TOWARDS INTERNATIONAL STANDARDS FOR EVALUATING MACHINE LEARNING

Frank Rudzicz[1,2,3,4], P Alison Paprica[1,3], Marta Janczarski[5]

1 UNIVERSITY OF TORONTO

2 St. Michael's
Inspired Care.
Inspiring Science.

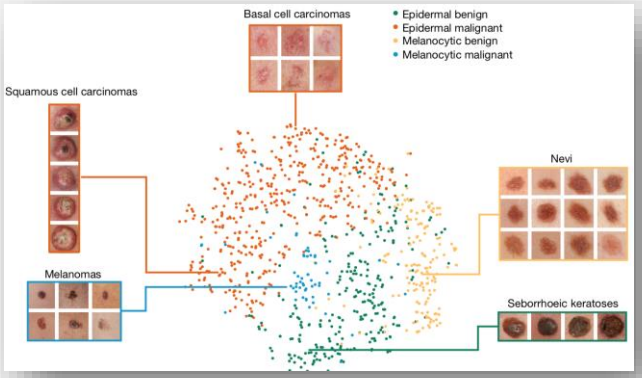
3 VECTOR INSTITUTE

4 SURGICAL SAFETY TECHNOLOGIES

5 Standards Council of Canada
Conseil canadien des normes

INTRODUCTION

- **International standards** are a primary method to ensure the safety of a process or product.
- Machine learning (ML) has been *ad hoc*, susceptible to trends, and relatively undirected.
  - Without careful consideration to **empirical methodology**, the truth and generalizability of our own experimental results may be suspect.
- Consider ML that identifies malignant skin in images.
  - Not accounting for *skin colour* may aversely affect under-represented populations.



Esteva A, Kuprel B, Novoa RA, *et al.* (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**:115-118

THE DAWN OF AI STANDARDS

- The following study groups within **ISO/JTC1 SC 42** were formed in 2018:
  - **Computational approaches and characteristics** includes specialized AI systems (e.g., NLP or computer vision) to understand and identify their underlying computational approaches, architectures, and characteristics.
  - **Trustworthiness** concerns approaches to establish trust in AI systems, e.g., through *transparency*, *verifiability*, *explainability*, *controllability*. Typical threats and risks, their mitigation techniques, and approaches to *robustness*, *accuracy*, *privacy*, and *safety* will also be investigated.
  - **Use cases and applications** focuses on application domains for AI (e.g., social networks and embedded systems) and the different context of their use (e.g., health care, smart homes, and autonomous cars).

ISO

STANDARDS FOR EVALUATING ML MODELS

- When comparing the performance of two or more models, the following aspects must be carefully controlled and reported:
  - **Implementation** E.g., if an algorithm can be accelerated in such a way that can affect outcomes, then this must be made explicit.
  - **Hyper-parameter** optimization **s**hould not favor one model over another.
  - **Preprocessing** will not unjustly favour one model over another. E.g., removing outliers, incomplete data, or noise should not unfairly affect performance.
  - **Training and testing data** should be ecologically valid, statistically indistinct, or otherwise similar to data expected to be observed in deployment.
  - **Appropriate baselines** Any classifier should be compared against ≥1 representative, appropriate baseline. Trivial baselines should not be considered.
  - **Limiting channel effects** incl. characteristics of the *manner* in which data were recorded, in addition to the nature of the data themselves. Some strategies explicitly factor out channel effects.
  - **Appropriate statistical tests of significance** must be undertaken, when possible.

STANDARDS FOR EVALUATING ML MODELS

- Ensuring the **safety** and **maintenance** of AI systems will be the subject of various standardization efforts, incl. explainable models, unintended biases, human-machine interaction, and scalable oversight.
- It is essential to objectively establish ML performance correctly, consistently, and with expected minimal levels of reporting, otherwise claims should not be trusted.
- Agreeing upon a **minimal set of standards for evaluation**, across sectors, may require a broad cultural change. Indeed, a meta-analysis showed, while the ML and NLP communities are driven by experimental results, statistical significance testing is ignored or misused most of the time.
- **Trusting the objective quantitative performance of our systems is itself a safety concern.**