



Improvement of recurrent deep neural networks algorithm by feature selection methods and its usage of automatic identification system data evaluated as time series

Yunus Doğan*^{ID}

Department of Computer Engineering, Dokuz Eylül University, Buca/Izmir, 35390, Turkey

Highlights:

- Proposing novel deep learning models evaluating AIS data as a time series for vessels route and collision prediction (RCP).
- Developing a new approach on AIS data using feature selection methods for improving accuracy values of models.
- A new clustering approach to explore the types of uncertain vessels in AIS, and not being ignored of these vessels in the analyzes.

Keywords:

- Big data
- Deep learning
- AIS data
- Feature selection
- Clustering

Article Info:

Research Article
Received: 18.01.2020
Accepted: 22.03.2020

DOI:

10.17341/gazimmfd.676862

Acknowledgement:

I would like to thank Prof. Dr. Selçuk Nas from Dokuz Eylül University Maritime Faculty for his help in providing the AIS data required for the testing of this study.

Correspondence:

Author: Yunus Doğan
e-mail: yunus@cs.deu.edu.tr
phone: +90 232 301 7419

Graphical/Tabular Abstract

It is observed in Figure A that the model having the lowest Mean Absolute Percentage Error (MAPE) value is Deep Recurrent Neural Network (DRNN) using Information Gain (IG) Feature Selection (FS) method with 6.2% for the route prediction of the vessels in the Marmara Sea. Nevertheless, DRNN models with the other FS methods, Relief (RF), Pearson's Correlation (PC) and Gain Ratio (GR), have satisfying low MAPE values. As a result, this study is claimed that DRNN with FS methods improves the prediction results.

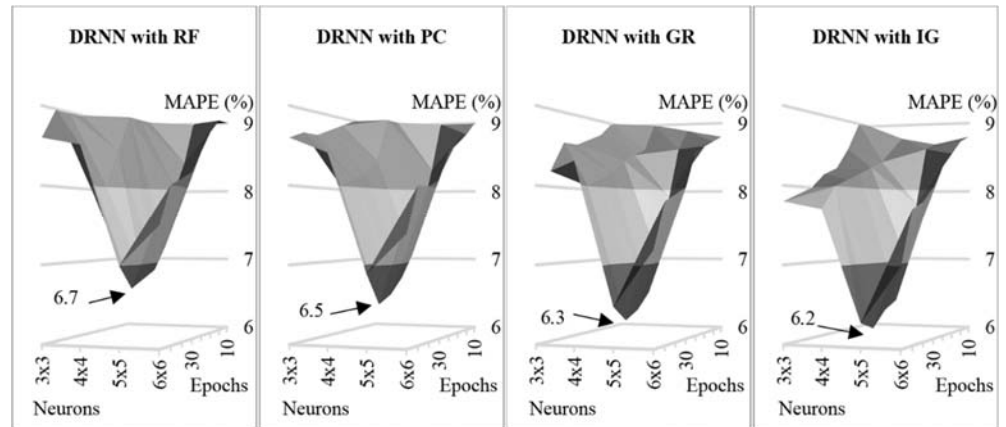


Figure A. For the AIS data of the Marmara Sea, the MAPE values of improved DRNN model using RF, PC, GR and IG methods, and having different neuron and epochs sizes

Purpose: This study focuses on the vessels route and collision predictions on AIS data evaluating as Time Series data. The main purpose was to produce novel approaches giving lower MAPE and higher accuracy values for these predictions than the other approaches in the literature. The other purpose was to produce a novel approach not to ignore the vessels with uncertain types in the study.

Theory and Methods:

New models as Deep Recurrent Neural Networks having a many-to-many architecture and a Long-Short Term Memory neuron structure have been developed. These models have been enforced with FS methods for improving the MAPE and the accuracy and f-score values. Also, SOM++ clustering algorithm has been used not to ignore the vessels with uncertain types. Linear Estimation (LE), Autoregressive Integrated Moving Average (ARIMA) and Multilayer Perceptron (MLP) approaches in the literature have been implemented for the comparisons. LE and ARIMA models have not contain FS improvement due to their conservative structures; however, in MLP and DRNN models, the other vessel features besides of the coordinate values have been considered and FS methods could be used.

Results and Conclusion:

The tests have been performed using AIS data of the Çanakkale Strait and the Marmara Sea separately due to their having different moving characteristics of vessels. For the Çanakkale Strait data, while the LE and ARIMA models have given as 6% MAPE approximately, the MLP model with PC and the DRNN model with GR have given as 4.5% and 3.6% respectively. However, for the Marmara Sea data, while the LE and ARMA models have given as 28% MAPE approximately, the MLP model with IG and the DRNN model with IG have given as 9.8% and 6.2% respectively. In addition, for the Marmara Sea data, the DRNN with IG has given the highest accuracy and f-score values as 99% and 96.5% respectively at the collision tests.



Kendini tekrarlayan derin sinir ağlarının öznitelik seçim yöntemleri ile iyileştirilmesi ve zaman serisi olarak ele alınan otomatik tanımlama sistemi verilerinde kullanımı

Yunus Doğan*^{ID}

Dokuz Eylül Üniversitesi, Bilgisayar Mühendisliği Bölümü, Buca/İzmir, 35390, Türkiye

Ö N E Ç I K A N L A R

- Gemilerin güzergah ve çarpışma tahmini için AIS verilerini zaman serisi olarak değerlendiren yeni derin öğrenme modellerinin önerilmesi
- Modellerin doğruluk değerlerini artırmak için özellik seçimi yöntemlerini kullanarak AIS verileri üzerine yeni bir yaklaşım geliştirmek
- AIS'deki belirsiz gemi türlerini keşfetmek için yeni bir kümeleme yaklaşımı ve analizlerde bu gemilerin göz ardı edilmemesi

Makale Bilgileri

Araştırma Makalesi
Geliş: 18.01.2020
Kabul: 22.03.2020

DOI:

10.17341/gazimmfd.676862

Anahtar Kelimeler:

Büyük veri,
derin öğrenme,
AIS verisi,
öznitelik seçimi,
kümeleme

ÖZET

Otomatik Tanımlama Sistemi (AIS), deniz taşımacılığının, çarpışma, yangın ve tehlikeli veya kirletici maddelerin dökülmesi gibi risklere sahip olması nedeniyle günümüzde zorunlu hale gelmiş gözlem ve analiz sistemidir. Literatürde, bu tehlikeli durumların önceden tespitinin yapılması, gemilerin kontrollü ve güvenli seyahatlerini gerçekleştirmeleri için AIS verilerinin kullanıldığı temel matematiksel modellerin, istatistiksel modellerin ve makine öğrenmesi algoritmalarının uygulamalarını görebilmekteyiz. Bu çalışmada AIS verileri zaman serileri bakışıyla ele alınmış ve geleneksel rota tahminleme modeli yanında; Bütünleşik Otopregresif Hareketli Ortalama, Çok Katmanlı Algılayıcı (ÇKA) ve Kendini Tekrarlayan Derin Sinir Ağları (KT-DSA) ile farklı modeller oluşturularak doğruluk karşılaştırmaları yapılmıştır. Ayrıca ÇKA ve KT-DSA modellerinde, öznitelik seçim tekniklerinden yararlanılarak nitelikler ağırlıklandırılmış ve bu iyileştirilmelerle yeni algoritmalar önerilmiştir. Öznitelik seçimlerinden Relief, Pearson'nun Korelasyonu, Kazanım Oranı ve Bilgi Kazanımı (BK) metotları kullanılmış ve verdikleri rota ve çarpışma tahminlemelerinin doğrulukları karşılaştırılmıştır. Bu doğruluk testlerinde kullanılmak üzere veri seti olarak belirli zamanlara ait Çanakkale Boğazı ve Marmara Denizi AIS verilerinden faydalanılmıştır. Sonuçlara bakıldığında Çanakkale Boğazı'ndaki gemilerin doğrusal bir hareket yapısına sahip olmasından dolayı tüm yaklaşımların birbirine yakın ve yüksek doğruluklara sahip olduğu gözlemlenirken, düzensiz yapısından dolayı Marmara Denizi'nde en iyi sonucu veren yaklaşımın BK ile iyileştirilmiş KT-DSA olduğu sonucuna varılmıştır.

