# Call for machine learning guidelines for precision livestock farming

M. Pastell[1], H. Fred[1], T. Norton[2], B. Aernouts[3], M. Taghipoor[4], P. J. De Temmerman[5], J. Maselyne[5]

[1]Luke, [2,3] KU Leuven, [4] INRAE, [5] ILVO

# Why do we need ML guidelines?



- Machine learning (ML) models have powerful predictive capabilities but are prone to overfitting – **general AI hype can also leak to scientific language**
- The use of ML in precision livestock farming (PLF) has gotten easier with increasing availability of software libraries and **with relative ease of modelling making mistakes is also easy**
- Inconsistency in the reporting of the methods and a tendency to make quite strong conclusions based on limited datasets can be observed
- Common guidelines could increase the quality and consistency of modeling and reporting of research

# Do we need own ML guidelines for PLF?

- General ML guidelines already exist, do we need to make PLF specific guidelines or push for adoption of existing ones?
- REFORMS guidelines from 2024 are well suited for PLF – developing PLF based reporting standards could make the adoption easier
- This presentation aims to address PLF specific questions: sensor-based classification of behavior, welfare, resilience and health, detection models and prediction models

SCIENCE ADVANCES | REVIEW

RESEARCH METHODS

## REFORMS: Consensus-based Recommendations for Machine-learning-based Science

Sayash Kapoor[1,2]*, Emily M. Cantrell[3,4], Kenny Peng[5], Thanh Hien Pham[1,2], Christopher A. Bail[6,7,8], Odd Erik Gundersen[9,10], Jake M. Hofman[11], Jessica Hullman[12], Michael A. Lones[13], Momin M. Malik[14,15,16], Priyanka Nanayakkara[12,17], Russell A. Poldrack[18], Inioluwa Deborah Raji[19], Michael Roberts[20,21], Matthew J. Salganik[2,3,22], Marta Serra-Garcia[23], Brandon M. Stewart[2,3,22,24], Gilles Vandewiele[25], Arvind Narayanan[1,2]

