Published online: 2020-08-21

Appendix: Content Summaries of Selected Best Papers for the 2020 IMIA Yearbook, Section 'Public Health and Epidemiology Informatics'

Bruzelius E, Le M, Kenny A, Downey J, Danieletto M, Baum A, Doupe P, Silva B, Landrigan PJ, Singh P

Satellite images and machine learning can identify remote communities to facilitate access to health services

J Am Med Inform Assoc 2019;26(8-9):806-12

In low-income countries, a promising strategy for improving care access among remote rural population is via the expansion of community health worker (CHW) programs. In settings where census data is missing and vital registration systems are weak, a persistent barrier of the expansion of CHW programs has been the difficulty to accurately enumerate population catchment areas. The authors used satellite-based neural network methods to automate the identification of communities in very rural areas.

Training data came from the publicly available SpaceNet corpus and a rural satellite image dataset specifically built for this project. External validation data was provided by a geographic information system dataset identifying all known Liberian communities within the health service catchment area of Last Mile Health, a non-profit organization. Community geolocation data was obtained by sending a team into the field with handheld GPS devices to collect community locations. Then 26,180 candidate rural images were labeled for this project and split into training and testing sets using an 80:20 ratio. The community prediction approach involved recognition of individual buildings from satellite imagery with TensorBox that output a set of coordinates describing the bounding box of each building. In a second phase, a clustering method was used to identify groups of densely connected buildings indicative of a community. The source code of their program is published.

Compared with existing health system community census data, the study method detected 75% of registered communities and identified an additional 167 building groupings that had not previously been identified. This new method for identifying communities in rural and remote settings using satellite imagery and deep learning has the potential to facilitate greater targeting of health services in low-income countries.

Feldman J, Thomas-Bachli A, Forsyth J, Patel ZH, Khan K

Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise

J Am Med Inform Assoc 2019;26(11):1355-9

Rapid onset of infectious disease epidemics can significantly reduce cases and deaths. Online media reports can facilitate timelier identification. The huge volume of media reports and the different languages make the identification of disease activity very challenging. The authors collected media records from the Global Database of Events Language and Tone (GDELT), that monitors the world's broadcast, print, and web news from nearly every country. Its global coverage and its updates every 15 minutes make it an invaluable source.

The authors used Google Translate to translate every media report they found into English. A dictionary containing a curated list of disease names was created. If an article didn't contain a disease name in its title, the article was deemed irrelevant. To distinguish articles talking about general infectious disease information and about disease activity, a supervised classification model was trained on 8,322 manually labeled articles. Finally, a user interface was built to allow clinical experts to verify articles clustered by disease, location, and time. The authors compared their GDELT-derived feed to the WHO disease Outbreak News reports from July 2017 to June 2018.

Their classification model achieved a F1 score of 0.87. On the study period, 37 outbreaks were reported by the WHO. Out of the 37 outbreaks, 89% were covered by online news outlets before the WHO reported the outbreak and the system correctly detected 94% of these events before reported by the WHO with a mean of 43.4 days earlier. Since it takes time for health authorities to investigate and confirm a disease, outbreak media reports can provide timelier information, but news reports fail often to distinguish between suspected and confirmed cases and are prone to false positive errors.

Combining natural language processing, machine learning, and human expertise, the authors created an international and near real-time event-based infectious disease activity database.