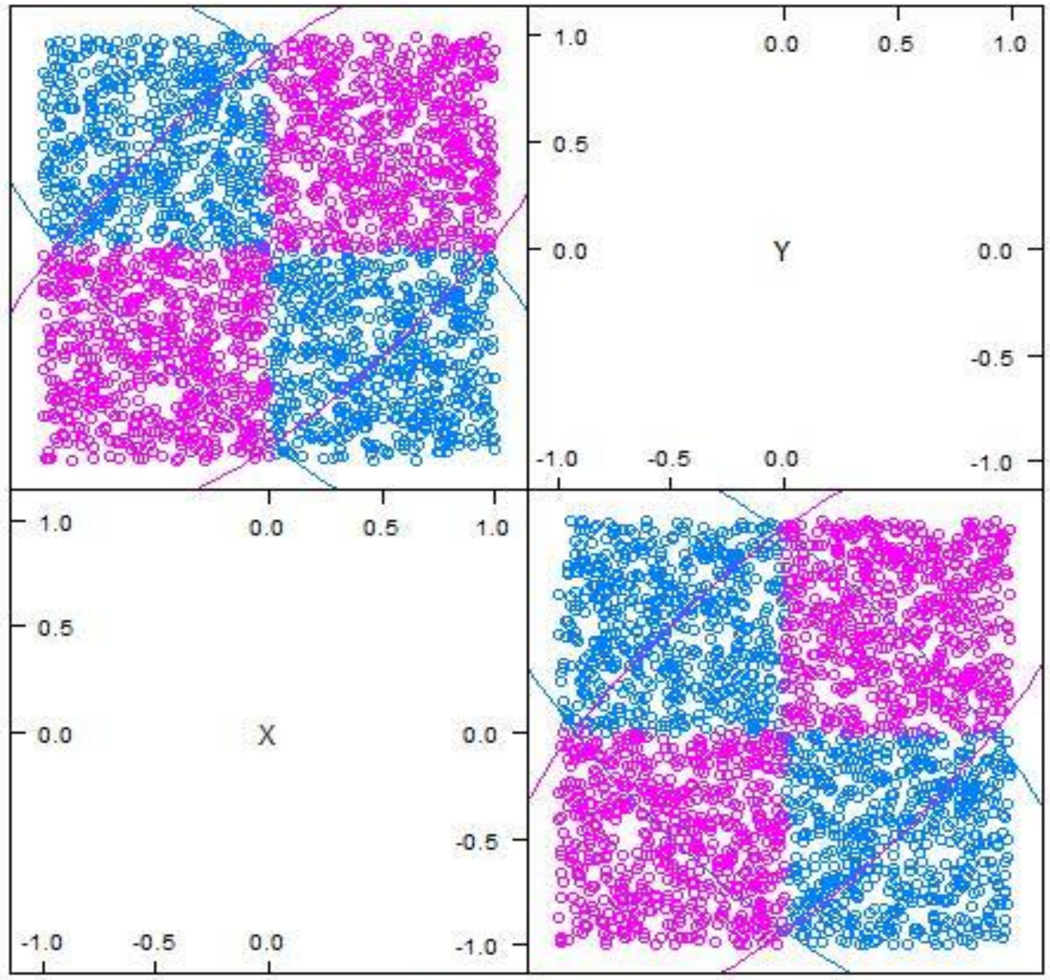
Example 1: 1D Normal Distributed Data in Two Classes

$$\begin{cases} Class\ 1:\ x(t) \sim N(3,1) & t_0 < 1000h \\ Class\ 2:\ x(t) \sim N(5,1) & t_0 > 1000h \end{cases}$$

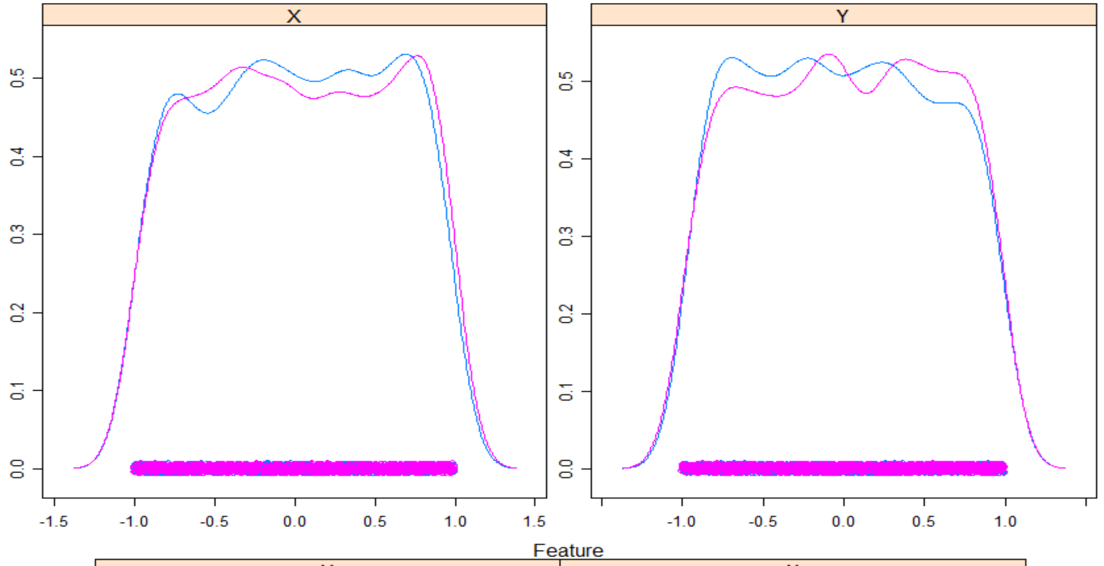
`	Kolmogorov-Smirnov	Kuiper	Anderson-Darling	Wasserstein	DTS	True Accuracy (Mean)	True Accuracy (Min)
LDA	0.9125000	0.8375000	0.9885137	0.8764794	0.9595329	0.9691176	0.9375000
CART	0.9110729	0.8405049	0.9896898	0.8801418	0.9620993	0.9569853	0.8750000
KNN	0.9053426	0.8356562	0.9868039	0.8771581	0.9608425	0.9691176	0.8750000
SVM	0.9053426	0.8356562	0.9868039	0.8771581	0.9608425	0.9569853	0.8750000
RF	0.8530239	0.7897328	0.9686393	0.7933787	0.9346339	0.9386029	0.8235294

