

What is the performance of machine learning models for multi-hazard disaster prediction and their feasibility of application in humanitarian forecasting?

Ewa Smura¹, Marijn Roelvink¹, Cynthia Liem¹

¹ Delft University of Technology, Delft, Netherlands
Correspondence: ewa.smura@gmail.com



1. Terminology and background

- **Humanitarian forecasting** means predicting humanitarian crises before they occur
- Machine learning **models** are used to predict natural disasters
- **Multi-hazard** means accounting for hazard interrelations
- Machine learning for hazard forecasting is a rapidly **growing field**, and last survey was done in 2019
- Decided to execute **a literature survey** with a focus on how well the models perform, what their actual feasibility of use is, and the **potential for cross-application**.

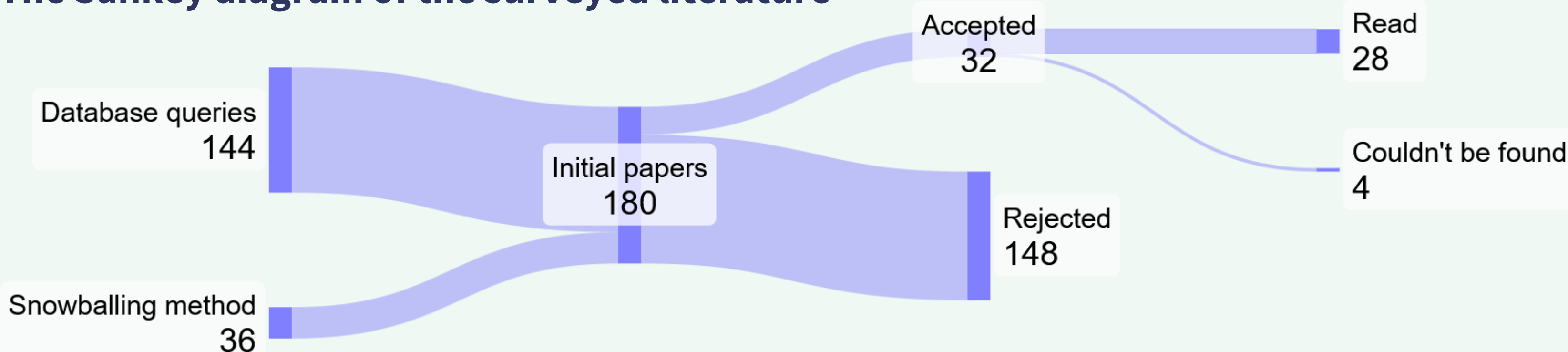
2. Sub-questions

- Group 1: What are the machine learning **models** used in the papers? What are the **metrics** used to report on the performance of the models? How does the choice of metric depend on model and domain?
- Group 2: What is the **performance** per metric, domain and model?
- Group 3: How to define and **judge feasibility** of practical application? What are the intended **practical applications** of the models? What factors influence the feasibility of their intended use? What is the feasibility of practical application of the models?
- Group 4: What are the possible **cross-applications** of the models in humanitarian forecasting?

3. Methodology

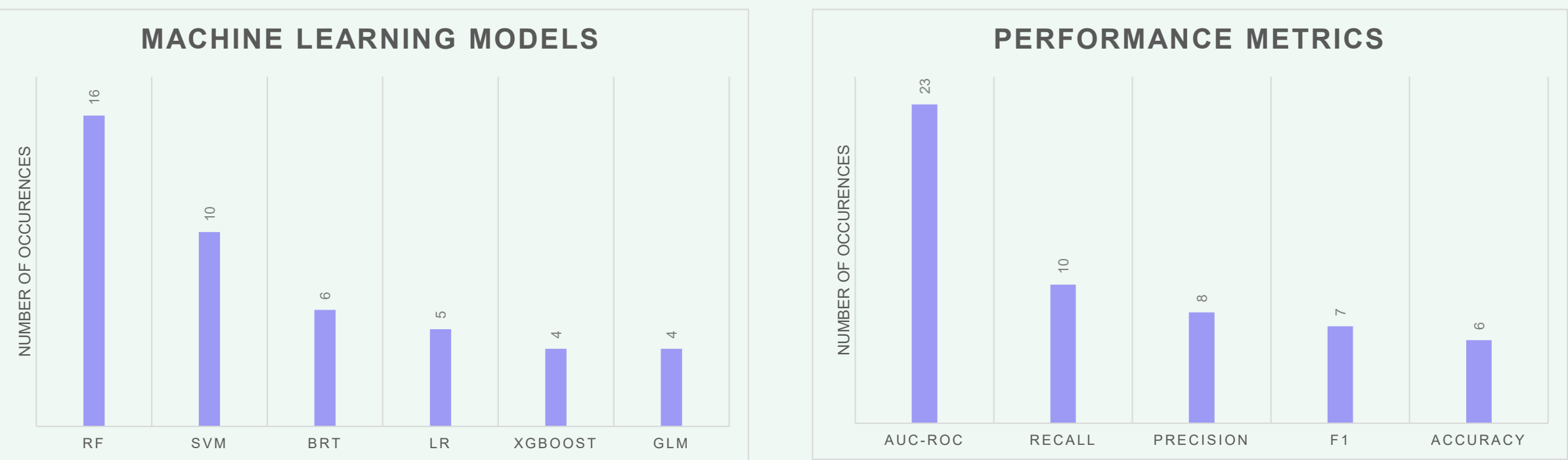
- Used the **SALSA method**: Search, Appraisal, Synthesis and Analysis
- Literature written in English and published between 2019 and 2025
- The **inclusion criteria** were using a machine learning model, reporting on its performance, and utilizing multi-hazard forecasting
- Two passes over the papers: a first pass to gather information and a second pass to judge feasibility
- The gathered data was **compiled into tables**, aggregated and analysed to find patterns

