Classifying Zoonotic Viruses

8/24/23

About Me

James McLaughlin

https://www.linkedin.com/in/james-mclaughlin-wm/

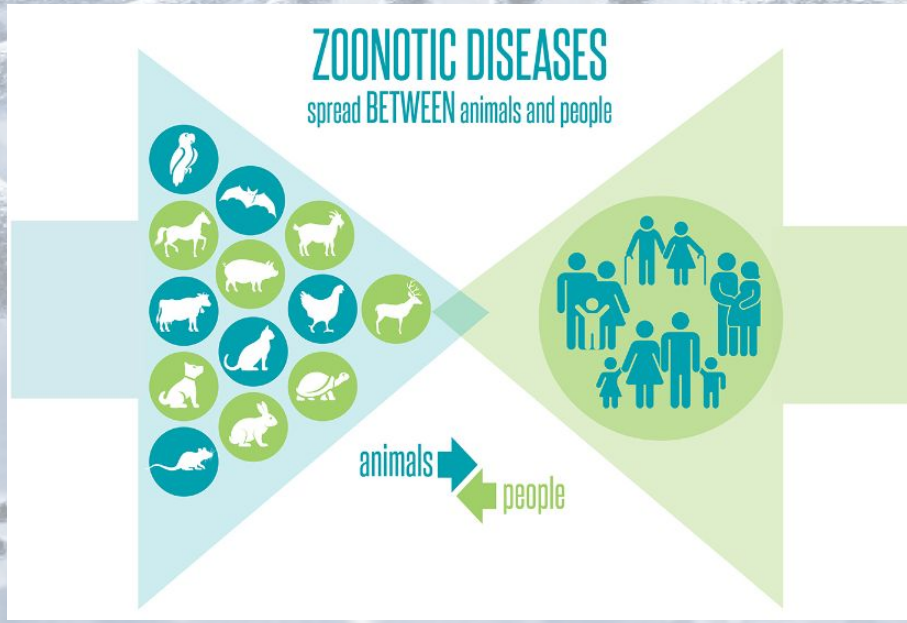
https://github.com/jmclaughlin712

- 3 years experience in multifamily real estate research and leasing
- Bachelor's degree in statistics
- Graduating Flatiron School in August 2023

What are Zoonotic Viruses?

- Caused by germs that spread between animals and people

- Can spread through:
  - Direct contact
  - Indirect contact
  - Vectors
  - Food
  - Water



https://www.cdc.gov/onehealth/basics/zoonotic-diseases.html

Bottom Line

- **Model can correctly identify associated animal about 95% of the time**
- **Quick way to classify a large number of genetic sequences**
- **Useful supplement to BLAST**

Why it Matters

- Key for vaccine development and preventing future outbreaks
- 3 out of 4 emerging infectious diseases come from animals
- Identifying the animal origins of COVID is an ongoing search

https://www.theatlantic.com/science/archive/2023/03/covid-origins-research-raccoon-dogs-wuhan-market-lab-leak/673390/

