VILNIAUS UNIVERSITETAS

Julius

VENSKUS

Dalinai prižiūrimų ir neprižiūrimų mašininio mokymosi metodų tyrimas jūrų eismo anomalijoms aptikti

DAKTARO DISERTACIJOS SANTRAUKA

Technologijos mokslai
Informatikos inžinerija (T 007)

VILNIUS 2021

Disertacija rengta 2016 - 2020 metais Vilniaus universitete.

Mokslinis vadovas:
doc. dr. Povilas Treigys (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija - T 007).

Mokslinis konsultantas:
prof. dr. Arūnas Andziulis (Klaipėdos universitetas, technologijos mokslai, informatikos inžinerija - T007).

Gynimo taryba:
Pirmininkas – prof. dr. Julius Žilinskas (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – T 007).

Nariai:
prof. dr. Vitalij Denisov (Klaipėdos universitetas, technologijos mokslai, informatikos inžinerija – T 007),
prof. dr. Kęstutis Dučinskas (Vilniaus universitetas, gamtos mokslai, informatika – N 009),
doc. dr. Gracia Ester Martin Garzon (Almerijos universitetas, Ispanija, technologijos mokslai, informatikos inžinerija – T 007),
prof. dr. Dalius Mažeika (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – T 007).

Disertacija ginama viešame Gynimo tarybos posėdyje 2021 m. birželio mėn. 30 d. 12 val. Vilniaus universiteto Duomenų mokslo ir skaitmeninių technologijų instituto 203 auditorijoje.
Adresas: Akademijos g. 4, LT-04812 Vilnius, Lietuva.

Disertacijos santrauka išsiuntinėta 2021 m. gegužės 28 d. Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU interneto svetainėje adresu:
`https://www.vu.lt/naujienos/ivykiu-kalendorius`

VILNIUS UNIVERSITY

Julius

VENSKUS

Semi-supervised and Unsupervised Machine Learning Methods for Sea Traffic Anomaly Detection

SUMMARY OF DOCTORAL DISSERTATION

Technological Sciences
Informatics Engineering (T 007)

VILNIUS 2021

This dissertation was written between 2016 and 2020 in Vilnius University.

Academic supervisor:
Assoc. Prof. Dr. Povilas Treigys (Vilnius University, Technological Sciences, Informatics Engineering - T 007).

Academic consultant:
Prof. Dr. Arūnas Andziulis (Klaipėda University, Technological Sciences, Informatics Engineering - T 007).

This doctoral dissertation will be defended in a public meeting of the Dissertation Defence Panel:

Chairman – Prof. Dr. Julius Žilinskas (Vilnius University, Technological Sciences, Informatics Engineering – T 007).

Members:
Prof. Dr. Vitalij Denisov (Klaipėda University, Technological Sciences, Informatics Engineering - T 007).
Prof. Dr. Kęstutis Dučinskas (Vilnius University, Natural Sciences, Informatics – N 009).
Assoc. Prof. Dr. Gracia Ester Martin Garzon (Almeria University, Spain, Technological Sciences, Informatics Engineering – T 007).
Prof. Dr. Dalius Mažeika (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering - T 007).

The dissertation shall be defended at a public meeting of the Dissertation Defence Panel at at 12:00 p.m. on 30^{th} of June, 2021 in Room 203 of the Institute of Data Science and Digital Technologies of Vilnius University.
Address: Akademijos str. 4, LT-04812, Vilnius, Lithuania

The summary of the doctoral dissertation was distributed on the 28th of May 2021.

The text of this dissertation can be accessed at the library of Vilnius University, as well as on the website of Vilnius University: `www.vu.lt/lt/naujienos/ivykiu-kalendorius`

Turinys

SANTRAUKA

1 ĮŽANGA

Jūrų logistikos sektorius yra vienas esminių pasaulinės prekybos ekonomikos komponentų, kurio apimčiai ir eismo intensyvumui augant, kyla ir jam keliami reikalavimai. 2019 m. nuo pirmojo iki trečiojo ketvirčio visuose pagrindiniuose 27-iuose Europos Sąjungos valstybių uostuose buvo perkrauta 2660 mln. tonų jūrų krovinių [1]. Tai yra 7 % daugiau lyginant su tais pačiais 2016 m. ketvirčiais. Iš viso, daugiau nei 90 % visų krovinių Europoje gabenama jūros keliais [2]. Ši pramonė yra pavojinga, tačiau kritiškai svarbi, o jos augimas kelia kontrolės ir saugumo iššūkius. Dėl didėjančio jūrų laivų eismo intensyvumo atsiranda vis didesnis poreikis incidentų valdymo sistemai, orientuotai į nepageidaujamų įvykių prevenciją. Neįprasto eismo aptikimas yra viena iš galimų šios sudėtingos valdymo sistemos technikų. Šis aptikimas pagrįstas laivo trajektorijos numatymu analizuojant navigacijos duomenų sekas ir ieškant nereguliarių, neteisėtų ir kitų anomalių trajektorijos / navigacijos duomenų [3]. Laiko sekas formuojantys jūros laivų trajektorijos navigacijos duomenys gali apimti laivo geografinę padėtį, eismo parametrus (pvz.: greitį, kryptį ir kt.), laivo objekto fizinius parametrus, identifikavimo numerius ir pagalbinius duomenis (pvz.: meteorologinius duomenis). Tokiame duomenų rinkinyje pateikiama plataus masto, sudėtinga duomenų struktūra, turinti visą reikalingą informaciją automatiniam laivų eismo prognozavimui ir prognozės įvertinimui norint nuspręsti, ar eismas stebimoje jūros zonoje yra normalus, ar neįprastas. Jūrų eismo stebėjimui yra skirtos automatizuotos duomenų rinkimo sistemos (pvz.: Automatinio iden-

tifikavimo sistema, AIS), kurios teikia didžiuosius laivų navigacijos vektorių duomenis. Didžiųjų duomenų anomalijų analizė yra praktiškai neįveikiama žmogaus kognityvinėms galimybėms [4], o aptikti laivų eismo anomalijų neįmanoma netaikant specializuotų programinės įrangos algoritmų. Norint išspręsti šią problemą, natūralus pasirinkimas yra mašininio mokymosi duomenų analizė ir duomenų gavybos metodai. Remiantis ankstesniais laivų trajektorijų duomenimis sukurti duomenų analizės modeliai galėtų padėti prognozuoti laivų judėjimą ir numatyti jį konkrečiomis eismo ir oro sąlygomis. Be to, laivai skirtinguose geografiniuose jūros regionuose ar uostuose elgiasi skirtingai, ir tai dar priklauso nuo laivo tipo. Tokia priklausomybė dar labiau sunkina prognozavimo ir anomalijų aptikimo užduotį. Šios prielaidos turi būti įtrauktos kuriant anomalijų nustatymo metodus.

Jūrų eismas yra dinamiška sistema - laivų eismo rodikliai keičiasi erdvėje ir laike. Eismo duomenų bazės struktūra sudaryta taip, kad atspindėtų daugelio laivų trajektorijas. Vieno laivo trajektorijos eismo duomenis sudaro laivo padėtis erdvėje ir kitos savybės, tokios kaip kryptis, kursas virš žemės ir kt. Paprastai jūrų laivų eismo duomenis renka ir struktūrą nustato AIS sistema. Į duomenis galima žiūrėti kaip į tam tikrų laivų duomenų vektorių, vaizduojantį laivo savybes, jo geografinę padėtį tam tikrais laiko momentais, ar, kalbant kitais žodžiais, juos galima vertinti kaip laiko ir erdvės eilutes. Erdvės ir laiko eilučių duomenų analizė yra sudėtinga klasikinių mašininio mokymosi metodų užduotis, nes elgesio modeliai turėtų reprezentuoti laivo padėtį tiek erdvės, tiek laiko atžvilgiu.

Neseniai paskelbtuose moksliniuose tyrimuose naudojami daugiasluoksniai LSTM (angl. Long Short Term Memory) neuroniniai tinklai, kurių pranašumai pasitelkiami norint išmokti erdvės ir laiko priklausomybes (žr. [5, 6, 7]). Šie tyrimai rodo galimas tolimesnio tyrimo perspektyvas.

1.1 Problemos aktualumas

Jūros situacijos informatyvumo koncepcija (angl. Maritime situation awareness (MSA)) buvo pristatyta NATO viršūnių susitikime Rygoje 2006 m. kaip Jūrų sektoriaus informatyvumo (angl. Marine Domain Awareness) tęsinys. Pora metų vėliau NATO pristatė MSA koncepcijos vystymo planą. Pagrindinis MSA koncepcijos tikslas yra įdiegti naują požiūrį į doktriną, organizavimą, mokymą, logistiką, lyderystę, personalą, infrastruktūrą ir sąveikumą (DOTMLPII) [8].

Pagrindinis MSA tikslas yra gauti išsamų vaizdą jūriniame sektoriuje gaunant informaciją iš keleto stebėjimo, sekimo ir žvalgybos sistemų, įskaitant žinių gavybos posistemes. Martineau ir Roy teigia, kad „niekas negali teigti, kad pilnas domimos srities situacijos informatyvumas yra pasiektas“ [9]. Kita vertus, akivaizdu, kad galutinė tokio tikslo būsena yra nepasiekiama dėl jūrų srities sudėtingumo ir kintamumo. Šį supratimą palaiko tie patys autoriai Martineau ir Roy teigdami, kad „tai būtų panašu į visažinystę, o jo pasiekimas būtų utopija“ [9]. Norint gauti aiškų situacijos vaizdą, reikia surinkti ne tik nuolatinius ir laiku gaunamus duomenis iš kelių šaltinių, bet ir atlikti atpažinimą bei išgauti tuos pačius duomenis. Žinių gavimas yra labai svarbi sistemos dalis praturtinant MSA.

Sauga ir saugumas jūrų srityje atlieka labai svarbų vaidmenį. MSA jūrų ir pakrančių valdžios institucijoms suteikia galimybę įvertinti galimą saugumo ir saugos riziką ir laiku imtis veiksmų šiai rizikai sumažinti [8]. Jūrų laivų elgsenos modelių išskyrimas ir saugos pavojų ar saugumo pažeidimo situacijos įvertinimas yra vieni svarbiausių MSA tikslų. Dėl didelio jūrų laivų eismo intensyvumo ir jų generuojamo duomenų kiekio, žmogaus pažintiniai gebėjimai negali efektyviai įvertinti situacijos. Duomenų rinkimo automatizavimas ir žinių gavimo metodai bei jų praktinis taikymas MSA gali padėti institucijoms siekti saugos ir saugumo tikslų [10]. Didelio kiekio įvairių duomenų rinkimas ir žinių gavimas padeda pakrančių valdžios institucijoms priimti pagrįstus sprendimus [11, 12, 13]. Vienas iš būdų pagerinti MSA yra neįprasto elgesio jūrų laivų eismo duomenyse nustatymas (anomalijų

aptikimas) [14, 15], tai tvirtai palaiko daugybė civilinių, karinių ir teisėsaugos institucijų visame pasaulyje [14].

Anomalijos apibrėžimas. „Anomalijų“ ir „anomalijų aptikimo“ sąvokų galima rasti įvairiose tyrimų srityse, tokiose kaip gedimų diagnostika, vaizdo stebėjimas, tinklo saugumas, žmogaus veiklos stebėjimas, priežiūra ir kt. [14]. Ekmanas ir Holstas teigia: „anomalijų aptikimas nieko nepasako apie aptikimo metodą ir iš tikrųjų nieko nepasako apie tai, ką aptikti“ [16]. Daugelyje mokslinių straipsnių „anomalijų aptikimas“ pateikiamas iš žmogaus suvokimo perspektyvos arba duomenų vertinimo perspektyvos. Anomalijos prasmėse yra tiek griežto, aiškaus apibrėžimo, tiek ir didelio neapibrėžtumo bei neaiškumo [14, 17].

Daugumoje duomenimis paremtų tyrimų anomalija yra apibrėžta kaip nuokrypis nuo normalumo. Tai aiškiai apibrėžia savo straipsnyje Portnoy *ir kiti*: „anomalijos aptikimo būdai sukuria įprastų duomenų modelį ir tada bando aptikti nuokrypį nuo normalaus modelio stebimuose duomenyse“ [18]. Moksliniai straipsniai, analizuojantys laivų eismą, aprašo anomalijas šiek tiek kitaip, bet didžiąja dalimi apibūdina kaip eismą be priežasties, nukrypimą nuo įprastų laivų eismo juostų, navigacinių kelių, trajektorijų, greičio ar kitų eismo parametrų [11, 19, 4].

