KU
THE UNIVERSITY OF KANSAS

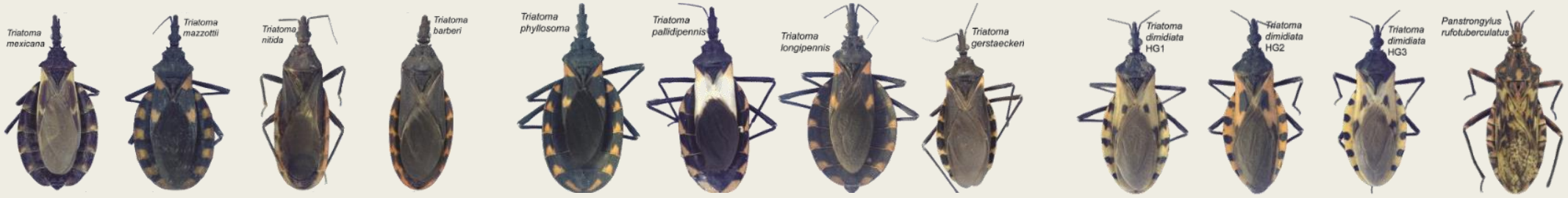
AUTOMATING IDENTIFICATION OF VECTOR INSECTS

Ali Khalighifar
Ed Komp
Town Peterson

Chagas Disease

https://goo.gl/images/m7rfq8

- The disease is caused by *Trypanosoma cruzi*
- Taxonomic impediment regarding vector species
- Most of the 150+ species of the family Triatominae feed on vertebrate blood
- Species identification is crucial because different species have different ecologies



VIRTUAL VECTOR LAB

Major support from the Office of the Provost @ KU
Spencer Art Museum, ITTC, Biodiversity Institute @ KU
Universidade de Brasília
Instituto Nacional de Salud Pública de Mexico

PeerJ

Automated identification of insect vectors of Chagas disease in Brazil and Mexico: the Virtual Vector Lab

Rodrigo Gurgel-Gonçalves[1], Ed Komp[2], Lindsay P. Campbell[3], Ali Khalighifar[3], Jarrett Mellenbruch[4], Vagner José Mendonça[1,5], Hannah L. Owens[3,6], Keynes de la Cruz Felix[7], A Townsend Peterson[3] and Janine M. Ramsey[7]

[1] Faculty of Medicine, Universidade de Brasília, Brasília, DF, Brazil
[2] Information and Telecommunication Technology Center, University of Kansas, Lawrence, KS, United States
[3] Biodiversity Institute, University of Kansas, Lawrence, KS, United States
[4] Spencer Art Museum, University of Kansas, Lawrence, KS, United States
[5] Centro de Ciências da Saúde, Universidade Federal do Piauí, Brazil
[6] Florida Museum of Natural History, University of Florida, Gainesville, FL, United States
[7] Centro Regional de Investigación en Salud Pública, Instituto Nacional de Salud Publica, Tapachula, Chiapas, Mexico

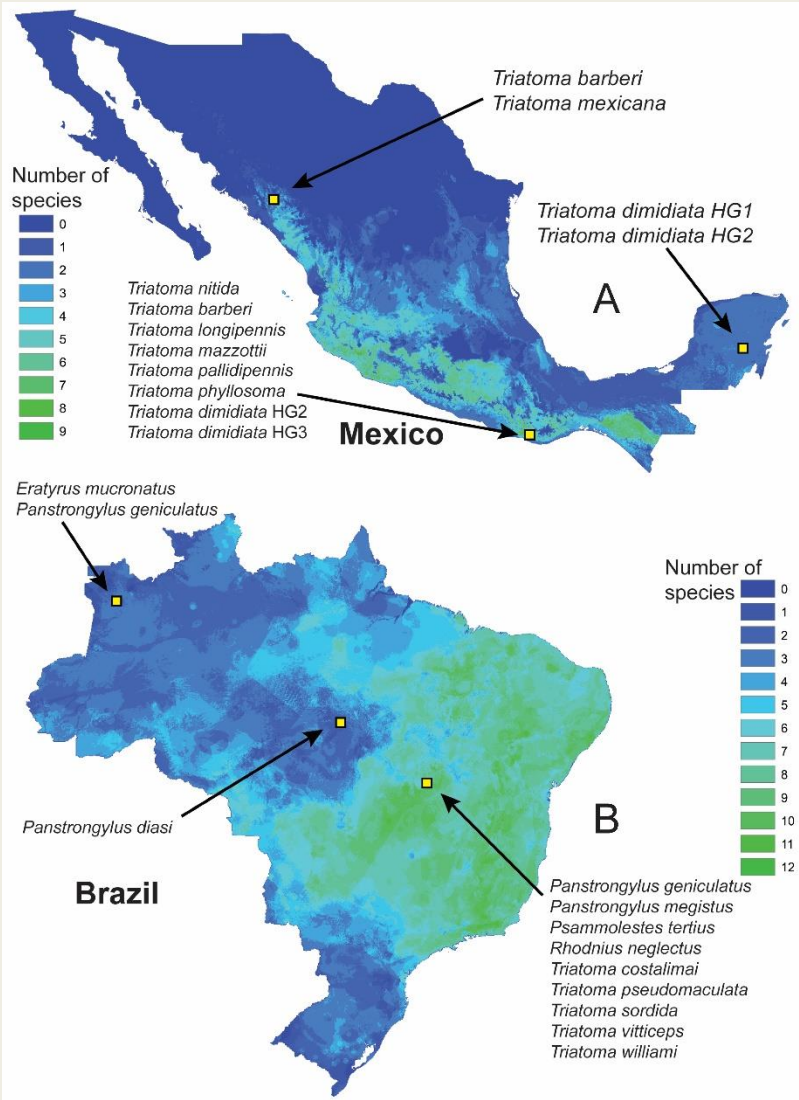
ABSTRACT

Identification of arthropods important in disease transmission is a crucial, yet difficult, task that can demand considerable training and experience. An important case in point is that of the 150+ species of Triatominae, vectors of *Trypanosoma cruzi*, causative agent of Chagas disease across the Americas. We present a fully automated system that is able to identify triatomine bugs from Mexico and Brazil with an accuracy consistently above 80%, and with considerable potential for further improvement. The system processes digital photographs from a photo apparatus into landmarks, and uses ratios of measurements among those landmarks, as well as (in a preliminary exploration) two measurements that approximate aspects of coloration, as the basis for classification. This project has thus produced a working prototype that achieves reasonably robust correct identification rates, although many more developments can and will be added, and—more broadly—the project illustrates the value of multidisciplinary collaborations in resolving difficult and complex challenges.

