Ataskaitinė informatikos inžinerijos krypties doktorantų konferencija

Ataskaita už 2019 – 2020 mokslo metus

Doktorantas: Julius Venskus

Vilnius, 2020

Bendra informacija

- **Disertacijos pavadinimas:** „Giliojo mašininio mokymo taikymas jūrų transporto eismo duomenims klasifikuoti“
- **Darbo vadovas:** dr. Povilas Treigys
- **Konsultantas:** dr. Arūnas Andziulis
- **Doktorantūros pradžia:** 2016 m.
- **Planuojama doktorantūros pabaiga:** 2020 m.

Informacija apie tyrimą

- **Tyrimo objektas:**
  - jūrų transporto priemonių eismo duomenys
- **Tyrimo tikslas:**
  - ištirti ir pasiūlyti būdus jūrų transporto duomenų klasifikavimui ir neįprasto eismo aptikimui.
- **Tyrimo uždaviniai:**
  - apžvelgti esamus jūrų transporto klasifikavimo ir neįprasto eismo aptikimo būdus AIS (automatinės identifikavimo sistemos) duomenyse;
  - ištirti klasikinius jūrų transporto klasifikavimo ir neįprasto eismo aptikimo AIS duomenyse būdus;
  - ištirti ir pritaikyti giliuoju mokymusi grindžiamus algoritmus transporto duomenims klasifikuoti;
  - atlikti eksperimentinį klasifikatorių vertinimą.
- **Planuojami rezultatai:**
  - jūrų transporto eismo klasifikavimo ir neįprasto eismo aptikimo algoritmus įgyvendinantis programų sistemos prototipas.

2019/2020 darbo planas

STUDIJŲ PLANAS

- **Mokslinių tyrimų publikavimas:**
  - Publikuoti bent vieną straipsnį, apibendrinantį disertacijos turinį, citavimo indeksą, turinčiame periodiniame, recenzuojamame leidinyje.

- **Dalyvavimas konferencijose, seminaruose, kitose doktorantų mobilumo veiklose:**
  - Disertacijos rezultatų pristatymas bent dviejų kitų universitetų kamieninių padalinių moksliniuose seminaruose.

2019/2020 darbo ataskaita (1)

- **Ataskaita už 2019/2020 mokslo metus**
  - Indeksuojami WOS periodiniai leidiniai:
    - J. Venskus, P. Treigys, and J. Markevičiūtė. “**Unsupervised Marine Vessel Trajectory Prediction using LSTM Network and Wild Bootstrapping Techniques**”. *Nonlinear Analysis: Modelling and Control*. (2020). Vilnius University. PRIIMTAS.
  - Seminarai:
    - Pranešimas moksliniame seminare. **Unsupervised Marine Vessel Trajectory Prediction using LSTM Network and Wild Bootstrapping Techniques**. Klaipėdos universitetas. Informatikos ir statistikos katedra.
    - Pranešimas moksliniame seminare. **Meteorologinių duomenų įtaka nustatyti trūkstamai informacijai apie jūrų laivo tipą, naudojant daugiasluoksnius LSTM neuroninius tinklus.** Vilnius University Institute of Data Science and Digital Technologies.

2019/2020 darbo ataskaita (2)

- Kitos konferencijos
  - Venskus, J.; Treigys, P.; Bernatavičienė, J.; Markevičiūtė, J. **Detecting Maritime traffic anomalies with long-short term memory recurrent neural network** // 11th international workshop on data analysis methods for software systems (DAMSS 2019), Druskininkai, Lithuania, November 28-30, 2019 / Lithuanian Computer Society, Vilnius University Institute of Data Science and Digital Technologies, Lithuanian Academy of Sciences. Vilnius : Vilnius University Press, 2019. ISBN 9786090703243. eISBN 9786090703250. p. 89. DOI: 10.15388/Proceedings.2019.8.

- Disertacijos rengimas.
  - Daktaro disertacijos juodraštis anglų kalba.

- Nuolatinis mokymasis
  - Stanford University. XCS224N, Natural Language Processing with Deep Learning. 10 CEU(s)

Už 2016/2020 mokslo metus (1)

- **Mokslinių tyrimų publikavimas:**
- **Indeksuojami WOS periodiniai leidiniai**
  - Venskus, Julius; Treigys, Povilas; Markevičiūtė, Jurgita. "**Unsupervised Marine Vessel Trajectory Prediction using LSTM Network and Wild Bootstrapping Techniques**". *Nonlinear Analysis: Modelling and Control*. (2020). Vilnius University. PRIIMTAS.
  - Venskus, Julius; Treigys, Povilas; Bernatavičienė, Jolita; Tamulevičius, Gintautas; Medvedev, Viktor. **Real-time maritime traffic anomaly detection based on sensors and history data embedding //** Sensors. Basel : MDPI. ISSN 1424-8220. 2019, vol. 19, no. 17, art. no. 3782, p. 1-10. DOI: 10.3390/s19173782. **Q1**
  - Venskus, Julius; Treigys, Povilas; Bernatavičienė, Jolita; Medvedev, Viktor; Voznak, Miroslav; Kurmis, Mindaugas; Bulbenkienė, Violeta. **Integration of a self-organizing map and a virtual pheromone for real-time abnormal movement detection in marine traffic** // *Informatica*. Vilnius : Vilniaus universiteto Matematikos ir informatikos institutas. ISSN 0868-4952. 2017, Vol. 28, No. 2, p. 359-374.

Už 2016/2020 mokslo metus (2)

- **Mokslinių tyrimų publikavimas:**
- **Recenzuojami kiti leidiniai**
  - Venskus, Julius; Treigys. **Meteorological Data Influence on Missing Vessel Type Detection Using Deep Multi-Stacked LSTM Neural Network** // XII International Conference Computer Data Analysis & Modeling 2019 Stochastics & Data Science. Proceedings of the XII International Conference. September, 18-22, Belarusian State University, 2019 Minsk, Belarus, ISBN-978-985-566-811-5
  - Venskus, Julius; Treigys. **Preparation of training data by filling in missing vessel type data using deep multi-stacked LSTM neural network for abnormal marine transport evaluation**. Proceedings of Abstracts. ITISE 2019 International Conference on Time Series and Forecasting. Granada, Spain, September, 25-27, 2019, ISBN 978-84-17970-79-6

Už 2016/2020 mokslo metus (3)

- **Pranešimai tarptautiniuose konferencijose**
- **Žodiniai:**
  - Venskus, Julius; Treigys. Meteorological Data Influence on Missing Vessel Type Detection Using Deep Multi-Stacked LSTM Neural Network // XII International Conference Computer Data Analysis & Modeling 2019 Stochastics & Data Science. September, 18-22, Belarusian State University, 2019 Minsk, Belarus
- **Stendiniai:**
  - Venskus, Julius; Treigys. Preparation of training data by filling in missing vessel type data using deep multi-stacked LSTM neural network for abnormal marine transport evaluation. ITISE 2019 International Conference on Time Series and Forecasting. Granada, Spain, September, 25-27, 2019
  - Venskus, Julius; Treigys, Povilas; Bernatavičienė, Jolita; Retraining strategies of modified SOM for abnormal marine traffic detection; Materials, Methods & Technologies 2018 : 20th International conference. Elenite, Bulgaria, June 26-30, 2018. International Scientific Events

Už 2016/2020 mokslo metus (4)