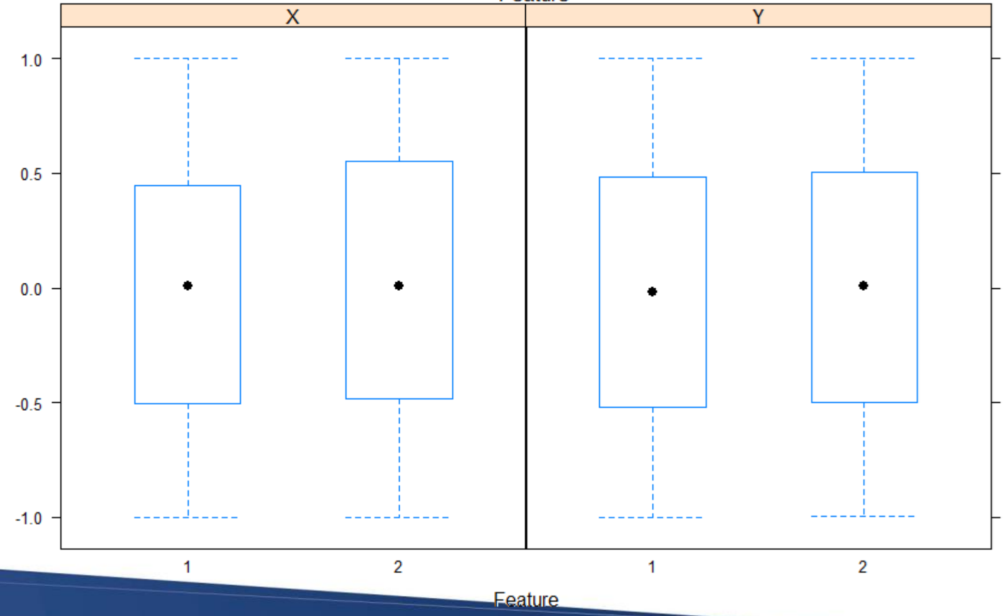
	Difference with True Accuracy (Min)				
	Kolmogorov-Smirnov	Kuiper	Anderson-Darling	Wasserstein	DTS
LDA	0.025000	0.100000	0.051014	0.061021	0.022033
CART	0.036073	0.034495	0.11469	0.005142	0.087099
KNN	0.030343	0.039344	0.111804	0.002158	0.085843
SVM	0.030343	0.039344	0.111804	0.002158	0.085843
RF	0.029495	0.033797	0.145110	0.030151	0.111105
Max Difference	0.036073	0.100000	0.145110	0.061021	0.111105

