Problems in data for **forecasting disease outbreaks**

"What are the areas of improvement for data available for the development of disease outbreak forecasting ML models?"

Humanitarian crises suffer from late and reactive responses. Machine Learning offers the potential to forecast disease outbreaks early, enabling preemptive interventions. However, the quality of data is an important limiting factor for these algorithms.

METHOD

Collected 22 papers about model development

Papers categorised by: disease, scope, data type and problems reported

Problems were sorted into 5 categories and analysed

RESULTS

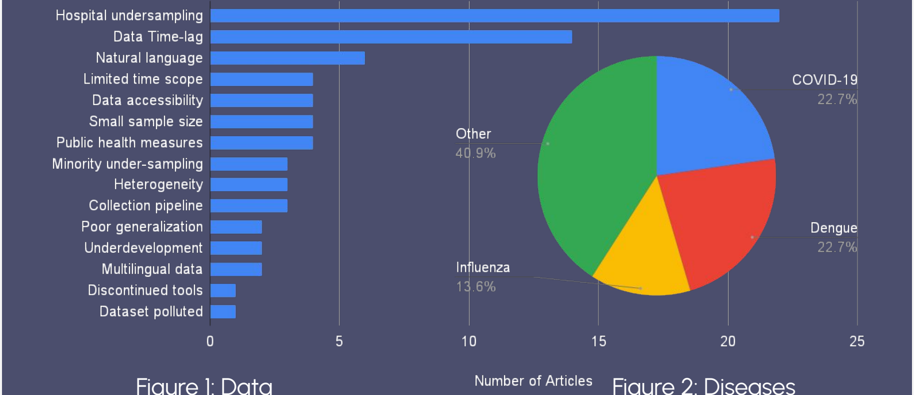

Figure 1: Data problem frequency

Figure 2: Diseases studied

CONCLUSIONS

Most models used hospital data, which inherently undercounts cases and has delays. This hurts real-time forecasting. Non-traditional sources like social media and google query data help but need heavy cleaning and are often ambiguous.

Problems were grouped into 5 types:

- **Structural** (e.g., hospital under-sampling, data lag)
- **Procedural** (e.g., short time scopes, poor data design)
- **Accessibility** (e.g., blocked or private data)
- **Logistical** (e.g., low resources in poorer regions)
- **Temporary** (e.g., early COVID-19 data gaps)

FULL PAPER

Few models used extra data like mobility or weather, though it often improves performance. Access to data is blocked by commercial interests or poor digitisation. Most models only work locally due to poor data standardization across countries and divergent disease behaviour.

Author: Matej Bavec
Supervisors: Marijn Roelvink, Cynthia Liem

TUDelft