Data Overview

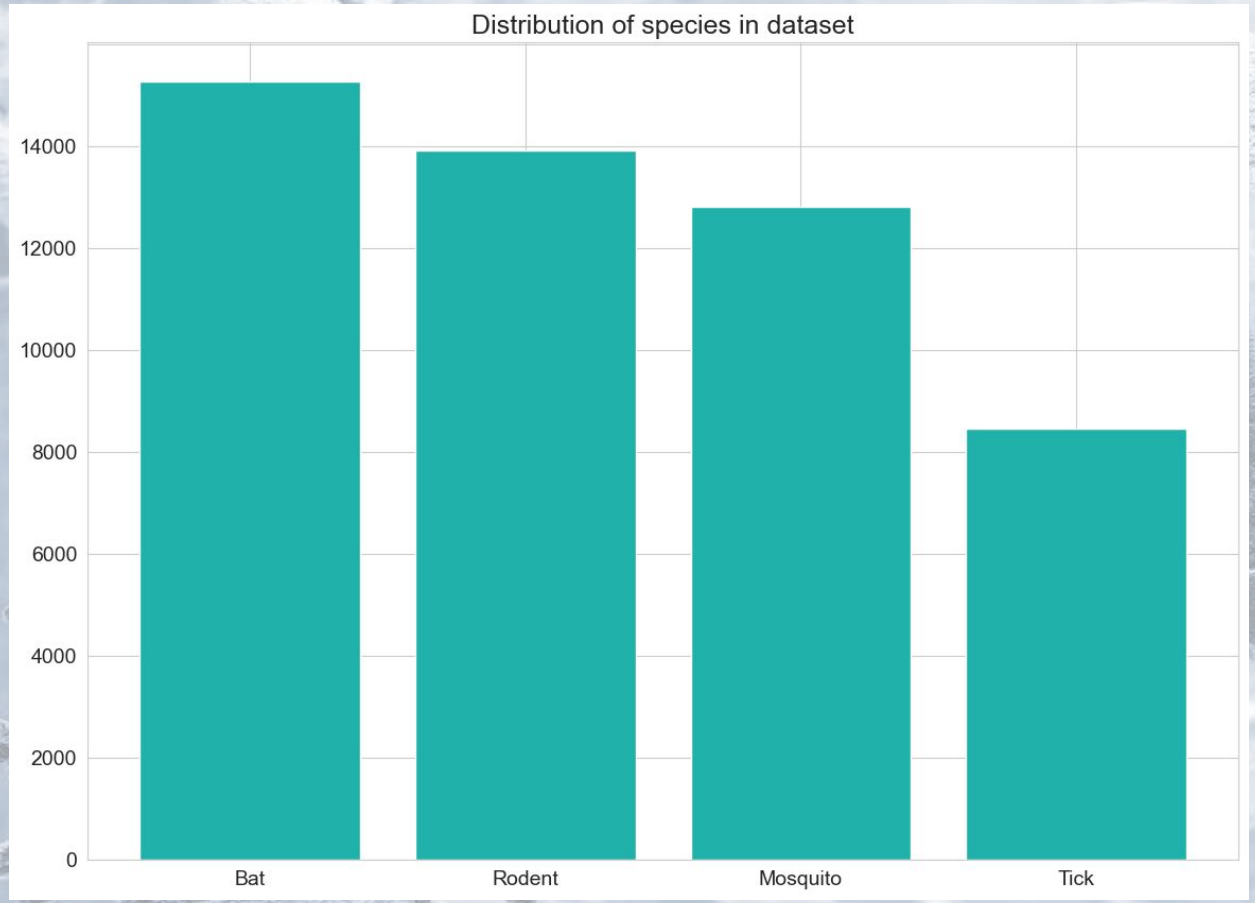
- Data comes from ZOVER
- ~50,000 unique viruses associated with bats, rodents, mosquitos, and ticks
- DNA and protein sequences




Data Limitations

- Incomplete sequences
- No numeric data
- Limited to 4 animals

Data Contained More Mammals than Insects



Where Did the Samples Come From?