Šioje disertacijoje anomalijos aptikimas tiriamas kaip jūrinio saugumo informuotumo tobulinimas. Informatikos inžinerijoje ir jūrų saugume terminas „anomalija“ yra naudojamas pakaitomis su kitais turinčias tą pačią prasmę sinonimais kaip: „neįprastas“, „anomalus eismas“ ir t.t.

1.2 Tyrimo objektas

Jūrų eismo anomalijų aptikimas AIS duomenyse.

1.3 Tyrimo tikslas ir uždaviniai

Tyrimo tikslas yra ištirti ir pasiūlyti modifikacijas dalinai prižiūrimo ir neprižiūrimo mašininio mokymosi metodams jūrų eismo anomalijoms aptikti.

Norint pasiekti šį tikslą būtina įgyvendinti šiuos uždavinius:

1. Apžvelgti susijusią mokslinę literatūrą.
2. Ištirti AIS duomenis ir pritaikyti duomenų paruošimo būdus duomenų valymui, požymių sudarymui didinant neįprasto eismo anomalijų aptikimo jautrumą.
3. Pasiūlyti metodą, skirtą spręsti trūkstamų duomenų problemą laivo tipo požymiuose.
4. Ištirti dalinai prižiūrimus, viena pozicija paremtus, modelių mokymosi metodus laivų eismo anomalijoms aptikti. Pasiūlyti metodo tobulinimą ir ištirti galimybę išnaudoti istorinius duomenis pritaikant algoritmą srautiniams AIS duomenims.
5. Ištirti neprižiūrimus, trajektorija paremtus, mokymosi metodus laivų eismo anomalijoms aptikti, sukurti neįprasto laivų eismo prognozės regiono apsimokymo metodą ir palyginti jį su statistiniais metodais.
6. Palyginti aptiktas anomalias trajektorijas, gautas dalinai prižiūrimo mokymosi metodais ir neprižiūrimo apsimokymo metodais, naudojant AIS duomenis jūrų regionuose.
7. Pasiūlyti duomenų paruošimo metodus skirtingos kilmės AIS duomenims, įtraukiant: duomenų struktūrizavimą, duomenų valymą, skaitmeninimo dažnio mažinimą, trūkstamų reikšmių užpildymą, naujų požymių kūrimą, trūkstamų laivų tipų klasifikatorių, duomenų skaidymą į sekas, kad būtų paruošti duomenys būsimam laivų eismo anomalijų aptikimui.
8. Išvystyti ir ištestuoti laivų tipų klasifikatorių, siekiant išspręsti trūkstamų laivų tipų duomenų problemą.

1.4 Tyrimo metodai

1. Literatūros apžvalga atliekama pagal naujausius mokslinius straipsnius, siekiant nustatyti, pasirinkti ir įvertinti pažangiausius algoritmus nurodytai problemai spręsti.

2. Kiekybiniai ir kokybiniai informacijos rinkimo metodai naudojami sukurti duomenų rinkiniams, kurie reikalingi eksperimentams atlikti.
3. Patvirtinančiai duomenų analizei atlikti buvo naudojami metodai, įskaitant statistinius metodus, tačiau jais neapsiribojant, užtikrinant duomenų patikimumą ir eksperimentinę sąranką.
4. Tiriamoji duomenų analizė: duomenų skirsniai, histogramos, duomenų vizualizavimas, poriniai požymių grafikai, neigiamos tikimybės kontūro diagramos.
5. Aprašomoji statistika: vienmatė ir daugiamatė duomenų analizė, Pearson's koreliacija, Cramer's V koreliacija; duomenų vidurkio ir variacijos mastelio keitimas, sintetinė mažumų perteklinės atrankos technika (SMOTE).
6. Modelio vertinimo metodai: klasifikacijos modelio supainiojimo matrica, klasifikacijos metrikos palyginimas, regresijos paklaidų vertinimas ir palyginimas, spėjimo regiono padengimo tikimybės vertinimas, spėjimo regiono normalizuoto pločio vertinimas, apsimokymo duomenų susirišimo su normaliniu skirsniu metodas (angl. wild bootstrapping).
7. Daugiamačiai klasterizavimo metodai: saviorganizuojantys neuroniniai tinklai (SOM), Ervinių laiko eilučių klasterizavimo metodai (Soft-DTW k-means);
8. Dirbtinių neuroninių tinklų metodai: ilgalaikio ir trumpalaikio intervalo atminties neuroniniai tinklai (LSTM); daugiasluoksnis perceptronas; autoenkoderiai; neuroninių tinklų nuoseklusis jungimas.
9. Taikomi konstruktyvūs tyrimo metodai siekiant pasiūlyti realaus pasaulio problemos sprendimų tobulinimus, taikomus gerinti jūrų saugumo informatyvumą.
10. Eksperimentinėje šio darbo dalyje buvo naudojami programinės įrangos kūrimo metodai, įskaitant jūrų laivų anomalijų aptikimą ir trajektorijos klasterizavimą.

1.5 Mokslinis indėlis ir praktinė tyrimo vertė

Šios disertacijos tyrimai prisideda prie jūrų laivų eismo anomalijų nustatymo vystymo siekiant tobulinti jūrinio saugumo informatyvumą. Pagrindinį šios disertacijos mokslinį indėlį bei praktinę vertę būtų galima apibūdinti taip:

1. Pasiūlytas naujas laivų eismo anomalijų aptikimo metodas, kuris gautas integruojant savior ganizuojantį neuroninį tinklą SOM su virtualiu feromonu. Modifikacija atliekama įtraukiant virtualius feromonų intensyvumo skaičiavimus paskutiniame modelio mokymosi etape. Šis metodas parodė geresnius klasifikavimo rezultatus taikant mažesniems jūrų laivų eismo duomenų rinkiniams. Pasiūlytas geriausios kaimynystės funkcijos ir SOM tinklelio dydžio parinkimo metodas.
   Šis metodas gerina jūrinės saugos informatyvumą laivų eismo tarnybai, kuri aptarnauja mažus uostus ar jūrų regionus.
2. Siūlomos naujos SOM pakartotinio mokymosi strategijos, leidžiančios pakartotinai apsimokinti anomalaus jūrų eismo modelį naujai gautais duomenimis reikšmingai sutrumpinant pakartotinio mokymosi laiką ir išlaikant aukštą preciziškumą ir jautrumą.
   Praktikoje tai ne tik padidina modelio pakartotinio mokymosi greitį, bet ir sutrumpina modelio pakartotinio mokymosi laiką. Sutrumpintas apmokymas ženkliai sutaupo kaštus, reikalingus modeliui pakartotinai apsimokyti kai gaunami nauji laivų eismo duomenys.
3. Pasiūlytas naujas jūrų laivų tipo nuspėjimo metodas, kad būtų užpildomos trūkstamos laivų tipų reikšmės duomenyse. Metodo pagrindą sudaro daugiamatės daugiasluoksnės lygiagrečiosios architektūros neuroninis LSTM tinklas. Modelio testavimo rezultatai parodė, kad klasifikavimo tikslumas ir jautrumas yra aukšti ir gali būti naudojami trūks-

tamų duomenų užpildymui.
Praktikoje šis metodas leidžia panaudoti daugiau duomenų apmokant neįprasto eismo aptikimo modelį.

4. Pasiūlyti du nauji neprižiūrimo mokymosi neįprasto eismo trajektorijų aptikimo metodai, paremti LSTM neuroniniais tinklais. Abu metodai aptinka anomalias trajektorijas tikrinant ar trajektorija yra spėjimo regione. Pirmasis metodas - tai LSTM spėjimo regiono apsimokymo metodas, kuris buvo sukurtas modifikuojant LSTM spėjimo intervalo mokymosi metodą, pritaikant jį LSTM autoenkoderiui ir daugiamačiams duomenims. Antras, LSTM „wild boostraping“ metodas, paremtas integruojant apsimokymo duomenų susirišimo su normaliniu skirsniu metodą (angl. wild bootstrapping) su daugiasluoksniu daugiamačiu LSTM autoenkoderiu, kurio pagalba formuojamos spėjimo regiono elipsės, skirtos nustatyti įprasto eismo modelį. Eksperimentiškai nustatyta, kad abu modeliai gali nustatyti įvairesnių trajektorijų formų anomalijas lyginant su SOM modeliais. Praktiniame taikyme šie metodai leidžia žymiai supaprastinti laivų eismo anomalijų aptikimo modelio apmokymą, kadangi neprižiūrimasis metodas leidžia išvengti didelio kiekio rankinio duomenų žymėjimo, kuris įprastai reikalingas dalinai prižiūrimiems SOM metodams. LSTM metodai gali būti naudojami didesniems ir intensyvesnio laivų eismo regionams, kur anomalių trajektorijų išankstinis žymėjimas sunkiai įmanomas. LSTM metodas siūlo didesnį jūrinio saugumo informatyvumo detalumą eismo tarnyboms aptinkant įvairesnes anomalias laivų trajektorijas.

1.6 Ginamieji teiginiai

1. Pasiūlytas SOM neuroninis tinklas su integruotu virtualiu feromonu tiksliau aptinka neįprasto laivų eismo atvejus mažesniuose duomenų rinkiniuose lyginant su SOM_GMM metodu. Nepaisant to, SOM_GMM metodas turėtų būti naudojamas didesniems laivų eismo duomenų rinkiniams.

2. SOM neuroninių tinklų metodai neįprastam laivų eismui aptikti gali būti pakartotinai apsimokomi reikšmingai trumpesniu laiku išlaikant minimalų klasifikavimo preciziškumo pokytį.
3. Sukurtas LSTM prognozavimo mokymosi metodas ir LSTM apsimokymo duomenų susirišimo su normaliniu skirsniu metodas geba aptikti laivų eismo trajektorijų anomalijas. Naudojant žymiai mažesnes mokymosi duomenų aibes LSTM prognozavimo metodo rodikliai aplenkia LSTM „wild boostraping“ metodą.
4. LSTM tinklo architektūra esant geriems klasifikavimo rodikliams nustatant laivo tipą, gali būti taikoma trūkstančioms reikšmės įterpti.
5. Taškiniai SOM_pheromone ir SOM_GMM metodai neaptinka laivų eismo trajektorijų, kurios yra aštraus posūkio arba sustojimo formos, bet LSTM metodai tokias formas aptinka.

2 Duomenų paruošimas

Siekiant užtikrinti vienodas sąlygas metodų tyrimui, duomenys yra paruošiami vienodomis sąlygoms. Šiame skyriuje aprašomi duomenų šaltiniai, duomenų struktūra, duomenų pertvarkymas, neapdorotų duomenų valymas, atranka, trūkstamų verčių užpildymas, požymių inžinerija, suskaidymas į sekas ir trūkstamo laivų tipo atpažinimas bei užpildymas.

Duomenų šaltinių aprašas. Tyrimuose naudoti trys duomenų šaltiniai. Pirmasis sudarytas iš jūrų laivų eismo duomenų, kurie gauti iš Danijos laivybos administracijos AIS sistemos ir yra saugomi šios organizacijos duomenų bazėje [20]. Iš šio šaltinio panaudoti Fehmarnbelt regiono navigaciniai duomenys. Antrasis

1.1 lentelė: Klaipėdos uosto duomenų aibės

Duomenų poaibis pagal laivo tipą	Navigaciniai vektoriai		
	Viso	Neįprasti	Įprasti
Cargo	138242	3362	134890
Pessenger	43879	2914	40965
Tug	50372	2306	48066

duomenų šaltinis yra vidutinio nuotolio orų prognozių Europos centro (angl. European Centre for Medium-Range Weather Forecasts (ECMWF)) meteorologiniai duomenys, gauti naudojantis World Weather Online serviso teikiamomis paslaugomis [21]. Ir trečioji duomenų aibė yra Klaipėdos jūros regiono AIS duomenys, gauti iš Klaipėdos saugios laivybos administracijos [22, 23]. Neįprasto eismo metodai tiriami pagal dviejų regionų AIS duomenis, sujungtus su meteorologiniais duomenimis.

Meteorologiniai duomenys sudaryti iš informacijos apie: vėjo kryptį ir greitį, bangavimo kryptį ir aukštį, dienos / nakties laiką, atoslūgio lygį. Ši informacija surinkta pagal meteorologinio tinklelio tankumą 0.5° trijų valandų periodiškumu.

