Reliability Assessment For DL Classifiers & An Assurance Framework for LES

Presenter: Xingyu Zhao

28/09/2021 at the SOLITUDE workshop

Agenda (10+10 minutes)

- Part 1: The Reliability Assessment Model (RAM)
  - For DL classifiers
    - Operational Profile (OP) info and Robustness evidence.
    - Uncovers inherent challenges of modelling reliability for DL software
  - AISafety21-workshop@IJCAI-21 (best paper award)
    - https://arxiv.org/pdf/2106.01258.pdf
- Part 2: The ``big-picture''---An assurance case framework
  - Probabilistic safety arguments based on the RAM
  - System-level safety requirements -> ML component level requirements
    - A chain of safety analysis methods: HAZOP, FTA, etc.
  - Challenges for assuring LES/autonomous systems

20. Littlewood, B., Strigini, L.: Software reliability and dependability: A roadmap. In: Proc. of the Conf. on The Future of Softw. Eng. pp. 175–188. ICSE'00 (2000)

PART 1---The Gist of the RAM

- A Reliability Assessment Model (RAM) for DL classifiers
  - First RAM for DL software that explicitly considers the Operational Profile (OP) Info and Robustness evidence.
- Why OP and robustness evidence matter?
  - Software reliability is a user-centric property [20].
  - DL is known to be unrobust.
- Output: reliability claims on *pmi*, e.g., confidence bounds, mean, variance
  - *pmi*: probability of misclassification per random input (e.g., image)

Definitions

20. Littlewood, B., Strigini, L.: Software reliability and dependability: A roadmap. In: Proc. of the Conf. on The Future of Softw. Eng. pp. 175–188. ICSE'00 (2000)
21. Musa, J.D.: Operational profiles in software-reliability engineering. IEEE Software **10**(2), 14–32 (1993)
26. Webb, S., Rainforth, T., Teh, Y.W., Kumar, M.P.: A statistical approach to assessing neural network robustness. In: ICLR'19. New Orleans, LA, USA (2019)
27. Weng, L., et al: PROVEN: Verifying robustness of neural networks with a probabilistic approach. In: ICML'19. vol. 97, pp. 6727–6736. PMLR (2019)

- delivered software reliability—a user centric property [20]
  - model the end-users' behaviours – OP [21]
  - defined by a probabilistic metric
    - *pmi*: probability of misclassification per random input

$$\lambda := \int_{x \in \mathcal{X}} I_{\{x \text{ causes a misclassification}\}}(x) Op(x) \, \mathrm{d}x \qquad (1)$$

- DL robustness
  - Prediction of the DL model is invariant against small perturbations.
  - Probabilistic robustness definition [26,27]

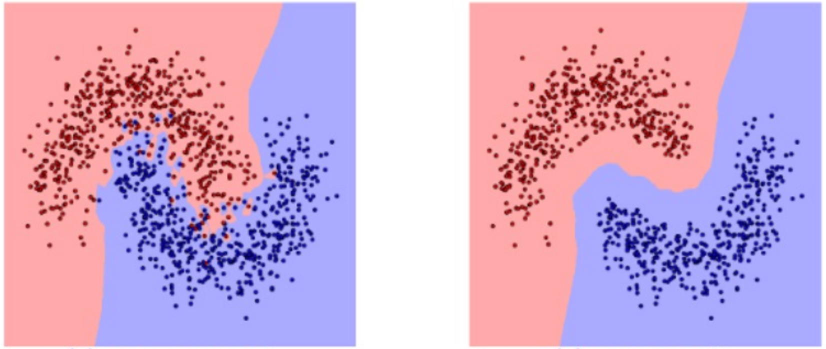
$$R_{\mathcal{M}}(\eta, y) := \sum_{x \in \eta} I_{\{\mathcal{M}(x) \text{ predicts label } y\}}(x) \times Op(x \mid x \in \eta) \qquad (2)$$

26. Webb, S., Rainforth, T., Teh, Y.W., Kumar, M.P.: A statistical approach to assessing neural network robustness. In: ICLR'19. New Orleans, LA, USA (2019)
27. Weng, L., et al: PROVEN: Verifying robustness of neural networks with a probabilistic approach. In: ICML'19. vol. 97, pp. 6727–6736. PMLR (2019)
28. Yang, Y.Y., Rashtchian, C., Zhang, H., Salakhutdinov, R., Chaudhuri, K.: A Closer Look at Accuracy vs. Robustness. In: NeurIPS'20. Vancouver, Canada (2020)