Improvement of recurrent deep neural networks algorithm by feature selection methods and its usage of automatic identification system data evaluated as time series

H I G H L I G H T S

- Proposing novel deep learning models evaluating AIS data as a time series for vessels route and collision prediction (RCP)
- Developing a new approach on AIS data using feature selection methods for improving accuracy values of models
- A new clustering approach to discover the types of uncertain vessels in AIS, and not being ignored these vessels in the analyzes

Article Info

Research Article
Received: 18.01.2020
Accepted: 22.03.2020

DOI:

10.17341/gazimmfd.676862

Keywords:

Big data,
deep learning,
AIS data
feature selection,
clustering

ABSTRACT

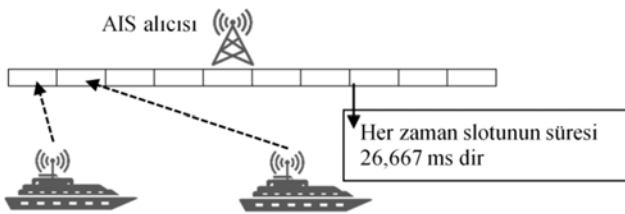
Automatic Identification System (AIS) is an observation and analysis system that has become compulsory nowadays due to the risks of maritime transportation such as collision, fire, and spillage of hazardous or polluting substances. In the literature, we can see the applications of basic mathematical models, statistical models and machine learning algorithms using AIS data in order to detect these dangers in advance and to make controlled and safe travel of ships. In this study, AIS data have been evaluated as time series, and accuracy comparisons have been made by being developed different models with Autoregressive Integrated Moving Average, Multilayer Perceptron (MLP) and Deep Recurrent Neural Networks (DRNN) beside traditional route estimation model. In addition, feature selection techniques have been weighted in MLP and RDNN models, and new algorithms have been proposed with these improving. Relief, Pearson's Correlation, Gain Ratio and Information Gain (IG) methods were used to compare the accuracy of the route and collision estimations. In order to be used in these accuracy tests, AIS data related into certain times of Çanakkale Strait and Marmara Sea were used. The results showed that all the approaches were close and high accuracy due to the linear movement of the ships in Çanakkale Strait. On the other hand, it has been observed that the best approach in the Marmara Sea was the improved DRNN with IG due to its irregular structure.

*Sorumlu Yazar/Corresponding Author: yunus@cs.deu.edu.tr / Tel: +90 232 301 7419

1. GİRİŞ (INTRODUCTION)

Dünya deniz taşımacılığındaki sürekli büyüme ile beraber gemi hareketlerinin izlenmesi, denizlerdeki durumsal farkındalık için çok önemli bir hale gelmiştir [1]. Küresel ticaretin çoğunluğunu temsil eden deniz taşımacılığının bu denli büyüme göstermesi, güvenliğin önemini de arttırmaktadır. Deniz taşımacılığı, çarpışma, yangın ve tehlikeli veya kirletici maddelerin dökülmesi gibi risklere sahiptir. Bu riskleri azaltmak için, deniz gözetimi verileri farklı kriterlerde toplanmaktadır. Otomatik Veri Tanımlama Sistemi (OTS), uluslararası adıyla Automatic Identification System (AIS), gemilerin geminin kimliğini, türünü, pozisyonunu, rotasını, hızını, seyir durumunu ve diğer güvenlikle ilgili bilgilerini elektronik olarak gönderimini ve böylece de gemilerin canlı verileri izlenmesini sağlamaktadır. Denizde Can Güvenliği Uluslararası Sözleşmesi (SOLAS) Bölüm 5 Yönetmeliği 19'a göre de AIS, gemiler için kademeli olarak zorunlu hale getirilmiştir [2, 3]. Gemilerde A sınıfı ve B sınıfı olmak üzere iki tip AIS bulunmaktadır. 300 Gross Ton (GRT)'nun üzerindeki ticari gemiler ve büyüklüğüne bakılmaksızın yolcu gemileri için A Sınıfı, daha küçük ticari gemiler, balıkçılık sektörü vb. için B Sınıfı AIS bulunmaktadır. Bu sınıflar arasındaki fark sinyallerin kapsama alanı ile ilgilidir. Yani gönderilen verilerin satır ya da kolon bazında boyutuyla ilgili bir fark yoktur [1, 4].

Kıyı istasyonlarında bulunan Deniz Aracı Trafik Servisleri (VTS- Vessel Traffic Services) sunucular, gemilerden gönderilen bu AIS bilgilerini kaydetmektedir. AIS ile depolanan bu veriler kısa sürede büyük veri problemine neden olmaktadır. Örneğin, Denizcilik Araştırma ve Deney Merkezi (CMRE-Centre for Maritime Research and Experimentation) şu anda birden fazla gemi ve istasyondan ayda yaklaşık 600 Milyon AIS verisi almaktadır ve bu sayı sürekli artmaktadır [5]. Daha derine inilirse; AIS sistemlerinin zaman dilimleri Şekil 1'deki gibidir ve AIS mesajı bir slot olarak gönderilmektedir [6]. Bir dakikada yaklaşık 2250 slot elde edilir ve bir saat boyunca AIS sistemleri tarafından 135.000 AIS bilgisi sırayla gönderilebilmektedir. Sunuculara kaydedilen AIS verileri, bir bölgedeki deniz trafiğini analiz etmek için yararlı olacaktır, ancak bir günlük AIS verileri bile analiz için standart yöntemlerle kullanılamayacak kadar fazla olmaktadır.



Şekil 1. AIS zaman dilimli çoklu erişim çalışma prensibi (AIS time division multiple access working principle)

Tablo 1'de, A sınıfı bir AIS'den farklı periyotlarda (P) istasyonlara çok fazla veri gönderildiği gösterilmiştir [6].

Örneğin, gemi hızının 10 knot olduğunu ve aynı rotada dolaştığını varsayarsak, bu gemi 10 saniyede bir slot veri gönderiyor demektir ve bir saat içinde diğer istasyonlara toplam $6 * 60 = 360$ AIS verisi göndereceği öngörülebilir. Dolayısıyla, yoğun trafiğe sahip bir bölgede, çok sayıda gemi, istasyonlara büyük miktarlarda veri göndereceği öngörülebilir. AIS teknolojisi ile meydana gelen bu büyük veriden anlamlı çıkarımlar yapabilmek için bilgisayar destekli sistemlerin ve algoritmaların giderek yaygınlaştığı görülmektedir.

Tablo 1. A sınıfı gemi kaynaklı mobil ekipman raporlama aralıkları

(Class A shipborne mobile equipment reporting intervals)

Gemi dinamiği koşulları	P(s: saniye)
Demirli ya da 3 knot'tan daha hızlı hareket etmeyenler	180 s
3 ile 14 knot arasında ve rotasından sapma göstermeden hareket edenler	10 s
3 ile 14 knot arasında ve rotasından sapma göstererek hareket edenler	3 $\frac{1}{3}$ s
14 ile 23 knot arasında ve rotasından sapma göstermeden hareket edenler	6 s
14 ile 23 knot arasında ve rotasından sapma göstererek hareket edenler	2 s
23 knot ya da yukarısında ve rotasından sapma göstermeden hareket edenler	2 s
23 knot ya da yukarısında ve rotasından sapma göstererek hareket edenler	2 s

Bilgisayar temelli uygulamalarda, başlarda herhangi bir istatistik ya da veri madenciliği temelli bir algoritma hesaba alınmadan AIS içinde yer alan yere göre hız (The Speed of Ground - SOG) ve yere göre rota (The Course Over Ground - COG) verilerinin kullanımı ile rota ve çarpışma tahminleme çalışmalarının yapıldığı görülmektedir [7, 8]. Daha sonraki çalışmalarda zaman serisi örnekleriyle karşılaşılmaktadır. Bunun sebebi AIS verisinin bulunduğu tarih ve saat alt alanları sebebiyle aslında bir zaman serisi verisi olmasıdır. Bütünleşik Otoregresif Hareketli Ortalama (Autoregressive integrated moving average-ARIMA) algoritması en temel zaman serisi algoritmasıdır ve geleneksel hali ile birçok alanda tahminleme amacıyla kullanılmıştır. Aynı yapıyı enerji tahminleme çalışmalarında geleneksel ARIMA tekniğinin Yapay Sinir Ağları'nın (YSA) gerisinde kaldığını Ortalama Mutlak Hata Yüzdesi (Mean Absolute Percentage Error-MAPE) değerleri ile ortaya koymuşlardır [9, 10]. Daha sonraları AIS verileri ile geleneksel ARIMA tekniğinin kullanıldığı çalışmalarda görülmeye başlanmış ve bu çalışmalardan [11-13]'de YSA'nın ARIMA'ya nazaran çoklu alt alanı desteklemesi ve MAPE değerlerinin daha düşük çıkması sebebiyle, YSA'nın tercih edildiğinden bahsedilmiştir. Aynı şekilde, veri madenciliği algoritmalarından Gürültülü Uygulamaların Yoğunluk Temelli Mekansal Kümelemesi (Density-based spatial clustering of applications with noise- DBSCAN) algoritmasının ARIMA'ya göre daha başarılı çıktığı