- **Pranešimai kitose konferencijose**
- **Stendiniai:**
  - Venskus, Julius; Treigys, Povilas; Bernatavičienė, Jolita; Markevičiūtė, Jurgita. **Detecting Maritime traffic anomalies with long-short term memory recurrent neural network** // 11th international workshop on data analysis methods for software systems (**DAMSS 2019**), Druskininkai, Lithuania, November 28-30, 2019 / Lithuanian Computer Society, Vilnius University Institute of Data Science and Digital Technologies, Lithuanian Academy of Sciences. Vilnius : Vilnius University Press, 2019. ISBN 9786090703243. eISBN 9786090703250. p. 89. DOI: 10.15388/Proceedings.2019.8.
  - Venskus, Julius; Treigys, Povilas; Bernatavičienė, Jolita; Andziulis, Arūnas. **Aspects of data collection for abnormal marine transport evaluation** // **DAMSS 2018** : 10th international workshop on "Data analysis methods for software systems", Druskininkai, Lithuania, November 29 - December 1, 2018 : [abstract book]. Vilnius : Vilniaus universitetas, 2018. ISBN 9786090700433. p. 88. Prieiga per internetą: <https://www.mii.lt/datamss/files/DAMSS_2018_1.pdf>.
  - Venskus, Julius; Treigys, Povilas; Bernatavičienė, Jolita; Medvedev, Viktor. **Retraining strategies of modified SOM for abnormal marine traffic detection** // 9th International workshop on Data Analysis Methods for Software Systems (**DAMSS 2017**), Druskininkai, Lithuania, November 30 - December 2, 2017. Vilnius : VIlniaus universitetas, 2017. ISBN 9789986680642. p. 54. Prieiga per internetą: <https://www.mii.lt/datamss/files/liks_mii_drusk_2017.pdf>.
  - Venskus, Julius; Kurmis, Mindaugas; Treigys, Povilas. **Modified SOM for abnormal marine traffic detection** // Data analysis methods for software systems : 8th international workshop on data analysis methods for software systems (**DAMSS 2016**), , Druskininkai, December 1-3, 2016. Vilnius : Vilniaus universiteto leidykla, 2016. ISBN 9789986680611. p. 66-67. Prieiga per internetą: <https://www.mii.lt/datamss/files/liks_mii_drusk_2016_abstract_el_v_.pdf>.

Už 2016/2020 mokslo metus (4)

- **Dalyvavimas seminaruose, kitose doktorantų mobilumo veiklose:**
  - **BigSkyEarth Training School 2018:** GPU-based analytics and data science. COST Action: TD1403, Training School Title: Third training school, Venue: Centro De Tecnologías De Interacción Visual Y Comunicaciones, Vicomtech, San Sebastian, Spain, Training School Dates: 2018-04-03 09:00:00 - 2018-04-09 18:00:00
  - Pranešimai: VU Duomenų mokslo ir skaitmeninių technologijų instituto pirmadienio seminaruose.

Ataskaitinė informatikos inžinerijos krypties doktorantų konferencija

Mokslinė ataskaita

Doktorantas: Julius Venskus
Vadovas: dr. Povilas Treigys
Konsultantas: prof. dr. Arūnas Andziulis

Vilnius, 2020

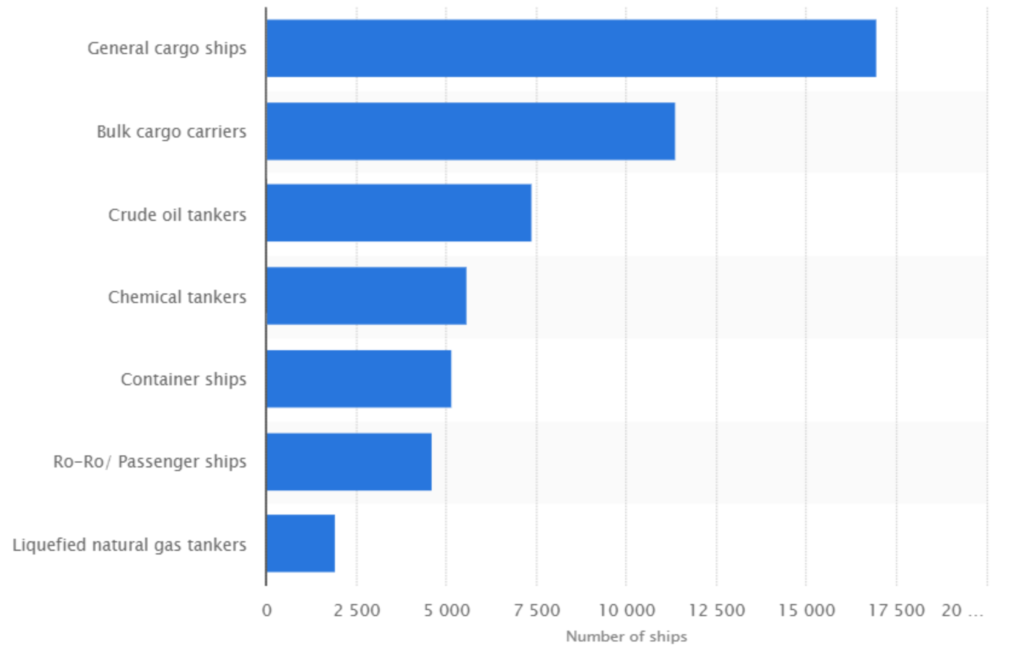
VILNIUS UNIVERSITY
INSTITUTE OF DATA SCIENCE AND DIGITAL TECHNOLOGIES

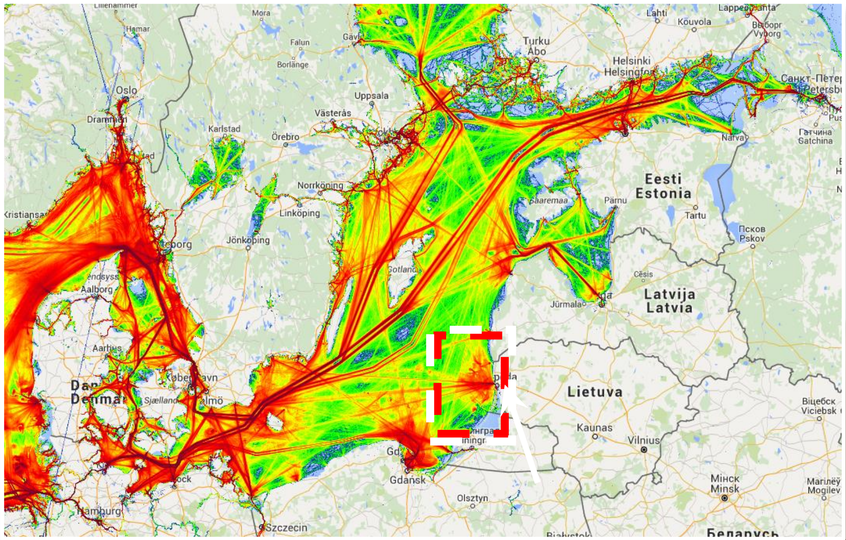
Unsupervised Marine Vessel Trajectory Prediction using LSTM Network and Wild Bootstrapping Techniques

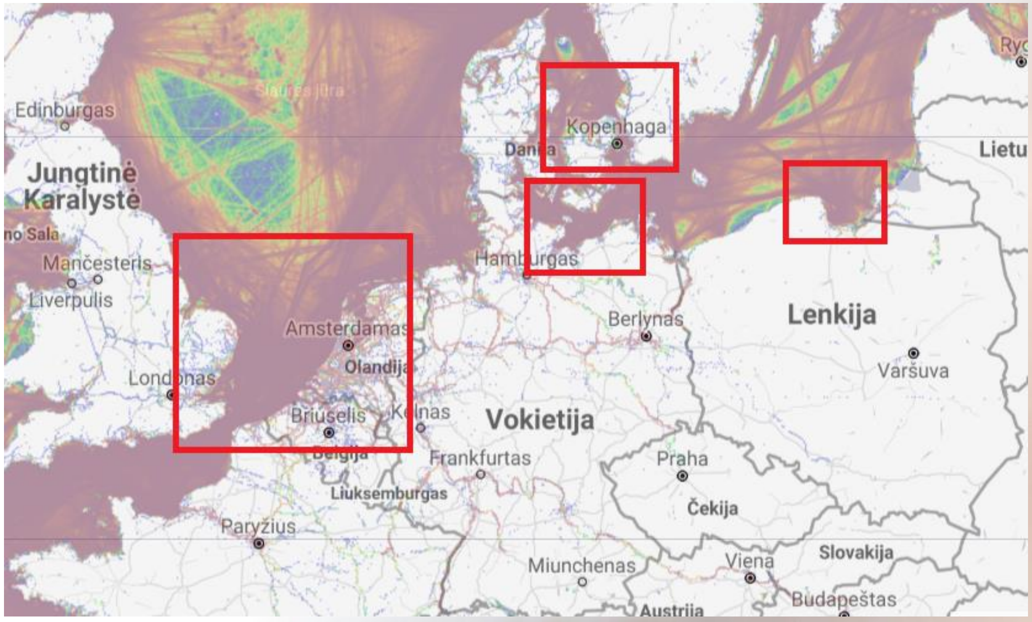
PhD candidate: Julius Venskus
Supervisor: doc. Dr. Povilas Treigys

Vilnius, 2020

Big data of marine vessel traffic







- Marine transport – 90% of total trade
- High risk of marine traffic incident
- Earthly detection of abnormal activity gives opportunity minimise accident’s risk
- The world's merchant fleet as of January 1, 2018, with a breakdown by type. Of the around 53,000 merchant ships trading internationally

Maritime traffic anomaly

Khatkhate et al.(2007) use the following definition within mechanical systems: *[a]n anomaly is defined as deviation from the nominal behavior of a dynamical system and is of-ten associated with parametric and non-parametric changes that may gradually evolve in time*."

