INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2022-2024

FGAI4H-P-020-A01

ITU-T Focus Group on AI for Health

Original: English

WG(s):	Plen	Helsinki, 20-22 September 2022

DOCUMENT

Source:	TG-Snake Topic Driver
Title:	Att.1 - TDD update (TG-Snake) [same as Meeting I]
Purpose:	Discussion

Contact:	Rafael Ruiz de Castaneda UniGe, Switzerland	Email: rafael.ruizdecastaneda@unige.ch
Contact:	Isabelle Bolon Institute of Global Health University of Geneva, Switzerland	Tel: + 33 6 72 94 03 84 Email: isabelle.bolon@unige.ch

Abstract: This project aims to determine how computer vision and crowdsourcing, and novel mobile technologies can support snake identification during snakebite events and help healthcare providers improve case-management and health outcomes of victims in endemic countries.
NOTE – Document from the Deliverables collection in the Collaboration site.

This version of the TDD is the same as seen in Meeting I (FGAI4H-I-020-A01), reproduced for easier reference as a Meeting N document.

We foresee further updates of this document as the project is evolving rapidly.

Table of Contents

1 Introduction

1.1 Document structure

As part of the work of the WHO/ITU Focus Group (FG) AI for health (AI4H), this document specifies a standardized benchmarking approach for AI-based applications for snakebites and snake identification.

1.2 Topic description

Snakebite envenoming is a major global health issue and neglected humanitarian crisis (https://www.who.int/snakebites/en/). Today, 5 million snake-bites occur globally every year causing 125,000 deaths and 400,000 victims of disability/disfigurement. In India, snakebite is responsible for five deaths/hour and 70% bitten are men and "bread winners" of the family (Menon et al, 2017). Snakebite exacerbates poverty in rural, and increasingly urban, communities in developing countries, which host high snake diversity and limited medical expertise (Gutierrez et al., 2017; Longbottom et al., 2018). Snakebite deaths are preventable using correct antivenoms, which are often expensive, scarce and can have side effects. Antivenom choice is critical and depends first on identifying the biting snake (Bawaskar and Bawaskar 2019). However, healthcare providers are not herpetologists (i.e. snake experts) and often struggle to identify the snake carcass brought by victims/bystanders/relatives or to interpret their description of the biting snake.

Categorization of the topic according to categorization guideline (currently C-0xx)

Although certain laboratory techniques (i.e. immunoassays and genetic techniques based on snake venom and cells) have been developed to identify certain snakes, they are not adapted to poor countries (i.e. expensive and require several hours and experienced professionals for often insensitive diagnosis). We propose the first mobile app to urgently and reliably (I) identify snakes from photos using AI and crowdsourcing (i.e. global network of herpetologists), and (II) support victims and healthcare providers in the management of snakebite in endemic countries (scenarios: https://unige.ch/medecine/isg/files/2315/2377/5239/Snapp_diagram.pdf). Our solution is direct, rapid and adapted, empowering communities and healthcare providers. It builds on:

- Rapid growth in smartphone/Internet use/access in developing countries (ITU 2018).
- Emerging smartphone-based practices in communities confronted with snakebite (e.g. snakes and/or their carcass are often photographed/filmed by snakebite victims/bystanders or healthcare professionals).

The current gold standard in handling the health topic of snakebite is identification of the biting snake by a herpetologist acting as a consultant for a medical professional who then integrates this taxonomic identification with best-practice medical treatment (e.g. knowledge about antivenom, supportive care). However, in practice this often does not happen because herpetologists are limited in their number and availability, and medical professionals normally lack expertise in herpetology and snake identification.

AI can make an impact by speeding up the process of suggesting an identification to a clinician. Snakebite is a health emergency requiring immediate action in the field and at the health centre (e.g. neurotoxic envenomation can produce generalized paralysis, respiratory arrest and death in 30 min to few hours). AI offers an unprecedented opportunity for high-speed snake identification to support snakebite victims and healthcare providers. Interestingly, AI-based animal identification exists for other groups (e.g. birds, Hernández-Serna et al., 2014) but not for snakes, offering an innovative application for a life-threatening situation. AI would be one more resource for clinicians aiming to make a snake identification, although consultation with an expert will remain the gold standard and

additional verification by crowds of online snake enthusiasts is an important component that is being explored, with promising results.