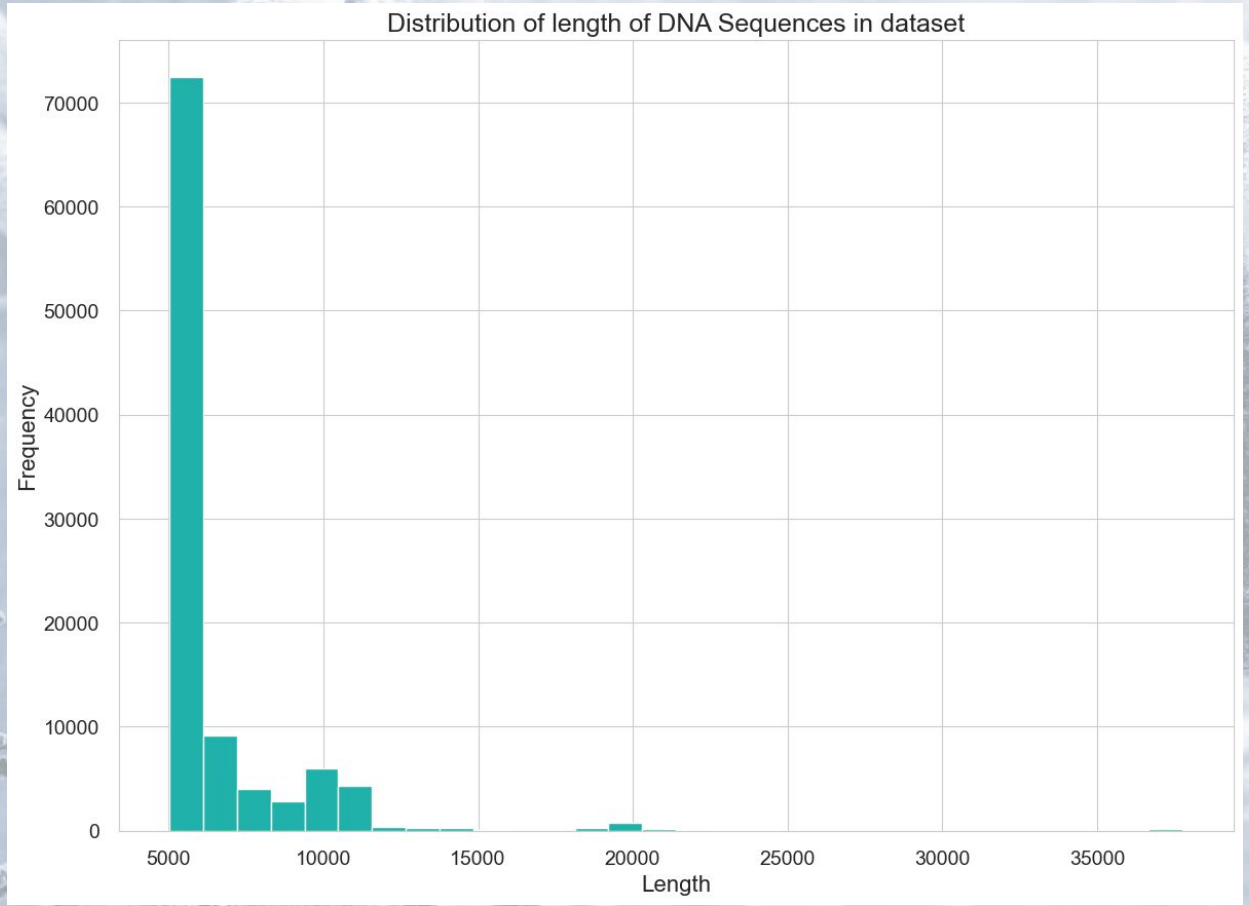
Bats

Rodents

Mosquitos

Ticks

Most DNA Sequences are Under 10000 Characters



Model

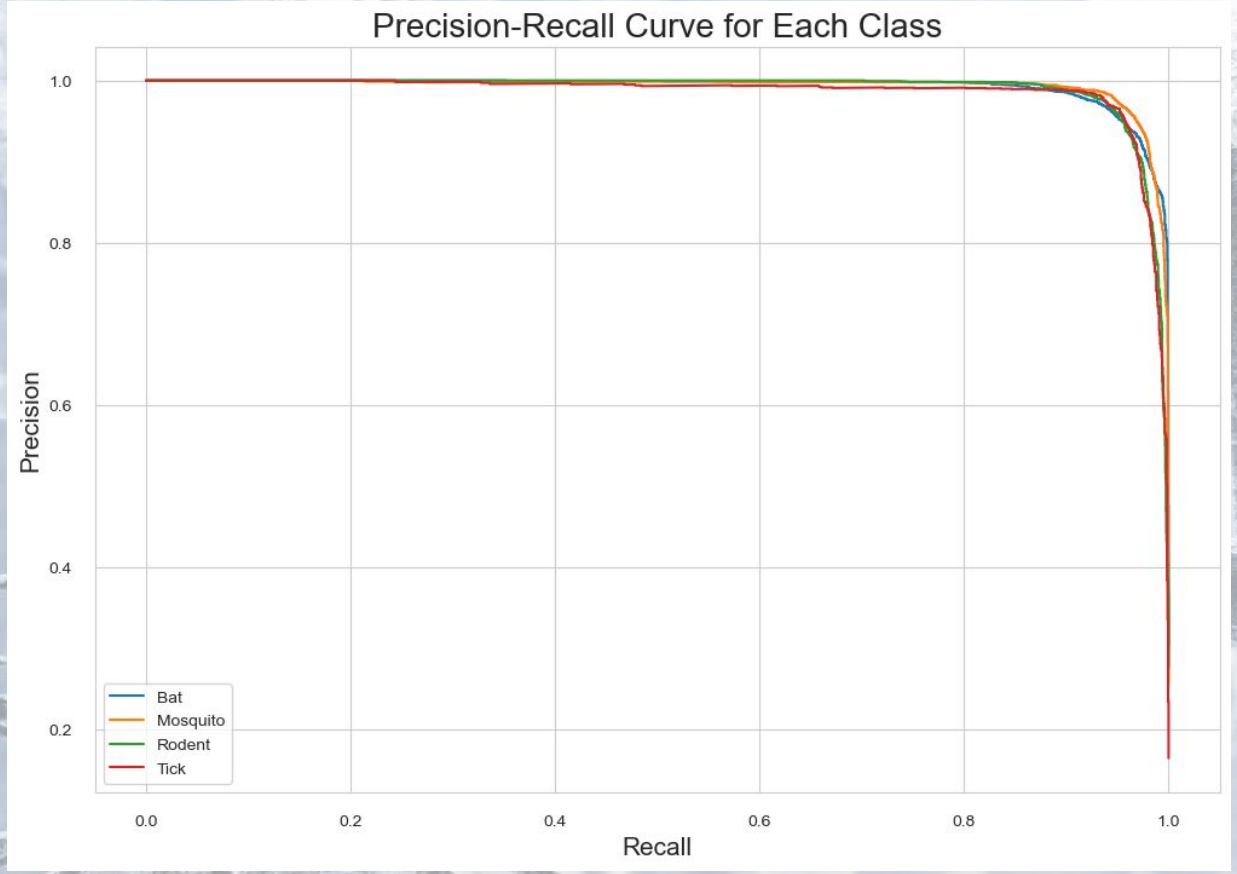
- **Voting classifier model that combines XGBoost and Neural Networks**
- **XGBoost has higher accuracy, less overfitting**
- **Neural Networks identify hidden features in data**