[14]'deki çalışmada vurgulanmıştır. AIS verisinin ve veri madenciliği algoritmalarının birlikte başarıyla kullanıldığını gösteren çalışmalardan; Gürültülü Uygulamaların Yoğunluk Temelli Mekansal Kümelemesi (DBSCAN) [15, 16], Yapay Sinir Ağı [17], Rastgele Orman (Random Forest) [18], Destek Vektör Makineleri [19], Pencere tabanlı rastlantı kümelemesi (Window-based encounter clustering) [20] ve Kmeans kümeleme [21] algoritmaları kullanılmış uygulamaları örnek gösterebiliriz. Literatürde büyük veri olarak AIS verilerini işleme tekniklerini öneren örnekler de yer almaktadır [22-24]. Bunun dışında görselleştirme ve simülasyon çalışmaları da yapılmıştır [25-27]. Son olarak, karar destek sistemleri ile ilgili örnekler de literatürde görülmektedir [28, 29]. Bunların yanında son zamanların popüler konusu derin öğrenme alanında da çalışmaların arttığı [30, 31] ve AIS üzerine de yeni yeni çalışmaya başlandığı görülmektedir [32-34]. Ancak bu çalışmalar, AIS verilerinin ve veri madenciliği algoritmalarının saf hallerinde kullanmış örnekleri olarak görülmektedir. Algoritmaların uygulanan veri seti üzerine doğruluğu açısından, ilgili algoritmanın çalışmasından evvel uygulanan ön işlem operasyonları oldukça önemlidir [35]. Bunun örneklerini birçok farklı alanda yapılmış çalışmada görebilmekteyiz. Örneğin: Twitter verisi üzerinden yapılmış doğal dil işleme [36, 37], tıp alanında yapılmış karar destek sistemi çalışmalarında [38, 39], elektrik ve hidroloji gibi enerji üzerine yapılmış veri madenciliği uygulamalarında, veri ön işlemenin veri analizindeki önemi ortaya konulmuştur [40-42]. Fakat, AIS verisi için bir ön işlem uygulaması örneği bulunmamaktadır. Ön işleme aşamasında, bir veri setindeki niteliklerin hepsinin eşit koşulları yakalaması için normalizasyon dediğimiz, standartlaştırma tekniği kullanılır, ancak bu durum tamamen bazı niteliklerin içinde yer alan sayıların göreceli büyüklüklerini kaybettirmek için diğer niteliklerdeki sayılarla aynı aralıklara denkleştirmek içindir [43]. Ama niteliklerin hepsinin baskınlıklarının analiz esnasında aynı olması beklenemez. Bunun için öznitelik seçimi denilen teknikler kullanılarak, her bir niteliğin analiz için ne kadar baskın olduğu tahlil edilir. Hatta, bu çalışmalarda bazı niteliklerin hiçbir baskınlığı olmadığı da tespit edilip veri setinden de çıkartılabildiği durumlar da olabilmektedir [44]. Niteliklere, baskınlıkları dahilinde skorlama yapan öznitelik seçimi tekniği, ağırlıklı sınıflandırma denilen algoritmalar için de kullanılmaktadır. Ağırlıklı sınıflandırmanın standart sınıflandırma algoritmalarından daha başarılı olduğunu birçok çalışmada görebilmekteyiz [45, 46]. Niteliklerin ağırlıklarının değişkenlik göstermesi prensibine dayanan bu algoritmaların başarılı sonuçlar verdiği örneklerin bazılarının tıbbi bilişim [47, 48], doküman ve resim verileri üzerine yapılan sınıflandırma çalışmaları [49, 50], ve network üzerine [51] olduğunu görebiliyoruz. Henüz, AIS üzerine yapılmış bir ağırlıklı sınıflandırma tekniklerinin kullanıldığı bir çalışmanın bulunmadığı gözlemlenmektedir.

Bu çalışmanın literatürdeki çalışmalara katkıları üç yönlüdür. İlki, AIS verisi üzerine ön işlem olarak öznitelik seçimi yapıp her niteliğin belirli ağırlıklandırmaya tabi tutulmasıdır. Böylece, rota ve çarpışma tahmini

modellerinde doğruluk analizi sonuçlarının iyileşmesi sağlanmıştır. İkinci katkı, bu önerilen ağırlıklandırma yöntemlerinin gemi tipleri için ayrı ayrı olması ve bazı gemi tiplerinin belirsiz olarak kaydedilmiş olması sebebiyle, bu belirsiz gemilerin tiplerini keşfedebilecek kümeleme algoritması kullanımı olmuştur. Böylece bu gemiler, rota ve diğer gemilerle çarpışma olasılık analizleri için göz ardı edilmemiştir. Üçüncüsü, AIS verisini bir zaman serisi verisi olarak nitelendirip, enlem (latitute- *lat*) ve boylam (longitude- *lng*) bilgilerinin aslına normalizasyon işlemi sırasında kaybetmeden girdi olarak kullanıp, çıktı olarak yine aslının elde edilmesini sağlayan yeni bir derin öğrenme modelinin geliştirilmesidir. Böylece normalizasyon sırasındaki bilgi kayıplarının önüne geçilmiştir.

Bu çalışma beş bölüm halinde detaylandırılmaktadır. Önerilen yöntemlerin veri hazırlama ve uygulama basamaklarının detaylı anlatımı Bölüm 2'de verilmiştir; Bölüm 3'te yapılan deneysel çalışmalar, karşılaştırmaları ve doğruluk sonuçları verilmiştir. Bölüm 4'te simgeler ve son olarak da Bölüm 5'te sonuç kısmı yer almaktadır.

2. DENEYSEL METOT / TEORİK METOD (EXPERIMENTAL METHOD / THEORETICAL METHOD)

Bu çalışmada 4 farklı yöntem kullanılmış ve testleri yapılmıştır. Bu yöntemler sırayla; Doğrusal Tahminleme (DT) Modeli, Otoregresif Entegre Hareketli Ortalama (OEHO) Modeli, Çok Katmanlı Algılayıcı (ÇKA) Modeli ve Kendini Tekrarlayan Derin Sinir Ağları (KT-DSA) Modeli'dir. Bu yöntemlerin her biri yazılım olarak Python dilinde yazılarak doğruluk testleri gerçekleştirilmiştir. Ayrıca, çalışmanın temel amacı olan ağırlıklandırılmış yaklaşımlarla da bu yöntemler iyileştirilmiştir. Tüm test sonuçları ve karşılaştırmaları sonraki bölüm olan Sonuçlar ve Tartışmalar başlığında verilmiştir.

2.1. Veri Tabanı Dizaynı (Database Design)

Bu çalışmada, 1-30 Haziran 2013 tarihleri arasındaki Çanakkale Boğazı'ndan ve 1-15 Eylül 2012 tarihlerinde Marmara Denizi'nden geçmiş gemilerin AIS verileri üzerine bir uygulama geliştirilmiş ve test edilmiştir. Bu tarihler arasında toplam 3.879 geminin, 19.263.121 hareket verisinin Çanakkale Boğazı'ndan, 1.990 geminin, 9.509.110 hareket verisinin Marmara Denizi'nden geldiği tespit edilmiştir. Bu büyük verinin kolay analiz edilebilir olması için veri tabanı yönetim sistemi (DBMS) işlemleri MSSQL 2019 üzerinden uygulanmıştır. Gemilerin sabit sayılabilecek özellikleri gibi niteliklerin tutulduğu `static_ship_data_table` ve her hareketin özellikleri gibi dinamik verilerinin tutulduğu `dynamic_ship_data_table` tabloları, sistemin temel iki ilişkisel tablosudur. Bu iki tablo Şekil 2'de görüldüğü gibi sabit nitelikler ve dinamik nitelikler olarak iki grup halinde oluşturulmuştur. Bu ayrılmanın nedeni, DBMS'lerde üçüncü normal formu (3NF) sağlayarak büyük AIS verilerinin kapasitesini en aza indirmektir [52].