The expected impact of the benchmarking is to evaluate the number of images per class necessary for training, in order to represent the large number of snake species with similar appearance, as well as morphological and colour pattern variation with age and from place to place. Given that some snake species are very rare and very few images of these species exist, a second goal is to evaluate model performance when trained on small data sets. Our AI model would consume taxonomically-identified wild snake images from across the globe and associated geolocation and timestamp metadata to reply to the zoological question: "What snake is that?". More specifically, we plan to test:

1. How the accuracy of the AI model is affected by the absolute number of training images of a species
2. How the accuracy of the AI model is affected by the ratio of training to testing images of a species
3. How well images that are difficult for humans to identify can be accurately identified by the AI model
4. Whether AI accuracy at identifying images to genus or family is better than identifying them to species, given the same amount of training data
5. How much AI accuracy improves when geographic location information are incorporated (e.g. in India there are about 300 snake species out of 3,700 globally)

Ultimately, the AI model would provide species- and context-specific recommendations on snakebite management for both victims (i.e. first aid/pre-hospital care, most accessible snakebite treatment centre) and healthcare providers (i.e. expected clinical signs, antivenom choice). For the latter, the mobile app would be developed integrating existing snakebite management guidelines and collaborating closely with snakebite experts from Doctors Without Borders, WHO, and the Global Snakebite Initiative. Data on patients' evolution and outcomes would be integrated to reinforce and improve clinical recommendations given by the app (i.e. integrated learning system).

1.3 Ethical considerations

- ethical considerations on usage of AI
- ethical consideration of and benchmarking including its data acquisition
- All data must be of high quality and compliant with ethical and legal standards, including respecting the copyright of the photographers and their ownership of their images (e.g. their right to use them in their own publications, books, online resources in the future).

1.4 Existing AI solutions (includes datasets, systems and benchmarks)

Current AI-based systems available for animal identification include iNaturalist and HerpMapper, which both have AIs for animal ID. Both take an image as an input, output is a top1 or top5 class list, both have a bias in the data towards North American species, iNaturalist can also go to family/genus and both are starting to integrate geographic location as a filter on the output. Neither one is specific to snakes but both of them include snakes. Our group has access to all the training data from both sources.

Existing benchmarking consists of the AICrowd challenge, based on a subset of the data with only classes that have >500 images per class. This has a similar bias in the data but covers at least a handful of species on all continents. The current top algorithm has an F1 score of 0.83 and a logloss of 0.66.

1.5 Existing work on benchmarking

- Papers on existing attempts to benchmark solutions on the topic – none at this time, although we are currently drafting one
- Clinical evaluation attempts, RCT, etc. – none at this time, but clinicians are involved in the process as end users
- Existing numbers: we currently have > 420,000 training images covering 2,460 species classes, and 22,653 testing images covering 937 species classes (894 in common with training data). We are actively recruiting additional image sources to increase both species class diversity and number of images per class. The ratio of training:testing data ranges from 3,768:1 for the species class with the most training images, to 1:168 for the species class with the fewest training images. 700 species classes have ratios >1 and 82 species classes have ratios <1.

1.6 AI4H topic group

- Topic group structure
    - Subtopic 1: What is the ideal ratio of training to testing images per species class?
    - Subtopic 2: What additional sources of testing data can we create or access?
        - Natural history museums are interested in helping by digitizing slides with verified identification and which can be tied to physical preserved specimens (in alcohol) that could later be used to sort out future taxonomic changes
        - Gold standard data set of testing images should be representative and long-term.
        - Ensuring taxonomic consistency long-term (could this be relevant to other TGs? e.g. classification system for cancers could change in the future)
- Topic group participation
- Tools/process of TG cooperation
- TG interaction with WG, FG
- Current topic group and topic status
- Contributors so far
- Next meetings
- Next steps for the work on this document:
    - Benchmarking AICrowd round 3 solution
    - Launch of AICrowd round 4 in May 2020, which will include more images/species and geography at the continent level
    - AICrowd direct ingestion of algorithms & live leaderboard top algorithm available via API
    - SnakeCLEF 2020 challenge to be launched in May 2020
    - CERN collaboration to develop better algorithms and annotate images using techniques for segmentation, including watersheding and image segmentation (e.g. the head, body, and tail of the snake)

2 Method

2.1 AI Input Data Structure

Inputs for benchmarking include the public dataset available on AICrowd (https://www.aicrowd.com/challenges/snake-species-identification-challenge/), consisting of data from Flickr, iNaturalist, and HerpMapper. We intend to soon add Twitter data scraped using CrowdBreaks platform using a list of keywords and crowdsourced to tag each image with the appropriate species class. In November 2019, we launched a second snake identification challenge based on 7,809 untagged photos from Twitter, and we asked online herpetologists to tag them. Our

preliminary results show that, after tagging the photos by the community , Twitter can be a valuable source of snake photos to keep feeding our snake photo dataset and improving computer vision snake classification algorithms.

