UVA DATA SCIENCE

Detecting Illegal Fishing with Automatic Identification Systems and Machine Learning

Samuel Brown, Danielle Katz,
Dana Korotovskikh, and Stephen Kullman

1 May 2024

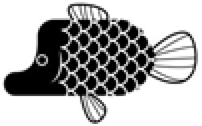
Table of Contents

Stakeholders

Sponsor: GA-CCRi

- Tara Valladares
- Rebecca DeSipio

Mentor: Heman Shakeri

Introduction 

Illegal, Unreported, and Unregulated (IUU) Fishing

- Widely defined – fishing outside of local or international regulations
- Critical issue impacting health of marine ecosystem and global security
- Estimated to cost $10-23 billion annually

Automatic Identification Systems (AIS)

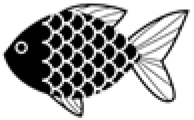
- Standardized tracking systems outfitted on all vessels
- Continuously transmits user data
- Utilized for maritime monitoring and collision avoidance
- Very available and easy to obtain

Motivation

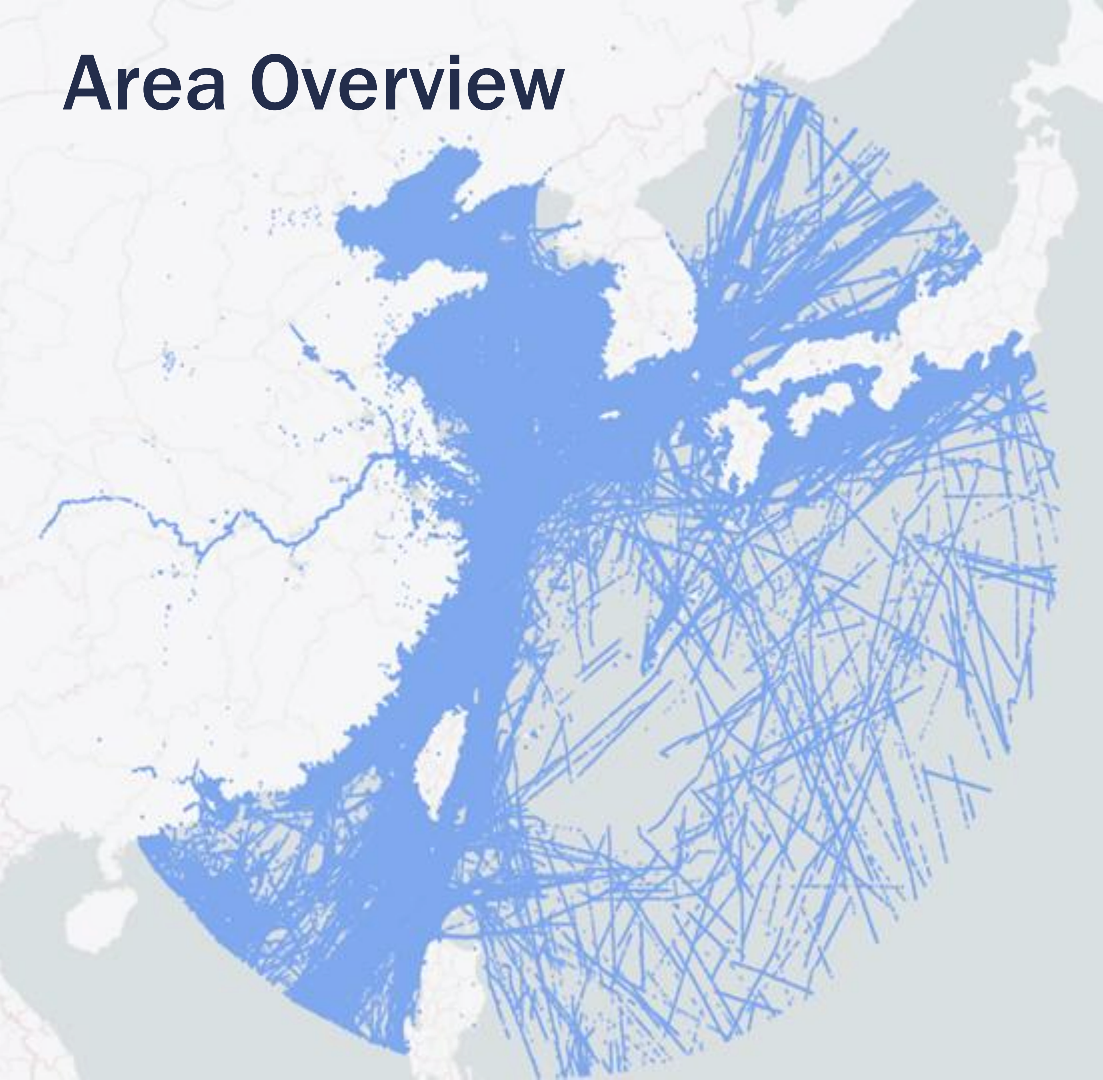
- AIS devices are commonly illegally exploited on fishing nets and buoys to protect large hauls
- Increased need to reform AIS regulations