The Sankey diagram of the surveyed literature



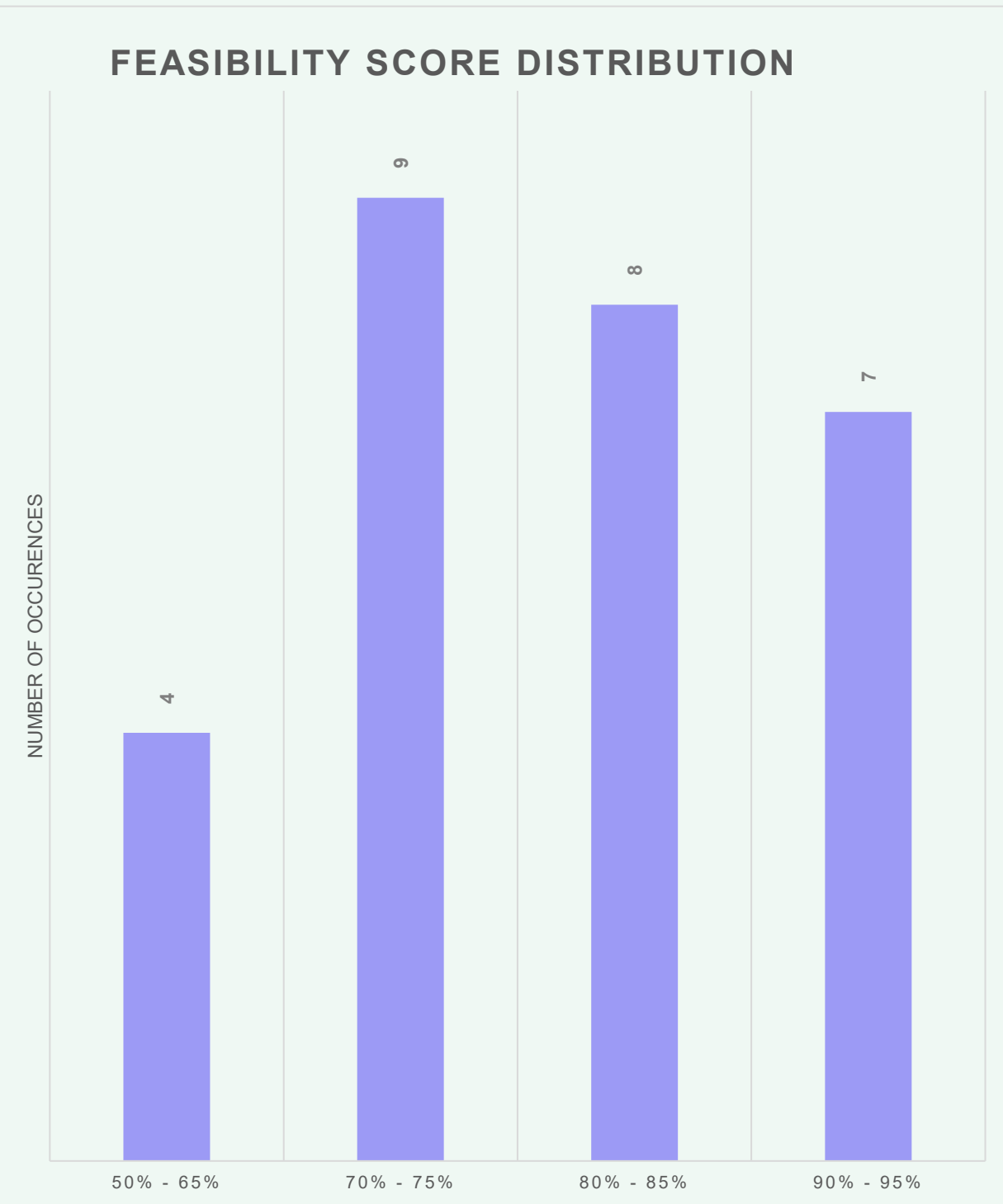
4.1. Results: Models and metrics

- Wide variety of both: 36 models, 20 metrics
- Common individual models are RF and SVM
- Boosted models and deep learning rising in popularity
- AUC-ROC is the dominant metric, followed by confusion matrix ones