- ontologies, terminologies: classes are species of snakes; there are 3,763 described species worldwide and this number is continuously growing. For AICrowd challenge round 1&2 we used 82,601 images from 45 species, and for round 3 we used 187,720 images from 85 species
- data format – JSON or CSV with image URLS, country, photographer, image source

2.2 AI Output Data Structure

- outputs to benchmark – F1, logloss
- ontologies, terminologies: see above
- data format: AICrowd now supports direct ingestion of algorithms (as a superior replacement to previous "solution files" containing metrics but not the algorithm itself) will soon support the implementation of the top algorithm available live from the leaderboard via an API

2.3 Test Data Labels

- label types – species (class), country (see above)
- ontologies, terminologies: see above
- data format: JSON or CSV

2.4 Scores and Metrics:

- which metrics & scores to use for benchmarking – F1 and logloss (see above)
- considering relation to parameters stakeholders need for decision making – ideal F1 for clinicians is 1, because both false positive and false negative misidentifications can have serious consequences and must be avoided
- considering scores that providers use
- considering the scope providers designed their solutions for
- considering the state of the art in RCT, statistics, AI benchmarking etc.
- considering bias transparency: a plot of the dataset bias is available on the AICrowd challenge page

2.5 Undisclosed Test Data Set Collection

- raw data acquisition / acceptance – already about 7000 images in hand from experts + 5000 from IndianSnakes.org
- test data source(s): availability, reliability,
- labelling process / acceptance – all training data are already labelled, and we built a citizen crowd tagging platform to verify IDs and to tag new images from Twitter & Facebook (CrowdBreaks)
- bias documentation process – by species, by global region
- quality control mechanisms – citizen crowd tagging platform (see above)
- discussion of the necessary size of the test data set for relevant benchmarking results – will never approach the size of the training data set, see above for consideration of the ideal ratio
- specific data governance derived by general data governance document (currently C-004)

2.6 Benchmarking methodology and architecture

- technical architecture - AICrowd

- hosting (IIC, etc.)
- possibility of an online benchmarking on a public test dataset – will soon be available on AICrowd
- protocol for performing the benchmarking (who does what when etc.)
- AI submission procedure including contracts, rights, IP etc. considerations

2.7 Reporting methodology

- Report publication in papers or as part of ITU documents
- Online reporting is accomplished through a public online leaderboard (https://www.aicrowd.com/challenges/snake-species-identification-challenge/leaderboards)
- Credit-Check like on approved sharing with selected stakeholders
- Report structure including an example
- Frequency of benchmarking: new AICrowd round once per month

3 Results

Reports of the different benchmarking runs: the top AICrowd leading algorithm resulting from rounds 1&2 has an F1 score of 0.861 and a log-loss of 0.53. We benchmarked the identification accuracy of the submitted algorithms at different taxonomic levels using an undisclosed test dataset, made up of 42,688 images of snakes submitted to iNaturalist between 1 January 2019 and 2 September 2019 (range 23 - 6,228 images per class). The top algorithm achieved an F1 score of 0.826 with a log-loss of 0.66 on these undisclosed test data. The top AICrowd leading algorithm resulting from round 3 has an F1 score of 0.9 and a log-loss of 0.39. We have not yet benchmarked this solution.

AICrowd challenge results

	Round 1&2	Round 3	Round 4
Training dataset			
Images	82 601	187 720	290 742
Species	45	85	788
Threshold images per species	500	100	20
Geography	No	Yes	Yes
Winner			
Name	Gokul Moorthy	Lukas Picek Miroslav Valan	
F1	0.861	0.9	
Logloss	0.532	0.383	
Benchmark on undisclosed dataset			
F1	0,826		
Logloss	0.658		

4 Discussion

- Discussion of the insights from executing the benchmarking on
  - external feedback on the whole topic and its benchmarking
  - technical architecture: AICrowd (see above), CrowdBreaks, pybossa
  - data acquisition: see above
  - benchmarking process: see above
  - benchmarking results: see above
  - field implementation success stories: coming soon!

Annex A:
Declaration of conflict of interest

We hereby declare that each contributor to this document has no relevant conflicts of interest.