Submitted 26 September 2016
Accepted 28 January 2017
Published 18 April 2017

Corresponding author
A Townsend Peterson, town@ku.edu

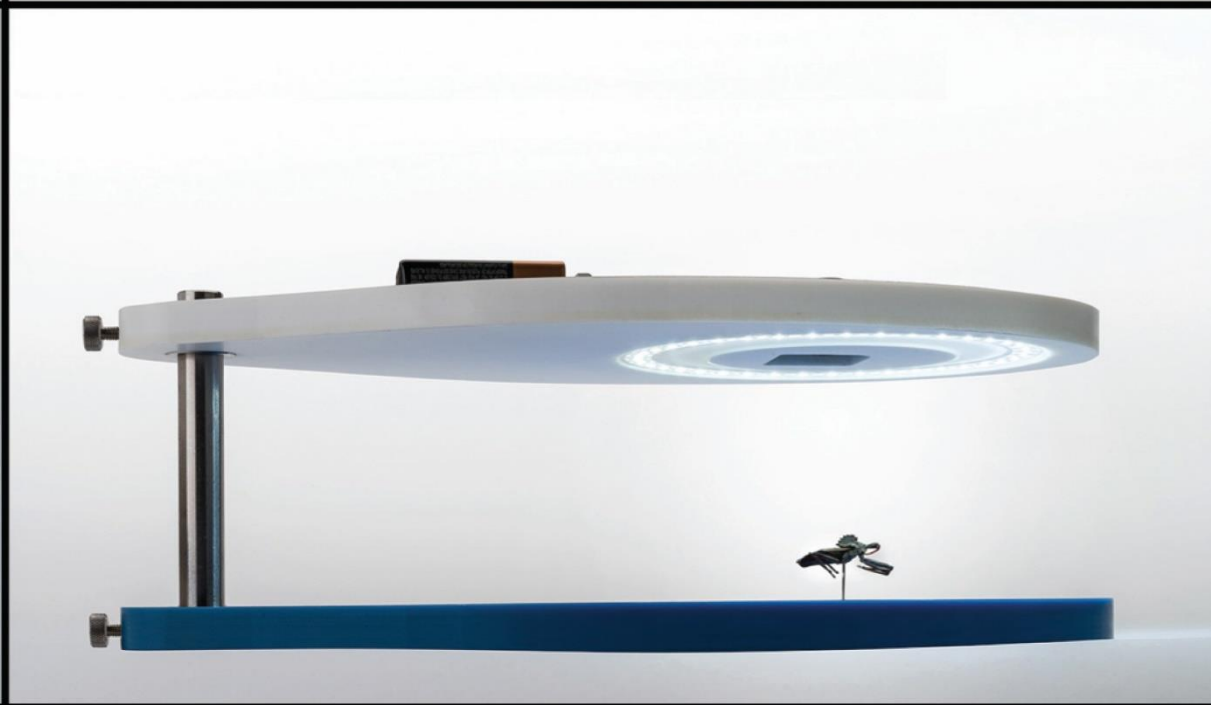
Subjects Entomology, Computational Science
Keywords Identification, Chagas disease, Triatominae, Automation, Primary occurrence data