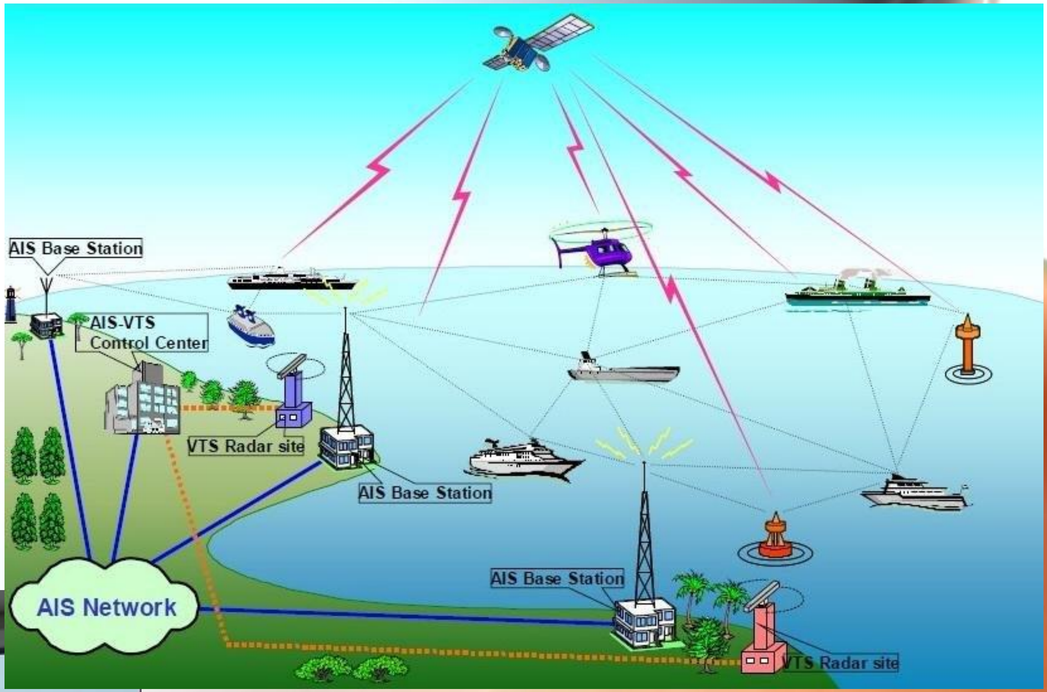
<u>Goals of detection:</u>

- Vessels collision.
- Vessel traffic into shallow water or other obstructions
- Vessel malfunction
- Vessel hijack
- Smuggling
- Espionage or reconnaissance
- Piracy
- Illegal fishing
- Military manoeuvres
- Territorial violations

AIS,VTS and missing data

VTS – Vessel traffic service. Control and navigational services. Rescue initiation.

AIS – Automatic identification system. Vessel ID, position, speed, heading, status, port call, vessel type and etc.



Missing a vessel type information in marine traffic data

Depending of data source, missing

Data preparation

The raw data is stored in a flat structure, where each record consists of the vessel navigational data at certain time. The data structure can be represented by:

$$\Psi = \{\Psi_1, \Psi_2, \ldots, \Psi_o\},$$

$$\Psi_i = \{\psi_1, \psi_2, \ldots, \psi_j, \ldots, \psi_f\},$$

where Ψi, i = 1, 2, ...o is a single vessel navigational data record that consist j, ψi = 1, 2,...*f* parameters as: vessel unique identifier MMSI, 'latitude', 'longitude', SOG, COG, Ship type, Timestamp of the data being received. Whole data set contains o number of records, where each record is set of f parameters of vessel navigational vector.

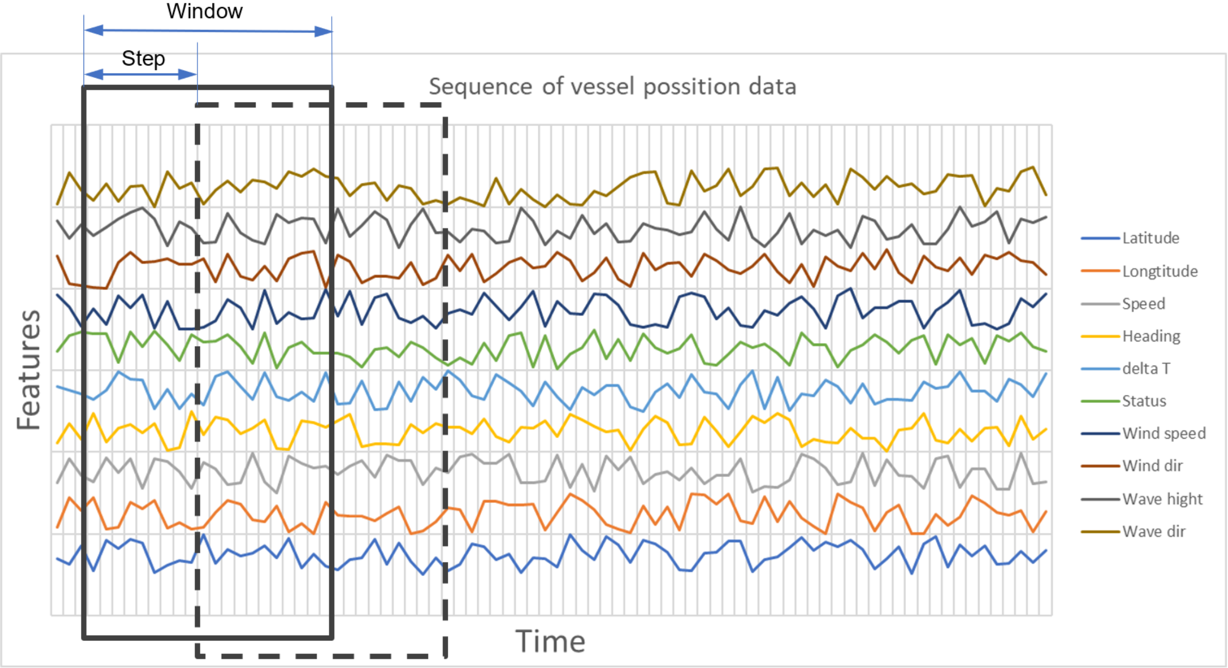
Data preparation

To achieve that, we restructure data per unique vessel, vessel's navigational data is grouped by unique vessel identifier MMSI and ordered by the timestamp:

$$S(\Psi) = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_v \end{bmatrix} = \begin{bmatrix} \{\Psi_{1,1}, \Psi_{1,2}, \ldots, \Psi_{1,b_1}\} \\ \{\Psi_{2,1}, \Psi_{2,2}, \ldots, \Psi_{2,b_2}\} \\ \vdots \\ \{\Psi_{v,1}, \Psi_{v,2}, \ldots, \Psi_{v,b_v}\} \end{bmatrix},$$

where $S(\Psi)$ is a data restructuring function that restructures the navigational data vectors of each ship into matrix rows according to the ship's MMSI parameter *MMSI*. Each row of the matrix consists of a set of navigational data vectors $s_1, s_2, \ldots, s_v$ of an individual vessel, where v is the number of distinct vessels.