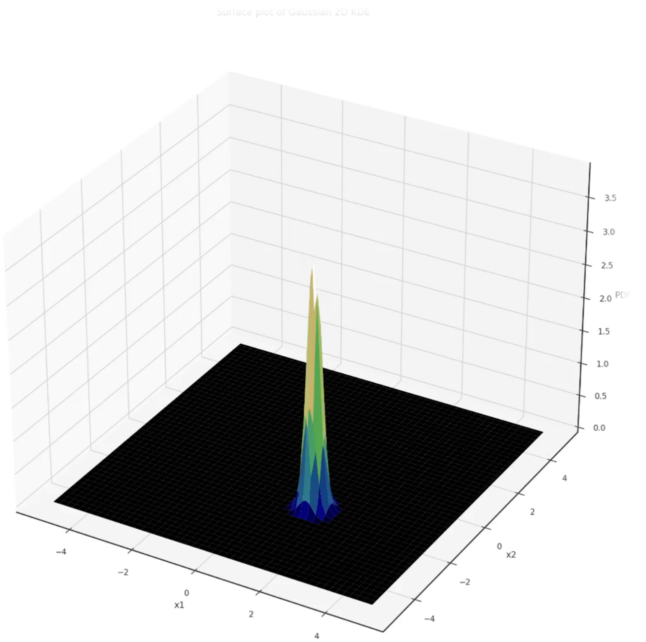
The RAM (omitting details, cf. the paper)

- Step 1: Partition the input space
  - Rule: cell size < *r*-separation [28]
  - Assumption: datapoints in a cell has one ground truth label
- Step 2: Approximation the OP
  - Estimate the PDF over the input domain
  - Kernel Density Estimation (KDE)
- Step 3: Cell robustness (to GTL) evaluation
  - 3rd party robustness estimators, e.g., [26,27]
- Step 4: "Assemble" cell-wise estimates

$$\lambda = \sum_{i=1}^{m} Op_i \lambda_i \qquad (4)$$



(a) not *r*-separated (b) *r*-separated



Experiments

- 5 Datasets + AUV case study:
  - 3 synthesized 2D datasets
  - MNIST, Cifar10
  - Testing accuracy, average cell (un)robustness, our reliability claims;
- Scalability issues by “the curse of dimensionality”
  - input pixel space -> latent feature space
  - sample *k* cells-> estimators for weighted-average
  - Efficient (multivariate) KDE and robustness estimators