Goal:

Explore how AIS data can be used to aid in IUU fishing net detection

Data 

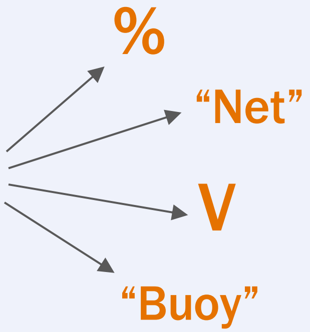
Area Overview



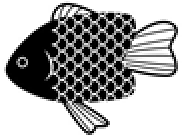
- Region of interest:
  - Southeast Asia
- Spatial, temporal, and user inputted data
- Training set
  - Around 4 days
  - September 1→5 2023
- Test set
  - Around 4 days
  - October 12→16 2023

Assumptions and Limitations

- Certain naming or positional conventions are suggestive of fishing nets



- Computational limitations led to restricted region of interest and timeframe

Methodology 

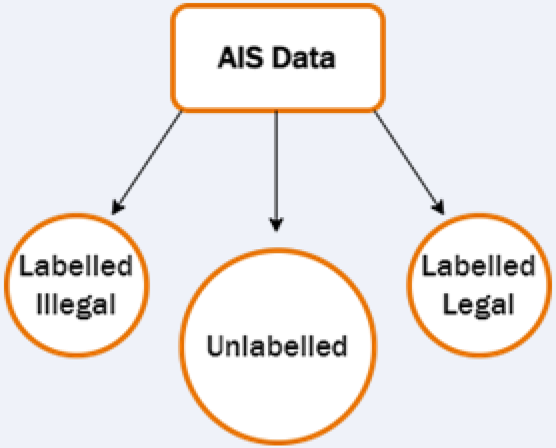
Pre-Processing

- Aggregate data by distinct device ‘trips’
  - Unbroken period of AIS transmission
  - Included scaled parameters for speed, heading, and positioning
- ***Red flag***: Indicators of non-standard naming conventions and movement
  - Score from 0→4
  - Used for modelling analysis

net_name	Names including a ‘V’, ‘%’, ’buoy’, or ’net’
mmsi_length	MMSI values not equal to 9 digits
spawn_offshore	Vessels whose first transmission is offshore (1 nautical mile off coastline)
spoof	Devices with unreasonably high calculated speeds (≥ 150 knots)

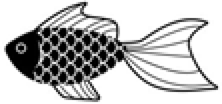
Pseudo-Labelling

- AIS dataset is unlabelled
  - Select models utilized pseudo-labels for training
- Pseudo-labelled confident points based on ‘red flag’ conditions
  - Illegal: Bad *net_name* and ≥ 3 total red flags
  - Legal: No red flags



Approaches

1. Unsupervised clustering
   - Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)
2. Semi-supervised classification
   - eXtreme Gradient Boosted trees (XGBoost) and Artificial Neural Network (ANN)
3. Supervised classification
   - ANN

Models 

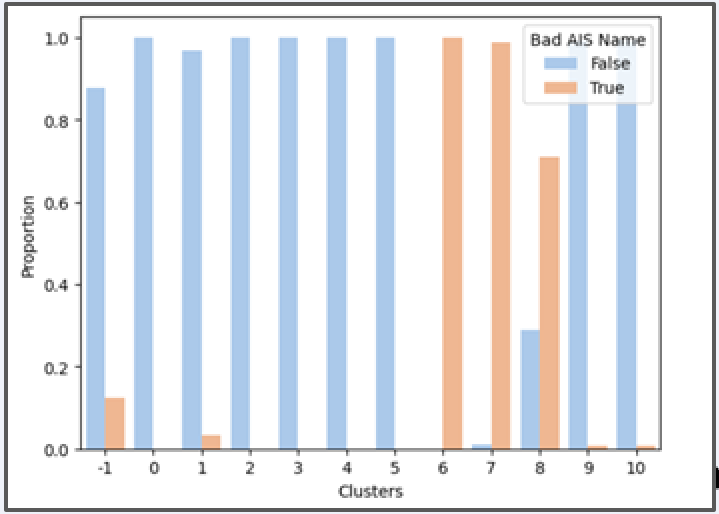
Unsupervised Clustering

- Treats dataset as truly unlabelled
- Clustering with HDBSCAN
  - 2 clustering iterations
- After first round new score is built:
  - Aggregates mean red flag score of each cluster onto respective observations