Column Name	Nullable	Data Type
Staticid	No	int
MMSI	Yes	nvarchar(50)
IMO	Yes	nvarchar(50)
Name	Yes	nvarchar(50)
CallSign	Yes	nvarchar(50)
ShipandCargo	Yes	int
Draught	Yes	float
Type	Yes	int
DimA	Yes	float
DimB	Yes	float
DimC	Yes	float
DimD	Yes	float

Column Name	Nullable	Data Type
Dynamicid	No	int
Staticid	Yes	int
NavigationStatus	Yes	int
Lat	Yes	float
Lng	Yes	float
ReportDate	Yes	date
SOG	Yes	float
COG	Yes	float
HDG	Yes	float
ROT	Yes	float

Şekil 2. Veritabanındaki AIS verilerini kaydetmek ve filtrelemek için oluşturulmuş tablolar
(In the database, tables created to save and filter AIS data)

2.2. Nitelik Ağırlıklandırma Teknikleri (Feature Weighting Techniques)

Bu çalışmada, veri setindeki nitelik değerleri, zaman serisi analiz modellerine orijinal değerlerinin dışında, 4 tip ağırlıklandırma yöntemi ile ön işlemden geçirilerek girdi olarak verilmiştir. Geleneksel yöntemlerin doğruluğu ve ağırlıklandırılmış değerlerle gerçekleştirilen zaman serisi analizlerinin doğrulukları ayrı ayrı karşılaştırılmalı sonuçları ile üçüncü bölümde detaylı olarak verilmiştir. Çalışmada tercih edilen ağırlıklandırma yöntemleri; Relief (RF), Pearson'nun Korelasyonu (PK), Kazanım Oranı (KO) ve Bilgi Kazanımı (BK) teknikleridir. Bu teknikler literatürde en çok kullanılan öznelitik seçim metotlarıdır [51]. AIS verisinin, gemilerin zamana bağımlı hareket verilerinden oluştuğunu bilmekteyiz. Bu yüzden hedef nitelik diye atfedilecek bir kolon verisi, verinin saf halinde bulunmamaktadır. Diğer taraftan da hedef nitelik kolonu, ağırlıklandırma metotlarının kullanımı için gereklidir. Bu nedenle, çalışmada, hedef nitelik kolonu için gemi türü kolonu kullanılmıştır. Çünkü, gemi türüne göre; boyut, hareket verileri, draftı (DR), SOG, COG ve hız gibi niteliklerde, keskin bir farklılık görülebilmektedir. Örneğin, tanker gemileri ile balıkçı gemilerinin SOG, COG ve hız nitelikleri açık bir şekilde birbirinden farklılık göstermektedir. Buna ek olarak; bu çalışmada sadece tanker, balıkçı ve yolcu gemisi gibi sabit türler bazında tüm verileri birlikte değerlendirip ağırlıklandırma yapılmamıştır. Tablo 1'deki gibi benzer bir yaklaşımla tüm türler için hareket verileri ayrı bir veri seti olarak değerlendirilip her tip gemi için ayrı ayrı kümeleme algoritması kullanılarak alt türler keşfedilmiştir. Bunun bir başka nedeni de veri tabanına kaydedilen gemilerin bazılarının tip niteliğinin belirsiz olarak kaydedilmiş olmasıdır. Bu gemileri de boyut, hareket verileri, SOG, COG ve hız gibi niteliklerde kümeleyerek belirsiz gemilerin rota tahmini ve diğer gemilerle çarpışma olasılık analizlerinin yapılabilmesinin sağlanmasıdır. Temelinde yapay sinir ağını bulandıran kümeleme algoritması Özdüzenleyici Haritalar (Self Organizing Map-SOM)'un iyileştirilmiş versiyonu olan SOM++, verdiği başarılı sonuçlardan dolayı, kümeleme için tercih edilmiştir [54, 55]. SOM++, geleneksel SOM'un nöronlarına ilk ağırlık değerlerini vermek için K-Means++ algoritmasından

yararlanarak SOM'un geliştirilmiş versiyonu olarak çalışmada uygulanmıştır. SOM++ algoritması, Eş. 1 ve 2 de verildiği gibi güncelleme fonksiyonu olarak Gauss fonksiyonunu kullanarak nöronların ağırlıklarını harita üzerinde eğitir.

$$h = \exp(-d^2/2\sigma^2) \quad (1)$$

$$\omega \leftarrow \sum_{i=1}^f \omega_i = \omega_i + h * \eta * (C_{c,i} - \omega_i) \quad (2)$$

Burada d , mevcut örneklem ile merkez noktalar arasındaki en kısa yol; σ , komşuluk genişliği parametresi; η , öğrenme oranı parametresi; h , komşuluk fonksiyonu; ω , tüm nöronlardaki her hücrede yer alan ağırlık değerleri kümesi; f , nitelik sayısı ve C , K-Means++' dan gelen ilk merkez noktaları olarak verilmiştir.

Belirli sayıda yinelemeden sonra ağırlık değerleri, merkez noktalarına göre eğitilmiş değerlerin durumuna gelir. Merkez noktalarını haritalamak için, tüm merkez noktalarının en yakın mesafeye sahip kazanan nöronları ayrı ayrı hesaplanır ve bu kazanan nöronların ağırlık değerleri, ilgili merkez noktalarının çok değişkenli değerlerine dönüştürülür (3). Son olarak, ağırlık değerleri olarak merkez noktaları, çok değişkenli değerlerini içererek döndürülür.

$$\sum_{i=1}^k \sum_{j=1}^f KK(i, j) = C(i, j) \quad (3)$$

Burada k , küme sayısı ve KK , C 'deki her bir eleman için kazanan hücreler kümesidir.

Hareket verilerinin kümelenebilmesinin ardından artık hedef nitelik sayılabilecek küme değerleri elde edilmiş olur. Verileri bu kabullenme ile değerlendirip, bir sonraki basamak olan öznelitik seçim teknikleri RF, PK, KO ve BK ile ağırlıklandırma kat sayılarının elde edilmesine geçilmiştir. Zaman serisi analizlerinde kullanılan veriler hem girdi hem de çıktı verileri olurlar. Bu çalışmada da lat ve lng değerleri iki ayrı model olacak ve önceki belirli zamanlardaki lat ve lng değerleri girdi olup sonraki lat ve lng değerleri çıktıyı oluşturacaktır. Bu konum değerleri dışında da diğer niteliklerin kullanılacağı zaman serisi modellerinde tüm verilerin ağırlıklandırma işlemi öncesi, sayısal değerlerden oluştuğu için normalizasyon yapılması gerekliliğini bilinmektedir. Ancak hedef lat ve lng değerlerinin de normalizasyon ile bozulması istenmez. Bu nedenle bu çalışmada Eş. 4'te gösterilen min-max normalizasyonu kullanılmış ve lat modeli için diğer nitelikler, lat kolonundaki minimum ve maksimum değer aralıklarına çekilerek lat değerleri değişime uğratılmamıştır. Aynı şekilde lng 'yi tahminleyen modeller için de bu yaklaşım uygulanmıştır.(Eş. 4)

$$F_i = \frac{F_i - \min_{F_i}}{\max_{F_i} - \min_{F_i}} x (\max_{lat} - \min_{lat}) + \min_{lat} \quad (4)$$

Burada F_i , F niteliğindeki i 'ninci değeri, F_i yeni değerini, \min_{lat} , lat kolonundaki en düşük değeri ve \max_{lat} en büyük değeri temsil etmektedir.

Sonraki adım öznelik seçim teknikleri ile ağırlık kat sayılarının belirlenmesidir. İlk olan RF , Eş. 5 ve 6'daki gibi elde edilir.

$$RF \leftarrow \sum_{i=1}^f RF_i = 0 \quad (5)$$

$$RF \leftarrow \sum_{i=1}^n \sum_{j=1}^f RF_j = RF_j - |I_{hj} - R_i| + |I_{mj} - R_i| \quad (6)$$

Burada R , rastgele bir veri örneğidir; I_h , R 'nin bulunduğu kümede R 'ye en yakın veridir ve I_m , R 'nin bulunmadığı başka bir kümede R 'ye en yakın veridir.

Burada elde edilen kat sayı değerleri yine lat ve lng 'nin değerlerini değiştirme olasılığına sahiptir. Bu nedenle bu öznelik seçim tekniklerinden elde edilen kat sayı değerleri lat ve lng için 1 olarak kabul edilip, diğer niteliklerinkini ona göre orantılı değiştirme işlemi gerçekleştirilmiştir. Böylece lat ve lng değerlerinin yine değişmesinin önüne geçilmiştir (Eş. 7 ve Eş. 8).

$$RF \leftarrow \sum_{i=1}^f RF_i = \begin{cases} RF_i / \min_{rf}, \min_{rf} > 0 \\ \max_{rf} / RF_i, \max_{rf} < 0 \\ \frac{RF_i - \min_{rf}}{\max_{rf} - \min_{rf}} \times 0.9 + 0.1, \text{diğer} \end{cases} \quad (7)$$

$$RF \leftarrow \sum_{i=1}^f RF_i = \frac{RF_i}{RF_{lat}} \quad (8)$$

Burada \min_{rf} , RF içindeki en küçük, \max_{rf} , RF içindeki en büyük ve RF_{lat} lat niteliği için RF kat sayı değeridir.