Example 2: 2D XOR Dataset



Scatter Plot Matrix





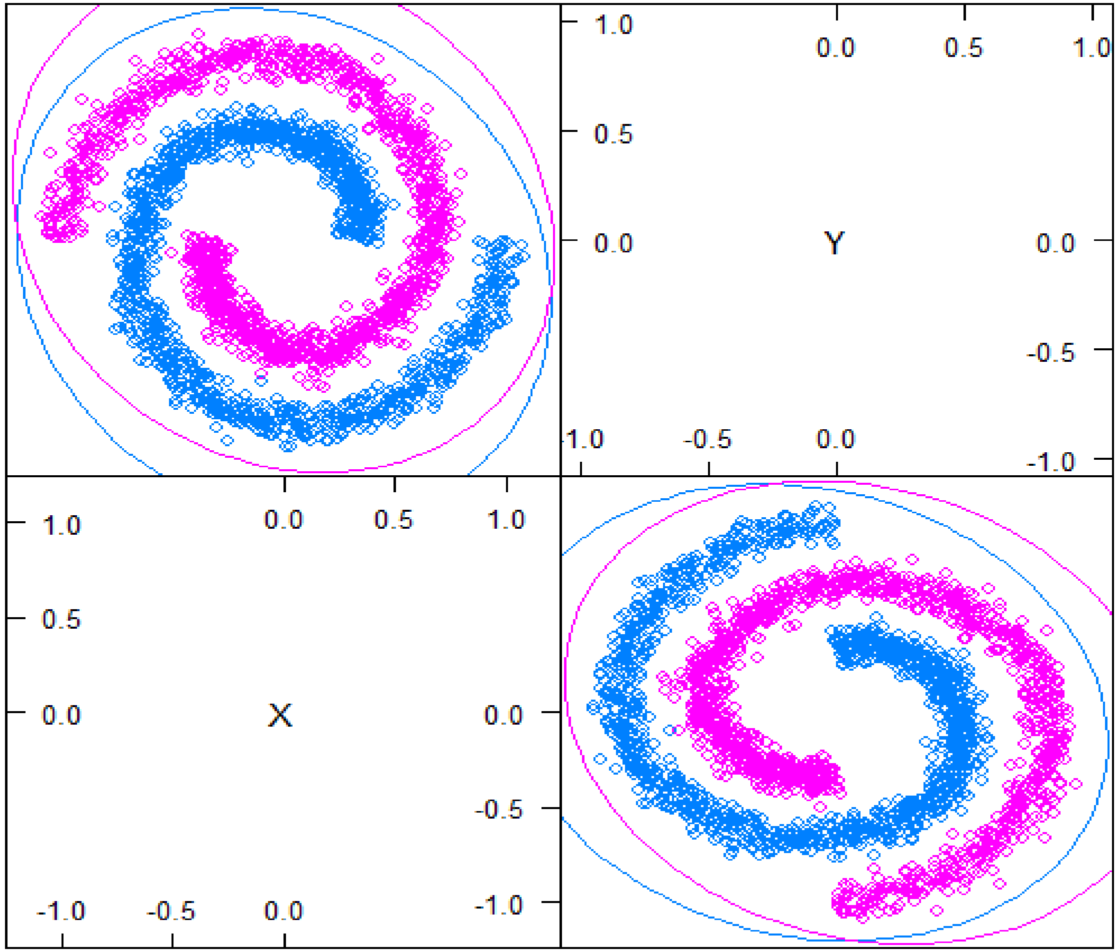
UNIVERSITY OF HULL

Example 2: 2D XOR Dataset

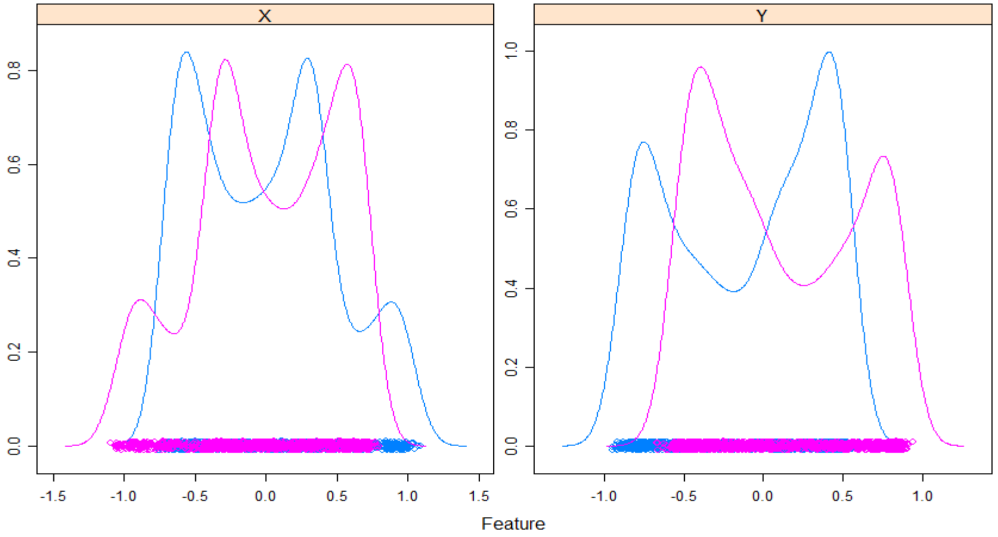
	Kolmogorov-Smirnov	Kuiper	Anderson-Darling	Wasserstein	DTS	True Accuracy (Mean)	True Accuracy (Min)
LDA	0.7722165	0.7706001	0.9028175	0.7550639	0.9856662	0.5912107	0.5083333
CART	0.9281788	0.9219821	0.9877216	0.9254581	0.9952106	0.9941579	0.9874477
KNN	0.9305751	0.9130628	0.9931512	0.9587683	0.9970757	0.9866649	0.9748954
SVM	0.9310446	0.9175864	0.9934891	0.9581909	0.997064	0.9879166	0.9791667
RF	0.9296264	0.9107489	0.9927418	0.9578211	0.9970175	0.9983333	0.9958333

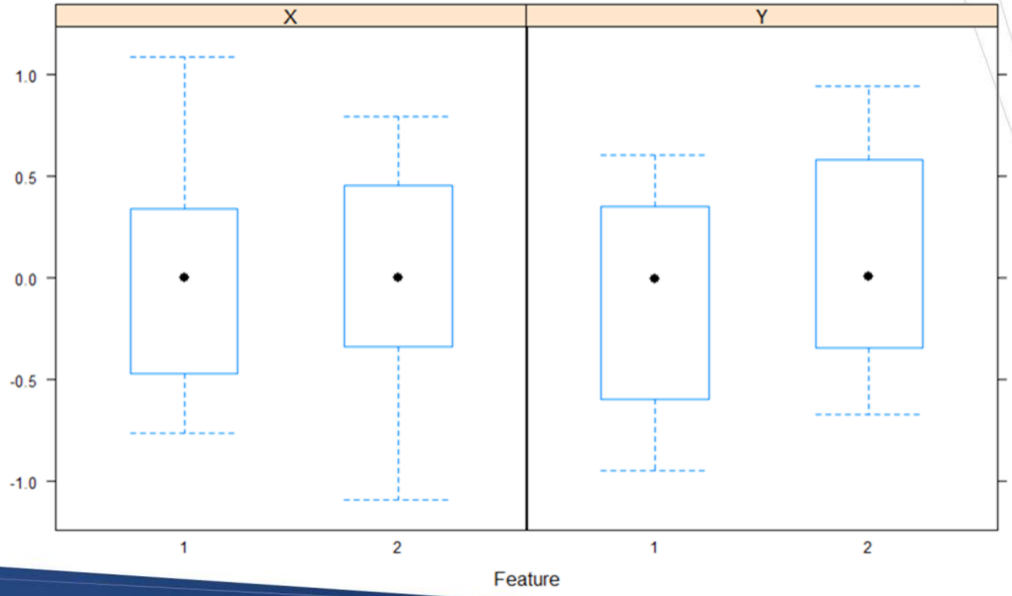
	Difference with True Accuracy (Min)				
	Kolmogorov-Smirnov	**Kuiper**	**Anderson-Darling**	**Wasserstein**	**DTS**
LDA	0.263883	0.262267	0.394484	0.246731	0.477333
CART	0.059269	0.065466	0.000274	0.06199	0.007763
KNN	0.04432	0.061833	0.018256	0.016127	0.02218
SVM	0.048122	0.06158	0.014322	0.020976	0.017897
RF	0.066207	0.085084	0.003092	0.038012	0.001184
Max Difference	0.263883	0.262266	0.394484	0.246730	0.477333