DURACELL


DURACELL

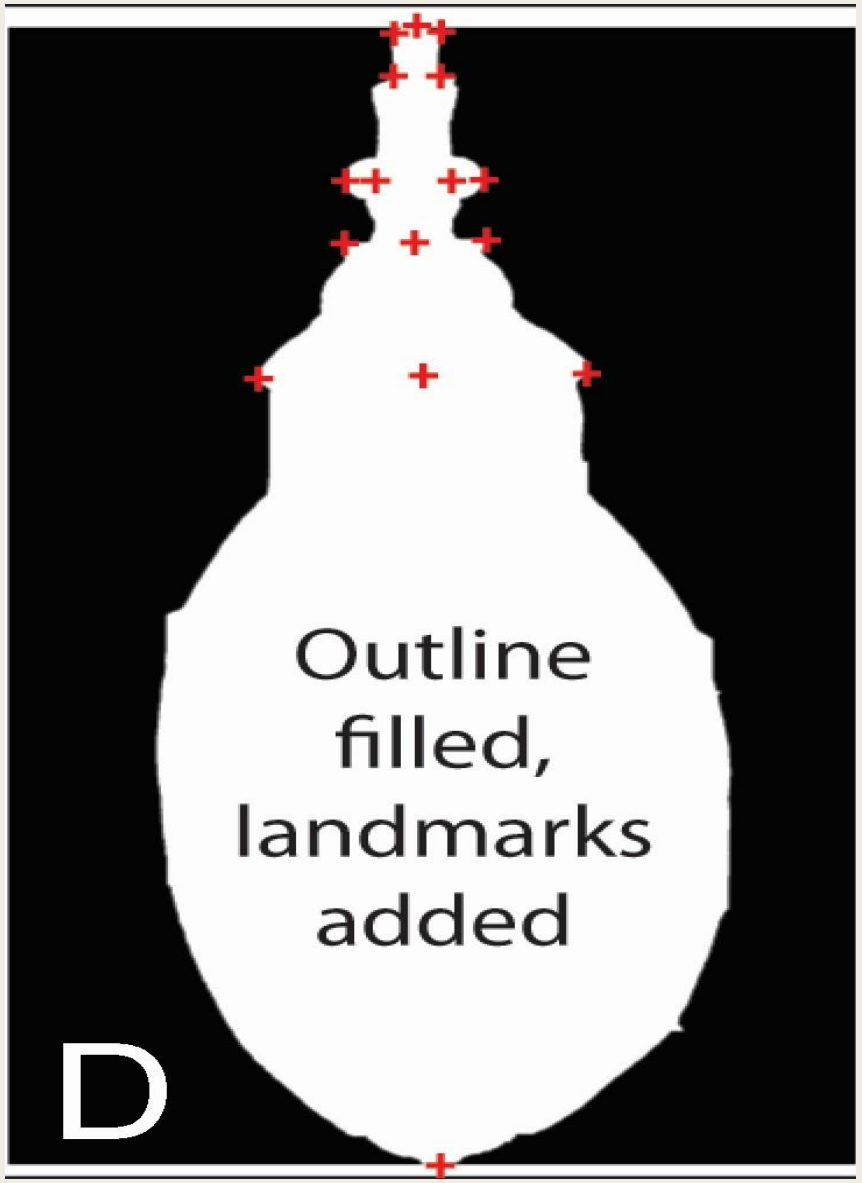

DURACELL

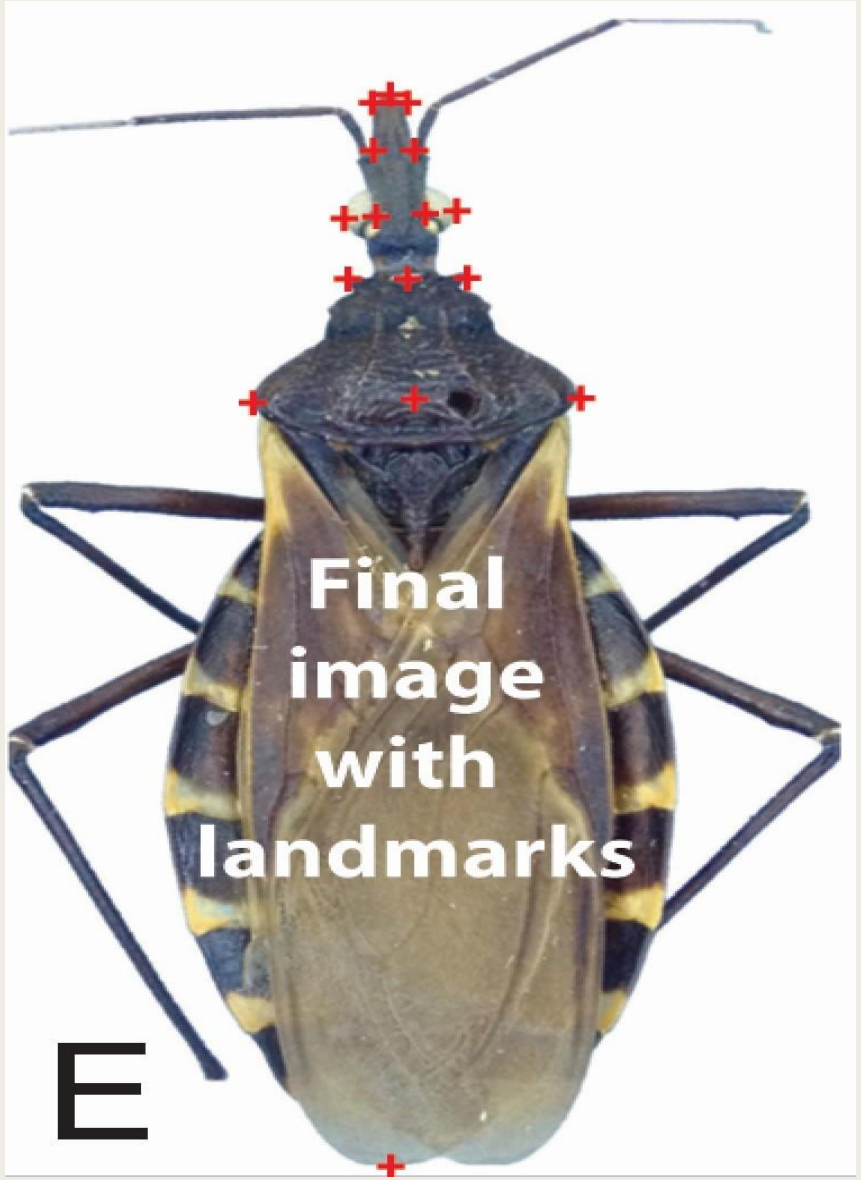



size
standard
Original
image
A


Binary
version
B

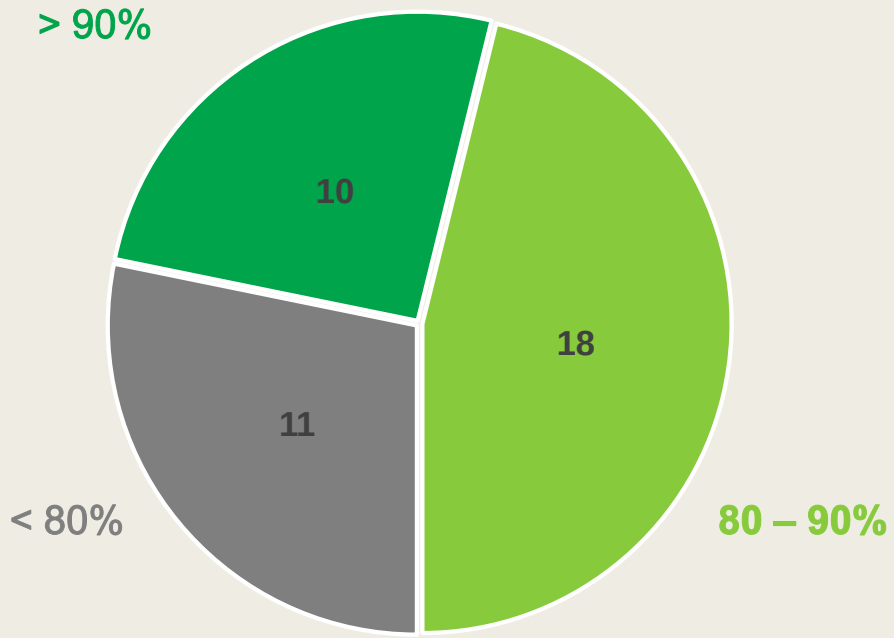

Legs and antennae removed, converted to outline
C


Outline
filled,
landmarks
added
D


Final
image
with
landmarks
E

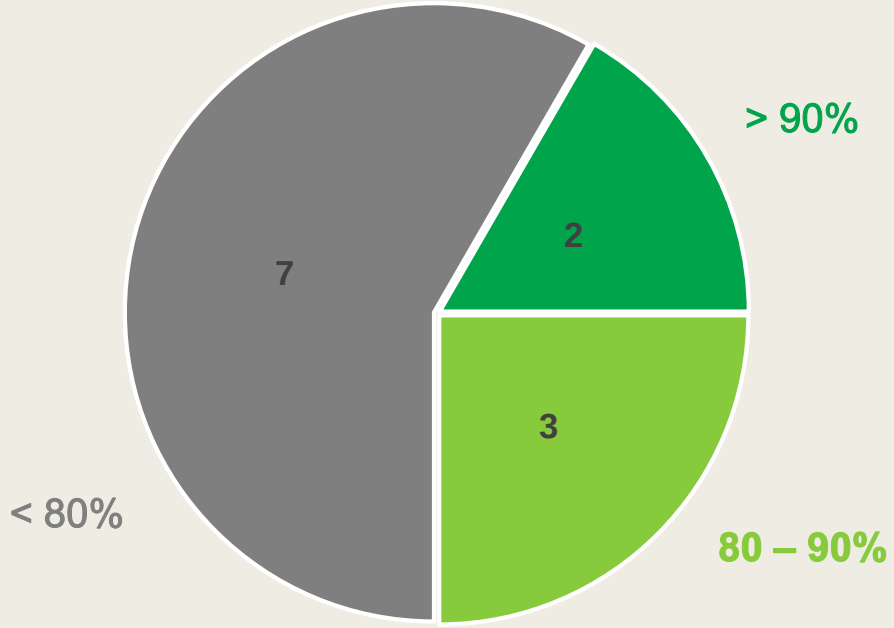