1.2 lentelė: Pagrindinė informacija apie Fehmarbelt duomenų aibę

Geografinio regiono pavadinimas	Fehmarnbelt
Laiko periodas	nuo 2019-01-01 00:00:00 iki 2019-03-31 23:59:59
Platumos intervalas, laipsniais	53.832833 to 54.998114
Ilgumos intervalas, laipsniais	9.97929 to 12.53451
AIS duomenų naudojama atmintis, megabaitais	73579.26
N' - viso AIS vektorių įrašų	98245370
f' - pradiniai AIS požymiai	11
V - viso unikalių laivų skaičius	3913
Meteorologinio tinklelio dydis	$4 \times 7 \times 0.5^{\circ}$

Duomenų paruošimo eiga. Disertacijoje naudojami duomenys paruošiami tokia tvarka:

- *Pradiniai duomenys.* Pradiniai neapdoroti duomenys saugomi paprastoje struktūroje, kur kiekvienas duomenų įrašas yra konkretaus laivo navigacinių duomenų vektorius tam tikru laiko momentu.

- *Duomenų valymas.* AIS duomenys, priimti iš AIS sistemos radijo bangomis, turi gausių triukšmų, kuriuos būtina išvalyti. Tai pat duomenys turi daug trūkstamų įrašų. Pasiremiant duomenų statistine analize, Pearson ir Cramer V koreliacijomis, duomenys išvalomi ir atrenkami požymiai. Taip pat yra sumažintas duomenų skaitmeninimo dažnis (angl. down-sampling) taip, kad laivų eismo duomenų intervalas būtų $\Delta T_{interval}$. Šis kalibruojamas parametras parinktas pagal MSA poreikį laivų eismo regionams ir disertacijoje jis naudotas $\Delta T_{interval} = 2min$.

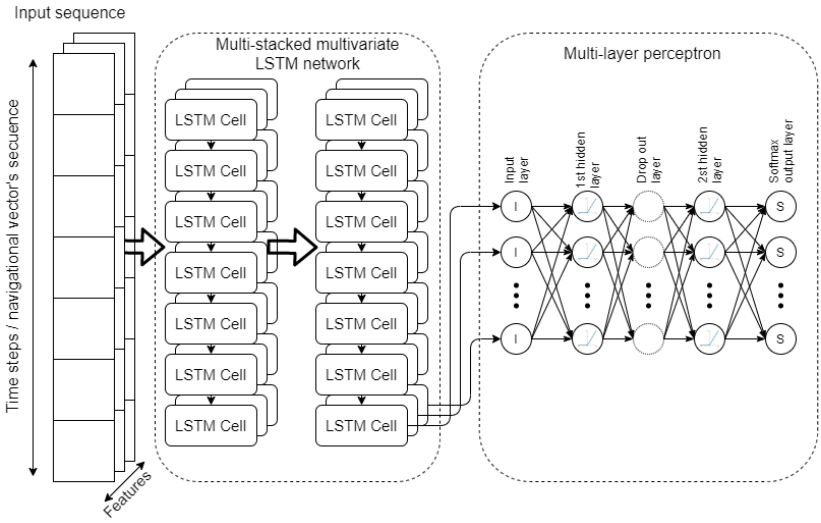
- *Požymių inžinerija.* Navigaciniai laivų vektoriai yra priimami skirtingais laiko intervalais, kas padidina nuplaukto atstumo ilgio variaciją per vieną laiko žingsnį. Norint padidinti prognozavimo stabilumą modeliui buvo įvesti papildomi požymiai: laiko intervalas tarp skirtingų vektorių, geografinės ilgumos ir platumos skirtumai per laiko vienetą.

- *Trūkstamų reikšmių užpildymas.* Užpildomos trūkstamos reikšmės vektoriuose. Priklausomai nuo duomenų kategorijos, naudojamos skirtingos strategijos: statiniams, dinaminiams ar maršruto tipo duomenims.

- *Vektorių grupavimas į sekas.* Naudojant slenkančio lango algoritmą laivų duomenys suskirstomi į sekas, grupuotas pagal individualų laivą ir vektorių gavimo eiliškumą.

- *Trūkstamų laivo tipo reikšmių užpildymas.* Apmokomas laivų tipo klasifikatorius ir jo pagalba užpildomos trūkstamos laivų tipo reikšmės.

Laivo tipo klasifikatorius. Dėl vektoriuose trūkstamų laivo tipo duomenų, pagal eismo trajektoriją buvo sukurtas laivų tipų klasifikatorius. Trūkstamų duomenų užpildymas leidžia padidinti



1.1 pav.: Laivų tipų klasifikatoriaus architektūros schema

duomenų kiekį apmokant neįprasto laivų eismo aptikimo modelį [24].

Modelis buvo apmokytas su žinomomis laivų tipo trajektorijomis ir ištestuotas ant testavimo aibės. Testavimo rezultatai pateikti lentelėje 1.3.

1.3 lentelė: Laivų tipo klasifikariaus rodikliai

Laivo tipas	Preciziškumas	Jautrumas	f1-score
Cargo	0.96727	0.97182	0.96954
Tanker	0.97004	0.98268	0.97632
Fishing	0.95735	0.96739	0.96234
Passenger	0.96518	0.96034	0.96275
Tug	0.85723	0.87490	0.86597
Military	0.96387	0.97838	0.97107
Sailing	0.95072	0.94405	0.94738
Dredging	0.96486	0.97852	0.97164
Pleasure	0.96388	0.96201	0.96295
SAR	0.99268	0.99299	0.99283
Pilot	0.91491	0.90887	0.91188
Towing	0.86471	0.88716	0.87579
Reserved	0.99896	0.99928	0.99912
Law_enforcement	0.97808	0.94057	0.95896
Towing_long_wide	0.95590	0.87545	0.91391
HSC	0.98834	0.99299	0.99066
Port_tender	0.94639	0.95798	0.95215
Diving	0.97736	0.99977	0.98844
Anti-pollution	0.99900	0.99860	0.99880
Spare_1	0.99991	0.99910	0.99950
WIG	0.99995	0.99973	0.99984
Vidurkis	0.96079	0.96060	0.96056
Tikslumas			0.96060

Skyriaus išvados. Šiame skyriuje aprašytos Klaipėdos ir Fehmarbelt regionų, bei meteorologinių duomenų aibės. Neapdorotos duomenų aibės sudarė tokius vektorių kiekius: Fehmarnbelt - 98245370, Klaipėda - 642541 ir meteorologiniai duomenys - 20608. Kad būtų išspręsta trūkstamų laivų tipų reikšmių vektoriuose problema, buvo sukurtas klasifikatorius, paremtas daugiamačiu daugiasluoksniu LSTM neuroniniu tinklu. Šis, tinklo pagrindu apmokytas modelis, klasifikuoja laivo tipą pagal jo eis-

1.4 lentelė: Paruošti galutiniai laivų eismo duomenys

Laivų tipai	Laivų skaičius	Sekos				Viso vektorių
		Mokymosi	Validavimo	Testavimo	Viso	
Cargo	1944	75625	25208	25209	126042	12604200
Tanker	645	22577	7526	7526	37629	3762900
Fishing	83	15748	5249	5250	26247	2624700
Passenger	80	47988	15996	15996	79980	7998000
Tug	71	9421	3140	3141	15702	1570200
Military	59	4743	1581	1581	7905	790500
Sailing	57	6692	2231	2231	11154	1115400
Dredging	42	4584	1528	1528	7640	764000
Pleasure	35	1965	655	655	3275	327500
SAR	32	3704	1235	1235	6174	617400
Pilot	23	6352	2118	2118	10588	1058800
Towing	14	522	174	175	871	87100
Reserved	14	436	146	146	728	72800
Law_ enforcement	13	4467	1489	1489	7445	744500
Towing_ long_wide	12	367	123	123	613	61300
HSC	9	32	11	11	54	5400
Port_tender	6	272	91	91	454	45400
Diving	5	103	35	35	173	17300
Anti-pollution	2	1094	365	365	1824	182400
Spare_1	1	15	5	6	26	2600
WIG	1	40	13	14	67	6700
Viso	3148	206749	68917	68925	344591	34459100

mo trajektoriją, o preciziškumas klasių atžvilgiu yra daugiau negu 0.96. Šio modelio panaudojimas leido padidinti apsimokymo duomenų aibę 4234160 navigaciniais vektoriais iš Fehmarnbelt duomenų aibės. Buvo pasiūlyta tinkama duomenų paruošimo schema: duomenų struktūrizavimas, neapdorotų duomenų valymas, atranka, trūkstamų verčių užpildymas, požymių inžinerija, suskaidymas į sekas ir trūkstamo laivų tipo atpažinimas bei užpildymas. Po duomenų valymo duomenų aibės sudarė: Fehmarnbelt - 34459100, Klaipėda - 232093.

Tyrimas vadovaujasi prielaida, kad trajektorijos anomalija analizuojama vidutiniame intervale, kai vidutiniškai laivo trajektorijos atkarpa užima 20% tiriamojo regiono.

3 Taškiniai dalinai prižiūrimo mokymosi metodai laivų eismo anomalijoms aptikti

Šiame poskyryje aprašyti taškais paremti dalinai prižiūrimo mokymo jūrų laivų eismo anomalijų aptikimo metodai ir algoritmai. Jų pagrindą sudaro SOM klasterizavimo algoritmas, integruotas su virtualiu feromonu [22], ir SOM, integruotas su Gauso mišinio modeliu [25]. Skyriuje detaliai aprašomi šie metodai, jų parametrų parinkimas. Šie tyrimai publikuoti Venskus et al. [22, 23].

Jūrų eismo anomalijos aptikimas naudojant į SOM tinklą integruotą virtualų feromoną. Disertacijoje sukuriamas naujas neįprasto eismo aptikimo metodas, paremtas SOM integracija su virtualiu feromonu. Integruoto SOM metodo mokymosi procesas yra toks pat kaip ir klasikinio SOM, išskyrus tai, kad virtuali feromono intensyvumo vertė įvedama pasibaigus paskutinei epochai. Atsižvelgiant į tam pačiam klasteriui priskirtų vektorių skaičių, galima įvertinti pradinę virtualaus feromono reikšmę. Ši reikšmė padeda įvertinti, kaip šis klasteris atstovauja duomenų daugumai.

Pradžioje, kiekvieno laimėjusio neurono feromono vertės intensyvumas yra lygus šio neurono klasterio dydžiui. Feromonų intensyvumo Q vertė apskaičiuojama taip: pasirinkus laimėjusį neuroną, konkretaus neurono feromono intensyvumas padidinamas vienetu.

Laivo eismas laikomas įprastu ar neįprastu priklausomai nuo to, į kokį SOM klasterį pateko ir kokia yra jo feromono reikšmė. Jei didesnė nei slenkstis, eismas klasifikuojamas kaip normalus, o jei mažesnis - tada kaip neįprastas.

Feromonų intensyvumo slenkstinė reikšmė, naudojama nenormaliam judesiui aptikti, apskaičiuojama naudojant validavimo duomenų rinkinį. Algoritmo tikslumo ir jautrumo maksimalias reikšmes galima derinti keičiant slenksčio vertę.

Jūrų eismo anomalijos aptikimas naudojant į SOM tinklą integruotą Gauso suminį mišinį. Šis anomalijų aptikimo metodas (SOM_GMM) yra SOM ir Gauso mišinio modelių (GMM) derinys. Šio modelio apmokymas apibūdinamas šias veiksmais:

- *Turimų duomenų rinkinių padalijimas.* Turimi laivo eismo duomenys iš dominančios srities yra suskirstyti į tris rinkinius: 50% - mokymosi duomenų rinkinys, skirtas SOM mokymui, 30% - validavimo / derinimo duomenų rinkinys, skirtas apskaičiuoti feromono intensyvumo slenkstį, ir 20% - testavimo duomenų rinkinys klasifikavimo rezultatams įvertinti.

- *Apsimokymo duomenų rinkinio normalizavimas.* Kiekvienas duomenų rinkinio požymis normalizuojamas į [0,1] intervalą naudojant Min-Max metodą.

- *SOM tinklo apmokymasis.* SOM mokymosi procesui įtaką daro keli parametrai: tinklelio forma yra stačiakampė, mokymosi greitis nustatytas į 0.5. Naudojama Gauso kaimynystės funkcija. Tiek pradinio kaimyninio spindulio, tiek spindulio slopinimo parametrai nustatyti –0.1.

- *Kovariacijos matricos skaičiavimas.* Kiekvienam SOM neuronui apskaičiuojama visų įvesties vektorių, atitinkančių konkretų klasterį, kovariacijos matrica.

- *Tikimybių skaičiavimas.* Kiekvienam SOM klasteriui buvo apskaičiuota n dimensijų Gauso tikimybės tankio funkcija. Kiekvienos tankio funkcijos vidurkis atitinka SOM neuronų vektoriaus svorį, o dispersija suteikiama pagal apsimokymo duomenis atitinkančius konkrečius klasterius.

- *GMM apskaičiavimas.* GMM apskaičiuojamas susumavus visus kiekvieno SOM klasterio Gauso skirstinius.