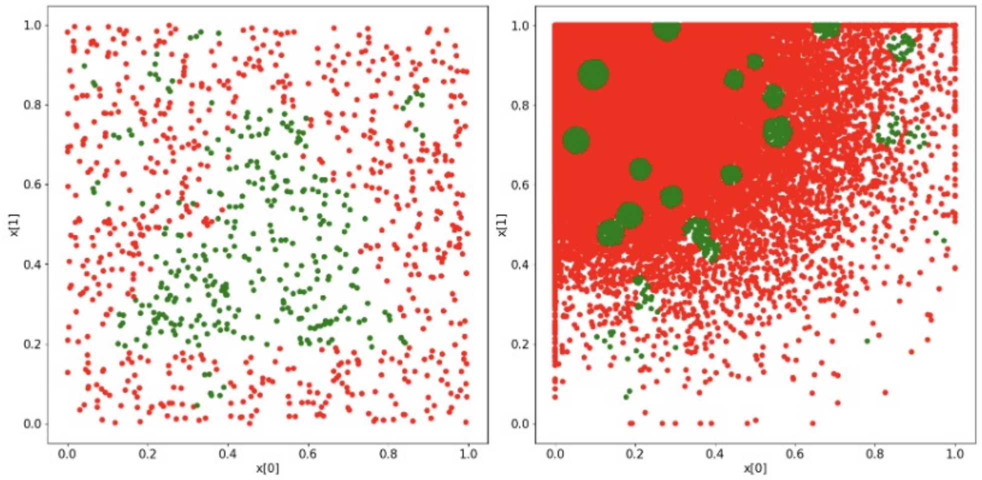
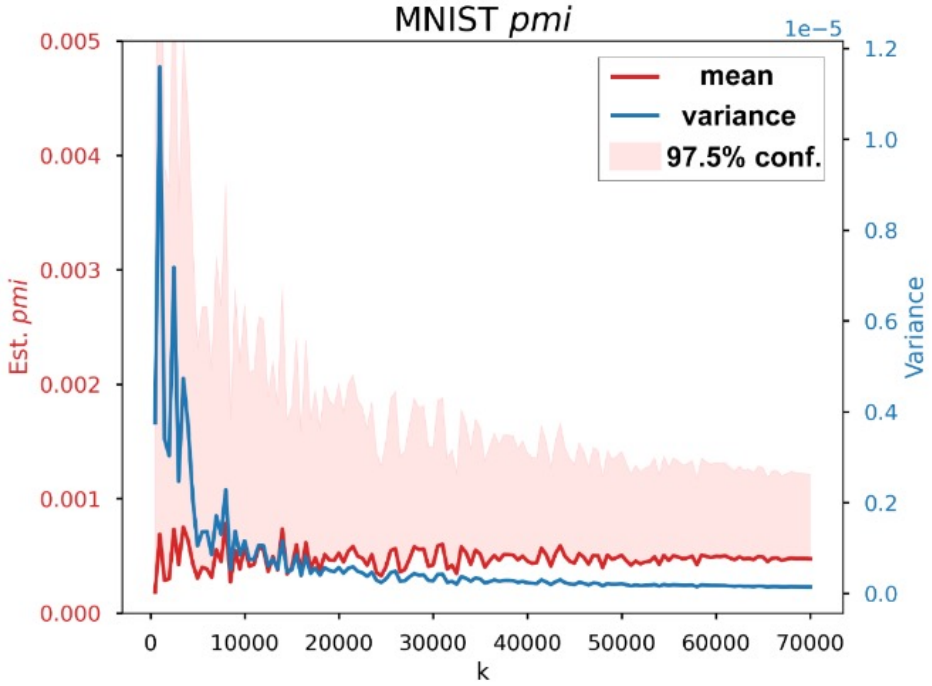


Figure 3: Synthetic datasets DS-1 (lhs) and DS-2 (rhs) representing relatively sparse and dense training data respectively.



Part 1---Discussion

- Discussions on 6 assumptions of our RAM.
  - Application specific knowledge/evidence to justify.
- Some inherent difficulties of assessing reliability for DL:
  - How to accurately build the OP in the high-dimensional input space with relatively sparse data? (domain expert knowledge + generative models)
  - How to build an accurate oracle?
    - e.g., by leveraging the existing human-labels in the training dataset?
  - What is the local distribution (conditional OP) over a small input region?
    - (random noise? Or natural variations of physical conditions?)
  - How to efficiently evaluate the robustness of a small region given AEs are rare events?
  - How to sample small regions from a large population (high-dimensional space) in an unbiased, uncertainty informed and efficient way?