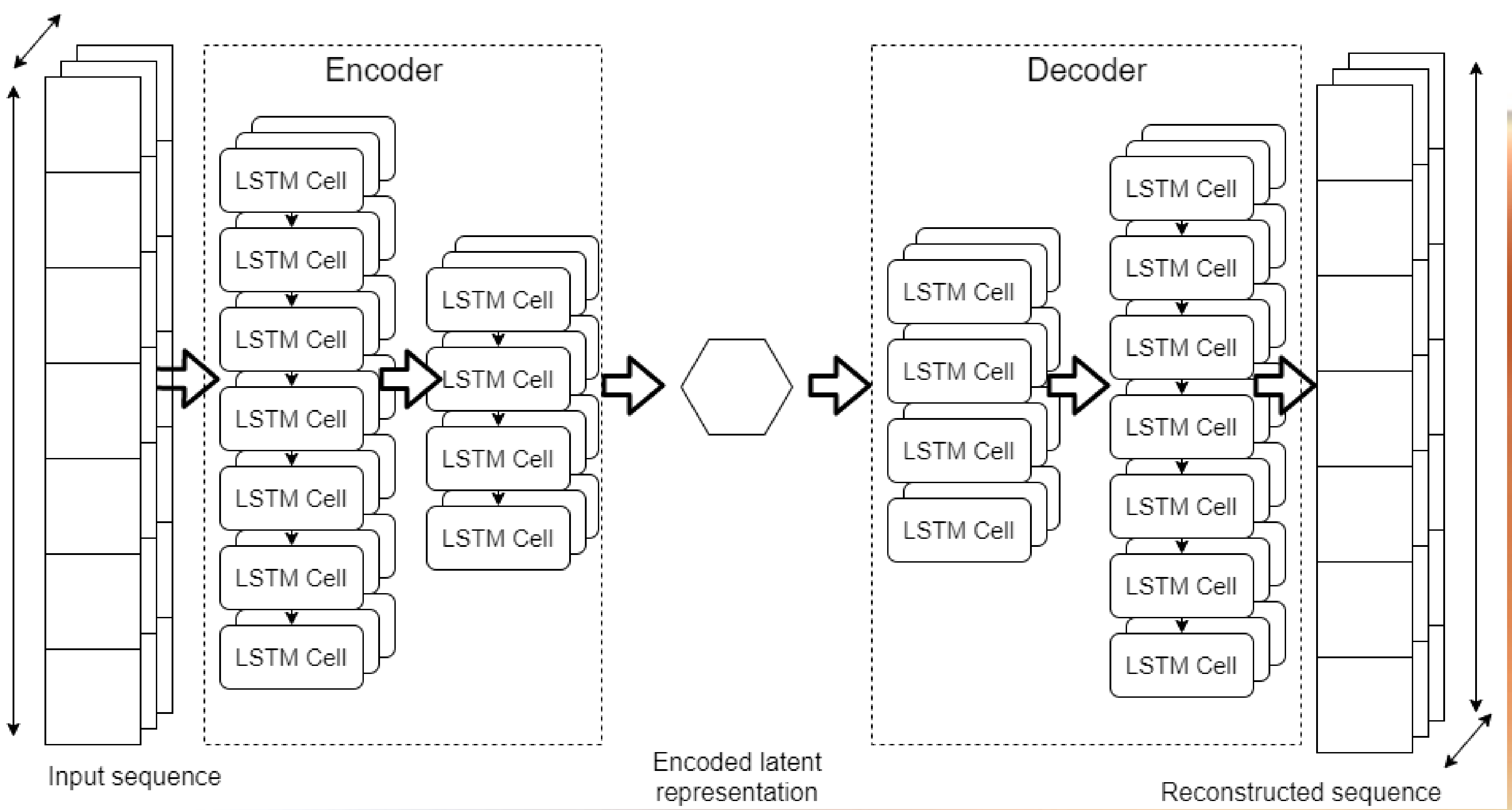
Data preparation, sliding window



$$X = \begin{bmatrix} X^1_{\tau-(n'-1)} & X^1_{\tau-(n'-2)} & \cdots & X^1_{\tau-1} & X^1_{\tau} \\ X^2_{\tau-(n'-1)} & X^2_{\tau-(n'-2)} & & X^2_{\tau-1} & X^2_{\tau} \\ \vdots & & \ddots & \vdots & \\ X^p_{\tau-(n'-1)} & X^p_{\tau-(n'-2)} & \cdots & X^p_{\tau-1} & X^p_{\tau} \end{bmatrix},$$

$$Y = \begin{bmatrix} Y^1_{\tau+1} & Y^1_{\tau+2} & \cdots & Y^1_{\tau+\tilde{n}-1} & Y^1_{\tau+\tilde{n}} \\ Y^2_{\tau+1} & Y^2_{\tau+2} & & Y^2_{\tau+\tilde{n}-1} & Y^2_{\tau+\tilde{n}} \\ \vdots & & \ddots & \vdots & \\ Y^p_{\tau+1} & Y^p_{\tau+2} & \cdots & Y^p_{\tau+\tilde{n}-1} & Y^p_{\tau+\tilde{n}} \end{bmatrix},$$

Marine vessel trajectory prediction LSTM auto-encoder



LSTM prediction region learning

A reconstruction (prediction) error of auto-encoder is obtained by:

$$e_{l,i,j} = Y_{i,j} - \hat{Y}_{l,i,j},$$
$$l = \{upper, crisp, lower\}, \quad i = 1, 2, \ldots, \widetilde{n}, \quad j = 1, 2, \ldots, f,$$

For crisp version of network a classical mean squared error (MSE) loss function is used:

$$L_s^l = \frac{1}{N\widetilde{n}f} \sum_{g=1}^{N} \sum_{i=1}^{\widetilde{n}} \sum_{j=1}^{f} e_{l,g,i,j}^2, \quad l = \{upper, crisp, lower\},$$

where L_s^l is the loss function for l (upper, crisp, or lower) type of model, N - number of training sequences in the training data set. $\widetilde{n}$ – length of sequence, f – number of features

LSTM prediction region learning

The specific loss function for upper and lower bounds is defined as follows:

$$L_\ell^{upper} = \frac{1}{N\widetilde{n}f} \sum_{g=1}^{N} \sum_{i=1}^{\widetilde{n}} \sum_{j=1}^{f} (ReLU(e_{upper,g,i,j}))^2,$$

$$L_\ell^{lower} = \frac{1}{N\widetilde{n}f} \sum_{g=1}^{N} \sum_{i=1}^{\widetilde{n}} \sum_{j=1}^{f} (ReLU(-e_{lower,g,i,j}))^2,$$

$$ReLU(x) = \begin{cases} 0, & \text{for} \quad x < 0 \\ x, & \text{for} \quad x \geq 0. \end{cases}$$

The overall loss function is defined as the weighted sum of the MSE and the region loss functions for upper and lower bounds respectively:

$$L_{total}^{upper} = L_s^{upper} + \lambda L_\ell^{upper},$$

$$L_{total}^{lower} = L_s^{lower} + \lambda L_\ell^{lower},$$

$$L_{total}^{crisp} = L_s^{crisp},$$

LSTM prediction region learning

The prediction region coverage probability (PICP) that quantify the number of measured values that fall within the region defined by the model and modified to support multi-variate features and multi-step predictions:

$$PICP = \frac{1}{N\tilde{n}f} \sum_{g=1}^{N} \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{f} (\delta_{g,i,j})$$