- P(H = normalus) tikimybės vertės parinkimas validavimo duomenų rinkinyje.

SOM pakartotinio mokymosi strategijos. Šioje disertacijoje pateikiamos dvi neuroninio tinklo pakartotinio mokymosi strategijos. Rezultatų tyrimą ir palyginimą su standartine neuroninio tinklo modelio eksperimentinio tyrimo procedūra (vadinamąja strategija I) Venskus *et al.* pateikia straipsnyje [23].

- *Strategija I* pateikia duomenų grupavimą ir algoritmų mokymą taip, kad kai tik atsiranda papildoma duomenų aibė, SOM apmokomas nuo pradžių, tarsi nebūtų buvęs apmokytas anksčiau. Tai yra bendras požiūris į neuroninių tinklų mokymą / validavimą / testavimą. Jis naudojamas kaip atspirties taškas norint palyginti II ir III pakartotinio mokymosi strategijas, kurios buvo pristatytos straipsnyje [23].

- *Strategija II* pateikia tokį metodą, kai SOM tinklas yra apmokomas naudojant anksčiau apmokytus modelio parametrus su ankstesniais duomenims. Naudojant šiuos pradinius parametrus tinklas pakartotinai mokinasi naudojant naujai gautus duomenis.

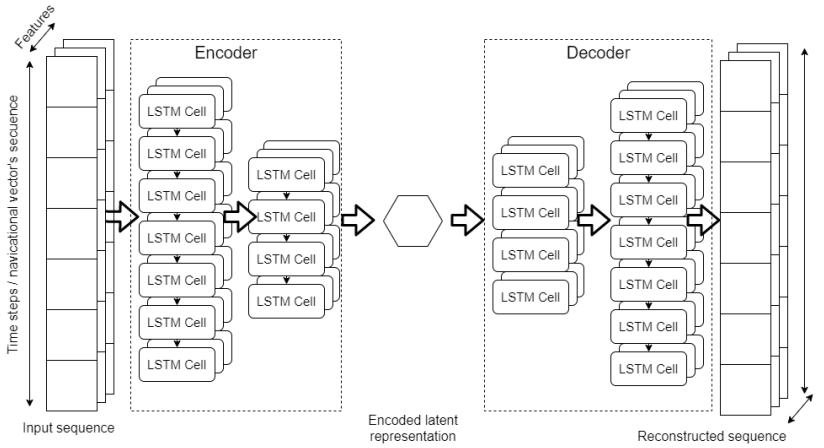
- *Strategija III* pateikia skirtingus duomenų rinkinių maišymo būdus, gautus skirtingais laikais. SOM tinklas apmokomas naudojant anksčiau apmokytus neuronų svorius. Apmokymas vykdomas su nauja duomenų imtimi, į kurią įmaišomą dalis anksčiau gautų duomenų.

Visos trys strategijos ištirtos dėl mokymosi greičio parametrų įtakos modelio tikslumui ir apmokymo trukmei. Iš laivų perduodamus duomenis galima vertinti kaip nuolatinį srautą, kuris pateikiamas tinklui pakartotiniam apsimokymui pagal faktinį jų gavimą.

4 Nepriž iūrimo mokymosi anomalių laivo trajektorijų aptikimo metodai

Jūrų laivų eismo trajektorijos prognozavimas. Laivo trajektorijos prognozavimui taikomas nepriž iūrimo mokymosi gilusis neuroninis tinklas. Giliojo neuroninio tinklo įvestis yra ankstesnio konkretaus laivo navigacijos trajektorijos / sekos duomenys, po to tinklas apskaičiuoja tolesnės laivo padėties / trajektorijos prognozę. Jei algoritmo gautas spėjimas patenka į nustatytą ribą, laivo eismas klasifikuojamas normaliu, kitu atvejų - neįprastu.

Šioje disertacijoje atlikti laivų eismo trajektorijos prognozavimą siūloma naudojant daugiamatį daugiažingsnį LSTM autoenkoderį, kurio architektūra pateikta paveiksle 1.2. Pagrindinės siū-



1.2 pav.: LSTM autoenkoderio architektūra

lomo LSTM autoenkoderio dalys yra: įvesties sluoksnis, enkoderio sluoksniai, užkoduotas latentinis vektorius, dekoderio sluoksniai ir išvesties sekos sluoksnis. Įvesties sluoksnis gauna struktūrizuotas navigacijos vektorių sekas χ ir grąžina prognozuojamas sekas $\hat{Y}$. Kodavimo ir dekoderio dalys susideda iš LSTM ląstelių, kurios yra nuosekliai sujungtos tam tikro bloko viduje ir yra

lygiagrečios tarp viršuje sukrautų sluoksnių.

Autoenkoderio prognozavimo klaida apskaičiuojama:

$$e^{(l)}_{(g,r,j)} = Y_{(g,r,j)} - \hat{Y}^{(l)}_{(g,r,j)};$$
$$l \in \{upper, crisp, lower\}; \quad g \in \{1, 2, \ldots, N\};$$
$$r \in \{1, 2, \ldots, \widetilde{n}\}; \quad j \in \{1, 2, \ldots, f\}; \tag{1.1}$$

kur $e^{(l)}_{(g,r,j)}$ yra prognozavimo klaida vieno navigacinio vektoriaus požymiui, $Y_{(g,r,j)}$ - tikroji navigacinio vektoriaus požymio reikšmė, $\hat{Y}^{(l)}_{(g,r,j)}$ - prognozuojama navigacinio vektoriaus požymio reikšmė, l - tinklo tipas: crisp, lower, upper (aprašytas kitame paragrafe), g - navigacinio vektorių sekos indeksas duomenyse, r - navigacinio vektoriaus padėties indeksas sekoje, $\widetilde{n}$ - prognozuojamos išvesties sekos ilgis, j - j^{th} navigacinio vektoriaus požymis, f - požymių skaičius.

Nuostolių funkcija apibrėžta formulėje 1.2 naudojama „crisp“ tipo modelyje:

$$L^{(l)}_s = \frac{1}{N\widetilde{n}f} \sum_{g=1}^{N} \sum_{r=1}^{\widetilde{n}} \sum_{j=1}^{f} (e^{(l)}_{(g,r,j)})^2, \quad l = \{upper, crisp, lower\}, \tag{1.2}$$

kur L^l_s yra bendroji nuostolių funkcija naudojama ir „upper“, „crisp“ ir „lower“ tipo modeliuose, N - apsimokymo sekų skaičius, s - indeksas žymintis bendrąją nuostolio funkcijos dalį.

LSTM spėjimo regiono apmokymas. Trys skirtingai sukonfigūruoti LSTM autoenkoderiai naudojami norint išmokyti modelius daugiamačio prognozavimo regiono ribų. „Crisp“ ($l = \{crisp\}$) modelio tipas prognozuoja tikslias geografines laivo trajektorijos koordinates. Apatinės ribos ($l = \{lower\}$) modelio tipas numato apatinę „crisp“ tipo modelio prognozavimo regiono ribą. Viršutinės ribos ($l = \{upper\}$) modelio tipas prognozuoja viršutinę „crisp“ tipo modelio prognozavimo regiono ribą. Kartu apatinių ir viršutinių ribų modeliai prognozuoja „crisp“ tipo modelio prognozavimo regioną.

Specifinė viršutinių ir apatinių ribų nuostolių funkcija apibrėžiama taip:

$$L_{\ell}^{(upper)} = \frac{1}{N\widetilde{n}f}\sum_{g=1}^{N}\sum_{r=1}^{\widetilde{n}}\sum_{j=1}^{f}(ReLU(e_{(g,r,j)}^{(upper)}))^2, \tag{1.3}$$

$$L_{\ell}^{(lower)} = \frac{1}{N\widetilde{n}f}\sum_{g=1}^{N}\sum_{r=1}^{\widetilde{n}}\sum_{j=1}^{f}(ReLU(-e_{(g,r,j)}^{(upper)}))^2, \tag{1.4}$$

kur $L_{\ell}^{(upper)}$ ir $L_{\ell}^{(lower)}$ yra specifinės nuostolių funkcijų dalys, atitinkamai viršutinei ir apatinei riboms, ℓ - indeksas žymintis specifinę nuostolio funkcijos dalį, $ReLU$ yra ištaisyta linijinio vieneto funkcija, apibrėžta:

$$ReLU(x) = \begin{cases} 0, & \text{for} \quad x < 0 \\ x, & \text{for} \quad x \geq 0. \end{cases} \tag{1.5}$$

Pilnutinės nuostolių funkcijos apibrėžiamos:

$$L_{total}^{(upper)} = L_{s}^{(upper)} + \lambda L_{\ell}^{(upper)}, \tag{1.6}$$

$$L_{total}^{(lower)} = L_{s}^{(lower)} + \lambda L_{\ell}^{(lower)}, \tag{1.7}$$

kur $L_{total}^{(upper)}$ yra pilnutinė viršutinės ribos nuostolio funkcija, $L_{total}^{(lower)}$ yra pilnutinė apatinės ribos nuostolio funkcija, λ yra derinamas parametras, kuris nustato specifinės funkcijos įtakos pilnajai funkcijai lygmenį [26].

„Crisp“ modelio pilnoji nuostolio funkcija:

$$L_{total}^{(crisp)} = L_{s}^{(crisp)}, \tag{1.8}$$

Su šiomis nuostolių funkcijomis pasiekiamas prognozavimo regiono minimizavimas. Jei funkcijos netaikomos, prognozavimo srities nuostolių funkcijos (L_i) padidina regiono plotą ir sukuria kompromisą tarp taškų, patenkančių į regioną, skaičiaus ir jo srities, kurią galima reguliuoti keičiant parametrą λ (1.6) ir (1.7).

λ pasirinkimas atliekamas iteratyviai didinant iki pageidaujamos reikšmės.

Siekiant įvertinti prognozavimo regiono kokybę, buvo naudojami du rodikliai. Pirmoji yra prognozavimo srities aprėpties tikimybė (PICP), kuri kiekybiškai įvertina išmatuotų verčių, patenkančių į modelio apibrėžtą regioną, skaičių [26] ir yra modifikuota, kad būtų palaikomos daugiamatės ir daugiažingsnės prognozės:

$$PICP = \frac{1}{N\widetilde{n}f} \sum_{g=1}^{N} \sum_{r=1}^{\widetilde{n}} \sum_{j=1}^{f} (\delta_{(g,r,j)}) \tag{1.9}$$

$$\delta_{(g,r,j)} = \begin{cases} 1 & \text{if} \quad Y_{(g,r,j)} \in \left[\hat{Y}^{(lower)}_{(g,r,j)}, \hat{Y}^{(upper)}_{(g,r,j)}\right] \\ 0 & \text{if} \qquad \text{otherwise.} \end{cases} \tag{1.10}$$

Antroji metrika yra „Prediction Interval Normalized Average Width“ (PINAW), naudojama matuojant prognozės regiono plotą [26]. PINAW taip pat buvo modifikuotas daugiamatėms ir daugiažingsnėms prognozėms:

$$PINAW = \frac{1}{N\widetilde{n}fR} \sum_{g=1}^{N} \sum_{r=1}^{\widetilde{n}} \sum_{j=1}^{f} (\hat{Y}^{(upper)}_{(g,r,j)} - \hat{Y}^{(lower)}_{(g,r,j)}) \tag{1.11}$$

kur R yra maksimalus skirtumas tarp $\max(\hat{Y}^{(upper)}_{(g,r,j)} - \hat{Y}^{(lower)}_{(g,r,j)})$ duomenų aibėje [26], [27].

LSTM „wild bootstrapping“ spėjimo regionas. Vienas iš pagrindinių „wild bootstrapping“ nustatymo metodo privalumų yra tas, kad nereikia daryti jokių prielaidų dėl tiriamo duomenų rinkinio paskirstymo. Tradiciškai „bootstrap“ metodas pakartotinai ima pradinius duomenis, kad gautų daugiau duomenų pavyzdžių, kuriuos būtų galima naudoti pakartotiniuose eksperimentuose. Vis dėlto „wild bootstrapping“ technika yra šiek tiek kitokia. Užuot generavę sąryšio skirsnio atvejus, kurie susideda

iš naujo atrenkant pradinius duomenis ar liekanas, „wild bootstrapping“ elementai sujungia duomenis su atsitiktiniais kintamaisiais, paimtais iš žinomo skirsnio, kad sudarytų sąryšio skirsnio pavyzdį. Šio metodo naudojimą disertacijoje galima apibendrinti taip:

1. Duomenų paruošimas.
2. Kiekvieno duomenų rinkinio dispersijos apskaičiavimas.
3. Atsitiktinių kintamųjų generavimas išlaikant tą patį matmenį ir vidurkį lygų nuliui, o dispersija - tokia pati kaip įvesties duomenų.
4. Elementinis pradinių duomenų rinkinio sumavimas su naujai sukurtu rinkiniu, t. y. triukšmas pridedamas prie duomenų su vidurkiu ir dispersija, apskaičiuota pagal pradinį duomenų rinkinį.
5. Gautų duomenų normalizavimas į intervalą $[0, 1]$, siekiant geresnių LSTM apsimokymo rezultatų, išlaikant kiekvienos funkcijos mastelio koeficientus numatomam duomenų atkūrimo tikslui.
6. LSTM automatinio kodavimo tinklo mokymas.
7. LSTM tinklo prognozių skaičiavimas r-žingsnių į priekį, $r \in \{1, 2, \ldots, \widetilde{n}\}$ ($\widetilde{n} = 50$).
8. Prognozių mastelio atkūrimas, t. y. didinti numatomas vertes pagal išsaugotos funkcijos mastelio parametrus, aprašytus dviejuose aukščiau nurodytuose veiksmuose.
9. Kartojami žingsniai 3-8 k-kartų (disertacijoje tai atliekama 100 kartų).