**Table 2. Stages of ML-based science and corresponding checklist modules.**

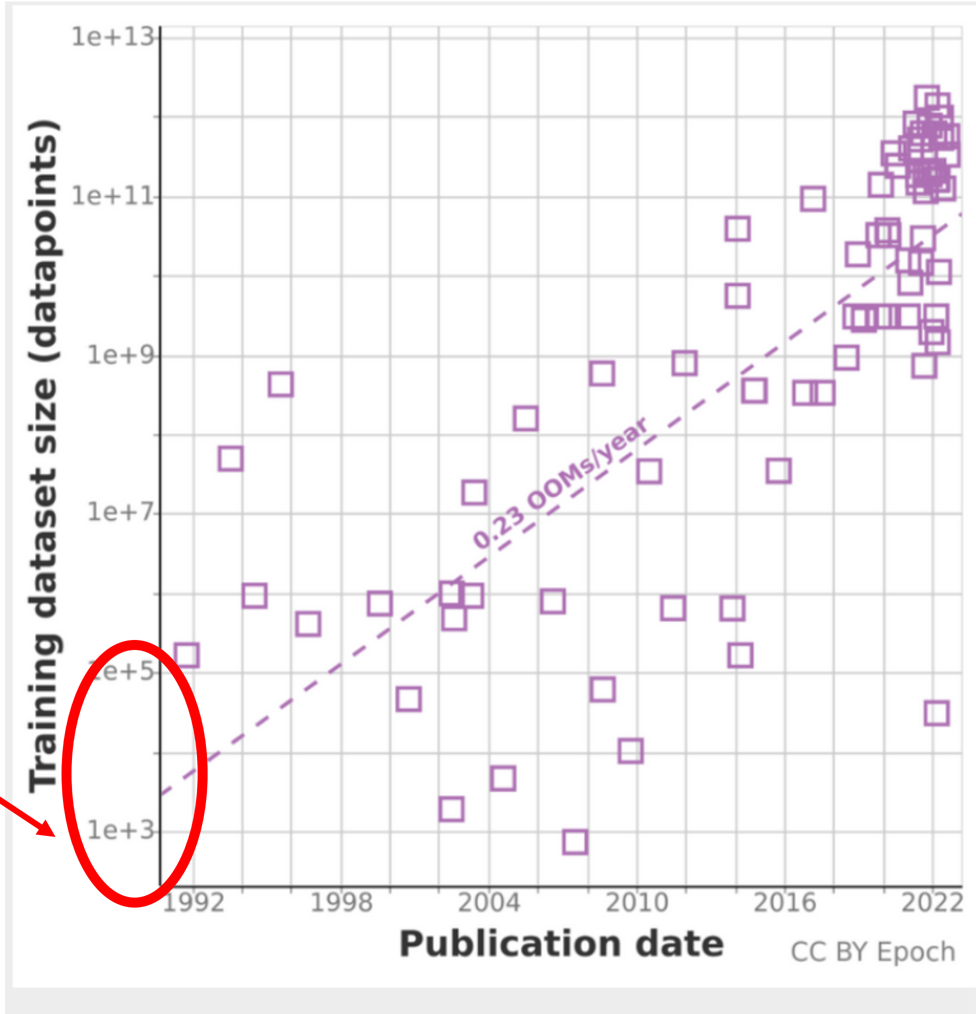
| Stage of scientific study | Section of the checklist |
|---|---|
| **Study design** | Study goals (Module 1) |
| | Computational reproducibility (Module 2) |
| **Data collection and preparation** | Data quality (Module 3) |
| | Data preprocessing (Module 4) |
| **Modeling** | Modeling decisions (Module 5) |
| **Evaluation** | Data leakage (Module 6) |
| | Metrics and uncertainty quantification (Module 7) |
| **Scope and limitations** | Generalizability and limitations (Module 8) |

Luke

# PLF models often deal with small data



- The general progress in AI models is largely based on increased training data
- It is not hard to find PLF papers using the "big data pitch" on small datasets <1000 animals, <1M data points
- The size of the used datasets in PLF studies hasn't increased radically in the past 20 years
  - The cost of data collection relative to research project size is fairly high
  - Sharing of research datasets is still limited (although increasing): cultural and actual reasons




Trends in Training Dataset Sizes, epoch.ai

# How much data do we need for good models?



RESEARCH Open Access

Use of machine learning to analyse routinely collected intensive care unit data: a systematic review

Check for updates

Duncan Shillan[1,2], Jonathan A. C. Sterne[1,2], Alan Champneys[3] and Ben Gibbison[1,4,5*]

- Sample size <1000 patients for machine learning studies provides overoptimistic results (overfitting)
- Model predictive accuracy increases with increasing sample size
- Highest accuracy reached with data from over 100 000 patients