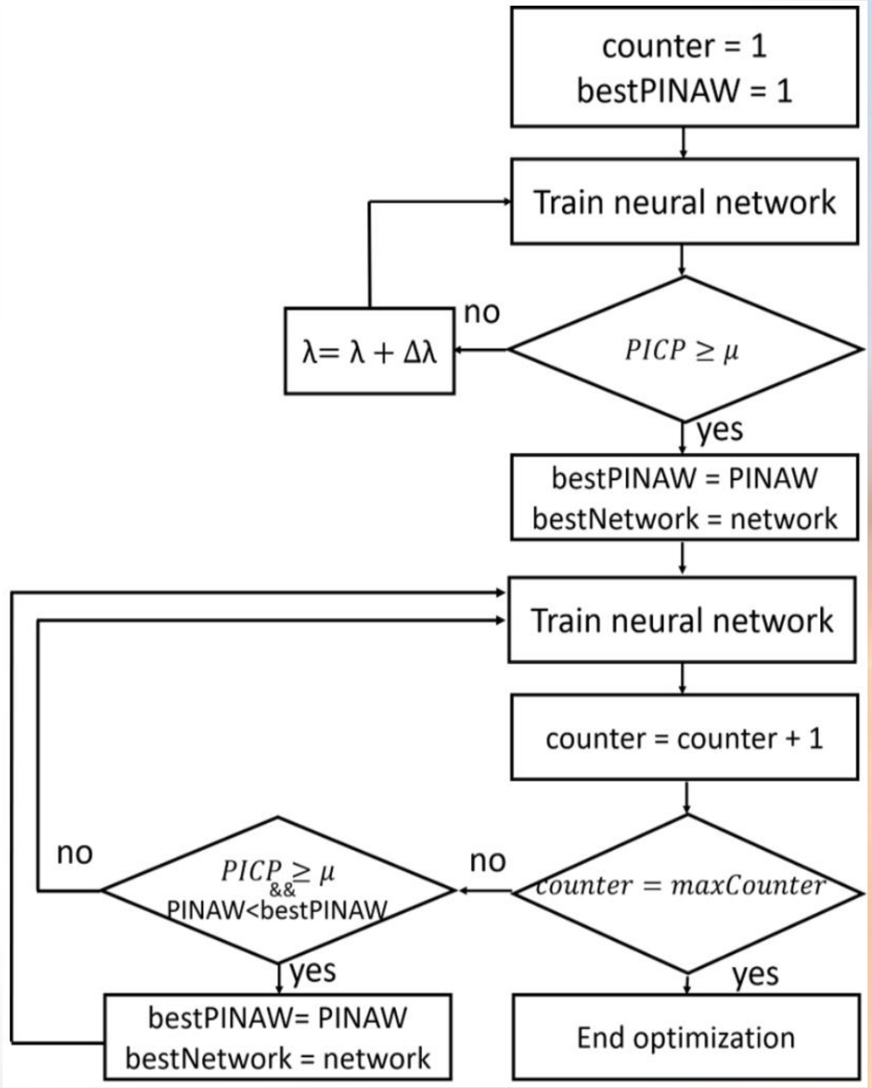
$$\delta_{g,i,j} = \begin{cases} 1 & \text{if} \quad Y_{g,i,j} \in \left[\hat{Y}_{g,i,j}^{lower}, \hat{Y}_{g,i,j}^{upper}\right] \\ 0 & \text{if} \quad \text{otherwise.} \end{cases}$$

The second metric is prediction region normalized average width (PINAW) that is used to measure the area of the region [8] also modified for multi-step and multi-variate features:

$$PINAW = \frac{1}{N\tilde{n}fR} \sum_{g=1}^{N} \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{f} (\hat{Y}_{g,i,j}^{upper} - \hat{Y}_{g,i,j}^{lower})$$

where R is the distance between the maximum and minimum values measurements $\max(\hat{Y}_{g,i,j}^{upper} - \hat{Y}_{g,i,j}^{lower})$ in the data set

LSTM prediction region learning



Wild bootstrap prediction region

1. Prepare data as described in subsection
2. Calculate data set's variance for every feature type.
3. Generate multi-variate normal random variables while keeping the same dimension, the mean equal to zero and the variance the same as that of the input data.
4. Element-wise sum the initial data set with the newly generated, i. e. add noise to the data with mean and variance calculated from initial data set.
5. Scale resulting data for better LSTM training results into interval [0; 1] while keeping each feature scaling factors for predicted data reconstruction purpose.
6. Train LSTM auto-encoder network
7. Calculate LSTM network predictions a-step ahead, a 2 f1; 2; ; : : : ; mg (authors chose m = 50).
8. Restore prediction scaling, i.e. up-scale predicted values according to the saved feature's scaling parameters described two steps above.
9. Repeat steps 3-8 k-times (authors repeated experiment 100 times.)

Wild bootstrap prediction region

After the application of the proposed above scheme, the matrix with predicted values is obtained. Then as point predicted value, the mean vector of *k* replicates is chosen for each feature and each prediction step. Thus 100(1 - α)% prediction region for the mean (average predicted value) of a *p*-dimensional normal distribution is the ellipsoid determined for unknown μ such that:

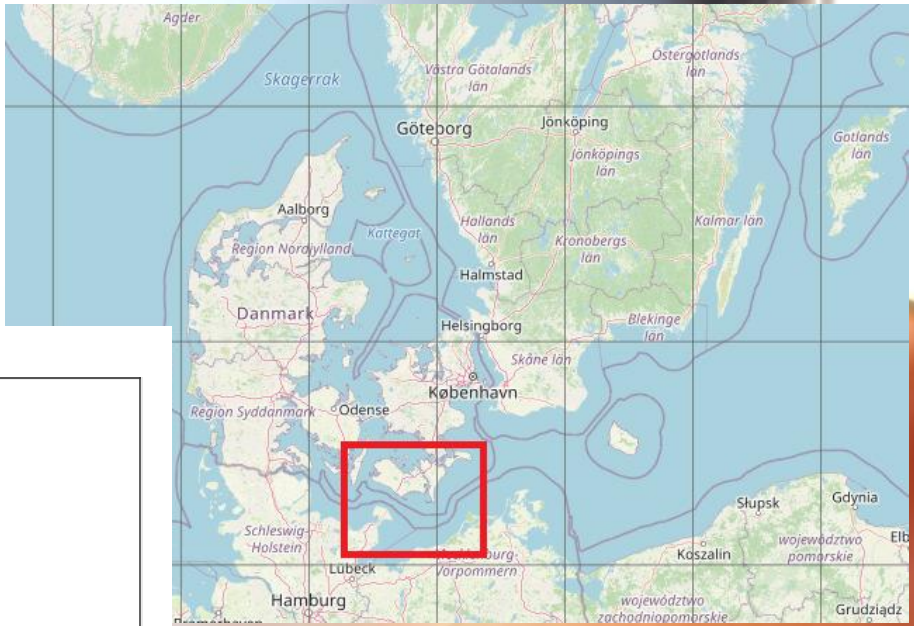
$$\frac{kr}{k+r}(\bar{x}_r - \mu)^T S^{-1}(\bar{x}_r - \mu) \leq \frac{(k-1)p}{k-p} F_{p,k-p}(1-\alpha), \tag{20}$$

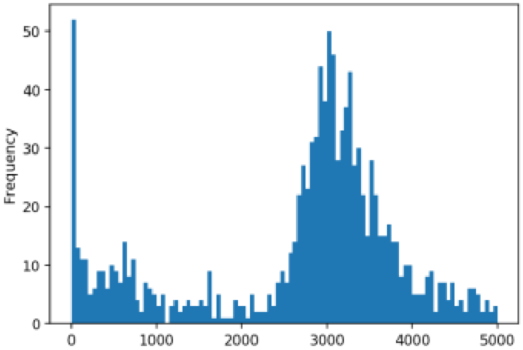
where

- $\bar{x}_r = \frac{1}{k}\sum_{i=1}^{k} x_{i,j,r}$ - the mean vector for each of the feature $j \in \{1 \dots f\}$ at each prediction step r,
- S - sample covariance matrix,
- $F_{p,k-p}(1-\alpha)$ is an $1-\alpha$-level critical value of a Fisher distribution with p and $k-p$ degrees of freedom.