Previous Results

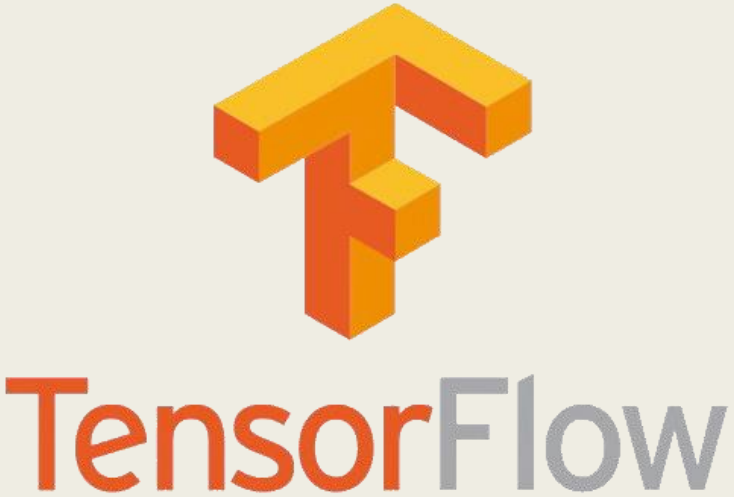
39 Brazilian Species ⟶ Overall Rate: 83.3%



Results achieved using diverse statistical classifiers and simple neural networks

12 Mexican Species ⟶ Overall Rate: 74.9%

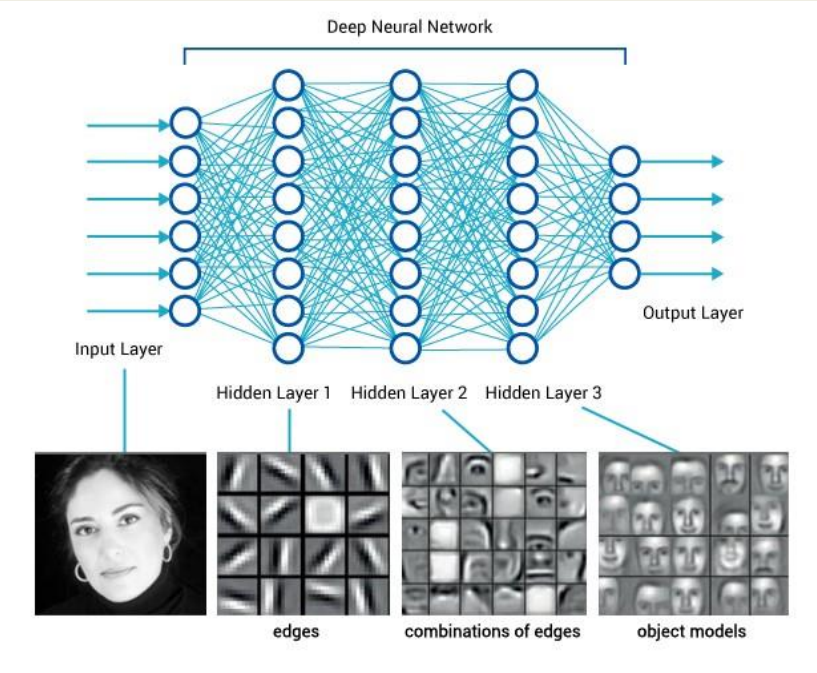



TensorFlow

Google Brain Team

Make machines intelligent. Improve people's lives.