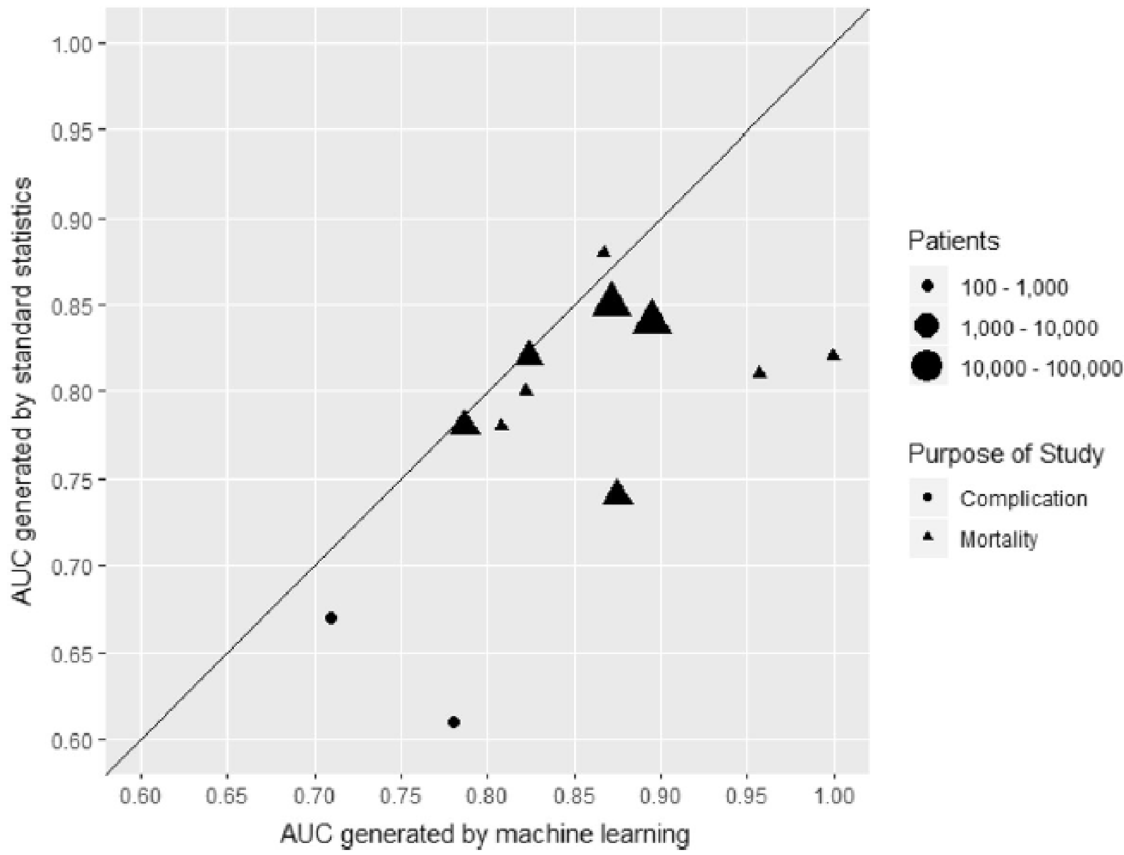



Figure 5. Comparison of AUC scores found in complication of mortality prediction papers according to the techniques used to produced them.

*"Methodological and reporting guidelines are needed … to increase confidence in reported findings "*

# What claims can be made? Generalizability

- Several PLF studies demonstrate poor performance of ML models in new environments (e.g. farms)
  - Adriaens et al. 2020 developed prediction models for resilience rank of dairy cows based on milk yield and activity sensor data from 5-year dataset from 27 farms and concluded that **individual models were needed for each farm** *"We could not find SF that* ***were commonly informative to predict RR over all farms****"*. Classification accuracy **varied between farms** 46% - 84%
  - Stygar et al. 2023 found that no common model for classifying welfare based on similar data across 6 farms was found, however **farm-specific models were more predictive**

- The strength of evidence increases as variability (number of animals, number of farms) in the test set increases but by how much?

# What claims can be made? Generalizability

- **Suggestions:**
  - All studies working with **data from single farm** should be reported as **case studies**
  - Small sample size should not prevent publication if the study is otherwise valid
  - Confidence intervals for models should be provided – while also understanding these may not hold on unseen data
  - If claims about external validity are made then evidence should be provided *"report quantitative evidence by testing their claims in out-of-distribution data … theoretical arguments about their expectations"* (REFORMS):

- Recommended reading: *ImageNet Large Scale Visual Recognition Challenge, Russakovsky et al. 2015.*

