

UNVEILING ONE OF THE PILLARS OF RELIABLE MACHINE LEARNING

Author:
Ahmed Ibrahim
A.Ibrahim-10@student.tudelft.nl

Supervisors:
Andrew Demetriou
Cynthia Liem

Evaluating Reporting Practices of Human Annotations within Top-Cited Papers in the IEEE Access Journal

1. Introduction

Machine learning (ML) impacts significant areas like medical diagnostics, emotion detection, and intelligent vehicle systems [1]. ML's success depends on 'ground truth', annotated by humans. Much like an unstable pillar risks a building's safety, poor annotations can undermine ML reliability.

This project reviews annotation practices across ML domains systematically.

DID YOU KNOW?

Childcare Benefit Scandal: Faulty ML algorithms led to 26,000 parents in debt and custodial losses [2].

2. Research Question

"What are the data collection and reporting practices of human annotations/labels in societally impactful applications of Machine Learning Research as reflected in top-cited papers from the IEEE?"

3. Methodology

4. Results

Table 1: Results of the systematic review

	Papers	Data Sources
Utilization of human annotations	70%	69%
Number of annotators reported	17%	31%
Formal annotation definitions and instructions reported	10%	27%
Training for annotators provided	3%	11%
Existence of multiple-annotator overlap	6%	22%
Inter-annotator agreement metric reported	50%	43%
All external data sources cited	68%	96%

Figure 1: Annotation Sources

5. Findings

- General Findings:**
- Papers rely on data sources
 - Diversity in data types
 - Superiority of larger datasets
 - Lack of annotation standards
 - Implications of transfer learning overlooked
- Domain-specific Findings:**
- Single annotator (pathologist) in medical diagnostics
 - Intelligent vehicle system domain needs enhancements
 - Human annotations in cybersecurity bypassed
 - Large dataset reliance in computer vision

Conclusion & Future Work

There is a **prevalent lack of formalization** in the **annotation process** across **all domains**.

Suggested future work:

- Establishment of a standard human annotation collection.
- Studying annotation quality's effect on ML algorithm.
- Examining how pre-trained model data affects classifier performance and adaptation to new data.

[1] H. Anahideh, A. Asudeh, and S. Thirumuruganathan, "Fair Active Learning," arXiv (Cornell University), Jan. 2020, doi: 10.48550/arxiv.2001.01796.

[2] E. Van Dam, "Martine raakte samen met 26.000 andere ouders in de knel door toeslagenaffaire," Omroep Gelderland, December 2020. [Online]. Available: <https://www.gld.nl/nieuws/6639460/martine-raakte-samen-met-26000-andere-ouders-inde-knel-doortoeslagenaffaire>.