PK , Eş. 9, Eş. 10 ve Eş. 11'deki gibi elde edilir. Ayrıca Eş. 12 ve Eş. 13'teki gibi normalize edilir.

$$PK1 = n \times \sum_{j=1}^n I_{ij}^2 - \left(\sum_{j=1}^n I_{ij} \right)^2 \quad (9)$$

$$PK2 = n \times \sum_{j=1}^n t_j^2 - \left(\sum_{j=1}^n t_j \right)^2 \quad (10)$$

$$PK \leftarrow \sum_{i=1}^f \frac{n \times \sum_{j=1}^n (I_{ij} \times t_j) - \sum_{j=1}^n I_{ij} \times \sum_{j=1}^n t_j}{\sqrt{PK1 \times PK2}} \quad (11)$$

Burada I , tüm veri setini; n , veri miktarını ve tr , hedef niteliği simgelemektedir.

$$PK \leftarrow \sum_{i=1}^f PK_i = \begin{cases} PK_i / \min_{pk}, \min_{pk} > 0 \\ \max_{pk} / PK_i, \max_{pk} < 0 \\ \frac{PK_i - \min_{pk}}{\max_{pk} - \min_{pk}} \times 0.9 + 0.1, \text{diğer} \end{cases} \quad (12)$$

$$PK \leftarrow \sum_{i=1}^f PK_i = \frac{PK_i}{PK_{lat}} \quad (13)$$

Burada \min_{pk} , PK içindeki en küçük, \max_{pk} , PK içindeki en büyük ve PK_{lat} lat niteliği için PK kat sayı değeridir.

KO , Eş. 14, Eş. 15 ve Eş. 16'daki gibi elde edilir ve Eş. 17 ve Eş. 18'deki gibi normalize edilir.

$$KO1 = \sum_{i=1}^b P(B_i) \log P(B_i) \quad (14)$$

$$KO2 = \sum_{i=1}^b \sum_{j=1}^g P(G_i) \sum_{i=1}^b P(B_i|G_j) \log P(B_i|G_j) \quad (15)$$

$$KO \leftarrow \frac{-KO1 + KO2}{-\sum_{j=1}^g P(G_i) \log P(G_i)} \quad (16)$$

Burada B_i , t 'deki i 'ninci alt kategoriyi; $P(B_i)$, i 'ninci kategorinin olasılığını; b , t 'deki alt kategori sayısını; G_j , bir nitelikteki j 'ninci alt kategorisini; $P(G_j)$, j 'ninci kategorisinin olasılığını; g , herhangi bir nitelikteki alt kategori sayısını ve $P(B_i|G_j)$, G_j teriminin, t 'deki i 'ninci alt kategoriye göre şartlı olasılığını temsil eder.

$$KO \leftarrow \sum_{i=1}^f KO_i = \begin{cases} KO_i / \min_{ko}, \min_{ko} > 0 \\ \max_{ko} / KO_i, \max_{ko} < 0 \\ \frac{KO_i - \min_{ko}}{\max_{ko} - \min_{ko}} \times 0.9 + 0.1, \text{diğer} \end{cases} \quad (17)$$

$$KO \leftarrow \sum_{i=1}^f KO_i = \frac{KO_i}{KO_{lat}} \quad (18)$$

Burada \min_{ko} , KO içindeki en küçük, \max_{ko} , KO içindeki en büyük ve KO_{lat} lat niteliği için KO kat sayı değeridir.

BK , Eş. 19, Eş. 20 ve Eş. 21'deki gibi elde edilir ve Eş. 22 ve Eş. 23'teki gibi normalize edilir.

$$BK1 = \sum_{i=1}^b P(B_i) \log P(B_i) \quad (19)$$

$$BK2 = \sum_{i=1}^b \sum_{j=1}^g P(G_i) \sum_{i=1}^b P(B_i|G_j) \log P(B_i|G_j) \quad (20)$$

$$BK \leftarrow -BK1 + BK2 \quad (21)$$

$$BK \leftarrow \sum_{i=1}^f BK_i = \begin{cases} BK_i / \min_{bk}, \min_{bk} > 0 \\ \max_{bk} / BK_i, \max_{bk} < 0 \\ \frac{BK_i - \min_{bk}}{\max_{bk} - \min_{bk}} \times 0.9 + 0.1, \text{diğer} \end{cases} \quad (22)$$

$$BK \leftarrow \sum_{i=1}^f BK_i = \frac{BK_i}{BK_{lat}} \quad (23)$$

Burada \min_{bk} , BK içindeki en küçük, \max_{bk} , BK içindeki en büyük ve BK_{lat} lat niteliği için BK kat sayı değeridir.

2.3. Doğrusal Tahminleme Modeli (Linear Estimation Model)

Bu çalışmada kullanılan ilk tahminleme modeli, hareket verilerindeki SOG , COG , lat ve lng niteliklerinden faydalanarak gemilerin sonraki rotayı tahminleme ve çarpışma olasılıklarını hesaplama üzerine olmuştur. Bu teknik en yaygın kullanılan geleneksel bir yöntemdir. Özellikle gemi hareketlerinin doğrusal ya da doğrusala yakın olduğu boğaz geçişlerinde başarı oranları oldukça yüksek çıkabilecek bir yaklaşımdır [56]. Testlerde bu nitelikler sadece oldukları gibi değerlendirilmiş, ağırlıklandırılarak test edilmemiştir. Çünkü aşağıdaki sözde kodlarda da verildiği gibi hesaplarda lat , lng , SOG ve COG , trigonometrik formüllerde değerlendirilmektedir. Bir

geminin bir sonraki noktasını hesaplayan sözde kod Şekil 3'teki gibidir.

```

COG = derecedenRadyanaDönüştür(COG)
yeni_x = (1,85 x SOG x cos(COG) x min) / 60
yeni_y = (1,85 x SOG x sin(COG) x min) / 60
yeni_lat = lat + (yeni_x / 6378) x (180 / 3,14)
yeni_lng = lng + (yeni_y / 6378) x (180 / 3,14) / cos(lng x 3,14 / 180)
target_coordinates = { yeni_lat , yeni_lng }

```

Şekil 3. Bir geminin bir sonraki noktasını hesaplayan sözde kod

(The pseudocode calculating the next point of a ship)

İki geminin buldukları *lat* ve *lng* bilgilerine göre aralarındaki mesafeyi hesaplayan sözde kod Şekil 4'teki gibidir.

```

metre_sabiti = 6371
φ1 = derecedenRadyanaDönüştür (lat1)
φ2 = derecedenRadyanaDönüştür (lat2)
Δφ = derecedenRadyanaDönüştür (lat2-lat1)
Δλ = derecedenRadyanaDönüştür (lon2-lon1)
sin_cos_carpim = cos(φ1) x cos(φ2) x sin(Δλ/2) x sin(Δλ/2)
sin_cos_degeri = sin(Δφ/2) x sin(Δφ/2) + sin_cos_carpim
arctan2_degeri = 2 x arctan2(sqrt(sin_cos_degeri),sqrt(1-sin_cos_degeri))
mesafe_farki = metre_sabiti x arctan2_degeri

```

Şekil 4. İki geminin arasındaki mesafeyi hesaplayan sözde kod

(The pseudocode calculating the distance between two ships)

2.4. Otoresif Entegre Hareketli Ortalama Modeli (Autoregressive Integrated Moving Average Model)

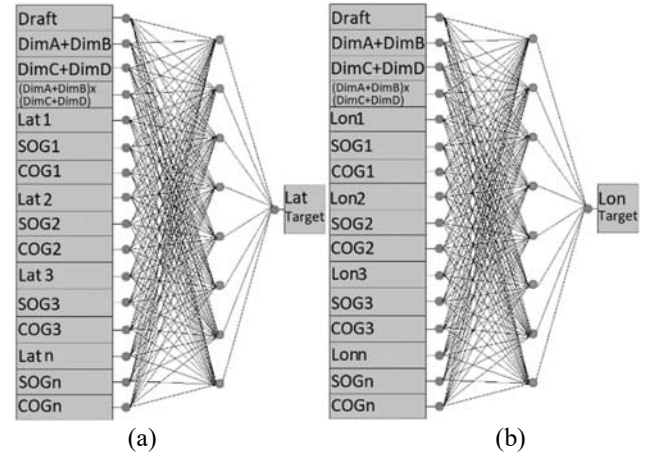
Üstel düzeltme ve ARIMA modelleri, zaman serileri tahmininde en yaygın kullanılan iki yaklaşımdır. Üstel pürüzsüzleştirme modelleri verilerdeki eğilim ve mevsimsellik açıklamasına dayanırken, ARIMA modelleri verilerdeki oto-korelasyonları tanımlamayı amaçlamaktadır [57]. Bu çalışmada mevsimsel bir değişim içermeyen bir veri karakteristiği bulunduğundan ARIMA yöntemi doğruluk testleri için kullanılan bir başka yaklaşım olmuştur.