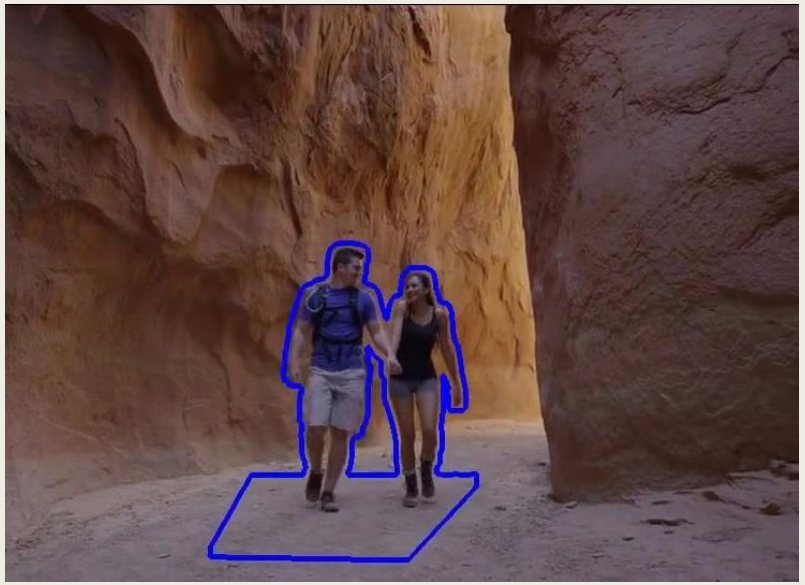
- Open source software
- Released in November 2015
- Provides a platform for deep learning technique



https://goo.gl/images/3vPoVh

Before

https://goo.gl/images/XP6Zsp



https://goo.gl/images/aAWKjE

After

VVL & TENSORFLOW

Previous Results

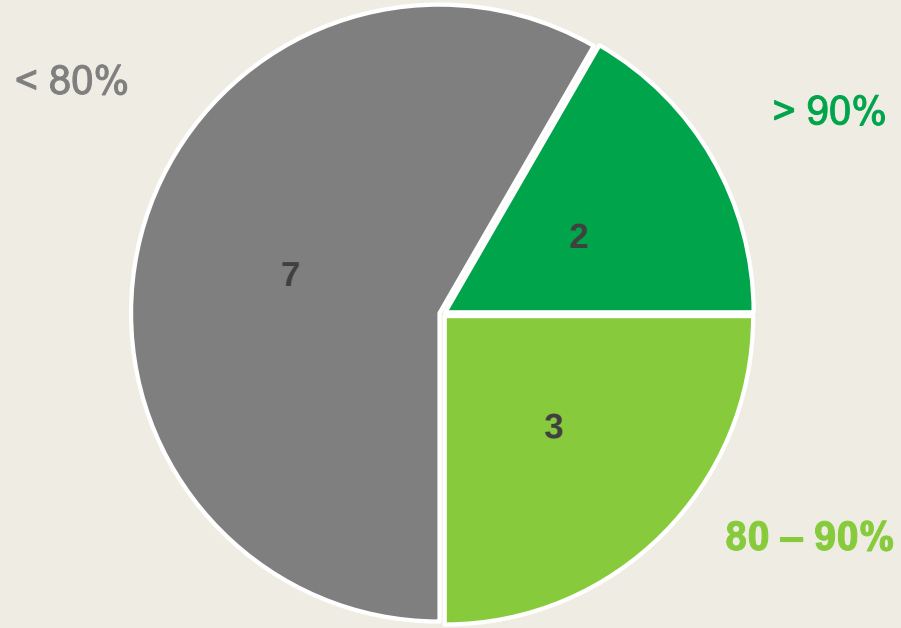
12 Mexican Species ⟶ Overall Rate: 74.9%

TensorFlow

12 Mexican Species ⟶ Overall Rate: 83.0%

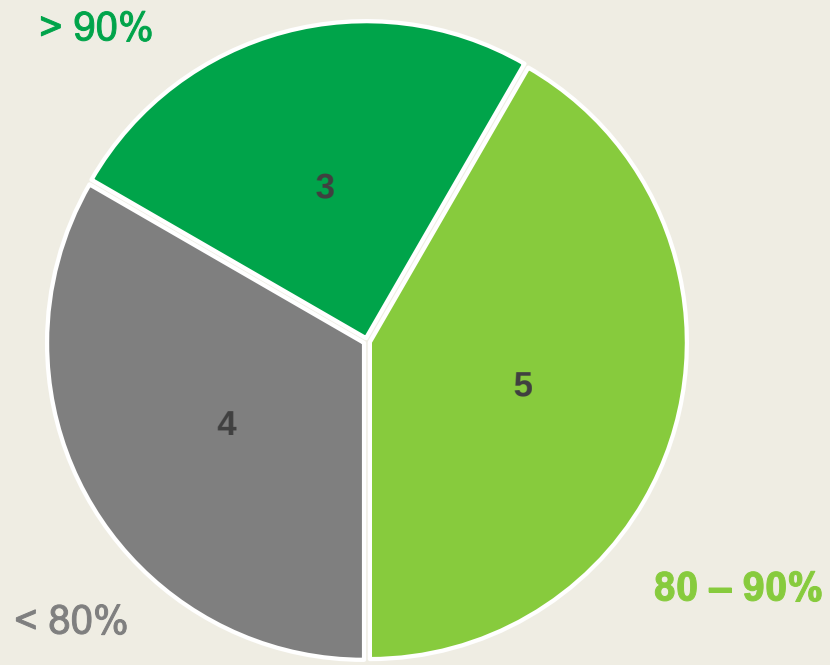
Previous Results

12 Mexican Species ⟶ Overall Rate: 74.9%



TensorFlow

12 Mexican Species ⟶ Overall Rate: 83.0%



Translation: an 8.1% improvement in overall correct ID rate over our previous VVL results