Pritaikius schemą, kaip siūloma aukščiau, gaunama matrica su numatomomis reikšmėmis. Tada prognozuojamo taško reikšmė kiekvienam požymiui, esančiame kiekvienam prognozavimo žingsnyje, parenkamas vidutinė prognozės reikšmė iš k pakartojimų

vektoriaus. Tai $100(1-\alpha)\%$ prognozavimo regionas, skirtas vidutiniam (vidutinei prognozuojamai vertei) p - dimensijos normalaus pasiskirstymo atžvilgiu yra elipsoidas, nustatytas nežinomam μ (žr. [28]):

$$\frac{kr}{k+r}(\bar{x}_r - \mu)^T \hat{S}^{-1}(\bar{x}_r - \mu) \leq \frac{(k-1)p}{k-p} F_{p,k-p}(1-\alpha), \quad (1.12)$$

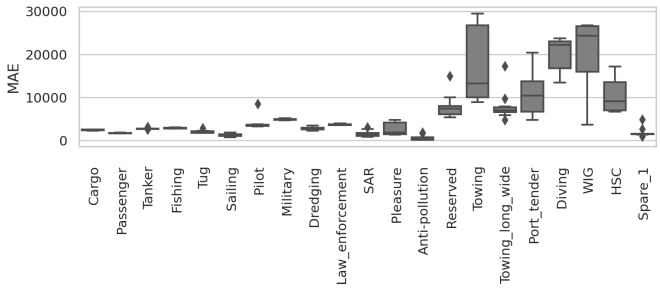
kur

- $\bar{x}_r = \frac{1}{k}\sum_{u=1}^{k} x_{u,r,j}$ - vidurkių vektorius kiekvienam požymiui $j \in \{1 \dots f\}$ kiekvienam spėjimo žingsniui r,
- $\hat{S}$ - kovarijacinė matrica,
- $F_{p,k-p}(1-\alpha)$ yra $1-\alpha$-lygmens kritinė reikšmė pagal Fisher skirsnį su p ir $k-p$ laisvės laipsniams.

5 Ekperimentai ir rezultatai

Šiame skyriuje pateikiama eksperimentų ir rezultatų serija, skirta palyginti siūlomų jūrų eismo anomalijų nustatymo metodų veikimą. Skyriuje aprašomi neprižiūrimo ir dalinai prižiūrimo mokymo anomalijų aptikimo algoritmų taikymo rezultatai, supažindinama su anomalių trajektorijų grupavimu ir aprašomos siūlomų metodų stipriosios ir silpnosios pusės.

LSTM spėjimo regiono mokymosi metodo vertinimas. Paveiksle 1.3 pavaizduotos LSTM „crisp“ tipo modelio prognozių klaidos. Paklaidos apskaičiuotos naudojant testavimo duomenų rinkinius. Pastebima, kad testavimo duomenų rinkinių paklaidos nėra žymiai didesnės negu validavimo duomenų imties. Tai rodo, kad modelio apibendrinimas yra tinkamas tolesniam naudojimui aptikti jūrų eismo anomalijas. Modelių paklaidos sudaro dvi grupes: vienose yra daugiau mažų paklaidų, kitose - reikšmingesnių paklaidų. Mažesnės paklaidos vertės yra „Antipollution“, „Cargo“, „Passenger“, „Tug“ ir kitų laivų tipų. Šie

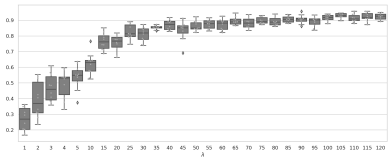


1.3 pav.: MAE paklaidų dispersijos skirtingiems laivų tipams

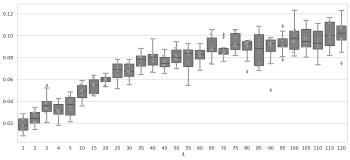
laivo tipo modeliai buvo mokomi naudojant didesnį navigacinių vektorių sekų skaičių (N), palyginti su modelių grupe, turinčia reikšmingesnių paklaidų, tokių kaip „Diving“, „HSC“, „Towing“, „WIG“, ir kt. (Paveikslas 1.3). Paveiksle 1.3 pavaizduota, kad ta pati grupė, turinti reikšmingesnių paklaidų, taip pat turi didesnę paklaidų dispersiją, kai modelis treniruojamas skirtingais atsitiktiniais pradiniais svoriais. Panaši priklausomybė matoma ir kitų tipų paklaidose.

Buvo apmokyta aibė viršutinės ir apatinės ribos modelių porų su skirtingomis λ reikšmėmis, kai $\lambda_{start} = 5, \Delta\lambda = 5, \lambda_{stop} = 120$. Visoje λ reikšmių aibėje kiekvienam laivo tipui buvo apmokyti atskiri neuroniniai tinklai. Kiekviena viršutinė / apatinė modelio pora buvo apkartotinai apsimokyta 10 kartų, kad būtų ištirta atsitiktinių pradinių tinklo svorių įtaka. Kiekvienam laivo tipui buvo sudarytos dvi stulpelinės diagramos: viena - λ įtakai PICP reikšmei, kita - λ įtakai PINAW reikšmei. Dvi laivo tipo stulpelinės diagramos pateikiamos paveiksluose 1.4a ir 1.4b. Dažniausiai PICP auga logaritmiškai artėjant link 1 reikšmės, kai λ auga tiesiškai. Konkrečiai λ reikšmei gaunamos skirtingos PICP reikšmės, kurios apibrėžiamos dispersija. Tai vyksta, nes neuroninis tinklas yra apmokomas skirtingomis pradinėmis neuronų reikšmėmis. Šis elgesys rodo, kad modelio anomalijai, turinčiai atitikti konkrečią PICP vertę, turime ieškoti atitinkamų λ iteratyviai.

Paveiksluose 1.5 pavaizduoti neįprasti laivų eismo atvejai (pa-



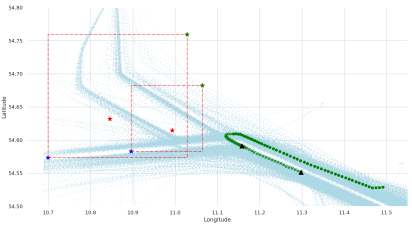
(a) PICP pagal λ



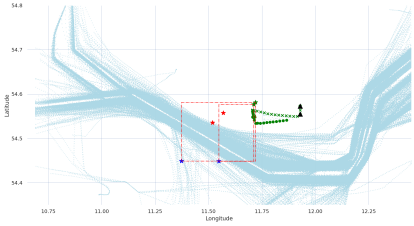
(b) PINAW pagal λ

1.4 pav.: λ reikšmės įtaka PICP ir PINAW naudojant „Cargo“ laivo tipo duomenis

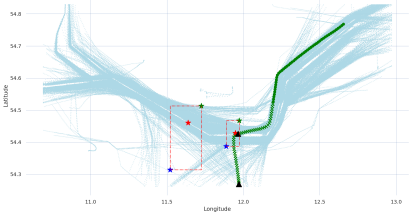
gal LSTM prognozavimo regiono mokymosi metodą). Paveiksle 1.5a pavaizduotas neįprastas atvejis, kai krovininis laivas netikėtai apsisuko, keisdamas kryptį 180 laipsnių, dėl sprendimo grįžti į uostą remontuoti variklių. Pirmieji 50 laivo navigacinių vektorių buvo naudojami kaip modelio įvestis, o modelis prognozavo spėjimo regioną, kur tikimasi tikrosios laivo padėties. Kadan-



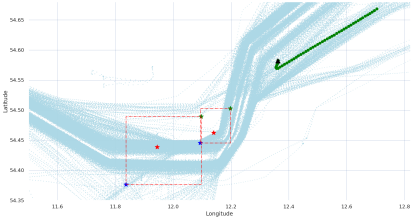
(a) Staigus krypties pakeitimas



(b) Eismas už farvaterio ribų



(c) Krypties pakeitimas



(d) Netikėtas sustojimas

1.5 pav.: Ivairių tipų neįprasto laivų eismo pavyzdžiai

gi laivas smarkiai pakeitė kryptį, tikrosios laivo pozicijos vertės

buvo už numatytų regionų ribų. Tas pats paveikslas 1.5a rodo, kad tikrosios 25 ir 50 laivų pozicijos (juodieji trikampiai) yra už prognozavimo regionų ribų pagal LSTM regiono prognozavimo metodą (raudoni stačiakampiai). Taigi tokį laivo judėjimą galima priskirti neįprastam, nes jis nepatenka į $PICP = 0.95$ prognozavimo regioną. Paveiksle 1.5b pavaizduotas dėl sugedusio variklio dreifuojantis krovininis laivas. Tikrieji jo navigaciniai vektoriai taip pat yra už prognozavimo srities ribų. Paveiksluose 1.5c ir 1.5d rodomi keli kiti neįprasto eismo atvejai: (c pav.) Parodytas netikėtas posūkis ir (d pav.) - neplanuotas sustojimas dėl variklio gedimo.

LSTM „wild bootstrapping“ metodo vertinimas. Po LSTM „wild bootstrapping“ modelio apsimokymo buvo įvertintos gautos PICP reikšmės pageidaujamam spejimo regiono dydžiui $100(1-\alpha) = 95.00\%$. Lentelėje 1.5 yra kiekvieno laivo tipo mokymosi ir testavimo duomenų rinkinių PICP rezultatai. Lentelėje parodyta, kad tokių laivų tipų kaip „Cargo“, „Dredging“, „Fishing“, „Law_enforcement“, „Military“, „Passenger“, „Pilot“, „Pleasure“, „SAR“, „Sailing“, „Tanker“, ir „Tug“ PICP reikšmės yra beveik tokios pačios kaip prognozavimo regionas $100(1-\alpha)$. Šis atvejis susijęs su tuo, kad šių laivų tipai, palyginti su kitais laivais, turi didesnes apsimokymo sekas. Kiti laivų tipai turi ženkliai mažesnę PICP reikšmę nei norimas prognozavimo regionas.

Taškinių dalinai prižiūrimo mokymosi metodų vertinimas SOM_pheromone ir SOM_GMM metodai ištirti naudojant dvi skirtingas laivų eismo duomenų imtis: Klaipėdos ir Fehmarnbelt [22, 23]. Pagrindiniai šių duomenų rinkinių skirtumai yra jūrų eismo intensyvumas ir sudėtingumas, o taip pat ir navigacijos vektorių kiekis, kuris turi būti išanalizuotas algoritmais. Po duomenų paruošimo jų kiekis yra: Klaipėdos „Cargo“ tipo laivai yra sudaro 138242, o Fehmarnbelt „Cargo“ - 12604200. Kitas svarbus aspektas yra tai, kad Klaipėdos duomenų aibėje yra ekspertų pagalba sužymėtos laivų eismo anomalijos. Šis žymėjimas naudojamas norint sureguliuoti SOM_pheromone β_{PPV}, β_{TPR} ir

1.5 lentelė: PICP reikšmės pageidaujamam $100(1-\alpha) = 95.00\%$ prognozavimo regionui

	Vektorių Sekos		$PICP, \%$	
Laivo tipas	Mokymosi	Testavimo	Mokymosi	Testavimo
Anti-pollution	1094	365	49.11	49.14
Cargo	75625	25209	94.78	97.51
Diving	103	35	0.01	0.00
Dredging	4584	1528	89.30	89.21
Fishing	15748	5250	94.34	89.66
HSC	32	11	3.49	0.91
Law_enforcement	4467	1489	81.30	81.62
Military	4743	1581	82.72	82.36
Passenger	47988	15996	96.59	96.39
Pilot	6352	2118	82.69	82.89
Pleasure	1965	655	87.21	86.81
Port_tender	272	91	0.01	0.00
Reserved	436	146	0.07	0.02
SAR	3704	1235	95.90	96.15
Sailing	6692	2231	97.93	98.14
Spare_1	15	6	27.50	7.42
Tanker	22577	7526	96.42	96.84
Towing	522	175	1.55	0.01
Towing_long_wide	367	123	0.61	0.01
Tug	9421	3141	94.55	94.54
WIG	40	14	1.11	0.00

SOM_GMM $P(H = normal)$ parametrus, siekiant maksimaliai padidinti anomalijų aptikimo tikslumą. Kita vertus, Fehmarnbelt duomenų rinkinys yra didžiulis ir visų tipų laivai sudaro 34459100 navigacijos vektorių, kurie priklauso visiems laivų tipams. Tampa akivaizdu, kad ekspertui anotuoti duomenis praktiškai neįmanoma. Dėl to, norint ištirti dalinai prižiūrimo mokymosi taškinius metodus naudojant Fehmarnbelt duomenis, šios duomenų aibės eismo anomalijos yra anotuojamos pritaikant LSTM prognozavimo regiono mokymąsi ir LSTM „wild bootstraping“ metodus fiksuotu anomalijos lygiu $(1-\alpha) = 0.95$ ir šios anomalijų anotacijos naudojamos kaip atsparos taškas modelių

tyrimui.