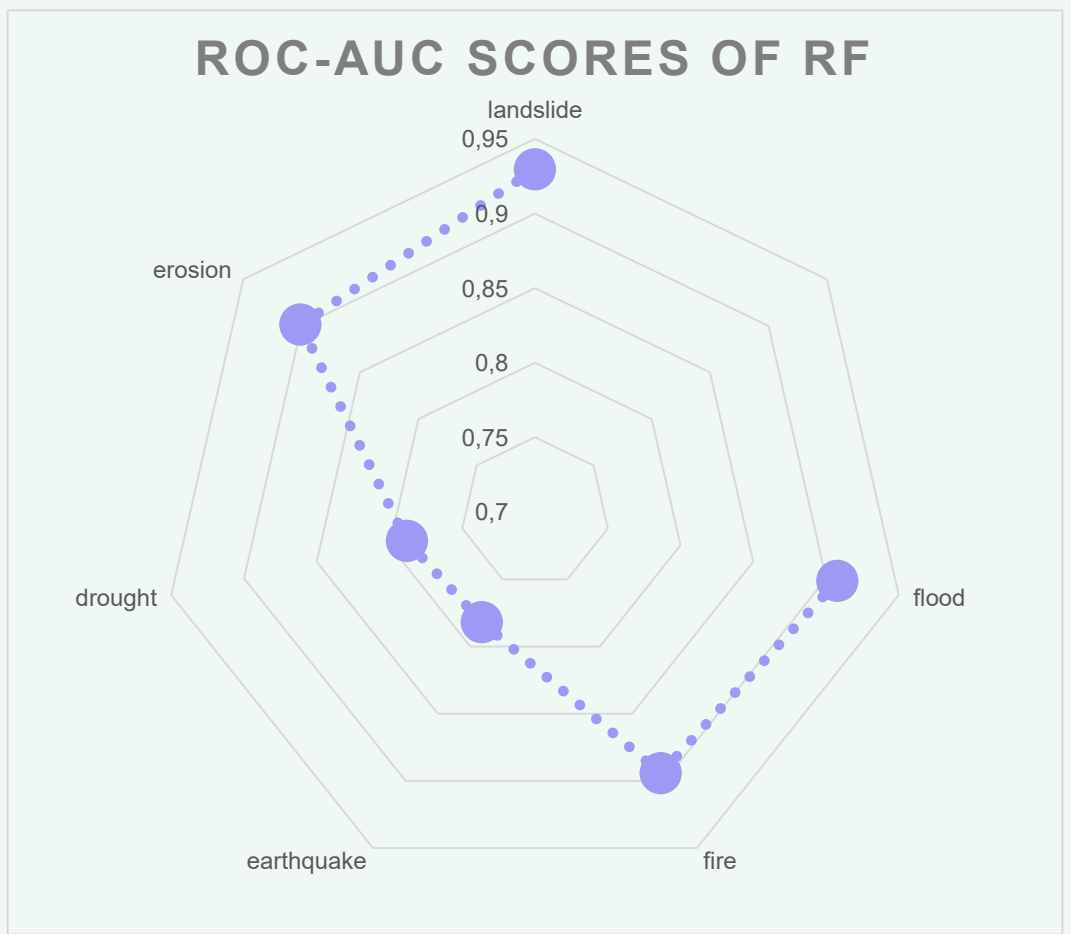
4.3. Results: Feasibility

- Defined as **readiness** to be applied in practice
- Common application is generating multi-hazard maps
- Use for land planning, disaster mitigation
- **Judged by:**
 - Development
 - Reliability
 - Performance
 - Detail
- General feasibility is overall good
- Best scores in performance and development