Example 2: 2D Spiral Dataset



Scatter Plot Matrix



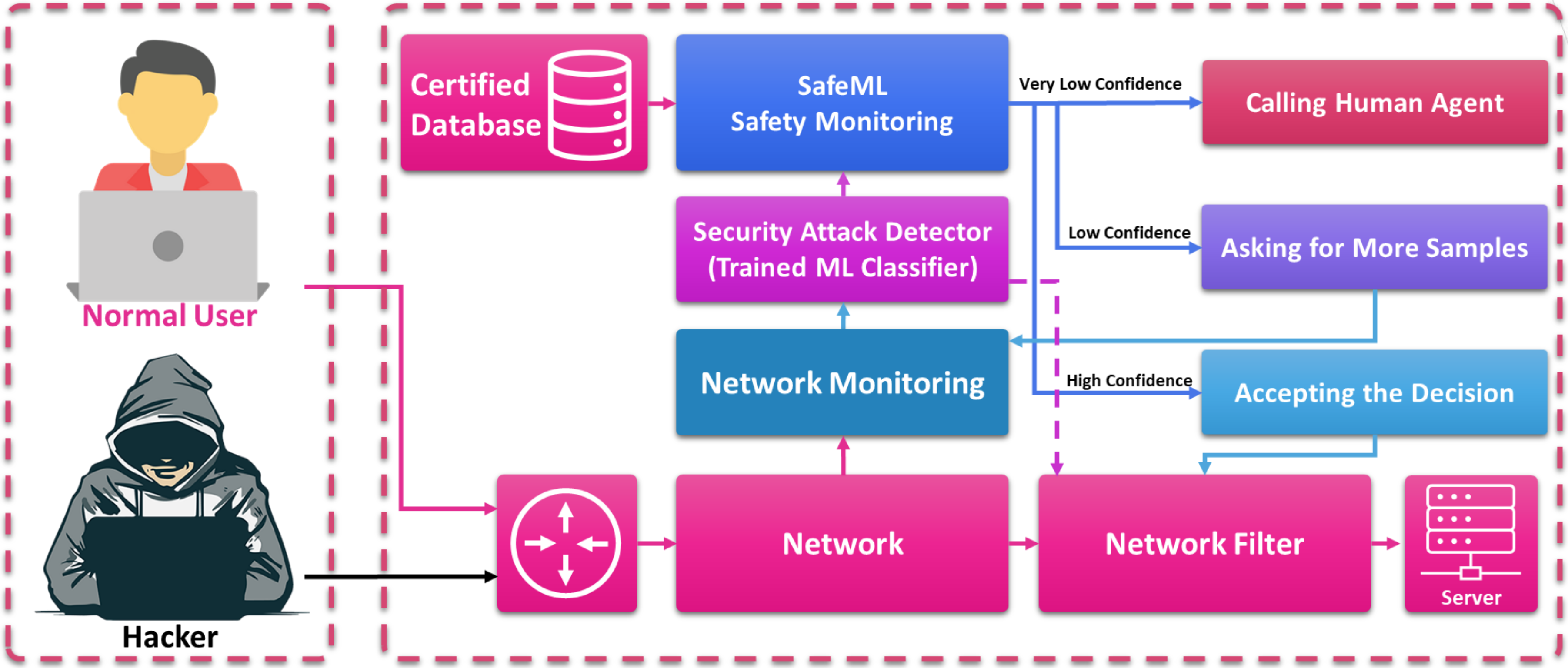


Example 2: 2D Spiral Dataset

	Kolmogorov-Smirnov	Kuiper	Anderson-Darling	Wasserstein	DTS	True Accuracy (Mean)	True Accuracy (Min)
LDA	0.950757	0.915021	0.998485	0.981323	0.998443	0.506250	0.454167
CART	0.964542	0.951250	0.998444	0.979598	0.998355	0.890833	0.837500
KNN	0.946680	0.933058	0.997262	0.966426	0.997802	0.999167	0.995833
SVM	0.947437	0.933940	0.997356	0.967035	0.997835	0.999167	0.995833
RF	0.946821	0.934418	0.997062	0.965102	0.997728	0.990833	0.979167

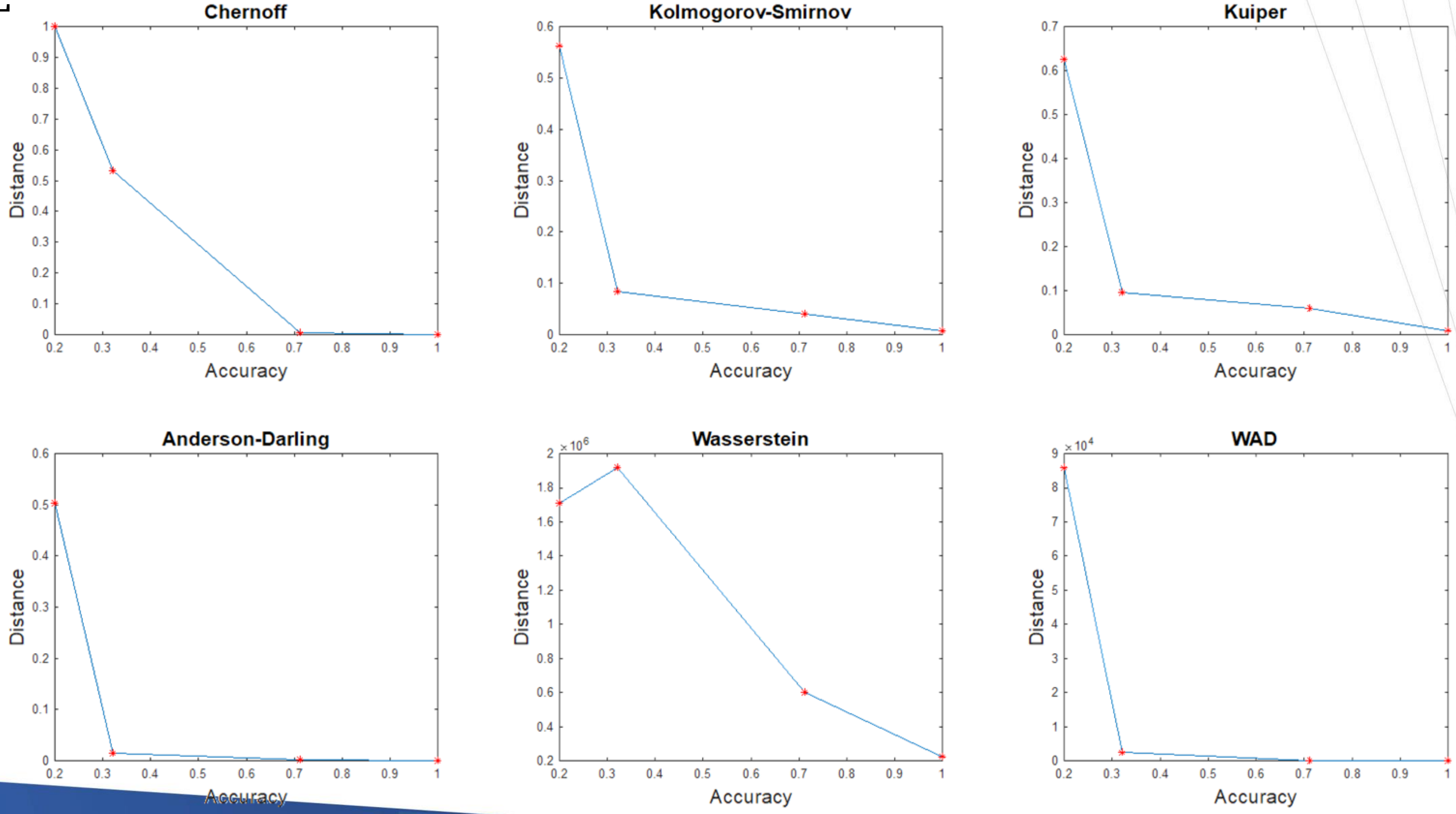
	Difference with True Accuracy (Min)				
	Kolmogorov-Smirnov	**Kuiper**	**Anderson-Darling**	**Wasserstein**	**DTS**
LDA	0.496590	0.460854	0.544319	0.527156	0.544277
CART	0.127042	0.113750	0.160944	0.142098	0.160855
KNN	0.049153	0.062775	0.001429	0.029407	0.001969
SVM	0.048397	0.061893	0.001523	0.028798	0.002002
RF	0.032346	0.044748	0.017896	0.014065	0.018562
Max Difference	0.0994468	0.0882515	0.2699748	0.2483959	0.528852