Lentelėje 1.6 rodomi galutiniai SOM_pheromone metodo eks-

1.6 lentelė: SOM_pheromone ir SOM_GMM eksperimentų rezultatai naudojant Fehmarnbelt duomenų imtį

	Vektoriai ,$\times 10^2$		LSTM prognozavimo regiono metodas						
				SOM-pheromone				SOM-GMM		
Laivo tipas	mokymosi	testavimo	PICP	Grid	NF	PPV	TPR	Grid	PPV	TPR
Anti-pollution	1094	365	0.941	40x40	MH	0.901	0.774	35x35	0.886	0.740
Cargo	75625	25209	0.960	70x70	CG	0.675	0.667	60x60	0.856	0.634
Diving	103	35	0.835	35x35	MH	0.914	0.869	30x30	0.899	0.896
Dredging	4584	1528	0.951	70x70	CG	0.667	0.640	55x55	0.844	0.601
Fishing	15748	5250	0.944	60x60	CG	0.626	0.656	60x60	0.849	0.631
HSC	32	11	0.846	35x35	MH	0.914	0.947	30x30	0.904	0.936
Law-enforcement	4467	1489	0.955	70x70	CG	0.638	0.555	55x55	0.857	0.600
Military	4743	1581	0.953	70x70	CG	0.667	0.700	60x60	0.842	0.586
Passenger	47988	15996	0.959	70x70	CG	0.671	0.617	60x60	0.806	0.550
Pilot	6352	2118	0.959	70x70	CG	0.658	0.683	40x40	0.859	0.648
Pleasure	1965	655	0.943	60x60	CG	0.660	0.601	50x50	0.846	0.635
Port-tender	272	91	0.487	35x35	MH	0.898	0.761	35x35	0.898	0.748
Reserved	436	146	0.947	30x30	MH	0.908	0.880	30x30	0.903	0.893
SAR	3704	1235	0.959	60x60	CG	0.612	0.605	55x55	0.848	0.619
Sailing	6692	2231	0.952	70x70	CG	0.662	0.602	60x60	0.855	0.629
Spare_1	15	6	0.903	30x30	MH	0.911	0.879	30x30	0.909	0.862
Tanker	22577	7526	0.946	70x70	CG	0.603	0.573	55x55	0.859	0.670
Towing	522	175	0.944	30x30	MH	0.890	0.788	30x30	0.901	0.844
Towing-long-wide	367	123	0.925	30x30	MH	0.913	0.885	30x30	0.907	0.892
Tug	9421	3141	0.955	60x60	CG	0.641	0.654	50x50	0.839	0.644
WIG	40	14	0.726	30x30	MH	0.916	0.937	30x30	0.911	0.940

Sutrumpinimai lentelėje: NF - kaiminystės funkcija; PPV - Preciziškumas; TPR - Jautrumas; MH - „Mexican hat“; CG - „Cut Gaussian“;

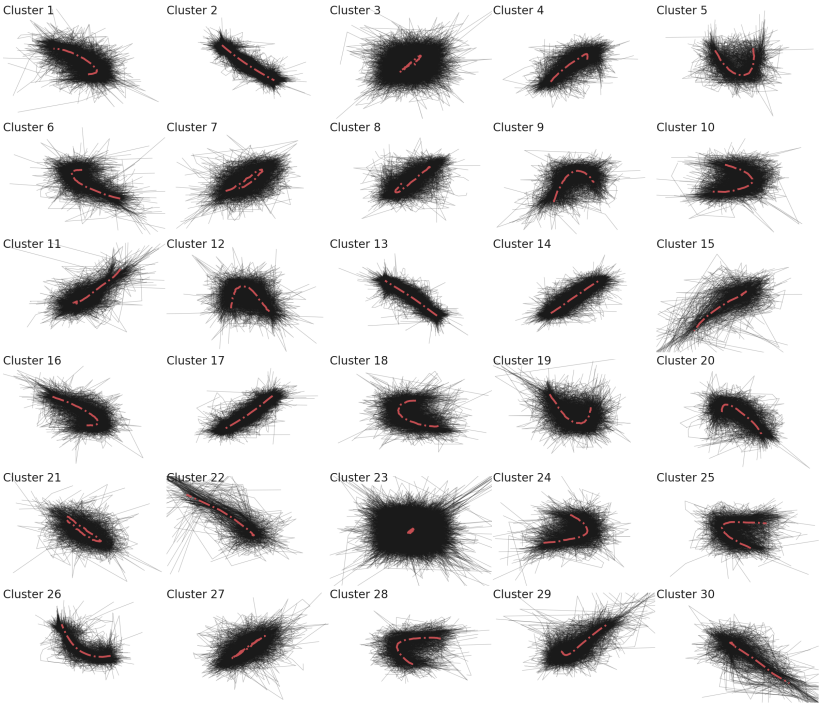
perimentų rezultatai, kurie buvo atlikti naudojant Fehmarnbelt duomenų rinkinį, sugrupuotą pagal konkretų laivo tipą. Tiriant SOM_pheromone ir SOM_GMM metodus buvo tiriama: tikslumas, jautrumas, tiksliausi kaimynystės funkcija ir SOM tinklelio dydis. Siūlomos pakartotinio mokymosi strategijos taip pat parodė gerus rezultatus. III strategija užėmė tik 0.671 I strategijos skaičiavimo laiko, o tikslumas sumažėjo vidutiniškai 0.007 ir 0.009. III strategijos mokymosi greičio parametras yra nuo 0.03

iki 0.04. Taigi Fehmarnbelt duomenų rinkinio rezultatai rodo, kad perkvalifikavimo strategijas galima pritaikyti siekiant sumažinti pakartotinio mokymosi laiką, išlaikant jautrumą ir tikslumą SOM_pheromone ir SOM_GMM algoritmų modifikacijoms.

Be to, pastebėta, kad didesniems duomenų rinkiniams, pvz., „Cargo“, „Passenger“ ir „Tanker“, reikia žymiai didesnių SOM tinklelių dydžių (70×70), palyginti su mažesniais duomenų rinkiniais, tokiais kaip „Diving“, „HSC“, „Towing“ ir kt. (35×35) (žr. lentelę 1.6). Be to, „Cut Gaussian“ kaimynystės funkcija geriau veikia didesniuose duomenų rinkiniuose, tačiau „Mexican hat“ kaimynystės funkcija geriau veikia esant mažesniems duomenų kiekiams. SOM_GMM reikia mažesnių SOM tinklelio dydžių nei SOM_Pheromone (žr. SOM_pheromone ir SOM_GMM Grid stulpelius lentelėje 1.6). Be to, matome, kad SOM_GMM yra mažiau jautrus tinklo dydžio pokyčiui ir yra tikslesnis didesniuose duomenų rinkiniuose. Tačiau SOM_pheromone metodas geriau veikia mažesnius duomenų rinkinius (pvz.: „Anti-pollution“, „Diving“). Taip pat galima pastebėti, kad abu SOM pagrįsti metodai turi mažas jautrumo vertes. Šis faktas rodo, kad tokie modeliai sukelia daug klaidingai neigiamų atvejų, kas rodo kad SOM pagrįsti metodai neaptinka tiek anomalijų, kiek LSTM metodai. Ši prielaida išsamiau patikrinta vėlesniame poskyryje.

Neįprasto eismo trajektorijų formų palyginimas. Šiame poskyryje atliekamas klaidingai neigiamų atvejų tyrimas. Kaip įrankis taikomas erdvės ir laiko eilučių klasterizavimo metodas. Cuturi ir Blondel [29] laiko eilučių duomenims klasterizuoti siūlo SoftDTW k-means algoritmą. Tyrime naudojamas siūlomo algoritmo daugiamatis variantas. Paveiksle 1.6 pavaizduoti klasterizuoti laivų navigacinių vektorių sekų rinkiniai iš klaidingai neigiamų anomalijų. Konkrečiam klasteriui priskirtos trajektorijos nuspalvintos juodai, o klasterio ašis - raudonai. Grupių skaičius buvo pasirinktas naudojant alkūnės metodą [30].

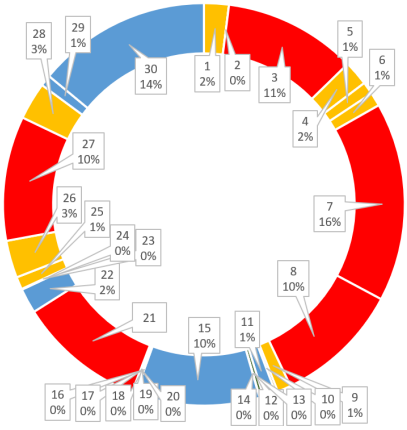
Paveikslėlyje 1.6 rodoma 30 „fishing“ tipo laivų trajektorijos klasterių. Galima pastebėti, kad raudonos linijos vaizduoja gana skirtingas trajektorijos formas. Be to, grupėms, kurių skaičius



1.6 pav.: Klasterizuotos žvejybinių laivų neįprasto eismo trajektorijos

yra 2, 13, 14 ir 17, būdingos tos pačios formos savybės, išskyrus skirtingas laivų judėjimo kryptis. Ši pastaba taikoma ir kitiems klasteriams. Taigi, atsižvelgiant į jūrinių laivų srities žinias ir formos pasisukimus, iš gautų 30 klasterių sudaromos 5 skirtingų tipų anomalių trajektorijų grupės:

- Tiesaus eismo formos grupė - sudaryta iš klasterių 2, 13, 14, 17 (Žalia).
- Sustojančio laivo trajektorijos formos grupė - 11, 15, 22, 29, 30 (Mėlyna).
- Švelnaus manevro trajektorijos formos grupė - 1, 4, 5, 6, 9, 10, 12, 16, 18, 19, 20, 24, 25, 26, 28 (Geltona).
- Aštraus manevro trajektorijos formos grupė - 3, 7, 8, 21,



1.7 pav.: Žvejybinių laivų anomalių trajektorijų klasterių grupavimas pagal SOM_GMM tariamai neigiamus rezultatus

27 (Raudona).

- Dreifuojančios trajektorijos formos grupė - 23 (Balta).

Paveiksle 1.7 parodytas klaidingai neigiamų trajektorijos formų pasiskirstymas tarp minėtų grupių. Galima pastebėti, kad „Aštraus manevro“ trajektorijos formos grupė (pažymėta raudonai) sudaro daugumą (57.00%). Antra pagal dydį yra „Sustojančio laivo“ trajektorijos formos grupė (pažymėta mėlyna spalva), kurios trajektorijos sudaro 27.9% nuo visų trajektorijų, o „Švelnaus manevro“ trajektorijos formos grupė sudaro 15.07%. Galiausiai „Tiesaus eismo“ ir „Dreifuojančios“ trajektorijų formų grupės sudaro atitinkamai 0.02% ir 0.005% klaidingai neigiamų atvejų. Panaši grupių tendencija pastebima didesniuose laivų tipuose, tokiuose kaip „Anti-pollution“, „Cargo“, „Dredging“, „Fishing“, „Law_enforcement“, „Military“, „Passenger“, „Pilot“, „Pleasure“, „SAR“, „Sailing“, „Tanker“ ir „Tug“.