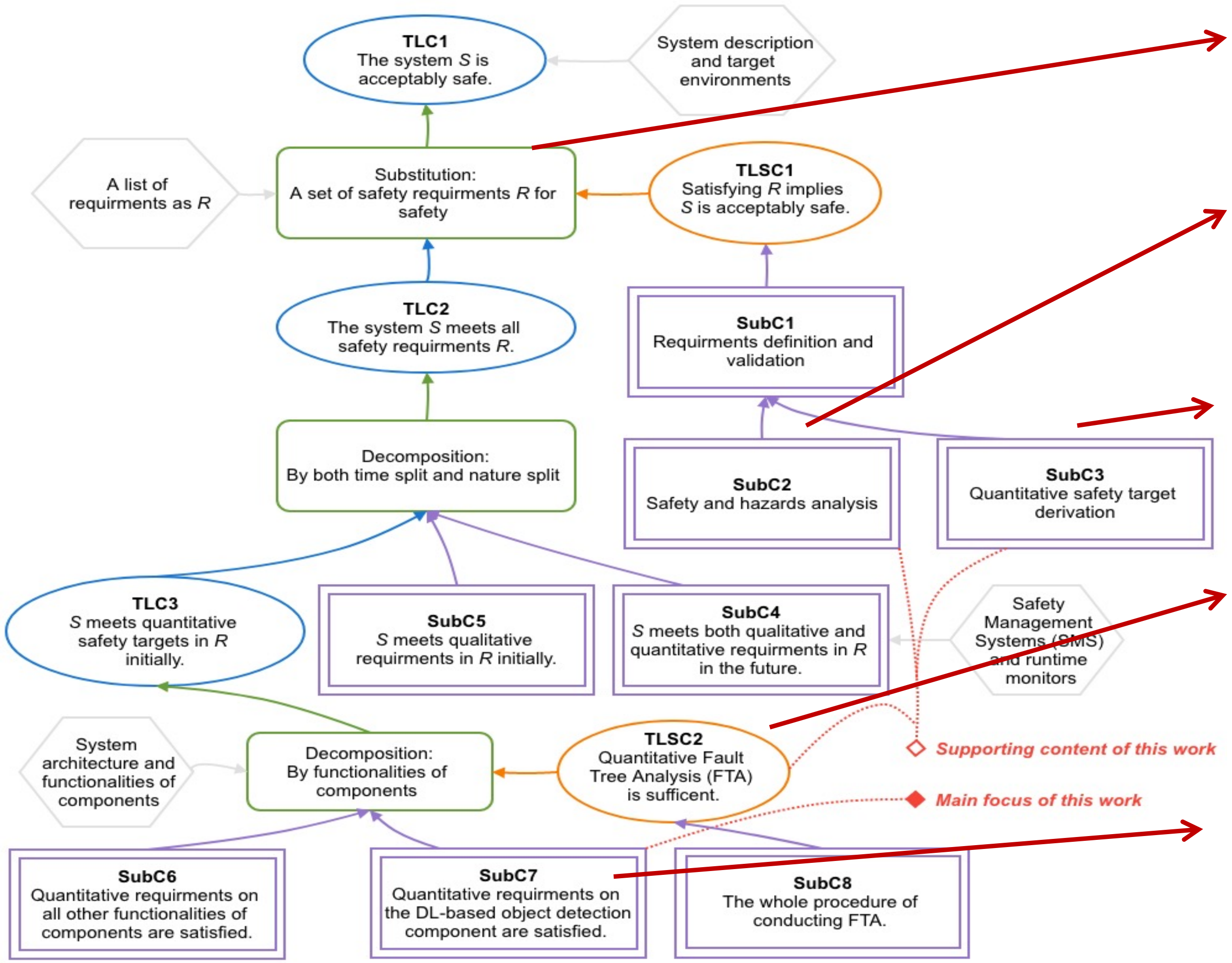
Part 1---Conclusion

- A conceptualized equation:

$$DL\ \ reliability = generalisability \times robustness.$$

  - How well it generalises to a new data-point, according to the future OP.
  - How good the local robustness is, around that new data-point.
- Our RAM advances in this research direction
  - First RAM for DL software that considers the OP and robustness evidence.
  - Compromised/practical solutions for scalability issues (for high-dimensional data)
  - Revealed inherent difficulties of DL reliability assessment

Part2---The Overall Assurance Framework



- Acceptably safe
  - satisfying all safety requirements (SRs)
- SRs are derived from safety/hazards analysis
  - HAZOP
  - Domain specific standards (missing for ML!)
- SRs are validated by regulation principles
  - ALARP, GALE, etc
- From system-level quantitative risk to component-level reliability requirements
  - …next slide…
- Our main focus
  - reliability claims on low-level DL-components