Unsupervised Clustering Results

- 3 primary clusters identified
  - Reflect 36.2% of total trips in dataset
  - Includes 1% of observations with valid *net_name*
- The developed features may be valid indicators of IUU activity that cannot be detected with naming conventions alone

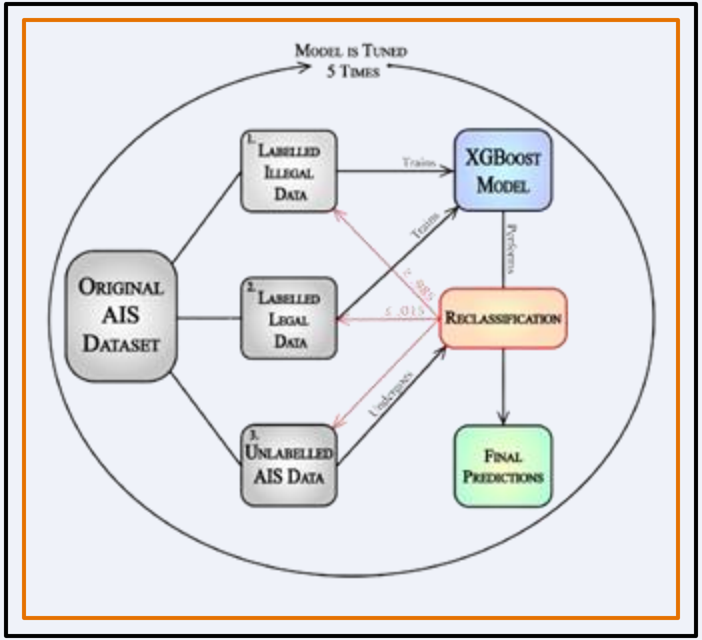
Cluster Results Analyzed by net_name



Semi-Supervised Classification

- Small set of confident pseudo-labelled points, other points are left unlabelled
- Model trained on sample of labelled AIS data
- Iteratively classifies unlabelled dataset if confident and retrains on updated labelled dataset

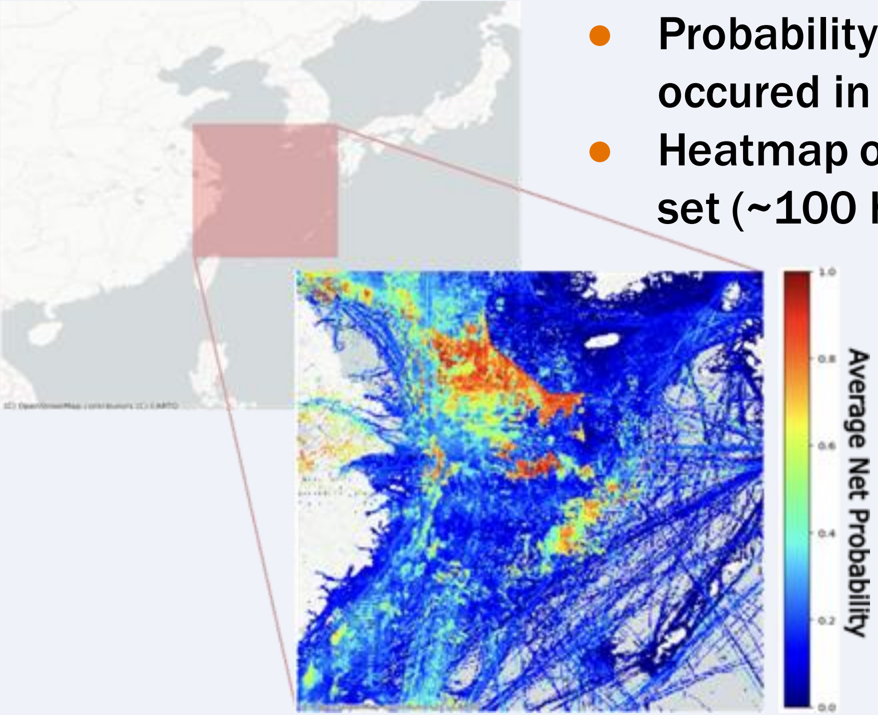
Workflow for XGBoost & ANN



Supervised Classification with ANN

- Uses pseudo-labelled training data once to build model
- Simple ANN predicts binary classification for each AIS trip
    - Same 4-layer ANN framework used for both semi and fully supervised models

Semi-Supervised and Supervised Classification Results



- Probability of 1 indicates illegal likely to have occured in region
- Heatmap of XGBoost prediction results on test set (~100 hours)

Model	Test Accuracy	TPR	TNR
**Semi-supervised XGBoost	0.891	0.915	0.848
Semi-supervised ANN	0.879	0.900	0.842
Supervised ANN	0.867	0.907	0.801

