# UNVEILING ONE OF THE PILLARS OF RELIABLE

MACHINE LEARNING

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

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

> **Supervisors: Andrew Demetriou Cynthia Liem**

## 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.

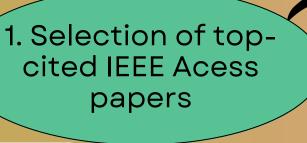
This project reviews annotation practices across ML domains systematically. DID

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?"





Accessible

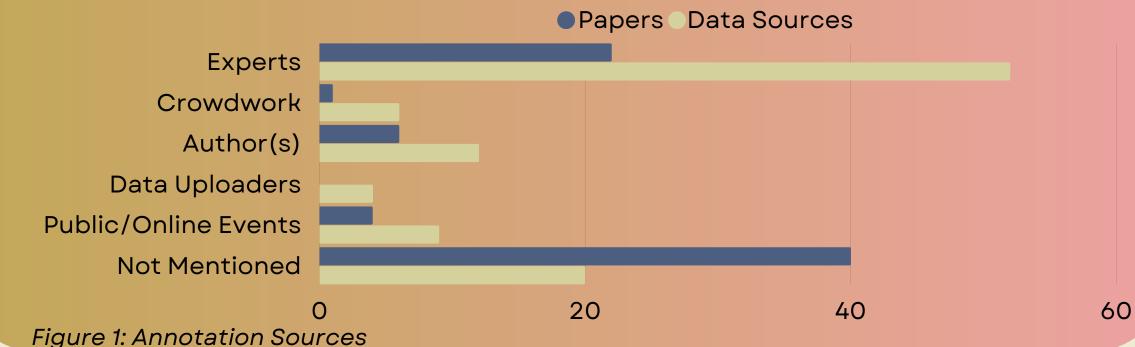
**Many domains** 

High h5-index

2. Conduct systematic review 3. Data analysis using Excel

### 4. Results

l able 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%
• Do	nore Dote Courses	



# 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.



YOU