Previous Results

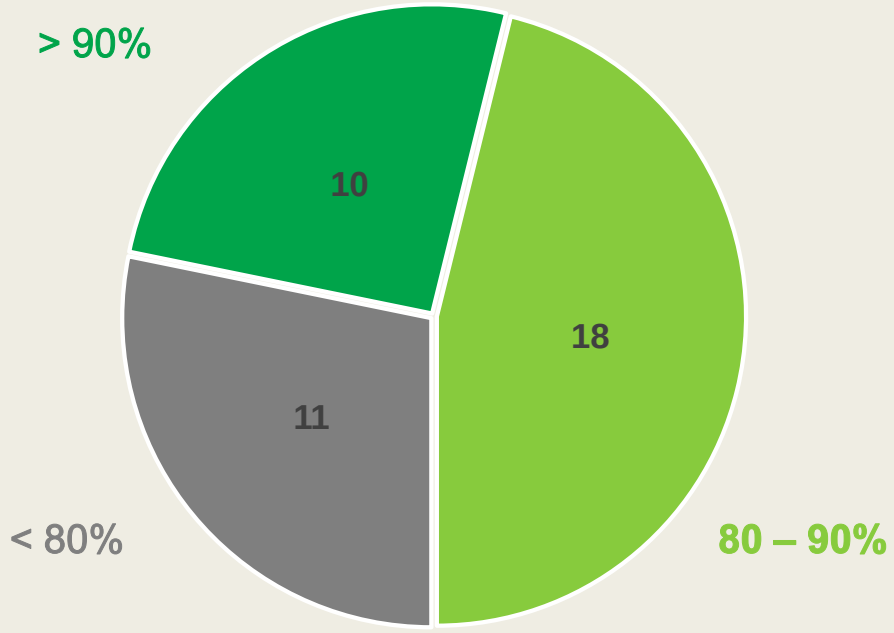
39 Brazilian Species ⟶ Overall Rate: 83.3%