Application of SafeML in Security

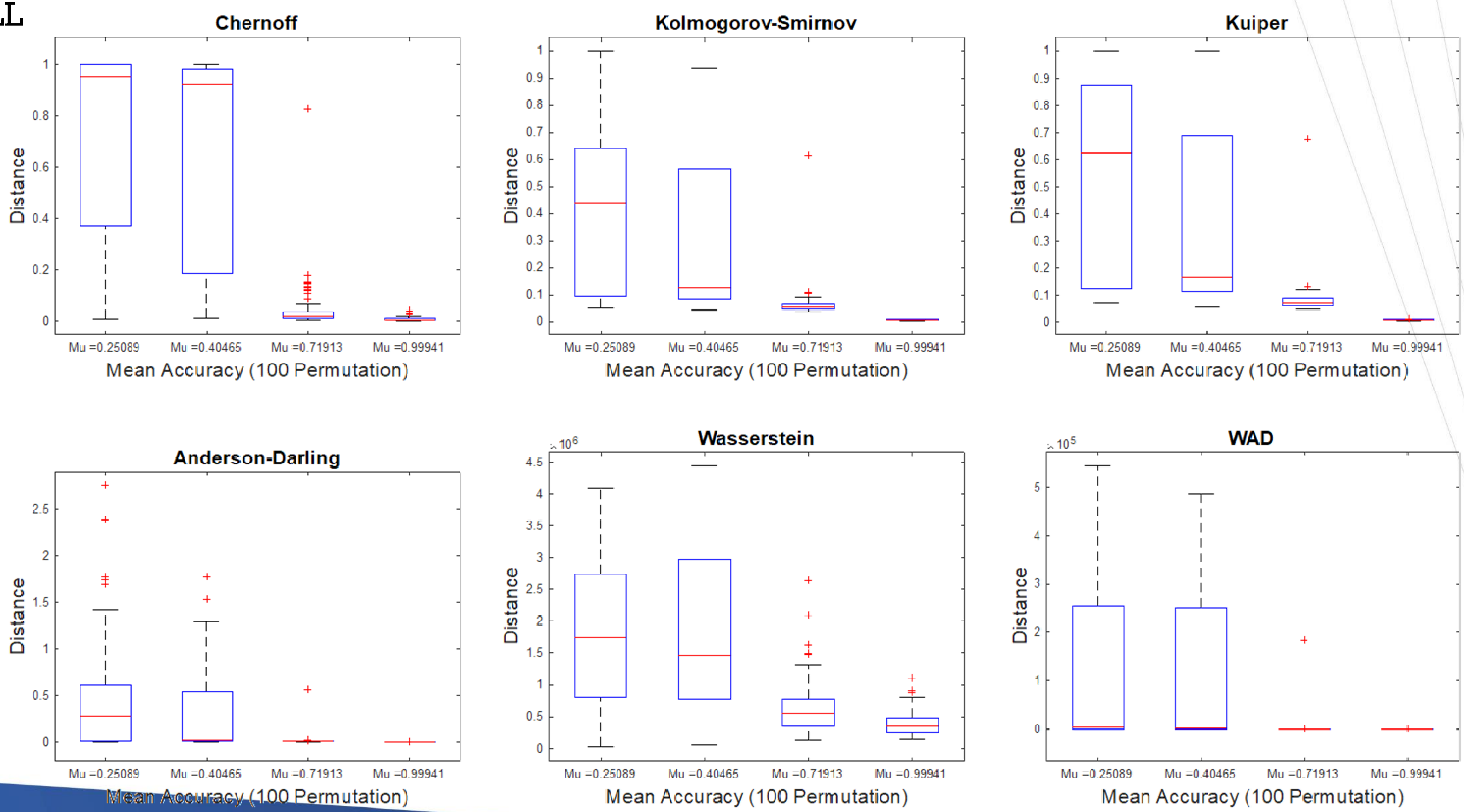


Example 4: Security Dataset

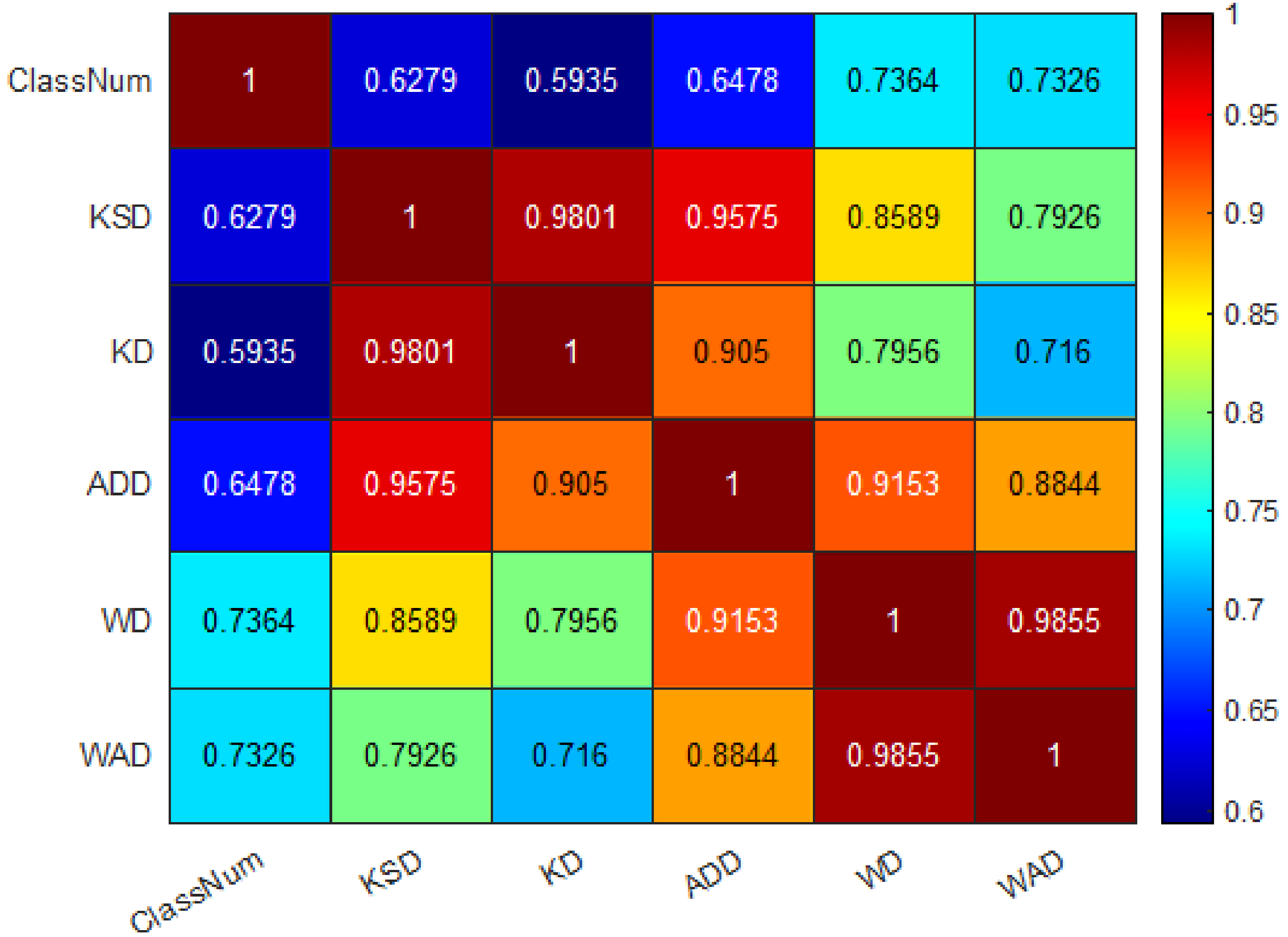
Intrusion Detection Evaluation Dataset (CIC-IDS2017)