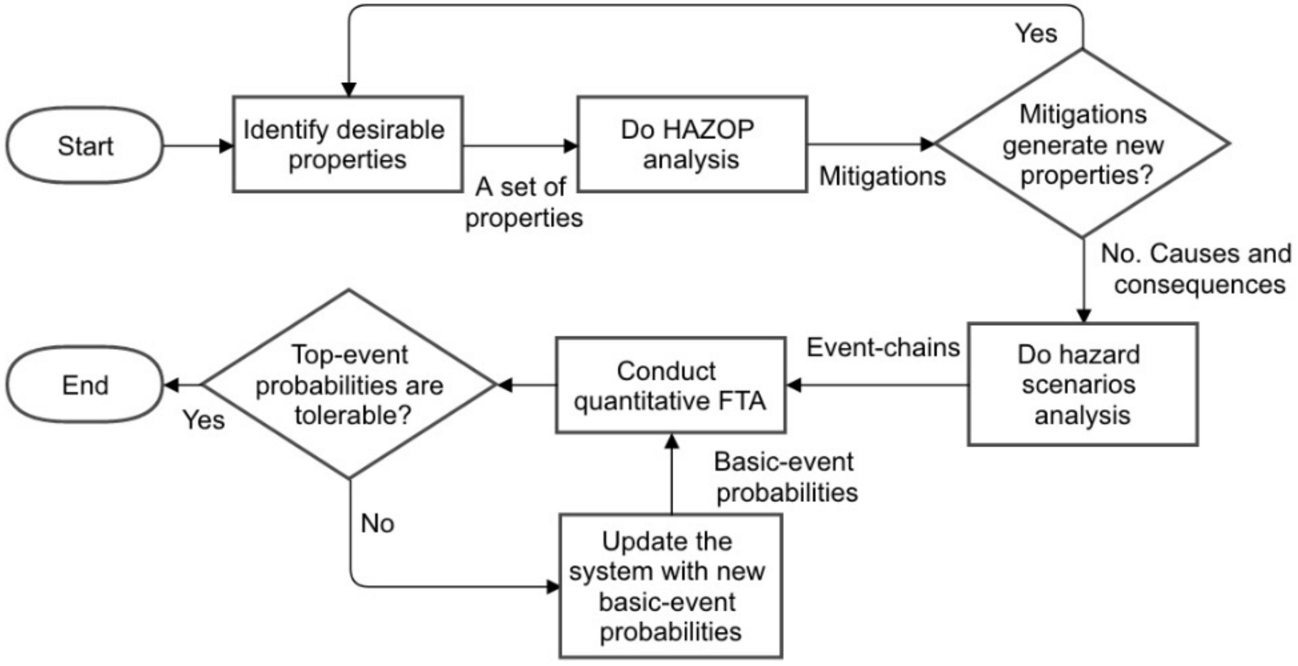
Safety Targets: System-level -> ML-Component level

- The chain of 3 safety analysis methods +2 loops
  - HAZOP
    - Given properties, identify hazards with causes, consequences, mitigations (potentially new properties)
  - Hazard scenarios modeling
    - link the hazard causes to their consequences by a chain of intermediate events
  - FTA
    - Expanding/combining the event-chains into tree structures.

- Basic events (BEs): misclassifications, wrong bounding box, H/W failure events.
- Top events (TEs): violation of system-level properties, e.g., fail to keep the safe distance to the asset.

- To answer (by iterations of what if calculations):

Given a tolerable/acceptable TE probability, what's the most practical combinations of BE probabilities?