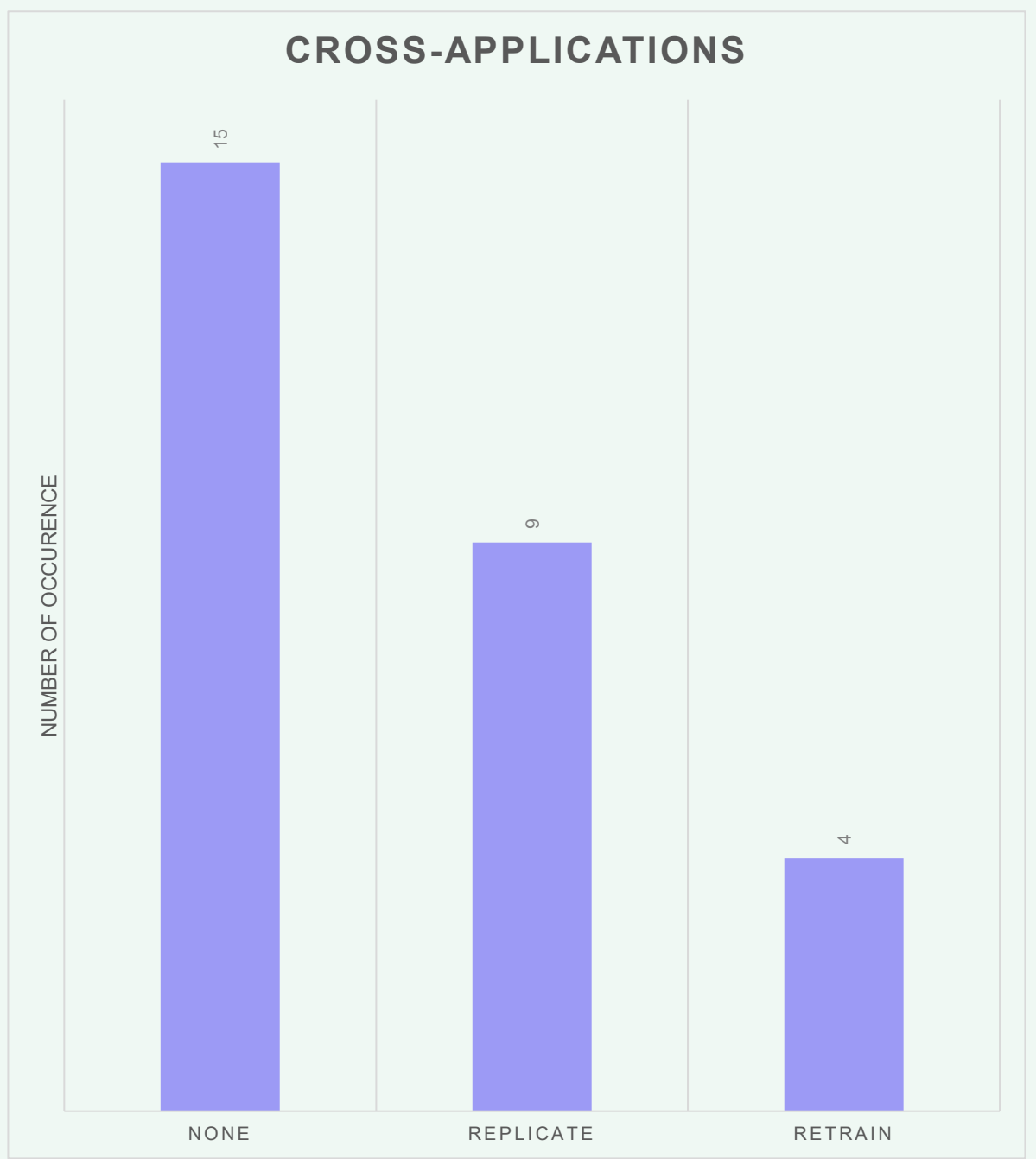
4.2. Results: Performance

- **Random forest** outperformed the other popular models
- Boosting models perform very well
- Articles tried multiple models, used the best one
- Earthquake and drought prediction seems challenging
- Landslide, floods and fire prediction got quite good scores



4.4. Results: Cross-application

- **Self-reported** by the articles
- Defined as ability to apply model in other places
- Fifteen don't mention it at all
- Nine recommend same methodology
- Four can be **retrained** on other data to work elsewhere
- One additionally says it would work on other hazards



5. Limitations and points of interest

- The hazards most commonly occurring in literature are **landslides and floods**, followed by earthquakes and wildfires/forest fires
- The number of hazard/metric/model combinations makes comparing performance very hard
- Geographical distribution of studies is uneven
- Mentioned stakeholders don't include humanitarian organizations specifically

Read the full paper



6. Conclusions

- The most common **models** are **RF and SVM**
- The most common **metric** is **ROC-AUC**
- No connections were found between hazard, model and metric choice
- RF and boosting models **perform best**
- Overall model performance is good, with ROC-AUC scores above 0.8, though comparisons are challenging due to variety
- **Feasibility** judged on development, reliability, performance and detail
- Overall feasibility was quite high
- **Cross-application** is not a wide consideration