Example 4: Security Dataset



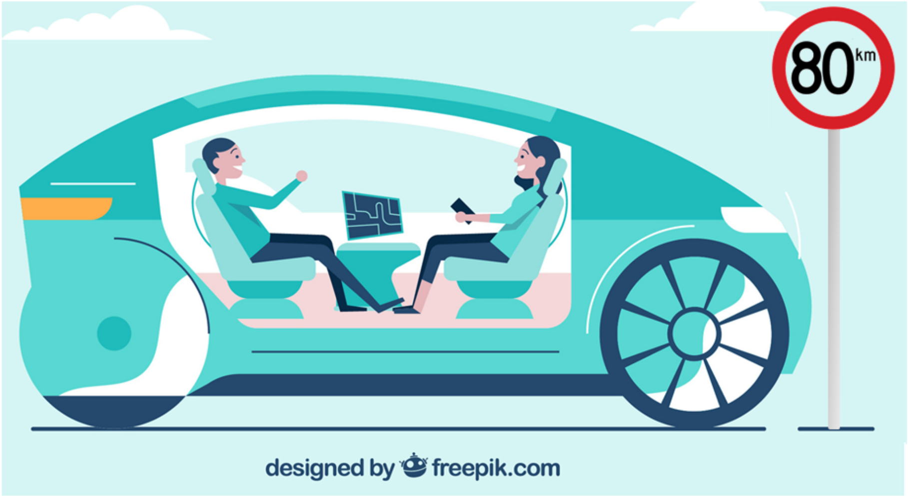
Example 4: Security Dataset

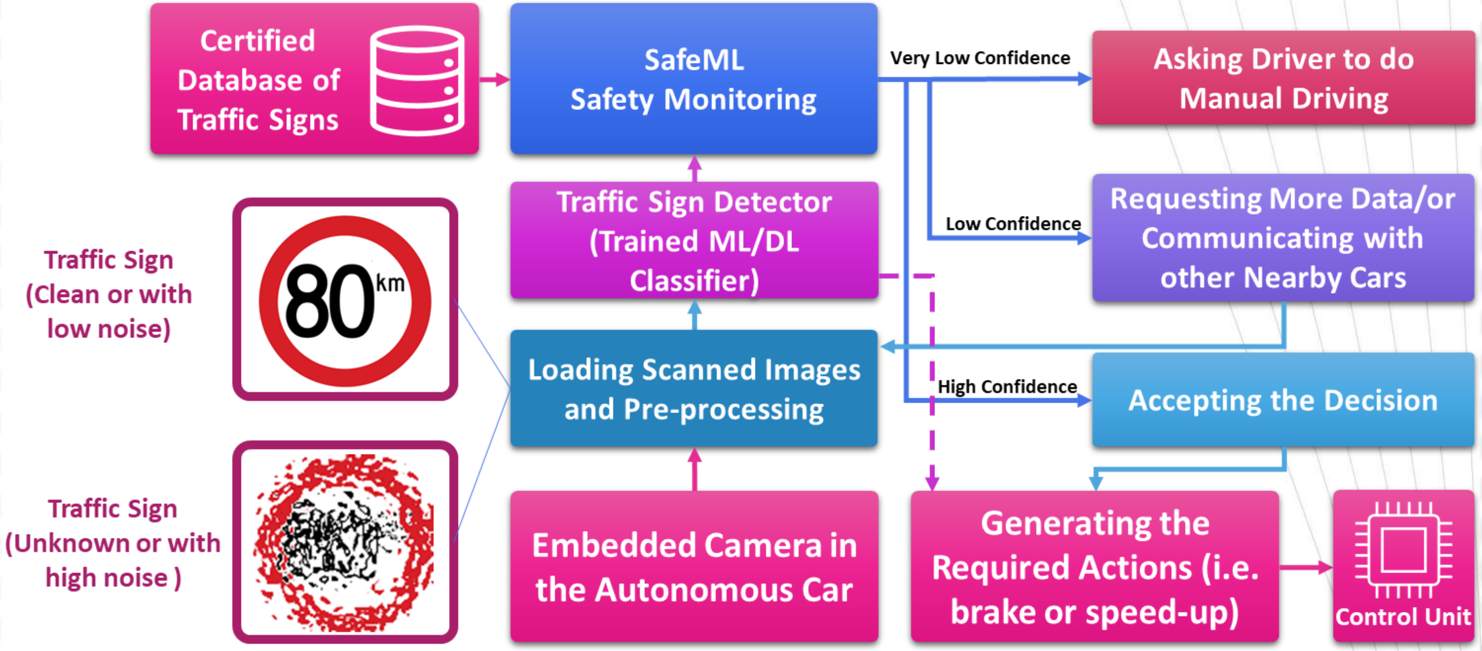


UNIVERSITY OF HULL

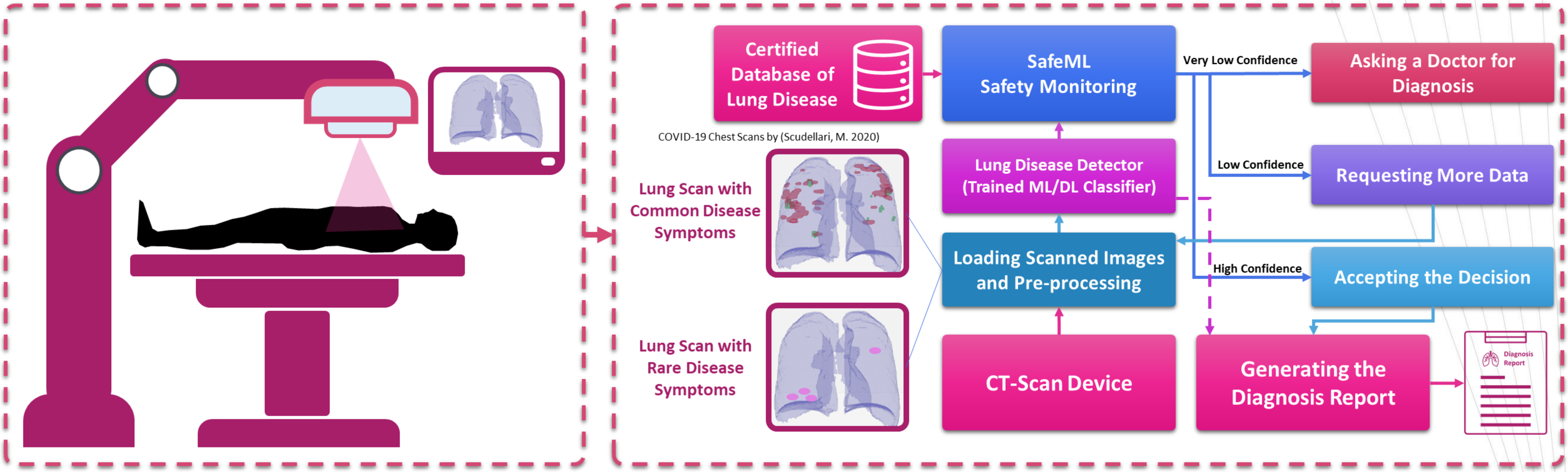
Applications of SafeML

Applications of SafeML

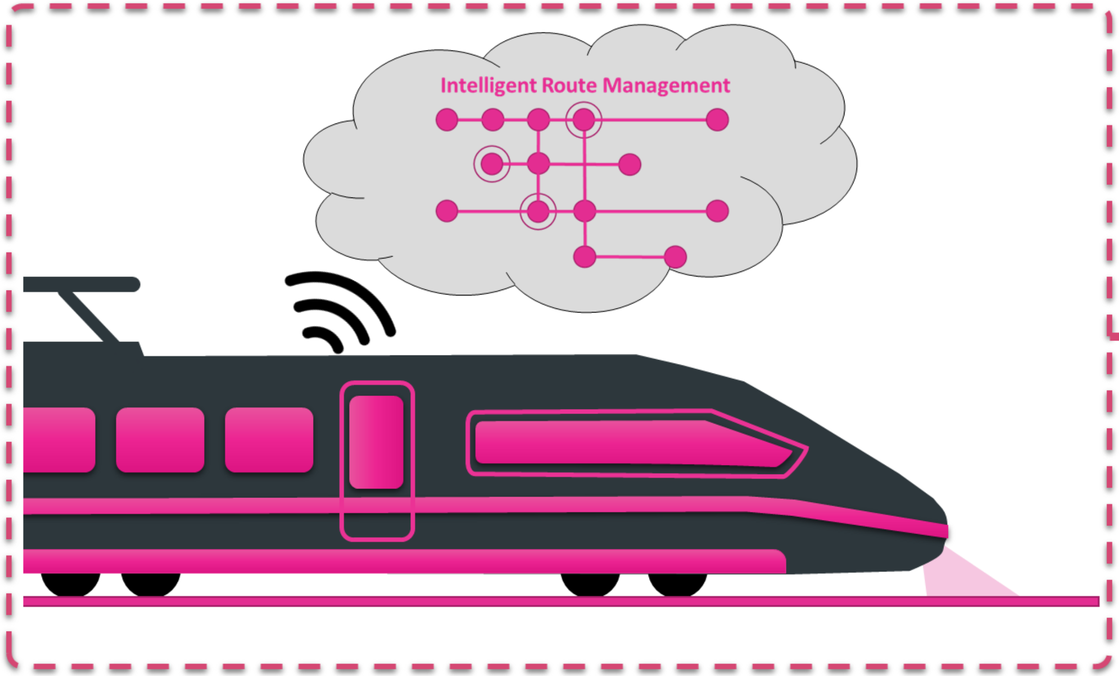


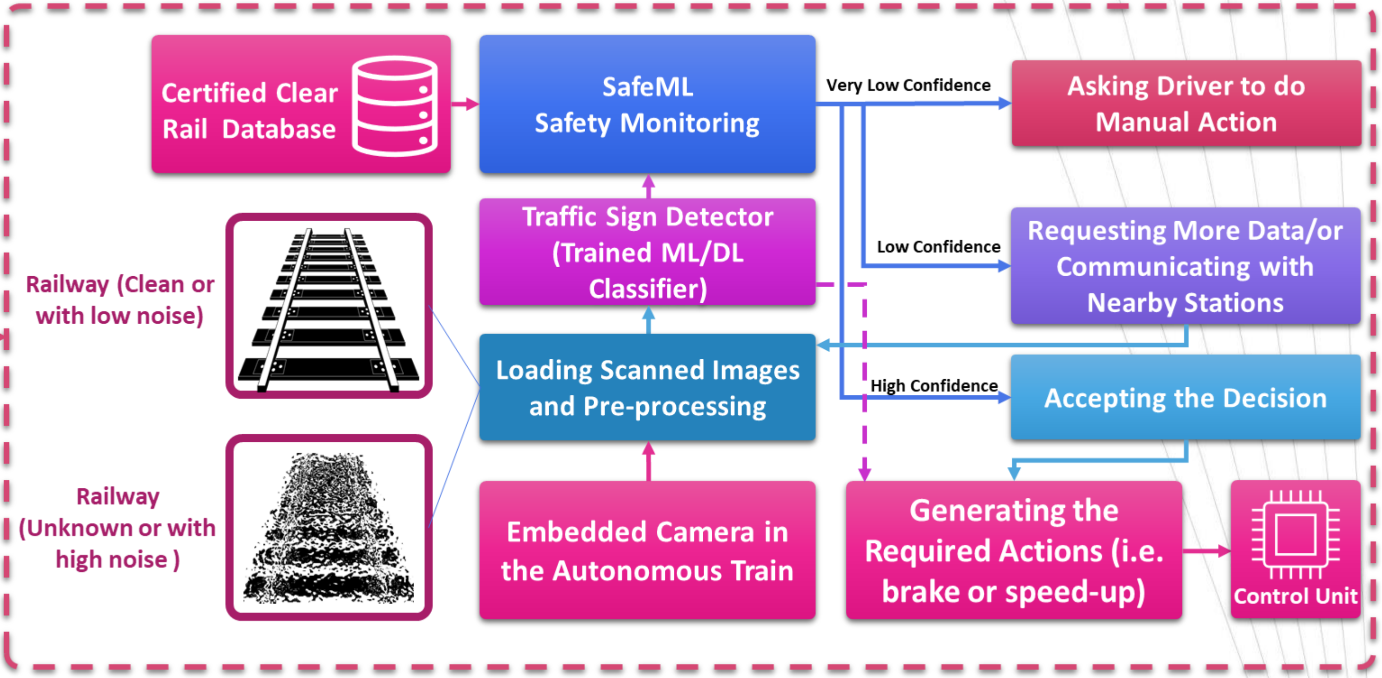


Applications of SafeML



Applications of SafeML

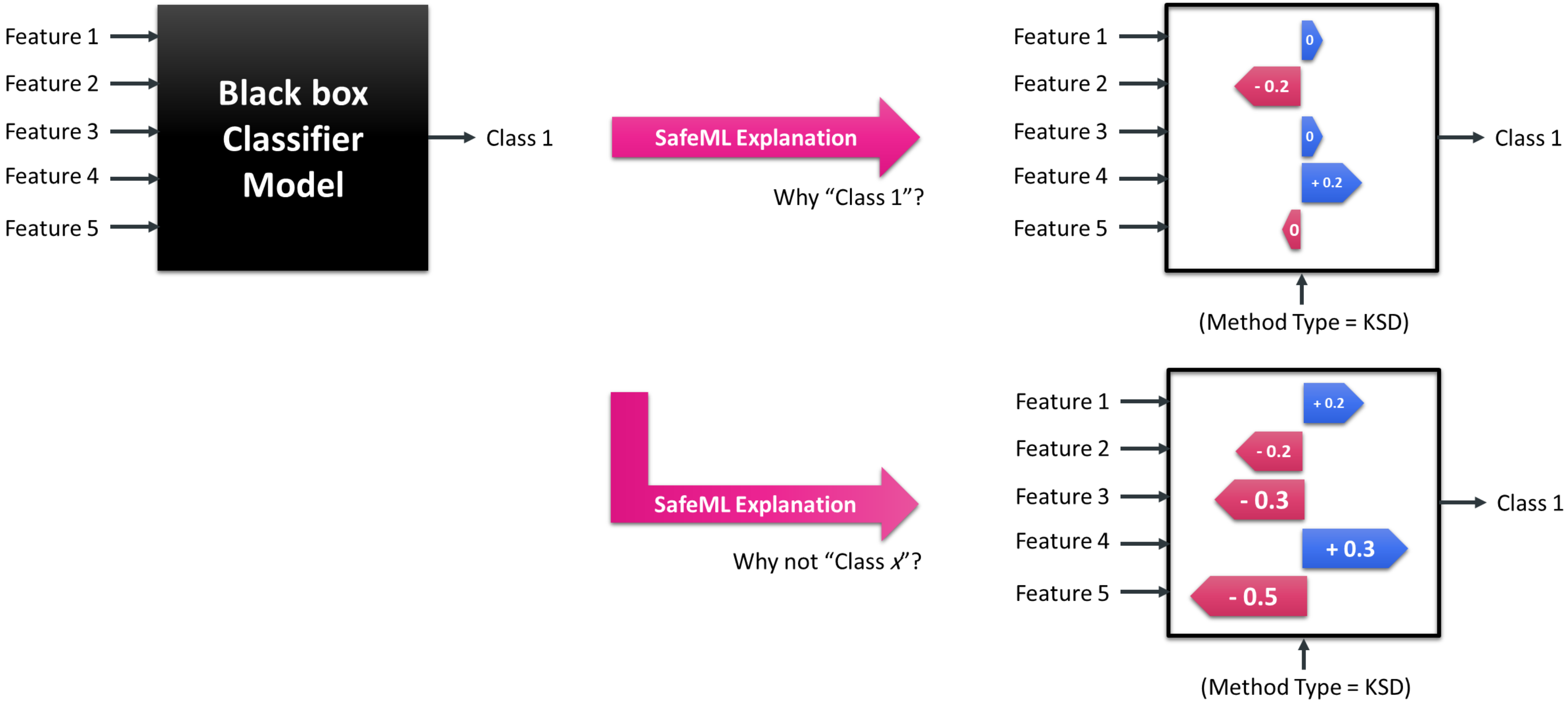




UNIVERSITY
OF HULL

SafeML Toward XAI

SafeML Toward eXplainable AI (XAI)

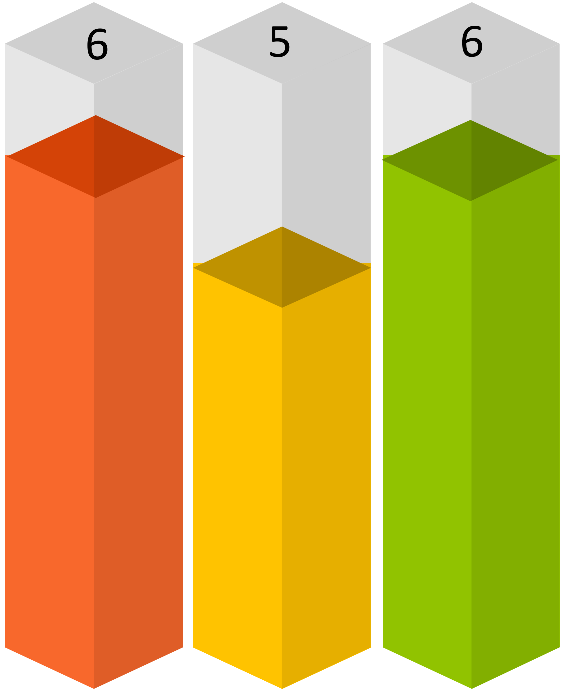


UNIVERSITY OF HULL

SafeML Reproducibility

GitHub **https://github.com/ISorokos/SafeML**

- **MATLAB Implementation**
- **Python Implementation**
- **R Implementation**



UNIVERSITY OF HULL

Conclusion

- Through modifying the existing statistical distance and error bound measures, the proposed method enables to estimate the accuracy bound of the trained ML algorithm in the field with no label on the incoming data.