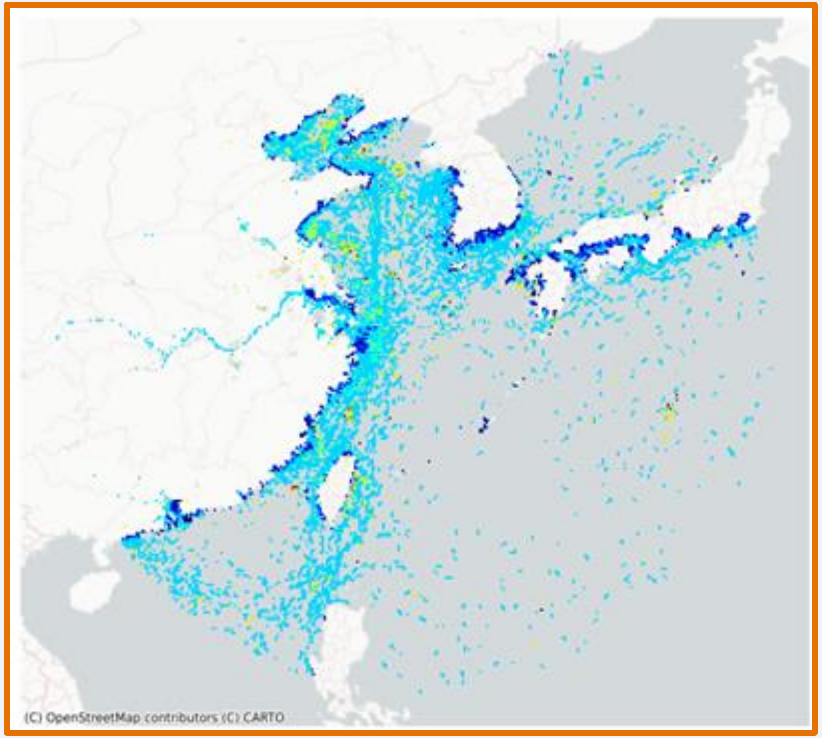
**XGBoost model was the top performer with test set

Regional Analysis

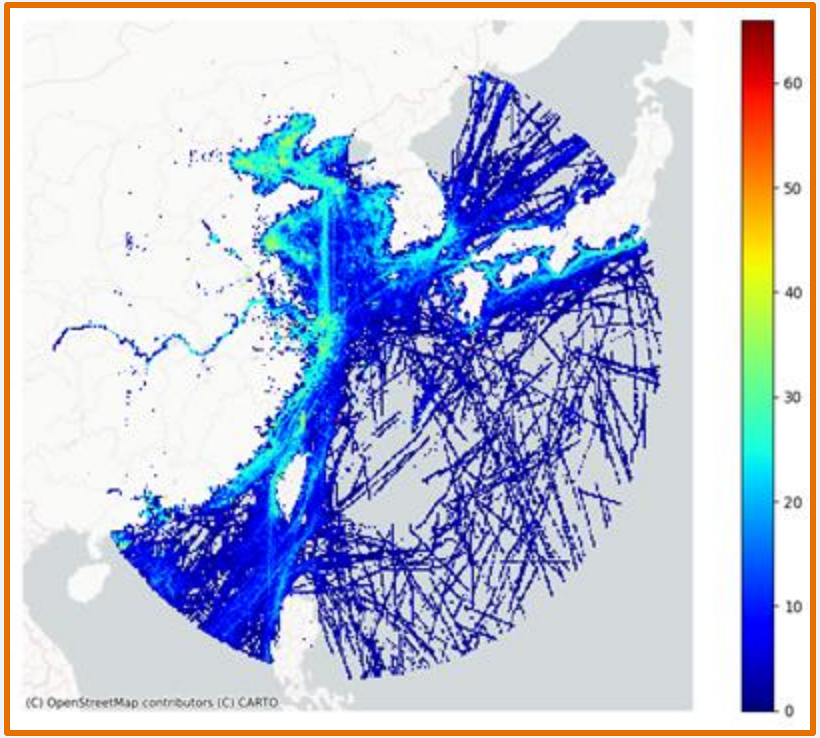
- Divided region into .1°x.1° cells
  - 36 mi^2
- ***hot_score:*** number of unique red flags divided by count of unique vessels
- Position each AIS signal within its corresponding grid cell
- Calculate total number of ‘red flags’ per unique vessel for each cell
  - Aggregate each hours’ score per cell to see which areas showed most signs of illegal activity per day

Regional Analysis Visual Results

September 1st, 1 hour



September 1st, 24 hours



UVA DATA SCIENCE

Acknowledgements

Special thanks to Tara Vallardes, Rebecca DeSipio, and Heman Shakeri for their expertise throughout! We could not have accomplished this without their help. ❤

Capstone Team

- Samuel Brown
- Danielle Katz
- Dana Korotovskikh
- Stephen Kullman