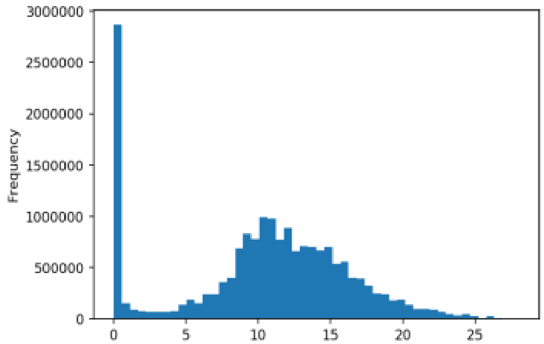
Experiments

Marine vessel region

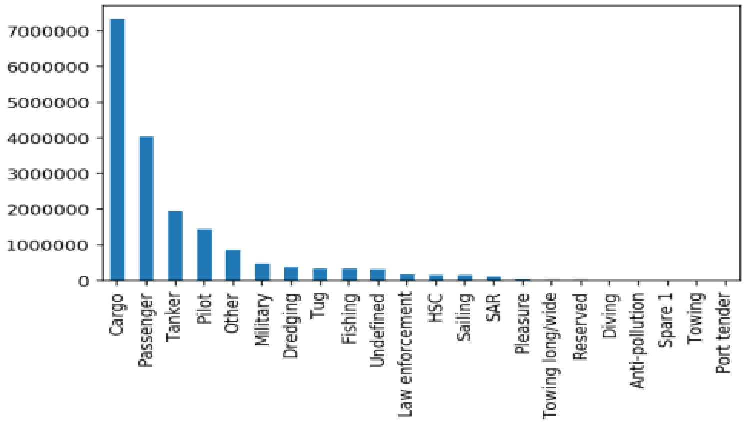




(a) Vectors per Vessel

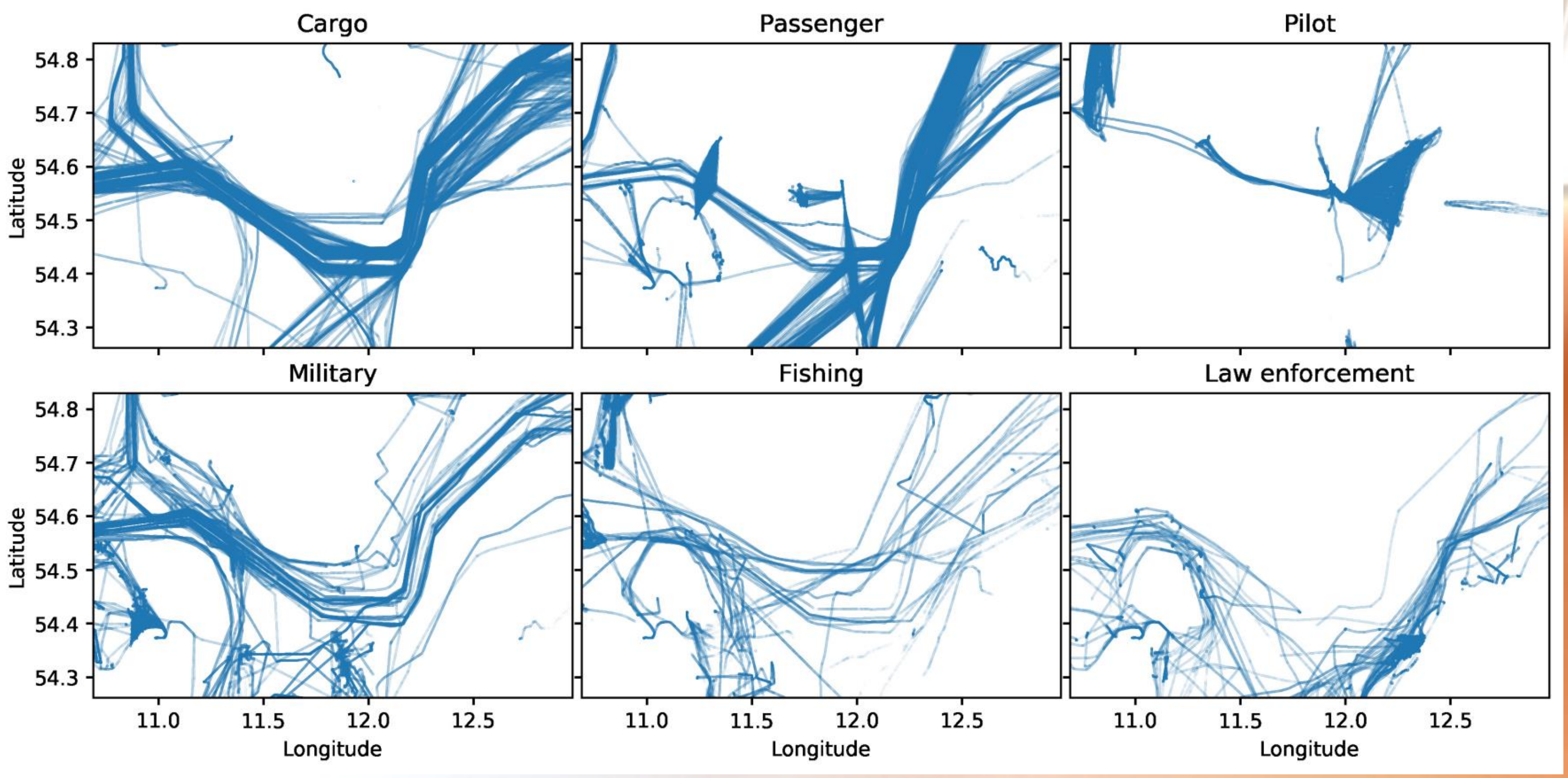


(b) Speed over ground

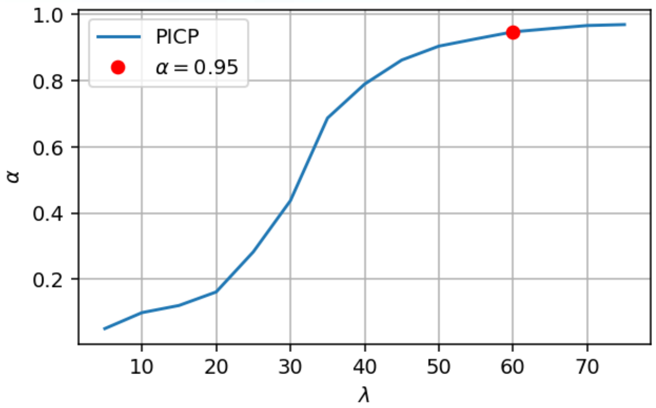


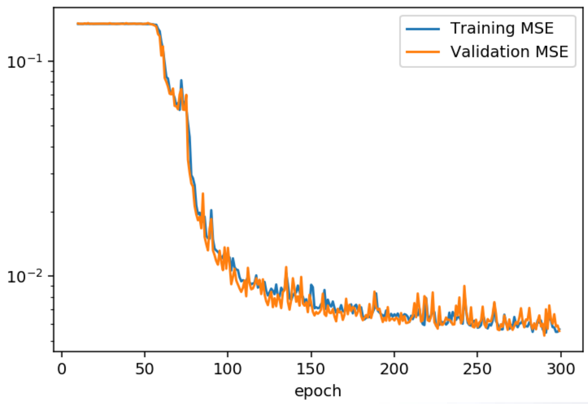
(c) Vessel types

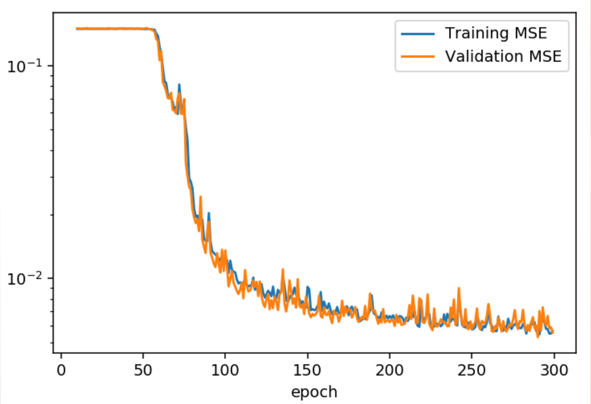
Vessel types

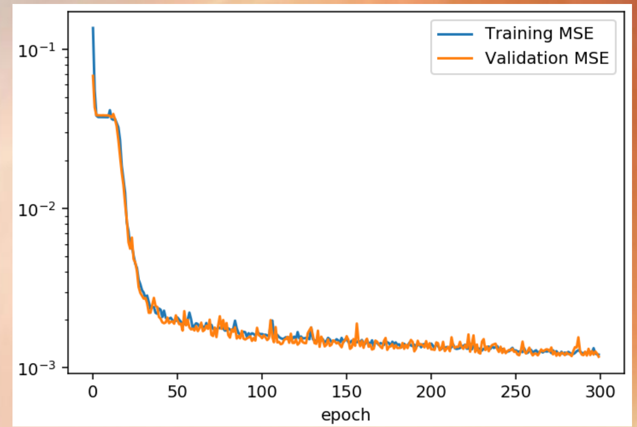


LSTM prediction region learning

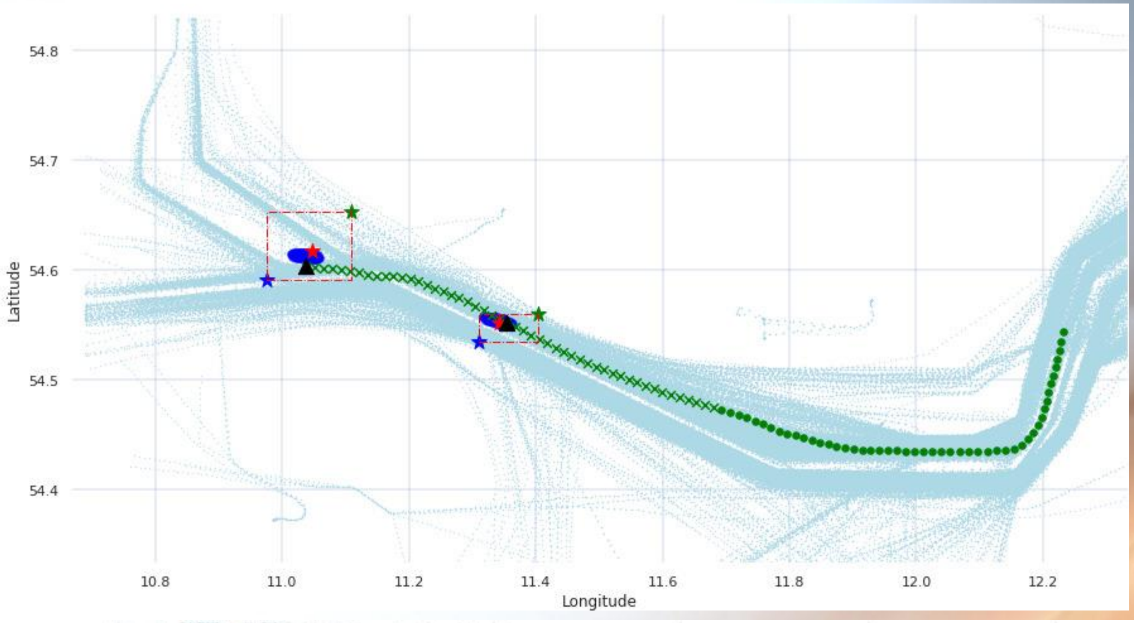


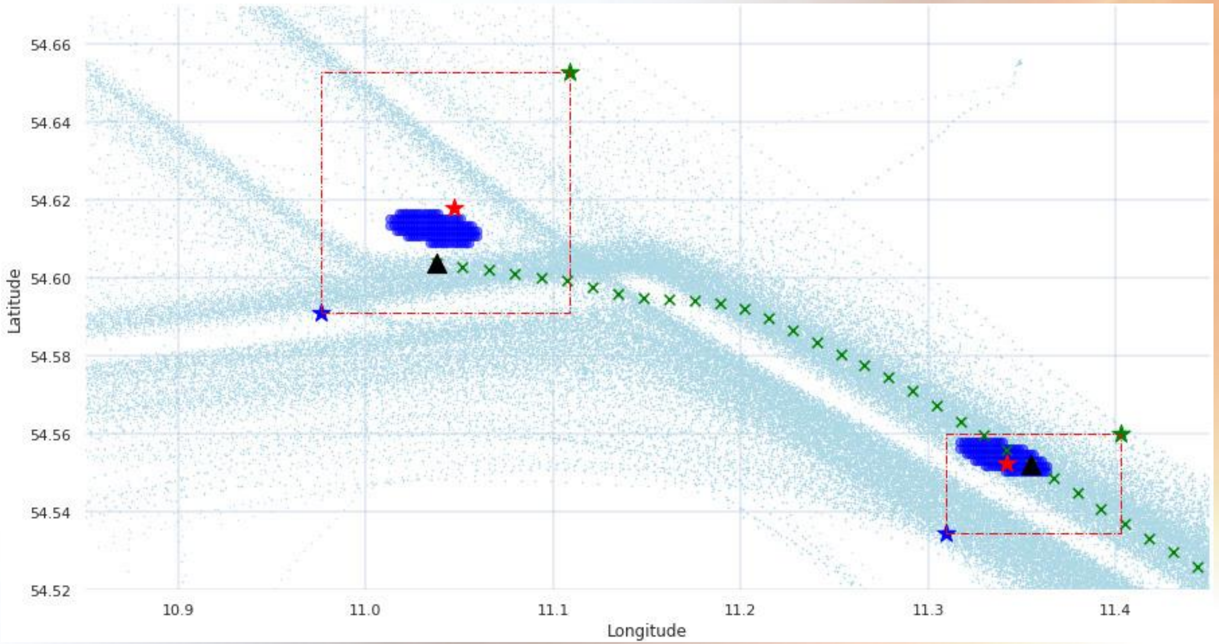




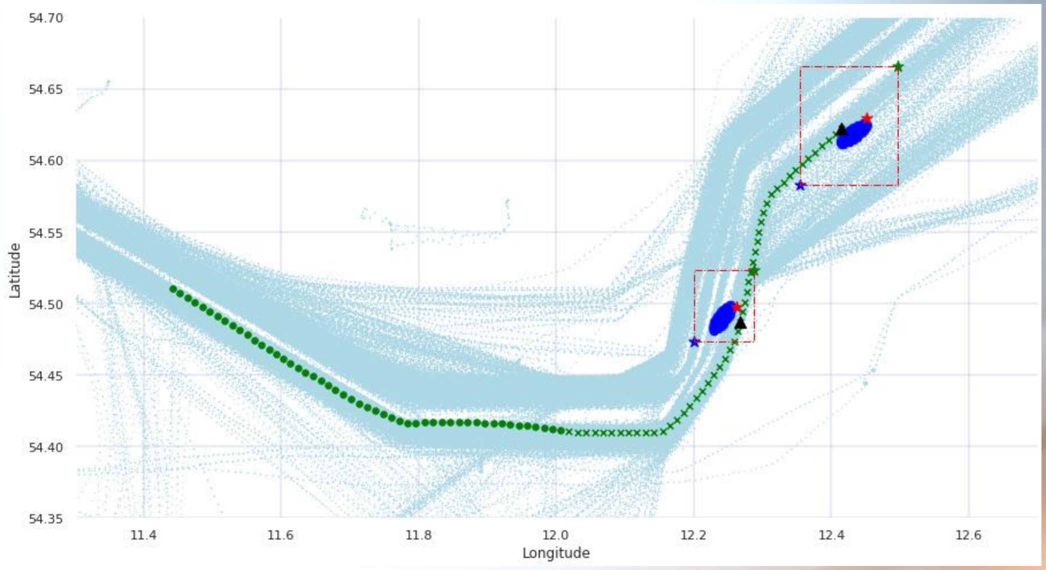


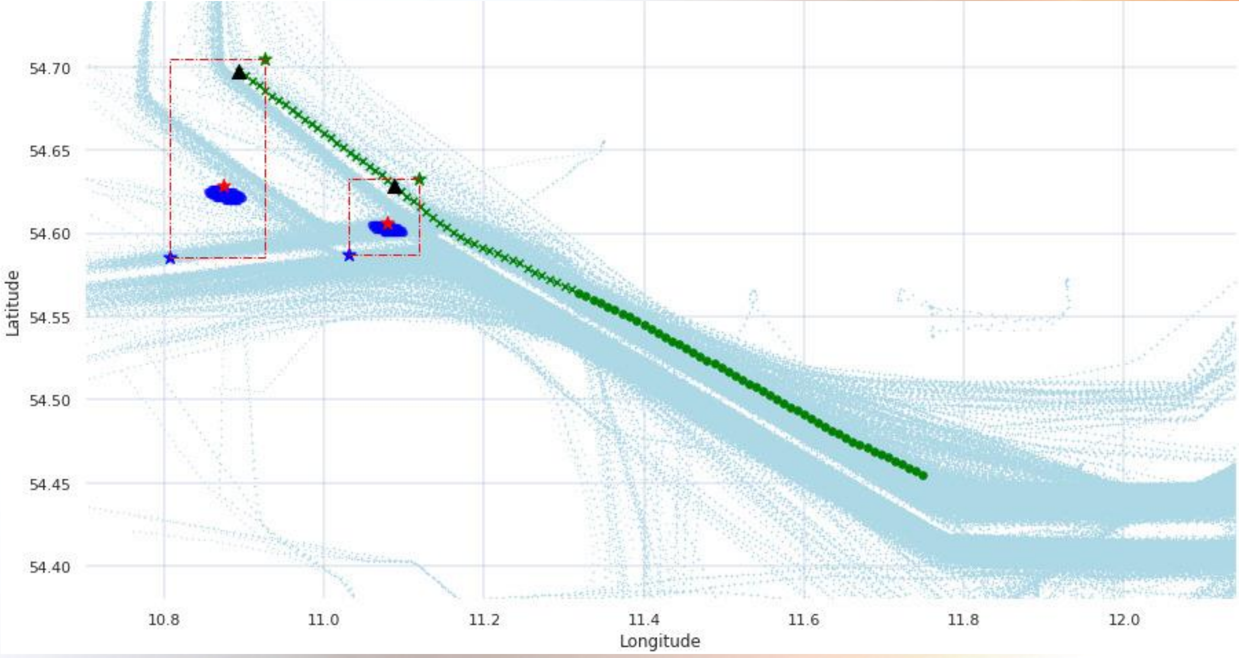
Normal traffic cases



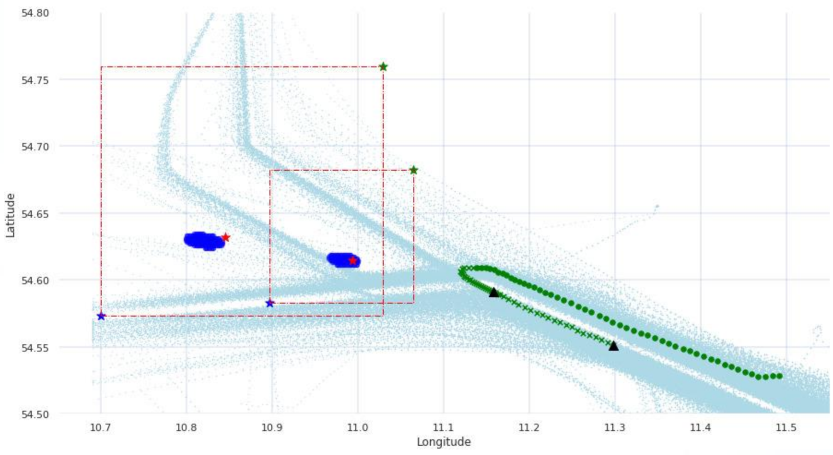


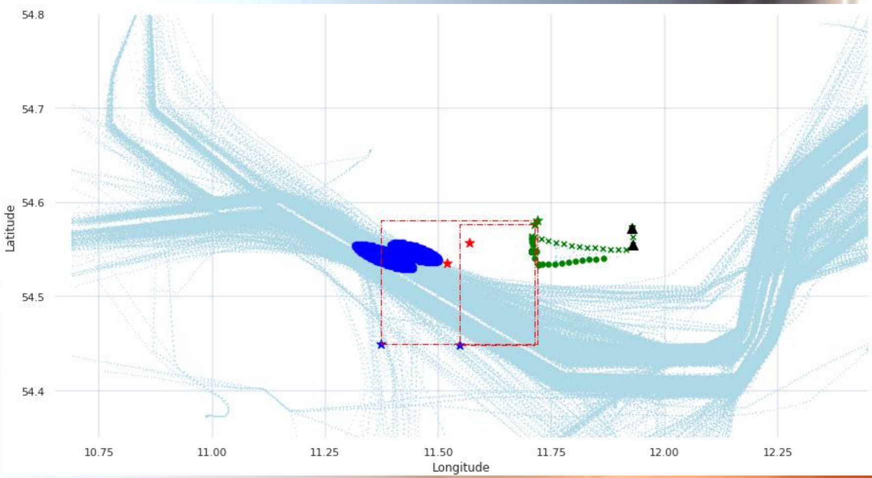
Normal traffic cases

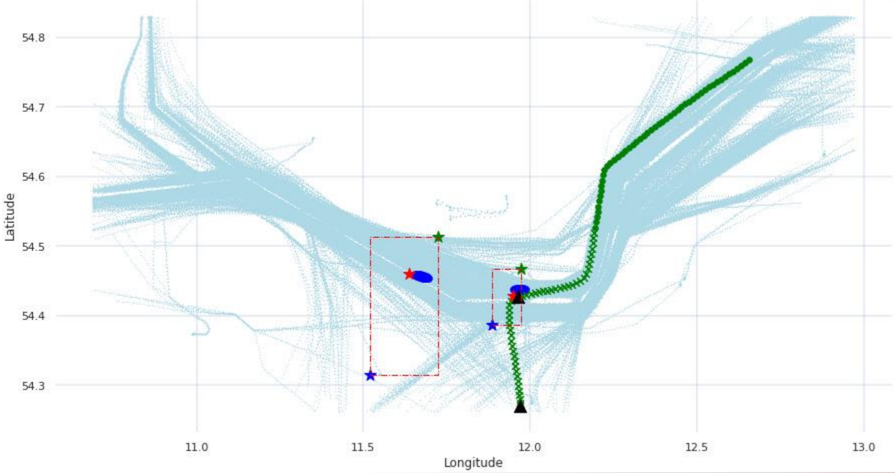


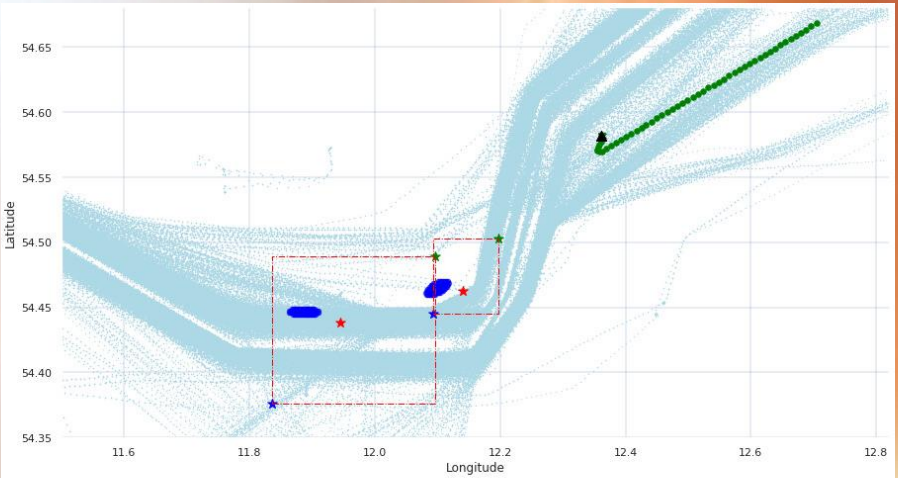


Abormal traffic cases









Results

Table 1: Detected abnormal traffic results

Method	Sequences		Ratios		MSE
	Normal	Abnormal	Normal	Abnormal	
Training data set					
LSTM prediction region	19922	1049	0,95	0,05	$1,21 * 10^3$
Wild bootstrapping	10066	10905	0,48	0,52	$1,22 * 10^3$