6 BENDROSIOS IŠVADOS

1. SOM_pheromone ir SOM_GMM metodų tikslumai priklauso nuo duomenų kiekio. Pastebėta, kad naudojant duo-

menis iki maždaug 140 000 navigacijos vektorių, SOM_pheromone metodas aptinka nenormalų laivų eismą tiksliau nei SOM_GMM metodas.
Eksperimentai rodo, kad rekomenduojami SOM_pheromone parametrai yra šie: „Mexican Hat“ kaimynystės funkcija ir SOM tinklelio dydis nuo 35×35 iki 40×40.
Mažesniuose duomenų rinkiniuose vidutinis preciziškumas yra 0.911, o jautrumas - 0.801, palyginti su SOM_GMM preciziškumu 0.886 ir jautrumu 0.740. Didesniuose duomenų rinkiniuose, kuriuose yra daugiau nei maždaug 140000 navigacijos vektorių, SOM_GMM lenkia SOM_pheromone metodą. SOM_GMM vidutinis preciziškumas yra 0.859, o jautrumas - 0.648, lyginant su SOM_pheromone vidutiniu preciziškumu 0.675 ir jautrumu 0.640. Rekomenduojami parametrai yra „Cut Gaussian“ kaimynystės funkcija ir tinklelio dydis nuo 55×55 iki 60×60.

2. Siekiant aptikti neįprastą jūrų eismo judėjimą, įmanoma žymiai sutrumpinti modelio mokymosi laiką, tuo pačiu išlaikant modelio preciziškumą ir jautrumą aukštose vertėse. Šiam tikslui pasiekti taikoma SOM paremtų metodų pakartotinio mokymosi strategija, kai ankstesni apmokytų neuronų svoriai naudojami kaip pradinė padėtis pakartotinai mokant tinklą su naujai surinktu duomenų pogrupiu, sumaišant jį su istoriniais duomenimis ir koreguojant mokymosi greitį. Gauti rezultatai rodo, kad SOM_pheromone ir SOM_GMM tinklai galėtų būti pakartotinai apsimokyti sutrumpinant laiką beveik perpus, išlaikant preciziškumą ir jautrumą beveik vienodai aukštomis vertėmis. Siūloma pakartotinio mokymosi strategija užtruko tik 67.1% skaičiavimo laiko, reikalingo tai atliekant klasikiniu metodu, o preciziškumas nukrito nuo 0.007 iki 0.009. Siūlomos strategijos pakartotinio mokymosi greičio parametras yra nuo 0.03 iki 0.04.

3. Abiejų LSTM metodų rezultatai rodo, kad didesniems duomenų rinkiniams, kuriuose yra daugiau nei maždaug 140

000 navigacijos vektorių, PICP vertė yra artima iš anksto nustatytai $100(1-\alpha) = 95.0\%$ vertei ir yra nuo 94.1% iki 97.5%. Mažesniuose duomenų rinkiniuose, kuriuose yra mažiau nei apytiksliai 140 000 navigacijos vektorių, LSTM prognozavimo regiono mokymosi metodas vis tiek sugebėjo išmokti, nors ir siauresnius prognozavimo regionus nuo 48.7% iki 84.6%. LSTM „wild bootstrapping“ metodui nepavyko išmokti mažesnių duomenų rinkinių prognozavimo regionų, tai rodo 0 PICP vertė. Abu LSTM metodai, skirti neprižiūrimam prognozavimo regionų įvertinimui, gali būti naudojami neįprastam jūrų eismui aptikti, kai mokymosi duomenų rinkiniai yra didesni nei maždaug 140 000 navigacinių vektorių. Mažesniems duomenų rinkiniams rekomenduojama naudoti LSTM prognozavimo regiono mokymosi metodą su siauresniais prognozavimo regionais.

4. Norint išspręsti laivų tipo duomenų trūkumą, buvo sukurtas daugiasluoksnis daugiamatis LSTM klasifikatorius. Siūlomas modelis veikia tiksliai, vidutinis preciziškumas yra 0.96079, vidutinis jautrumas yra 0.96060, o „f1-score“ yra 0.96056. Klasifikavimo metrika rodo geras apibendrinimo savybes, leidžiančias atlikti trūkstamų duomenų užpildymą ir gauti 4.28% procentų duomenų, kuriems trūko laivo tipo reikšmių (iš viso 4234160 navigacijos vektorių) duomenų rinkinyje Fehmarnbelt.

5. Kai buvo naudojami LSTM metodais sužymėti neįprasto eismo duomenys, SOM grindžiamų metodų bandymo rezultatai rodė mažas jautrumo vertes (nuo 0.555 iki 0.700) dėl didelių klaidingai neigiamų verčių (angl. false negative). Šių klaidingai neigiamų rezultatų analizė leidžia daryti išvadą, kad staigaus posūkio ir stabdymo trajektorijos manevrų linijų formos dominuoja gautose klaidingai neigiamuose rinkiniuose ir sudaro vidutiniškai 57.0 % ir 27.9 % didesnėje duomenų grupėje.

SUMMARY

1 Research Context

The maritime logistics industry is a crucial component of the global trade economy with expanding volume, traffic intensity, and requirements. In Q1-Q3, 2019, 2,660 million tons gross weight of seaborne goods were handled in EU-27 main ports [1]. That is 7% more in comparison with the same quarters in 2016. Totally, more than 90% of cargo is transported by sea [2] in Europe. The industry is a critical and hazardous area of human activity and its growth raises control and security challenges. Increasing intensity in maritime traffic creates an increasing requirement for better prevention-oriented incident management systems. One of the control techniques of this complex management system is the detection of abnormal vessel movement.

2 Statement of the Problem

Maritime Situational Awareness (MSA) concept was presented by North Atlantic Treaty Organization (NATO) in their summit in Riga in 2006 as an extension of Maritime Domain Awareness (MDA). [8]. The main goal of MSA is to obtain a complete picture in Marine Domain by receiving information from multiple monitoring, surveillance, and reconnaissance systems, including knowledge extraction subsystems. Martineau and Roy state that "all aspects of a situation of interest in a timely manner, one can then say that complete and continuous situational awareness has been achieved" [9]. On the other hand, the final state of such goal is unreachable due to complexity and variability of the maritime

domain. That understanding is supported by the same authors Martineau and Roy by stating it "would be akin to omniscience and achieving it would be a utopia" [9]. Continuous and timely data from multiple sources must be collected to obtain a clear picture of a situation. Additionally, pattern identification and extraction from the same data must be performed. Knowledge extraction is an essential system part of enriching MSA.

Safety and security have an essential role in marine domain. The MSA enables marine and coastal authorities to evaluate potential security and safety risks and take timely actions to mitigate these risks [8]. The high intensity of marine traffic and data generated by it makes it impossible for human cognitive abilities to be aware of situation. The data collection automation and knowledge extraction methods and their practical application in MSA might help authorities to pursue those goals [10]. Extraction of marine vessel behavioural patterns and evaluation hazardous situation of safety or security infraction are among the most important goals in MSA. Collection of large quantities of diverse data and knowledge extraction help coastal authorities to make well-founded decisions. [11, 12, 13]. One of the ways to enhance MSA is the identification of anomalous behaviour in marine traffic data (anomaly detection) [14, 15], that is strongly supported by multiple civilian, military, and law enforcement authorities around the world [14].

3 Research Object

Detection of marine traffic anomalies in AIS data.

4 Research Aim And Objectives

The aim of the research is to investigate existing approaches and solutions and to propose a complex systemic (or integrated) approach including improvement of ML algorithms for detection of marine vessel traffic anomaly in AIS data.

For this aim, the following objectives should be achieved:

1. To perform literature analysis in the research field to elaborate a research workflow, covering all necessary problem-solving stages.
2. To inspect the AIS data and apply data preprocessing techniques to propose an appropriate scheme for data preparation according to the different nature of AIS data. The schema includes data structuring, cleaning, down-sampling, missing values imputation, feature engineering, the missing vessel type classifier, and splitting to sequences of vessel navigational vectors, with the view to prepare the data for upcoming anomaly detection analysis.
3. To introduce a method that can solve the imputation problem of missing vessel type values in data. To develop and test vessel type classifier to cope with the issue in the real-world AIS data set.
4. To inspect semi-supervised (point-based) methods for anomaly detection, propose enhancement and explore the possibility to use historical vessel movement data to speed up the semi-supervised algorithm while analyzing streaming AIS data.
5. To inspect unsupervised (trajectory-based) methods for anomaly detection, define extraction technique for abnormal vessel movement region, and compare the obtained results using methods based on statistical techniques.
6. To perform a comparative analysis of abnormal trajectories obtained by applying semi-supervised and unsupervised methods on AIS data by investigating a region at two sea areas.

5 Research Methods

Research that was performed in this thesis is based on these scientific methods:

1. Literature review is performed on the latest scientific papers in the research field to identify, select and evaluate

state-of-the-art algorithms for solving the stated problem.

2. Quantitative and qualitative information gathering was performed to create data sets, which were used for experiments and experimental data describing the performance of the proposed solution or its components.
3. Methods including but not limited to statistical ones were used to perform confirmatory data analysis, ensuring the reliability of data and experimental setup.
4. EDA: Box plot, Histogram, Scatter plots, Pair plots, Negative likelihood contour plots.
5. Descriptive Statistics: Univariate Analysis, Multivariate Analysis, Pearson's correlation, Cramer's V correlation; Data mean variance scaling; SMOTE.
6. Model evaluation: classification confusion matrix, classification metric comparison, evaluation and comparison of regression errors, PICP and PINAW, Wild bootstrapping techniques.
7. Multivariate clustering techniques: SOM, Soft-DTW k-means.
8. Dimensionality reduction techniques: MDS.
9. Artificial neural network techniques: LSTM; MLP; Autoencoders; Neural network layers stacking.
10. Constructive research was used to propose improvements to the solution of the real-world problem and propose new methods to improve MSA.
11. Software development and parallel computation methods with GPUs and TPUs were used in the experimental part of this thesis, including the implementation of marine vessel anomaly detection and trajectory clustering.

6 Scientific Contributions and Practical Value of the Research

This thesis contributes to the development of marine vessel traffic anomaly detection as an extension to MSA. The main contribu-

tions of this thesis can be outlined as follows:

1. The point-based modified SOM algorithm for marine vessel movement data classification into normal and abnormal classes is proposed and investigated on two independent data sets. The modification is done by incorporating virtual pheromone intensity calculations at the last stage of model training. This method has shown better classification results on less intense (less than 140,000 navigational vectors) marine vessel traffic data sets. The procedure for selecting the best neighbourhood function and SOM grid size is introduced.
   From the practical point of view, it can improve MSA for VTS of relatively small ports with moderate traffic.
2. The retraining strategies for SOM point-based methods are proposed. Applying different SOM model retraining strategies while keeping the same data batch sizes substantially decreased the time for retraining the maritime traffic abnormal movement detection model sustains precision and sensitivity at very high values. The results obtained show that the SOM network could be retrained in half the time while keeping precision and sensitivity at almost the same high values.
   In practice, it can increase speed and shorten the time for model retraining by keeping the model updated with the most up-to-date data or significantly reduce the cost of hardware required for model training.
3. Vessel type prediction method is proposed for missing vessel type imputation by vessel trajectories using multi-stacked multivariate LSTM method. Such classification experiment has shown that classification precision and sensitivity are satisfactory and can be used for this purpose.
   In practice, it enriches the training data set with additional training samples.
4. Two LSTM based methods were proposed for unsupervised detection of abnormal marine vessel trajectories. Both met-

hods detect anomalies by checking trajectories in the prediction region. First, the LSTM prediction learning method was created by modification of univariate LSTM interval learning to learn multivariate prediction region. Second, the LSTM wild bootstrapping method based on the integration of statistical wild bootstrapping technique was adapted to LSTM multi-stacked multivariate auto-encoder to create prediction region ellipses for normal movement model. Both methods show the ability to detect a broader range of anomalous trajectory line shapes compared to SOM based methods.

In practice, it could simplify the anomaly detection models' training by avoiding vessel trajectory labelling for anomalous traffic cases, which is usually required for tuning semi-supervised models based on semi-supervised SOM methods. LSTM method could be used for larger areas or sea areas with substantial traffic, where labelling of abnormal trajectories is unfeasible. The wider range of detected anomalous trajectories improve MSA for VTS.

7 Defensive Claims

The following claims are defended in this thesis:

1. Proposed SOM neural network with integrated virtual pheromone for detection of vessel traffic anomaly performs better on smaller data sets than Self-Organizing Map (SOM) with integrated Gaussian Mixture Model (GMM) (SOM_GMM). However, the SOM_GMM should be used for the larger sets.
2. SOM neural networks can be retrained for anomaly detection tasks in a shorter time with a minor change in precision compared to classical training workflow.
3. The proposed LSTM prediction region learning and LSTM wild bootstrapping methods can detect vessel trajectory anomalies. The LSTM prediction region learning outper-

forms LSTM wild bootstrapping method on quite small data sets.