Bir ARIMA modelinde tahminleme işlemi, (*og*, *fs*, *ho*) parametrelerine bağlıdır. AR (Otomatik Gerileme) terimlerinin sayısı (*og*): AR terimleri yalnızca bağımlı değişkenlerin gecikmeleridir. Örneğin, eğer *og*, 10 ise, *x*(*t*) için tahminler *x*(*t* - 1)... *x*(*t* - 10) olacaktır. MA (Hareketli Ortalama) terimlerinin sayısı (*ho*): MA terimleri, tahmin denklemindeki tahmin hatalarında geciktirilir. Örneğin, eğer *ho* 5 ise, *e*(*t*) için tahminler, *e*(*t*-1)... *e*(*t*-5) olacaktır; burada *e*(*t*), o andaki hareketli ortalama ile gerçek değer arasındaki farktır. Farkların Sayısı (*fs*): Bunlar, mevsimsel olmayan farkların sayısıdır, yani bu durumda birinci dereceden farkı aldık. Böylece ya bu değişkeni geçip *fs* = 0 koyabiliriz ya da orijinal değişkeni geçip *fs* = 1 koyabiliriz. Her ikisi de aynı sonuçları üretecektir. Bu yöntemle tek değişken üzerinden

tahminleme yapılabilmektedir. Bu yüzden bu çalışmada, *lat* ve *lng* için ayrı ayrı ARIMA modelleri oluşturup yeni rota ve çarpışma tahminlemesi yapılmıştır. Bu yöntemde de DT modelinde olduğu gibi ağırlıklandırma kullanılmamıştır. Çünkü veri setindeki *COG* ve *SOG* gibi diğer nitelikler ARIMA tekniği gereği göz ardı edilmiş, sadece *lng* veya *lat* değerlerinden oluşan veri setleri kullanılmıştır. Ancak tahminlemelerin doğruluğunu artırma adına gemi tipleri için ayrı ayrı *lng* ve *lat* veri setleri oluşturularak ayrı ayrı ARIMA modelleri elde edilmiştir.

2.5. Çok Katmanlı Algılayıcı Modeli (Multilayer Perceptron Model)

Yapay Sinir Ağları ile zaman serisi tahminleri oldukça yaygın bir tekniktir. Bu çalışmada Çok Katmanlı Algılayıcı (ÇKA) modeli üzerinden gemi rota ve çarpışma tahminlemeleri yapılmıştır. Çalışma kapsamında gerçekleştirilen ARIMA modellerinde olduğu gibi, ÇKA modellerinde de *lat* ve *lng* tahminleri her bir gemi tipi için ayrı ayrı gerçekleştirilmiştir. ARIMA modeli tek değişkenli bir yapıya sahip olmasından dolayı, sadece *lat* veya *lng*'ye bağımlı iki ayrı modele izin vermekteydi. ÇKA modellerinde ise *lat* ve *lng* niteliklerine ek olarak geminin iskele/sancak uzunluğu *iu* (Eş. 24), iskele ile sancak arasındaki kemer uzunluğu *ku* (Eş. 25), çevre uzunluğu *çu* (Eş. 26), *DR*, *SOG* ve *COG* ÇKA modellerine girdi olarak verilmiştir (Şekil 5).



Şekil 5. Örnek bir ÇKA modeli (a) *lat* tahminleme için (b) *lng* tahminleme için

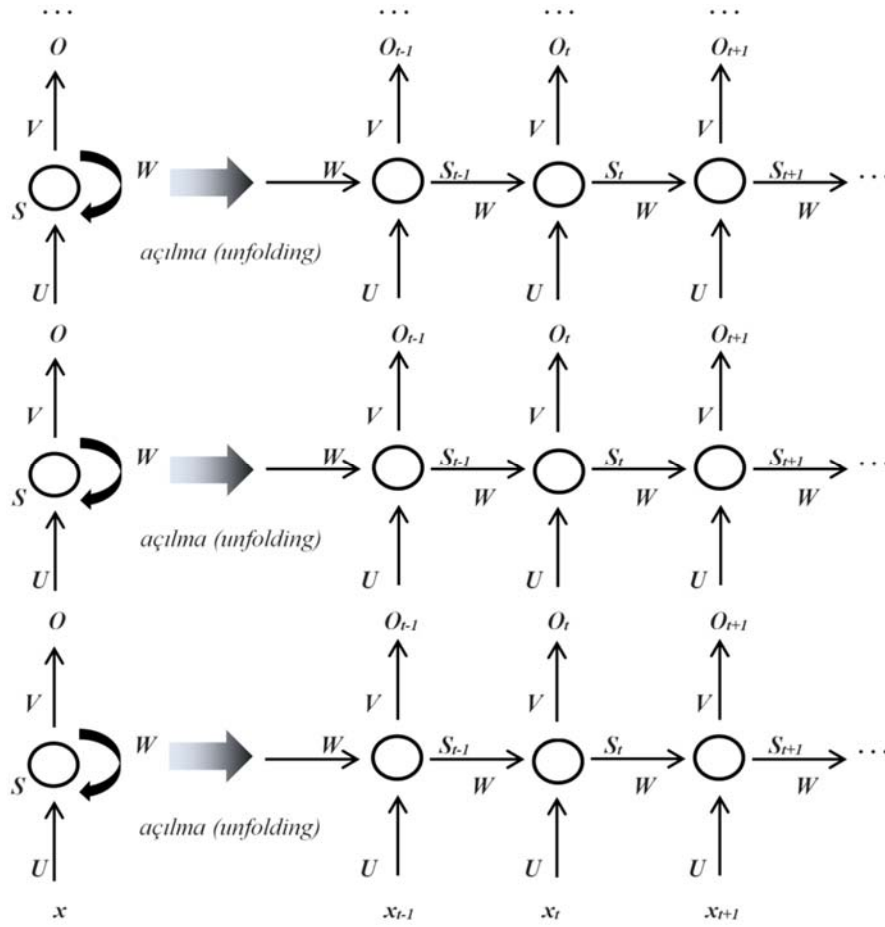
(An example Multilayer Perceptron Model (a) for *lat* estimation (b) for *lng* estimation)

$$iu = DimA + DimB \quad (24)$$

$$ku = DimC + DimD \quad (25)$$

$$çu = 2 x (iu + ku) \quad (26)$$

Burada *DimA*, pruvadan AIS vericisine kadar olan uzunluğu; *DimB*, AIS vericisinden pupaya kadar olan uzunluğu, *DimC*, iskeleden AIS verisine kadar olan uzunluğu ve *DimD*, AIS verisinden sancağa kadar olan uzunluğu temsil etmektedir (Şekil 6).



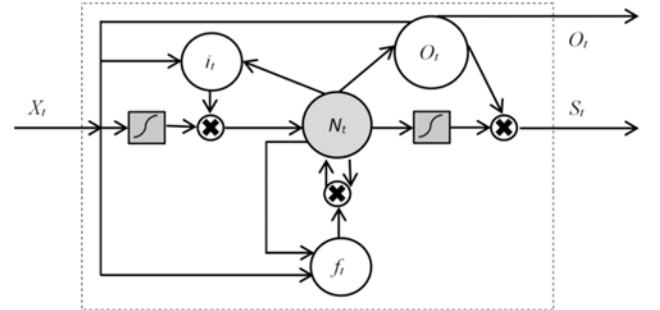
Şekil 7. Çoklu çoklu modeline sahip DRNN (Many-to-Many Deep Recurrent Neural Network Model)

ve *lat* değil, zamandan bağımsız gemilerin *iu*, *ku*, *çu*, *DR*, *SOC* ve *COG* değerleri de tahminlemede kullanıldığı için bu teknik tercih edilmemiştir. LSTM'ler, kısa vadeli ve uzun vadeli bağımlılıkları öğrenebilen özel bir RNN türüdür. GRNN'lerin aksine, LSTM'ler uzun vadeli bağımlılık problemlerinden kaçınmak için tasarlanmıştır. LSTM ağı, zaman içinde geri yayılım kullanılarak eğitilir ve yok olan gradyan sorununun üstesinden gelir. LSTM birimleri ardışık katmanlar arasında bağlanmış bellek bloklarına sahiptir. Her blok, bloğun ve çıkışın durumunu işleyen geçitler içerir. Bir LSTM biriminde üç tip kapı vardır: unutmaya, giriş ve çıkış [62]. Unutmaya kapısı, belirli koşullara dayanarak hangi bilgiyi bloktan atacağını belirler. Giriş kapısı, belirli koşullara göre bellek durumunu güncellemek için girilen değerleri belirler. Çıkış kapısı, belirli koşullara bağlı olarak giriş ve bloğun belleğine göre çıktıyı belirler.