Luke

# PLF methods: shared data, open code and reproducibility



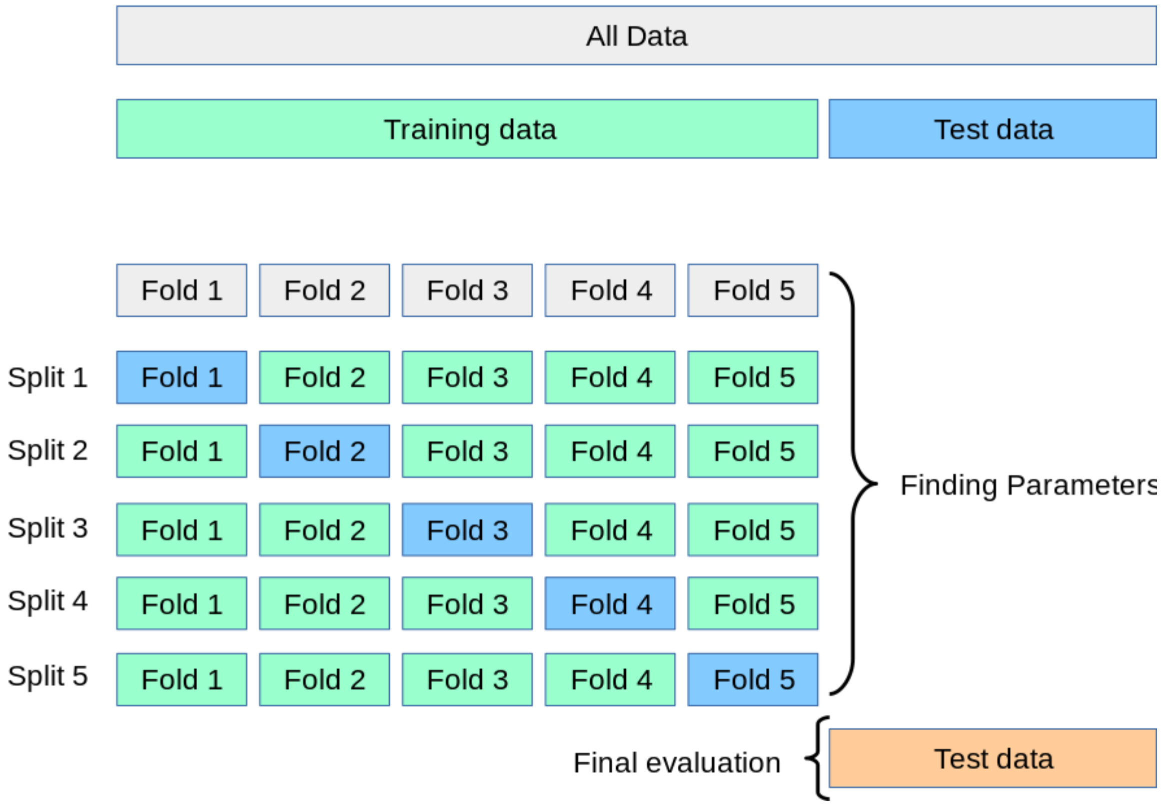
- Open source models and open/shared datasets have been very important for the progress of deep learning
  - Code and model sharing for PLF is not yet the norm – but it should be
  - We are moving to right direction – especially image datasets are increasingly shared
  - We should look for ways to also share data from commercial farms – do we need standards for anomyous data or use federated learning?
  - Instances of broken links still surprisingly common – **use persistent repositories**
  - **Benchmark datasets to drive methodological progress should be established** – can we find a point where different types of models begin to generalize?