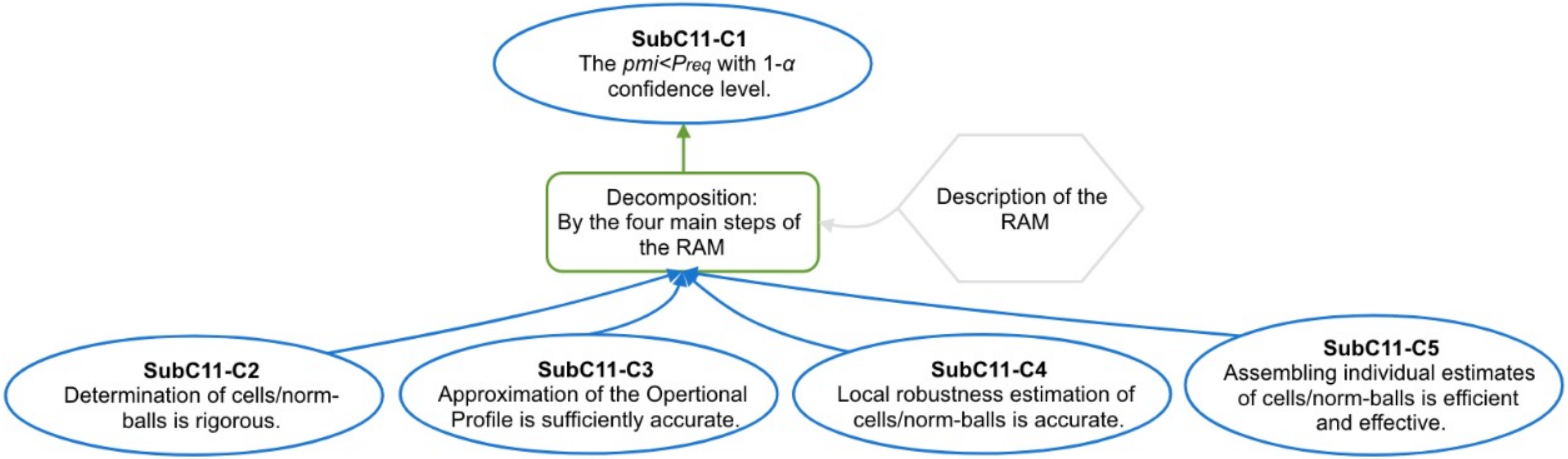


UNIVERSITY OF LIVERPOOL HERIOT WATT UNIVERSITY [dstl] The Science Inside ADELARD National Oceanography Centre ORCA HUB

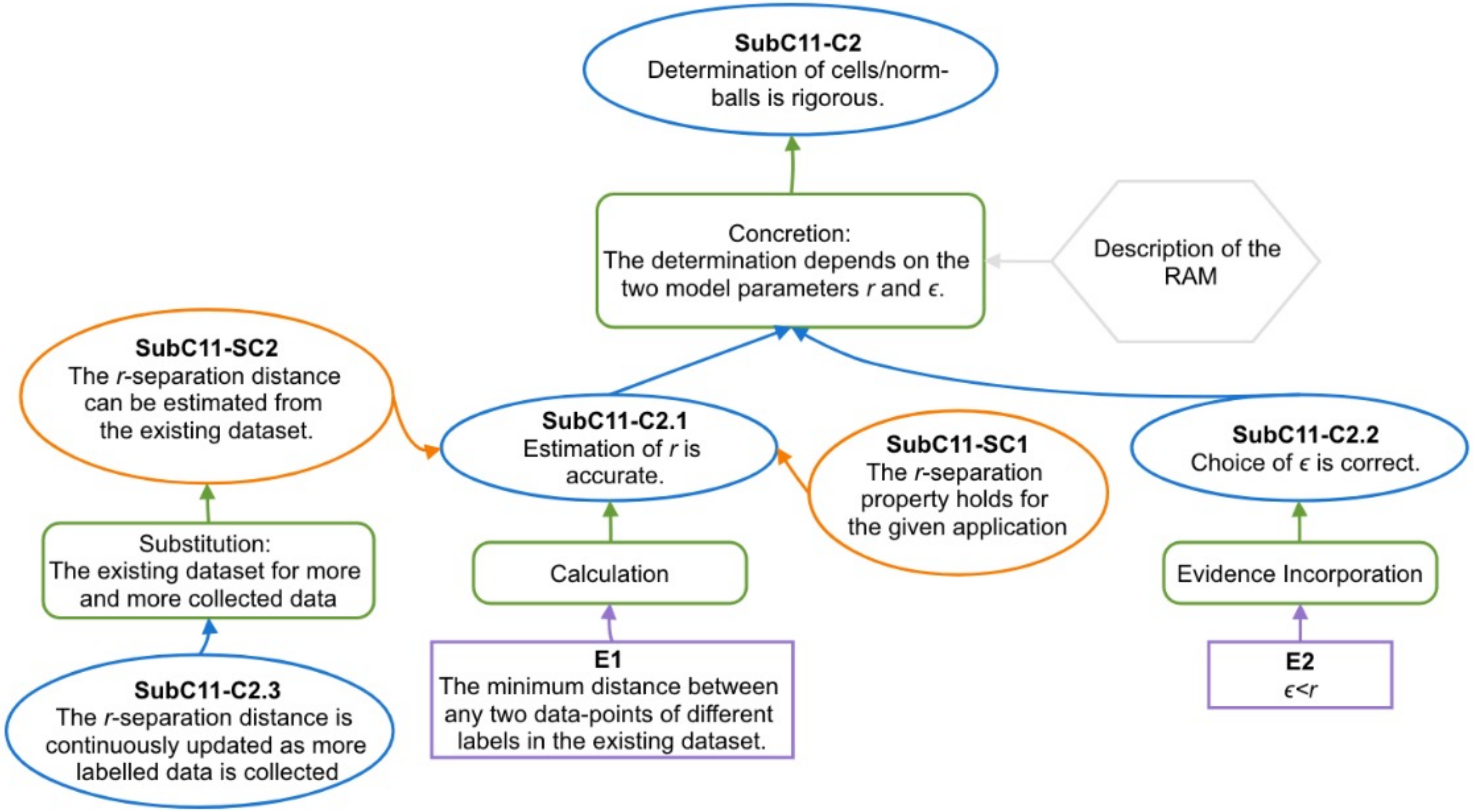
Probabilities of TEs and BEs?