Test data set					
LSTM prediction region	4980	263	0,95	0,05	$1,22 * 10^{-3}$

Conclusions and Future Works

- Vessel movement prediction and prediction evaluation techniques that can be applied for traffic abnormality detection.
- The literature review revealed that the most algorithms of trajectory prediction are supervised or semi-supervised. The authors of the paper proposes new unsupervised trajectory prediction and prediction regions at the arbitrary probability.
- The paper depicts two methods - the LSTM prediction region learning and the wild bootstrapping.
- The LSTM method is based on learning of prediction region with the view to reach required confidence level by learning the parameters of custom loss function. The prediction region is defined by multivariate LSTM models according to different loss functions for learning the upper and lower bounds, that produces prediction region of trajectory point in the shape of hyper-rectangle.
- By investigating experimentally, it was observed that the 95% LSTM prediction region is wider than that obtained by wild bootstrapping technique. And it is assumed that the traffic outside the prediction region is abnormal.
- The second proposed method is based on statistical wild bootstrap approach, that estimates 95% prediction region. Nevertheless, during the testing it was noticed that only 48% of true vessel trajectory point values are inside wild bootstrapping prediction region. Thus, the method provides narrower confidence regions than those obtained by LSTM. The approach is recommended to use where strict control of marine traffic is required such as sea ports, seaport surroundings, or other sea regions with limitations induced by designated geographical locations.
- Results shows that both the LSTM, and wild bootstrapping algorithms for estimation of prediction regions can be used for abnormal marine traffic detection. The experiments with the data shows that algorithms, by evaluating prediction region, can detect different types of abnormal marine traffic such as: vessel slowdown, turn around, sharp direction change, unplanned stop, and traffic not on seaway etc in an unsupervised manner.

Ačiū
Klausimai?