**Recommended reading:** *Reproducibility standards for machine learning in the life sciences, Heil et al. 2021*

Luke

# Data leakage - common issues

- Data leakage means that information leaks from training set to test set
  - Same **animals/groups/farms** are used in training and testing datasets and false claims about generalizibility are made
  - Calculating features from the entire dataset (e.g. mean normalization) and not independently for training and test sets
  - Including predictors which are not available in unseen data (e.g. data from future samples)

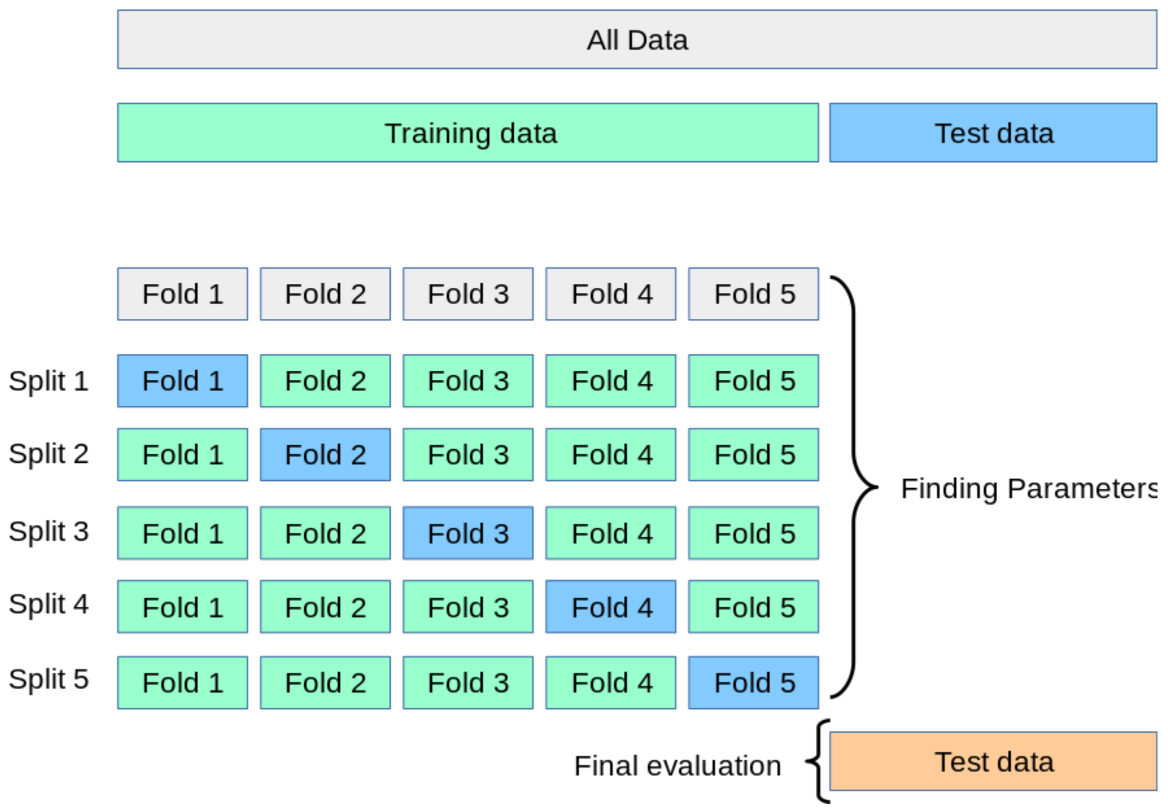





Luke

# Data leakage - recommendations

- Dataset should aim for independent split: e.g. no data from the same animals in training and testing data
- All feature engineering should be independent of the test / holdout sets
- An **independent holdout set** or **nested crossvalidation** should always be used




Luke

# Discussion

- Adoption of machine learning guidelines for PLF/Animal science in key journals could enchance consistency of reporting of studies and lead to improved science
- The guidelines **should be helpful** for new ML practioners and new animal science practioners and **not be a barrier to entry** to publishing
- Consistency of reporting of different metrics would make comparing studies easier *(Stygar et al. 2021)*
- Developing consensus based “own guidelines” can be a useful process

- **What do you think?**

Digi4Live

aWISH
ANIMAL WELFARE INDICATORS AT THE SLAUGHTERHOUSE

# Thank you!

luke.fi

Luke

Funded by
the European Union