- What is the tolerable/acceptable top-event (TE) probability?
    - (TE: violation of system-level properties)
  - Out the scope of an assurance framework;
    - (missing now, but eventually will be) Given by regulators/domain-standards?
  - Refer to human performance seems to be the trend…
    - AVs (human drivers’ metric on fatality per mile)
    - Cancer diagnostic (human doctors’ successful rate)
- How to demonstrate the basic-event (BE) probabilities are satisfied?
    - (BE: failure of component-level functionalities)
  - Our RAM is one way to demo. the probability of the BE on misclassification.
  - Bespoke RAMs are needed for each functionalities of ML components.

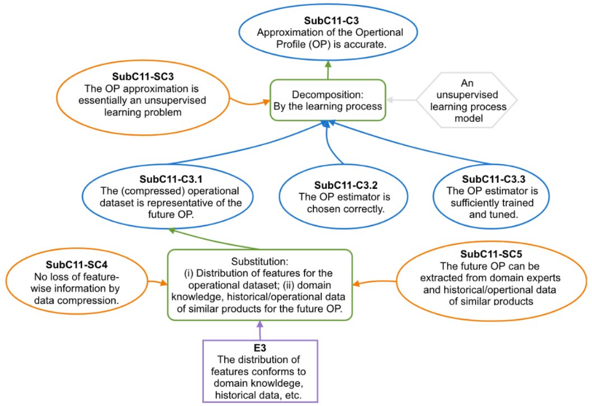
Present the RAM as Probabilistic Safety Arguments



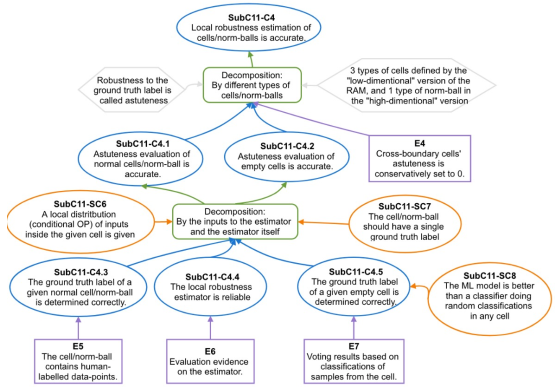
- Main steps:
  - Partition the whole input space into small “cells”;
  - Approximate the OP of cells;
  - Evaluate the robustness (w.r.t. the ground truth label) of cells;
  - weighted average on the robustness of a population of cells, based on limited samples from the population (weights are their OPs).