Şekil 8'de gösterildiği gibi, bir LSTM bloğu bir giriş dizisi alır ve daha sonra her kapı tetiklenip tetiklenmeyeceğine karar vermek için aktivasyon birimleri kullanır.

Bu işlem durum değişikliğini ve blok boyunca akan bilginin eklenmesini belirli koşullar altında yapar. Kapılar, eğitim aşamasında öğrenilebilecek ağırlığa sahiptir ve LSTM bloklarını klasik nöronlardan daha akıllı hale getirir. Her LSTM birimi, t zamanında N_t durumuna sahip bir hücre

içerir. Bu hücre bir bellek birimi olarak düşünülebilir. Bu hücrenin okunması veya değiştirilmesi, giriş kapısı i_t (sigmoid geçiti), unutmaya kapısı a_t ve çıkış kapısı O_t üzerinden kontrol edilir. LSTM ünitesi, her bir zaman adımında dört terminalin her birinde (yani, üç kapı ve girdi) iki harici kaynaktan girdi alır. Bu iki harici kaynak, onun X_t girdisi ve aynı katmandaki tüm LSTM birimlerinin önceki gizli katman değeri S_{t-1} 'dir.



Şekil 8'de gösterildiği üzere; kapılar, toplam girişlerini lojistik fonksiyonuna sokarak etkinleştirir (Eş. 33).

$$i_t = \sigma(UX_t + WS_{t-1} + VN_{t-1}) \quad (33)$$

Giriş terminalindeki toplam girdi, doğrusal olmayan *tanh* fonksiyonundan geçirilir. LSTM, sonuçta meydana gelen bu aktivasyon sonucunu, giriş kapısının aktivasyonu ile çarpır. Daha sonra, hücre durumunu, a_t 'nin aktivasyonu ile çarptıktan sonra, hücre durumu ile çarpmanın sonucunu toplar (Eş. 34 ve Eş. 35).

$$a_t = \sigma(UX_t + WS_{t-1} + VN_{t-1}) \quad (34)$$

$$N_t = a_t N_{t-1} + i_t \tanh(UX_t + WS_{t-1}) \quad (35)$$

Ardından, LSTM, güncellenmiş hücre durumunu *tanh* fonksiyonundan geçirir ve çıkan sonucu, LSTM birim S_t 'den nihai çıktısını belirlemek için, çıkış kapısının aktivasyonları ile çarpır (Eş. 36 ve Eş. 37).

$$O_t = \sigma(UX_t + WS_{t-1} + VN_{t-1}) \quad (36)$$

$$S_t = O_t \tanh(N_t) \quad (37)$$

DRNN modelleri yine ARIMA ve ÇKA modellerinde olduğu gibi tüm gemi tipleri için ayrı ayrı oluşturulmuştur. Ayrıca ÇKA modellerinde olduğu gibi *lat*, *lng*, *iu*, *ku*, *çu*, *DR*, *SOG* ve *COG* ÇKA modellerine girdi olarak verilmiş ve bu değerler hem oldukları gibi hem de ağırlıklandırılarak ayrı ayrı test edilmiştir.

3. SONUÇLAR VE TARTIŞMALAR (RESULTS AND DISCUSSIONS)

Çalışmanın testleri hem Marmara Denizi hem de Çanakkale Boğazı verileri ile gerçekleştirilmiştir. Çanakkale Boğazı'nda gemiler transit ve çoğu kez lineer bir rotada seyahat ederken, Marmara Denizi'ndeki gemi hareketlerinin daha düzensiz olduğu bilinmektedir. Birbirine bağlı bu iki bölge hem sahip oldukları bu farklı karakteristik özelliklerden hem de yoğun gemi trafiğinden dolayı analiz için tercih edilmiştir. AIS verilerinin büyük boyutta olmasından dolayı, algoritmaların kullanımının zorlaşacağı ön görülmüştür. Bu nedenle, önceki bölümde de anlatıldığı üzere önce, veri tabanında 3NF kuralına uygun 2 ayrı tabloda bu veriler tutularak boyutunun azaltılması sağlanmıştır.

İkinci olarak; çalışmada kullanılan niteliklerin ağırlıklandırma yaklaşımı gereği ve her geminin hareket davranışının farklılık göstermesinden dolayı, 9 ana gemi tipine göre tüm hareket verileri gruplandırılmıştır. Bu gemi tipleri; kargo, tanker, yolcu, römorkör/yedekleme, şahsi yat gemileri, yüksek süratli tekneler, balıkçı tekneleri, özel eğlence tekneleri ve yelkenlilerdir. Veri tabanındaki AIS verileri; Çanakkale Boğazı için 3.879 gemi ve onların 19.263.121 hareket verisi iken, Marmara Denizi için 1.990 gemi ve onların 9.509.110 hareket verisi olarak, toplamda 5.869 gemi ve 28.772.231 hareket verisidir. Gemi tiplerine göre gruplandırma yapıldığında bazı gemi tiplerin AIS tarafından tespit edilemediğinden dolayı gemi tipi belirsiz olarak sisteme kaydedildiği görülmüştür. Tablo 2'nin ilk 3 kolonunda gemi tiplerine göre adetleri ve hareket veri sayıları (HVS) verilmiştir. Eğer bu yüksek sayıdaki gemileri ve onların hareket verileri veri tabanından silinirse o gemilerin rota tahmini ve diğer tanımlı gemilerle çarpışma durumları göz ardı edilecekti. Bu nedenle bu tipi belirsiz gemilerin karakteristik özelliklerini verebilecek; boyutları ile ilgili olan nitelikler *iu*, *ku* ve *çu*, manevra ve hız kabiliyetleri ile ilgili olan diğer nitelikler *SOG* ve *COG*, ve konum bilgileri *lat* ve *lng* nitelikler baz alınarak SOM++ algoritması ile kümeleme işlemi yapılmıştır.

Kümeleme işleminden önce, bu büyük veri içindeki 10 saniyede bir, ondalık sistemde kaydedilmiş konum bilgileri olan *lat* ve *lng* verilerinin peşi sıra aynı geldiği veya göz ardı edilebilecek şekilde çok az değiştiği gözlemlenmiştir. Bu da hareket verilerinin belirli koordinatlarda gruplanabilmesine ve veri sayısının azalmasına olanak sağlamaktadır. Bu azaltma işlemi Tablo 1'deki gemi hızı ve AIS veri gönderme sıklıkları ile orantılı olarak yapılmıştır. Burada 5 farklı kategoride hıza sahip geminin olduğu görülmektedir. En hızlı gemi olan yüksek süratli gemilerin koordinat verileri virgülden sonra 6 basamak, 2. en hızlı gemi olan şahsi yat gemilerinkini 5 basamak, 3. en hızlı olan yolcu gemileri, özel eğlence tekneleri ve römorkör/yedekleme gemilerinkini 4, 4. en hızlı olan kargo gemileri ve balıkçı teknelerinkini 3 ve en yavaş olan tanker ve yelkenlilerinkini de 2 basamak alarak konum verisi kaybı olmadan HVS azaltılabilmektedir. Örneğin

Tablo 2. Tiplerine göre gemiler ve veri tabanındaki HVS değişimleri
(The vessels by their types and the variations of movement data dimensions in the database.)

Gemi tipi	İlk gemi adetleri	İlk HVS	BKİ sonrası HVS	SOM++ sonrası gemi adetleri	SOM++ sonrası HVS
Kargo	3.517	16.059.234	152.984	3.579	156.826
Tanker	921	5.638.956	12.312	984	13.098
Yolcu gemisi	432	1.975.231	198.962	484	221.763
Römorkör/yedekleme	408	1.883.274	190.298	452	212.007
Şahsi yat gemileri	112	456.129	207.629	143	262.409
Yüksek süratli tekneler	87	394.909	189.231	114	243.662
Balıkçı tekneleri	86	358.982	3.292	135	5.198
Özel eğlence gemileri	43	119.351	21.287	77	37.713
Yelkenli	36	91.829	2.313	51	3.208
Belirsiz	377	1.644.336	-	-	-
Toplam	5.869	28.772.231	978.308	5.869	1.155.884

40.5733614431152 kuzey enlemi ile 28.077220916748 doğu boylamındaki yüksek süratli geminin konum bilgileri, virgülden sonra 6 basamak alınarak yani 40.573361 ve 28.0772209 olarak güncellenmiştir. HVS, tipi belirsiz olanlar hariç, Tablo 2'nin 4. kolonunda verilmiştir.

Boyut küçültme işlemi (BKİ) sonrasında 377 tipi belirsiz geminin 1.644.336 adet hareket verileri, SOM++ algoritması ile 3x3 nörona yani 9 kümeye bölünmüş ve diğer gemi türlerinin verilerine, benzerlikleri dahilinde dağıtılarak eklenmiştir. Eklenirken yine BKİ yapılmış ve Tablo 2'nin 5. ve 6. kolonlarındaki gibi veriler analiz için son hallerini almıştır. Toplamda 5.869 gemi ve 1.155.884 hareket verisi ile analizler yapılmıştır.