Model Metrics

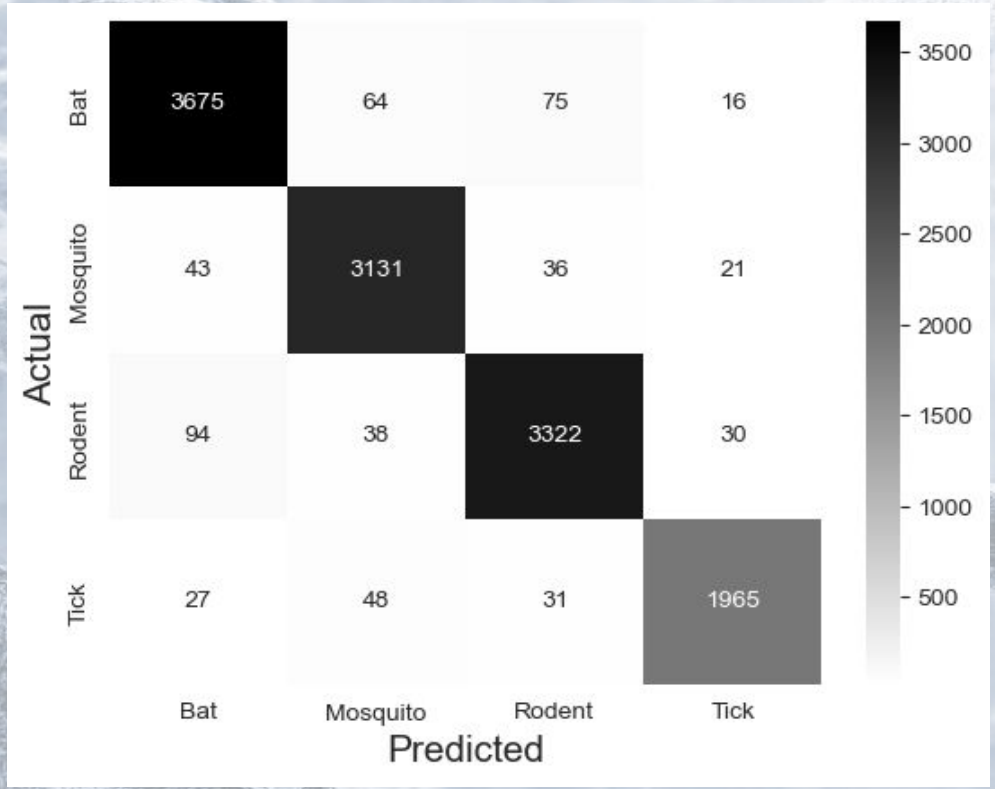
Accuracy

95.4%

Model Metrics



Model Metrics



Questions?

James McLaughlin

LinkedIn

GitHub


THANK YOU