4. LSTM architecture with good generalization properties can be applied for the detection of vessel type to perform an imputation of missing values.
5. Point-based anomaly methods SOM_pheromone and Self-Organizing Map (SOM) with integrated Gaussian Mixture Model (GMM) (SOM_GMM) do not detect anomalies in trajectories with sharp manoeuvres and stopping line shapes but LSTM methods do.

8 GENERAL CONCLUSIONS

1. The accuracy of SOM_pheromone and SOM_GMM methods depend on amount of data. It was observed that SOM_pheromone method detects anomalous traffic more accurately than SOM_GMM method does in data up to around 140,000 navigational vectors. Experiments show that the recommended SOM_pheromone parameters are: Mexican Hat neighbourhood function and grid size from 35×35 to 40×40. On smaller data sets, an average precision is 0.911 and sensitivity is 0.801 versus SOM_GMM's precision of 0.886 and sensitivity of 0.740. On larger data sets with more than approximately 140,000 navigational vectors, the SOM_GMM outperforms SOM_pheromone. The SOM_GMM's average precision is 0.859 and sensitivity is 0.648 in comparison with SOM_pheromone's average precision of 0.675 and sensitivity of 0.640. The recommended parameters are Cut Gaussian neighbourhood function and grid size from 55×55 to 60×60.

2. It is possible to substantially decrease model training time to detect abnormal maritime traffic movement when the model precision and sensitivity retain high values. Proposed SOM retraining strategy is applied to achieve this goal, where neurons' previous trained weights are used as

a starting position for retraining the network with newly gathered data subset mixing it with historical data and adjusting the learning rate. The obtained results show that the SOM_pheromone and SOM_GMM networks could be retrained in half the time while keeping precision and sensitivity at almost the same high values. The suggested retraining strategy took only 67.1% of computational time required by the classical method with the precision drop to the range from 0.007 to 0.009. The learning rate parameter for the proposed strategy is between 0.03 and 0.04.

3. The prediction region results of both LSTM methods show that for larger data sets with more than approximately 140,000 navigational vectors, the PICP value is close to the predefined $100(1 - \alpha) = 95\%$ value, when values are in range from 94.1% to 97.5%. On smaller data sets with less than approximately 140,000 navigational vectors, the LSTM prediction learning method was still able to learn with narrower prediction regions from 48.7% to 84.6%. The LSTM wild bootstrapping method was unable to learn prediction regions for smaller data sets, which is indicated by the PICP value of 0. Both LSTM algorithms for unsupervised estimation of prediction regions can be used for the detection of abnormal marine traffic when training data sets are larger than approximately 140,000 navigational vectors. For smaller data sets, it is recommended to use LSTM prediction region learning method with narrower prediction regions.

4. In order to solve the issue of missing vessel type data, a multi-stacked multivariate LSTM classifier was developed. The proposed model performs well, the average precision is 0.96079, the average sensitivity is 0.96060, and f1-score is 0.96056. Classification metrics show good generalization properties that allow to perform imputation and gain classes for the 4.28% percent with missing feature value (from a total of 4234160 navigational vectors) in the "Fehmarnbelt"

data set.

5. The test results of SOM methods, where LSTM output was taken as class reference, show low values of sensitivity (from 0.555 to 0.700) due to high values of false negatives. The analysis of these false negative results allows to conclude that sharp manoeuvre trajectory and stopping trajectory line shapes dominate in obtained false negatives set and constitute an average of 57.0% and 27.9% in the larger data set group.

Literatūros sąrašas

[1] The European Commission. *Maritime Transport Statistics-Short Sea Shipping of Goods*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Maritime_transport_of_goods_-_quarterly_data. (accessed: 21.08.2020).

[2] Z. Wan ir k.t. "Four routes to better maritime governance". *Nature* 540 (2016), p. 127–29.

[3] P. Fu ir k.t. "Finding Abnormal Vessel Trajectories Using Feature Learning". *Nature* 5 (2017), p. 7898–7909.

[4] J. Will, L. Peel ir C. Claxton. "In Proceedings of the IMA Maths in Defence Conference, Swindon, UK". *the IMA Maths in Defence Conference* 20 October (2011).

[5] Zhixiang He, Chi-Yin Chow ir Jia-Dong Zhang. "STNN: A Spatio-Temporal Neural Network for Traffic Predictions". *IEEE Transactions on Intelligent Transportation Systems* 1.1 (2020), p. 1–10. ISSN: 1524-9050. DOI: 10.1109/tits.2020.3006227.

[6] D. Li ir k.t. "Smoothed LSTM-AE: A spatio-temporal deep model for multiple time-series missing imputation". *Neurocomputing* 411 (2020), p. 351–363. ISSN: 18728286. DOI: 10.1016/j.neucom.2020.05.033. URL: https://doi.org/10.1016/j.neucom.2020.05.033.

[7] S. Tian ir k.t. "Spatio-Temporal position prediction model for mobile users based on LSTM". *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS* 2019-Decem (2019), p. 967–970. ISSN: 15219097. DOI: 10.1109/ICPADS47876.2019.00146.

[8] Centre Of Excellence For Operations In Confined And Shallow Waters ir k.t. *Maritime Situation Awareness*. 2015. URL: https://www.coecsw.org/fileadmin/content_uploads/projects/20150423_MSA_Study_Paper_-_Final.pdf.

[9] E Martineau ir J Roy. *Maritime Anomaly Detection: Domain Introduction and Review of Selected Literature*. 2011.

[10] Filipe Dias ir k.t. "Maritime Situational Awareness, the singular approach of a dual-use Navy". *Scientific Bulletin of Naval Academy* XXI (liepa 2018), p. 203–215. DOI: 10.21279/1454-864X-18-I1-033.

[11] A. Sidibé ir G. Shu. "Study of automatic anomalous behaviour detection techniques for maritime vessels". *J. Navig* 70 (2017), p. 847–858.

[12] V. Fernandez Arguedas, G. Pallotta ir M. Vespe. "Maritime Traffic Networks: From Historical Positioning Data to Unsupervised Maritime Traffic Monitoring". *IEEE Transactions on Intelligent Transportation Systems* 19.3 (2018), p. 722–732. ISSN: 15249050. DOI: 10.1109/TITS.2017.2699635.

[13] Virginia Fernandez Arguedas, Fabio Mazzarella ir Michele Vespe. "Spatio-temporal data mining for maritime situational awareness". *MTS/IEEE OCEANS 2015 - Genova, Italy*. Geg. 2015, p. 1–8. DOI: 10.1109/OCEANS-Genova.2015.7271544.

[14] M.J. Riveiro. "Visual analytics for maritime anomaly detection". Disertacija. Orebro universitet, 2011.

[15] Centre Of Excellence For Operations In Confined And Shallow Waters. *The Role and Relevance of the Maritime Domain in an Urban-Centric Operational Environment*. 2017. URL: https://www.coecsw.org/fileadmin/content_uploads/projects/Role_and_Relevance_of_the_Maritime_Domain_in_an_Urban-Centric_Operational_Environment.pdf.

[16] Jan Ekman ir Anders Holst. “Incremental stream clustering and anomaly detection”. *SICS Technical Report* 1 (saus. 2008), p. 55. ISSN: ISSN 1100-3154.

[17] N. Lu ir k.t. “Shape-Based Vessel Trajectory Similarity Computing and Clustering: A Brief Review”. *2020 5th IEEE International Conference on Big Data Analytics, ICBDA 2020* (2020), p. 186–192. DOI: 10.1109/ICBDA49040.2020.9101322.

[18] Leonid Portnoy, Eleazar Eskin ir Salvatore Stolfo. “Intrusion Detection with Unlabeled Data Using Clustering”. *In: Proceedings of ACM CSS Workshop on Data Mining Applied to Security*. Lapkr. 2001.

[19] R.O. Lane ir k.t. “Maritime anomaly detection and threat assessment”. *In Proceedings of the FUSION 2010 : 13th International Conference on Information Fusion, Edinburgh, UK* 26–29 July 2010 (2010).

[20] Danish Maritime Authority. *Historical AIS data*. https://www.dma.dk/SikkerhedTilSoes/Sejladsinformation/AIS/Sider/default.aspx. 2020.

[21] World Weather Online. *World Weather Online meteorological data of Danish waters region*. https://www. worldweatheronline.com/. 2020.

[22] J. Venskus ir k.t. “Integration of a Self-Organizing Map and a Virtual Pheromone for Real-Time Abnormal Movement Detection in Marine Traffic”. *Informatica (Netherlands)* 28.2 (2017), p. 359–374. ISSN: 08684952. DOI: 10.15388/Informatica.2017.133.

[23] J. Venskus ir k.t. “Real-time maritime traffic anomaly detection based on sensors and history data embedding”. *Sensors (Switzerland)* 19.17 (2019). ISSN: 14248220. DOI: 10.3390/s19173782.

[24] J. Venskus ir P. Treigys. “Meteorological data influence on missing Vessel type detection using deep Multi-Stacked LSTM neural network”. *Computer data analysis and modeling: stochastics and data science : proceedings of the XII international conference, Minsk, September 18-22, 2019*. Minsk: Minsk : Belarusian State University, 2019, p. 307–310. ISBN: 9789855668115.

[25] Maria Riveiro ir k.t. “Supporting maritime situation awareness using self organizing maps and gaussian mixture models”. *Frontiers in Artificial Intelligence and Applications* 173 (2008), p. 84.

[26] N. Cruz, L.G. Marin ir D. Saez. “Prediction Intervals With LSTM Networks Trained By Joint Supervision”. *IJCNN. International Joint Conference on Neural Networks. Budapest, Hungary* 14-19 July 2019 (2019).

[27] N. Cruz, L.G. Marın ir D. Saez. “Neural network prediction interval based on joint supervision”. *2018 International Joint Conference on Neural Networks (IJCNN)* July 2018 (2018), p. 1–8.

[28] V. Chew. “Confidence, Prediction, and Tolerance Regions for the Multivariate Normal Distribution”. *Journal of the American Statistical Association* 61.315 (1966), p. 605–617.

[29] M. Cuturi ir Blondel M. “Soft-DTW: a Differentiable Loss Function for Time-Series”. *Thirty-eighth International Conference on Machine Learning, ICML*. 2017.

[30] Cyril Goutte ir k.t. “On Clustering fMRI Time Series”. *NeuroImage* 9.3 (1999), p. 298–310. ISSN: 1053-8119. DOI: https://doi.org/10.1006/nimg.1998.0391. URL: http://www.sciencedirect.com/science/article/pii/S1053811998903913.

PUBLIKACIJŲ SĄRAŠAS

Publikacijos periodiniuose moksliniuose žurnaluose, indeksuojamuose Web of Science duomenų bazėse:

- J. Venskus, P. Treigys, and J. Markevičiūtė. "Unsupervised Marine Vessel Trajectory Prediction using LSTM Network and Wild Bootstrapping Techniques". Nonlinear Analysis: Modelling and Control. (2021). Vilnius University. ISSN 1392-5113 | eISSN 2335-8963. (ACCEPTED)
- 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.
- 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.

Straipsniai recenzuojamuose leidiniuose:

- Venskus, Julius; Treigys, Povilas. Meteorological data influence on missing Vessel type detection using deep Multi-Stacked LSTM neural network // Computer data analysis and modeling: stochastics and data science : proceedings

of the XII international conference, Minsk, September 18-22, 2019. Minsk : Belarusian State University, 2019. ISBN 9789855668115. p. 307-310.

UŽRAŠAMS

UŽRAŠAMS

Julius Venskus

DALINAI PRIŽIŪRIMŲ IR NEPRIŽIŪRIMŲ MAŠININIO MOKYMOSI METODŲ TYRIMAS JŪRŲ EISMO ANOMALIJOMS APTIKTI

DAKTARO DISERTACIJOS SANTRAUKA
Technologijos mokslai,
Informatikos inžinerija (T 007)
Redaktorė Agnė Lukošiūtė

Julius Venskus

SEMI-SUPERVISED AND UNSUPERVISED MACHINE LEARNING METHODS FOR SEA TRAFFIC ANOMALY DETECTION

SUMMARY OF DOCTORAL DISSERTATION
Technological Sciences
Informatics Engineering (T 007)
Editor Liutauras Bartašius

Vilniaus universiteto leidykla
Saulėtekio al. 9, III rūmai, LT-10222 Vilnius
El. p.: info@leidykla.vu.lt, www.leidykla.vu.lt
Tiražas 35 egz.