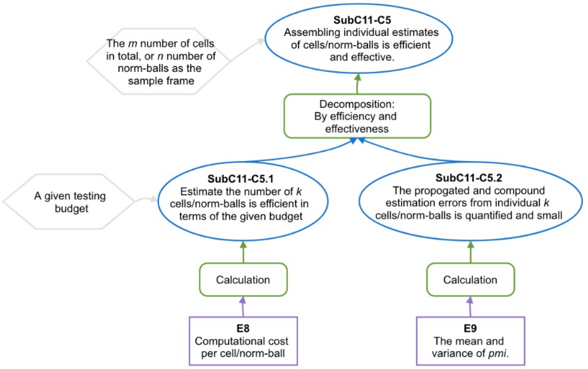
SubC11-C2
Determination of cells/norm-balls is rigorous.
Concretion:
The determination depends on the two model parameters r and ϵ.
Description of the RAM
SubC11-SC2
The r-separation distance can be estimated from the existing dataset.
SubC11-C2.1
Estimation of r is accurate.
SubC11-SC1
The r-separation property holds for the given application
SubC11-C2.2
Choice of ϵ is correct.
Substitution:
The existing dataset for more and more collected data
Calculation
Evidence Incorporation
SubC11-C2.3
The r-separation distance is continuously updated as more labelled data is collected
E1
The minimum distance between any two data-points of different labels in the existing dataset.
E2
ϵ<r



SubC11-C3
Approximation of the Opertional Profile (OP) is accurate.
SubC11-SC3
The OP approximation is essentially an unsupervised learning problem
Decomposition:
By the learning process
An unsupervised learning process model
SubC11-C3.1
The (compressed) operational dataset is representative of the future OP.
SubC11-C3.2
The OP estimator is chosen correctly.
SubC11-C3.3
The OP estimator is sufficiently trained and tuned.
SubC11-SC4
No loss of feature-wise information by data compression.
Substitution:
(i) Distribution of features for the operational dataset; (ii) domain knowledge, historical/operational data of similar products for the future OP.
SubC11-SC5
The future OP can be extracted from domain experts and historical/opertional data of similar products
E3
The distribution of features conforms to domain knowldege, historical data, etc.



SubC11-C4
Local robustness estimation of cells/norm-balls is accurate.
Robustness to the ground truth label is called astuteness
Decomposition:
By different types of cells/norm-balls
3 types of cells defined by the "low-dimentional" version of the RAM, and 1 type of norm-ball in the "high-dimentional" version
SubC11-C4.1
Astuteness evaluation of normal cells/norm-ball is accurate.
SubC11-C4.2
Astuteness evaluation of empty cells is accurate.
E4
Cross-boundary cells' astuteness is conservatively set to 0.
SubC11-SC6
A local distritbution (conditional OP) of inputs inside the given cell is given
Decomposition:
By the inputs to the estimator and the estimator itself
SubC11-SC7
The cell/norm-ball should have a single ground truth label
SubC11-C4.3
The ground truth label of a given normal cell/norm-ball is determined correctly.
SubC11-C4.4
The local robustness estimator is reliable
SubC11-C4.5
The ground truth label of a given empty cell is determined correctly.
SubC11-SC8
The ML model is better than a classifier doing random classifications in any cell
E5
The cell/norm-ball contains human-labelled data-points.
E6
Evaluation evidence on the estimator.
E7
Voting results based on classifications of samples from the cell.



The m number of cells in total, or n number of norm-balls as the sample frame
SubC11-C5
Assembling individual estimates of cells/norm-balls is efficient and effective.
Decomposition:
By efficiency and effectiveness
A given testing budget
SubC11-C5.1
Estimate the number of k cells/norm-balls is efficient in terms of the given budget
SubC11-C5.2
The propogated and compound estimation errors from individual k cells/norm-balls is quantified and small
Calculation
Calculation
E8
Computational cost per cell/norm-ball
E9
The mean and variance of pmi.

Part 2---Discussion

- Similar assurance case frameworks are emerging
  - Complement others from quantitative aspects
    - e.g., allocating quantitative safety targets, supporting reliability claims stated in some measure.
- Complete?
  - ``Vertically'', it is ``end-to-end'' (from the very top claim to evidence, chain of methods)
  - ``Horizontally'', it is incomplete with undeveloped claims (RAMs for other ML func.)
- System-level quantitative safety targets? Esp. the AUV case study…
  - lack of statistical data due to the novel applications of AUVs.
    - Human divers doing similar underwater tasks?
- Highly depends on domain-knowledge/engineering-experience
  - HAZOP, hazards scenarios modeling, FTA

Thank you!

- xingyu.zhao@liverpool.ac.uk
- https://x-y-zhao.github.io/

- Please refer to our SOLITUDE project website for more technical details, source code, DL models, datasets and publications.