- A novel proposed human-in-loop procedure is made to certify the ML algorithm in a real-time manner. The procedure has three levels of operation: I) runtime estimated accuracy, II) Lack of enough data and need for buffering more samples (it may cause a delay in decision-making), and III) No low runtime estimated accuracy and a human agent is needed.

- The proposed approach is easy to implement, and it can support a variety of distribution (Exponential and normal distribution families).

Future Works

- Extending the SafeML Idea for Machine Learning Regression and Prediction Algorithms
- Considering Recurrent Methods and Dealing with Time Series.
- Improving the method for adaptive and online-learning algorithms.
- Integrating the feature importance to the exiting algorithm.
- Implementing the SafeML XAI for Image classification.

Selected References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). ***Concrete Problems in AI Safety***. http://arxiv.org/abs/1606.06565

Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., & Porter, Z. (2020). **Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective**. *Artificial Intelligence*, *279*, 103201. https://doi.org/10.1016/j.artint.2019.103201

Davenport, T. H., Brynjolfsson, E., McAfee, A., James, H., & Wilson, R. (2019). ***Artificial Intelligence: The Insights You Need*** *from Harvard Business Review*. Harvard Business Review.

Fukunaga, K. (1992). ***Introduction to Statistical Pattern Recognition*** *(Second Edition)*. Academic Press.

Nielsen, F. (2018). **The Chord Gap Divergence and a Generalization of the Bhattacharyya Distance**. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2018-April*, 2276–2280. https://doi.org/10.1109/ICASSP.2018.8462244

Quiñonero-Candela, J., & Schwaighofer, A. (2009). ***Dataset Shift in Machine Learning***. MIT Press.

Schulam, P., & Saria, S. (2019). ***Can You Trust This Prediction? Auditing Pointwise Reliability After Learning***. http://arxiv.org/abs/1901.00403

Zahm, O., Cui, T., Law, K., Spantini, A., & Marzouk, Y. (2018). ***Certified dimension reduction in nonlinear Bayesian inverse problems***. http://arxiv.org/abs/1807.03712

Thank You

If you have any question, please feel free to ask