TensorFlow

39 Brazilian Species ⟶ Overall Rate: 86.7%

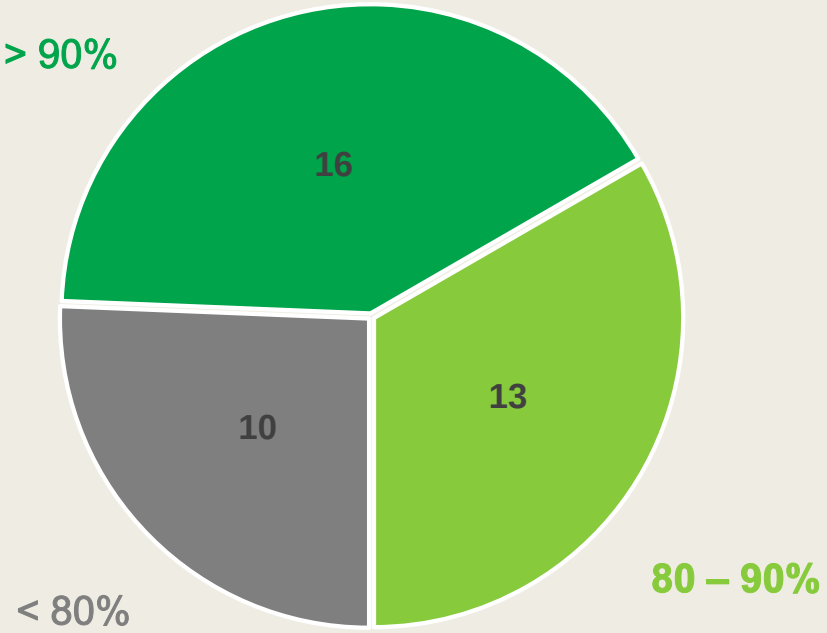
Previous Results

39 Brazilian Species ⟶ Overall Rate: 83.3%

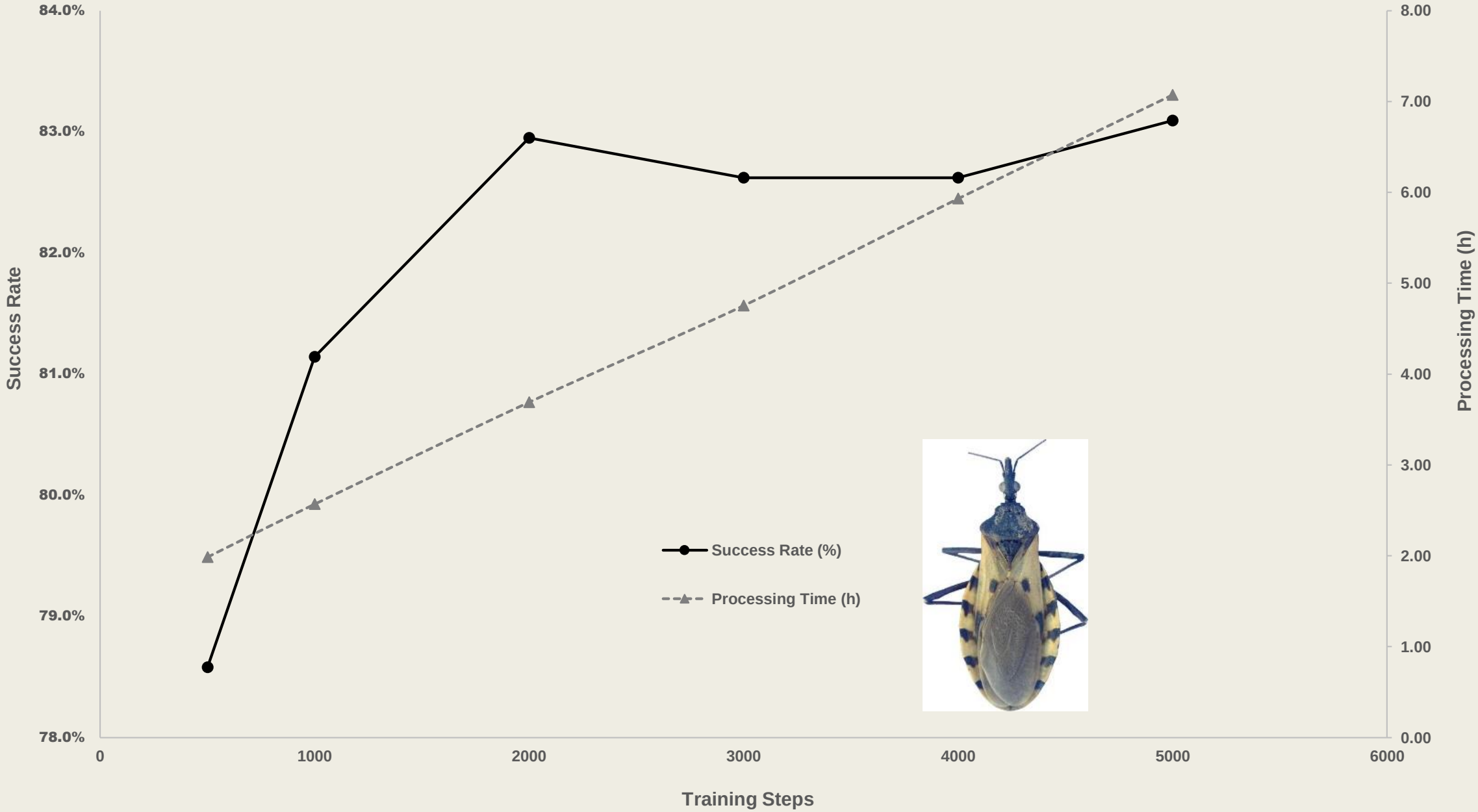


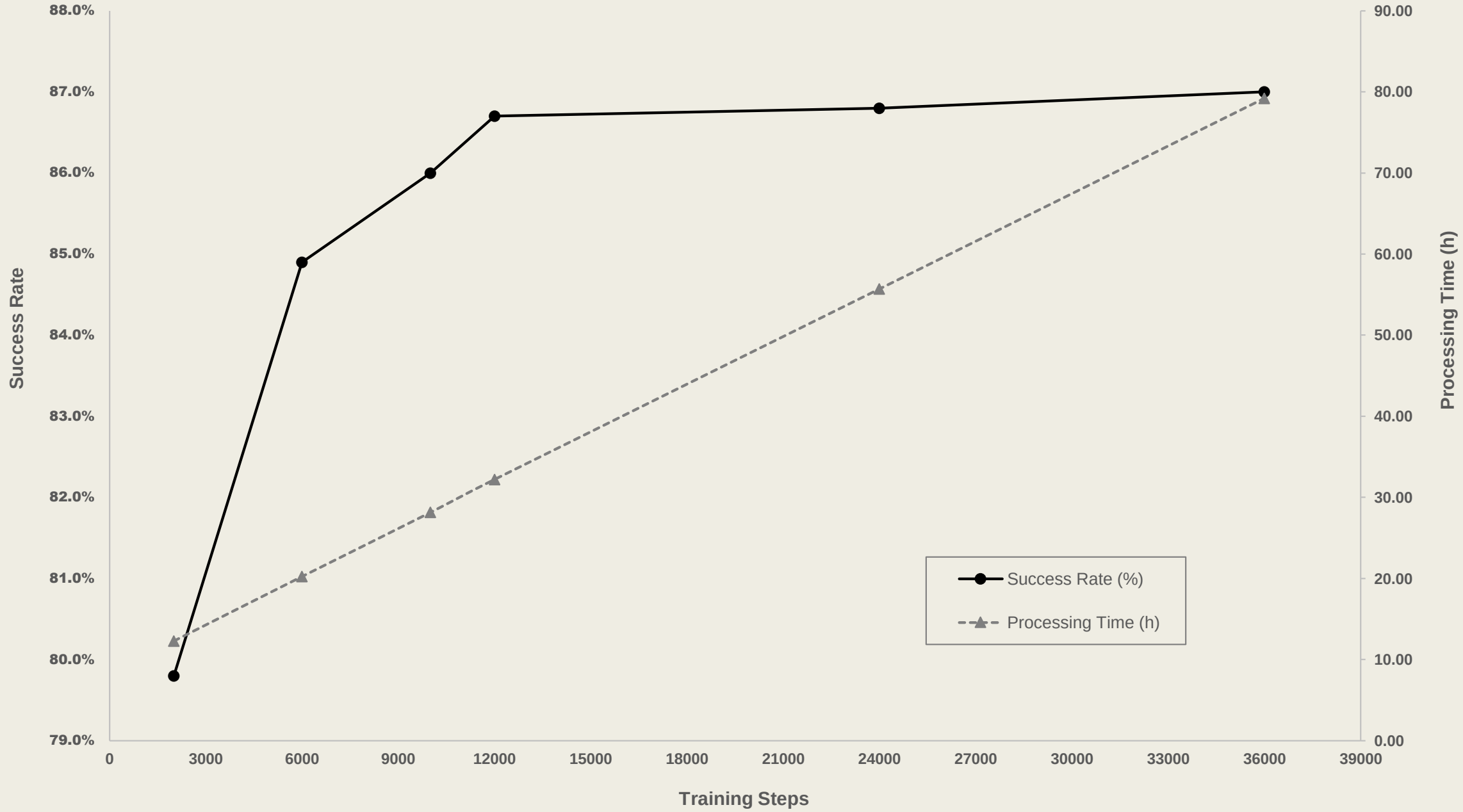
TensorFlow

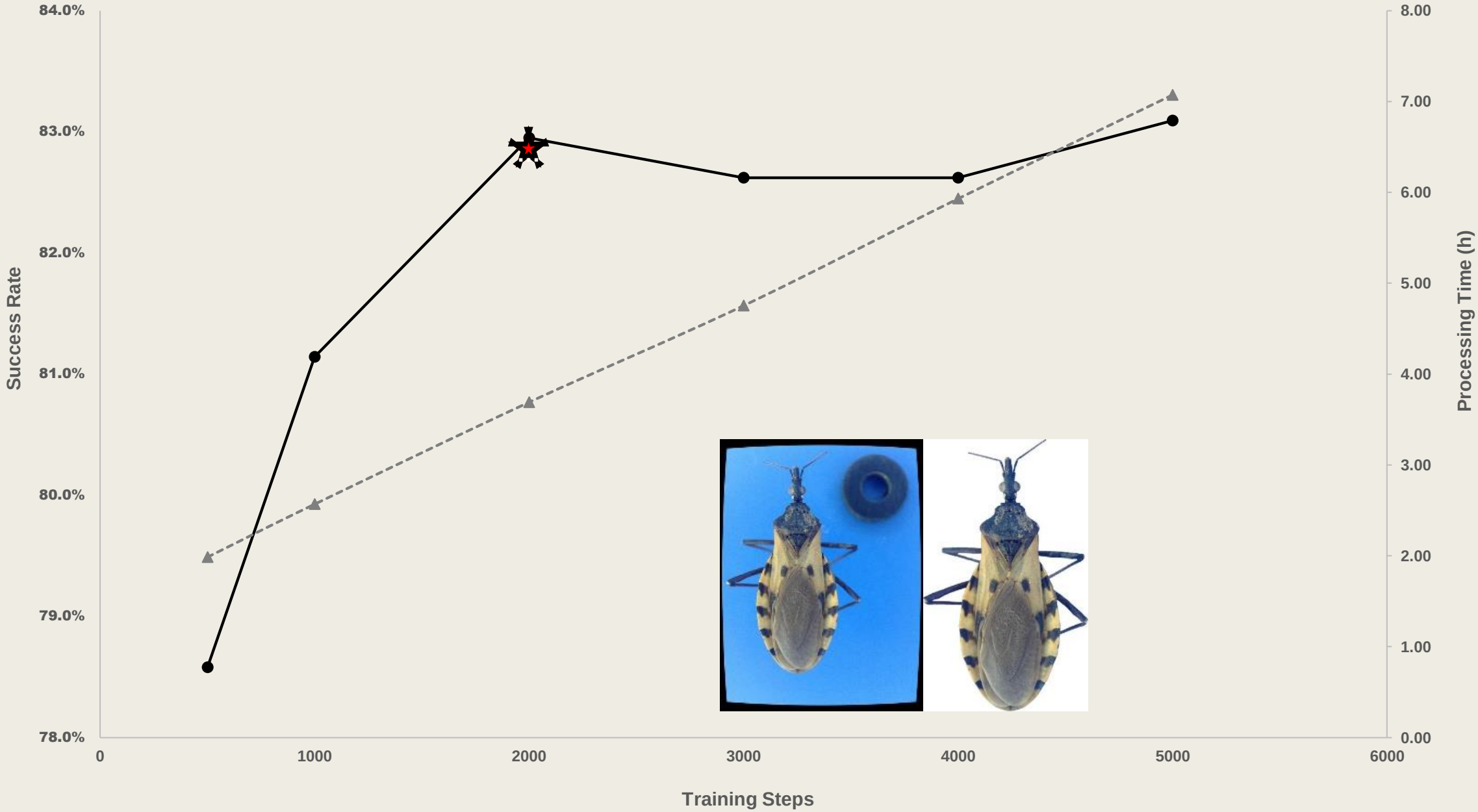
39 Brazilian Species ⟶ Overall Rate: 86.7%



Translation: a 3.4% improvement in overall correct ID rate over our previous VVL results


84.0%
83.0%
82.0%
81.0%
80.0%
79.0%
78.0%
Success Rate
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
Processing Time (h)
0
1000
2000
3000
4000
5000
6000
Training Steps
Success Rate (%)
Processing Time (h)


88.0%
87.0%
86.0%
85.0%
84.0%
83.0%
82.0%
81.0%
80.0%
79.0%
Success Rate
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
Processing Time (h)
0
3000
6000
9000
12000
15000
18000
21000
24000
27000
30000
33000
36000
39000
Training Steps
Success Rate (%)
Processing Time (h)


84.0%
83.0%
82.0%
81.0%
80.0%
79.0%
78.0%
Success Rate
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
Processing Time (h)
0
1000
2000
3000
4000
5000
6000
Training Steps

https://goo.gl/images/Bno76q

Conclusion

- Improved overall classification of triatomine species markedly over our previous results, particularly the total number of species with ID rate of above 80%
- Opened the possibility of using raw images in classification, without a need for pre-processing
- Confidence rate information provided by TensorFlow can inform us about when identification is uncertain
- Technical support from Google will update and improve versions, and offer web-based implementations

Thank you!

a.khalighifar@ku.edu