3.1. Rota Tahminleme Sonuçları (Route Estimation Results)

Çalışmada gemi rotalarının tahminlemesi, mevcut konumlarından sonraki bulunması beklenen konumun bulunması üzerinedir. Zaman serisi çalışmalarında yöntemlerin başarılarını karşılaştırmak için kullanılan Ortalama Mutlak Yüzde Sapması (Mean Absolute Percentage Error- *MAPE*) formülü Eş. 38'deki gibidir.

$$MAPE = \left(\sum_{i=1}^n \left| \frac{T_i - H_i}{H_i} \right| \times 100 \right) / n \quad (38)$$

Burada *T* verisi tahmin edilen ve *H* verisi de olması gereken hedefi simgelemektedir.

MAPE bu hali ile tek bir verinin tahminlemesi için kullanılmış yöntemin hata yüzdesini vermek için uygundur. Bu çalışmada testlerde iki boyutlu *lat* ve *lng* verileri üzerinden tahminleme yapıldığından dolayı, tahmin edilen koordinatın *lat* ve *lng* verileri için ayrı ayrı *MAPE* değerleri bulunup onların da ortalamaları alınarak modellerin hata yüzdeleri ortaya konulmuştur.

Tablo 3'te Çanakkale Boğazı ve Marmara Denizi verileri için DT ve ARIMA yöntemlerinin *MAPE* değerleri verilmiştir. Çanakkale Boğazı için bu değerlerin %6'nın

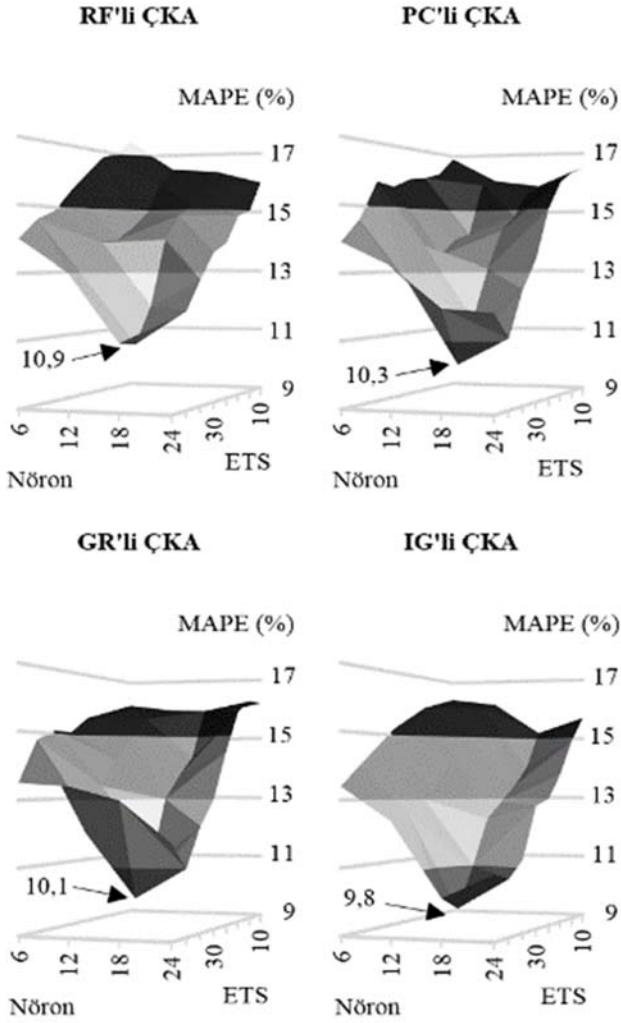
altında kalarak oldukça düşük çıktığı görülürken, Marmara Denizi verileri için *MAPE* değerlerinin %28 civarına yükselmiş olduğu görülmektedir. Bu değerler yorumlanırsa, Çanakkale Boğazı gibi geçiş yerleri için rota tahminlemesi yapılırken DT ve ARIMA tercih edilebilirken, Marmara Denizi gibi daha düzensiz hareketlerin olduğu bölgelerde tercih edilmemesi gerektiği söylenebilir.

Aynı testler ÇKA yöntemiyle, ağırlıklandırma yöntemleri kullanılarak, 6'dan 24'e kadar farklı sayıda gizli katman nöronuyla ve 10'dan 45'e kadar farklı sayıda eğitim turu sayısı (ETS) yapılmıştır. *MAPE* sonuçları tüm gemi türleri için ağırlıklı ortalamaları alınarak gerçekleştirilmiştir. Bu sonuçlar yorumlandığında, Çanakkale Boğazı verileri için en başarılı sonucu veren, *PC* ağırlıklandırma yöntemi, 12 adet gizli katman nöronu ve 30 ETS içeren ÇKA modelinin, %4,5 ile yine *MAPE* değerini oldukça düşük elde ettiği gözlemlenmiştir. Marmara Denizi için ise *IG* ağırlıklandırma yöntemi, 18 adet gizli katman nöronu ve 40 ETS içeren ÇKA modeli %9,8 *MAPE* değeri elde etmiştir. Böylece, DT ve ARIMA'ya göre gözle görünür şekilde *MAPE* değerini azalttığı söylenebilir. Şekil 9'daki grafiklerde Marmara Denizi verileri için sonuçlar verilmiştir.

Testler DRNN ile yapıldığında; ÇKA modellerinde olduğu gibi yine ağırlıklandırma yöntemleri kullanılarak, Şekil 7'deki çoka çok yapısındaki 3x3'ten 6x6'ya kadar farklı sayıda gizli katman nöronuyla ve 10'dan 45'e kadar farklı sayıda eğitim turu sayısı (ETS), oldukça başarılı sonuçların elde edildiği gözlemlenmiştir. Bu sonuçlar *MAPE* değerlerinin yine tüm gemi türleri için ağırlıklı ortalamaları alınarak gerçekleştirilmiştir. Bu sonuçlar yorumlandığında, Çanakkale Boğazı verileri için en başarılı sonucu veren DRNN modelinin, %3,6 *MAPE* değerini veren *GR* ağırlıklandırma yöntemi, 4x4 adet gizli katman nöronlu ve 35 ETS içeren model olduğu görülmüştür. Marmara Denizi için ise *IG* ağırlıklandırma yöntemi, 5x5 adet gizli katman nöronlu ve 40 ETS içeren DRNN modelinin olduğu söylenebilir. Çünkü %6,2 *MAPE* değeri ile diğer tüm modellerden daha az hatalı bir yapı sunmuştur. Şekil 10'daki grafiklerde Marmara Denizi verileri için sonuçlar verilmiştir.

Tablo 3. DT ve ARIMA için *MAPE* değerleri (*MAPE values for Linear Estimation Model and ARIMA*)

Gemi tipi	Çanakkale Boğazı – DT için <i>MAPE</i> değerleri (%)	Çanakkale Boğazı – ARIMA için <i>MAPE</i> değerleri (%)	Marmara Denizi – DT için <i>MAPE</i> değerleri (%)	Marmara Denizi – ARIMA için <i>MAPE</i> değerleri (%)
Kargo	3,33	3,17	17,82	17,54
Tanker	2,86	2,55	15,32	14,97
Yolcu gemisi	4,98	4,33	20,83	20,88
Römorkör/yedekleme	5,11	4,74	24,14	24,02
Şahsi yat gemileri	6,33	5,91	32,72	34,10
Yüksek süratli tekneler	9,19	8,88	39,58	37,18
Balıkçı tekneleri	10,1	9,86	40,19	39,46
Özel eğlence gemileri	6,73	6,47	33,26	31,94
Yelkenli	9,34	9,07	41,50	39,76
Ağırlıklı Ortalama	5,97	5,64	28,16	27,82



Şekil 9. ÇKA modellerinin Marmara Denizi verileri için *MAPE* sonuçları grafikleri
(*MAPE* results graphs of Multilayer Perceptron models for Marmara Sea data)

Sonuç olarak, rota tahminlemede Çanakkale Boğazı verileri için tüm modellerin %94 başarı barajını geçtikleri görülürken, Marmara Denizi verileri için yapay sinir ağları içeren modellerin daha başarılı oldukları söylenebilir. Özellikle LSTM içeren DRNN modelleri, verilere öznetelik seçim tekniklerinin uygulanması ve onların verdikleri kat sayılarla verilerin ağırlıklandırılması yöntemiyle, düzensiz bir yapıya sahip Marmara Denizi verileri için sahip oldukları %6,7 ve %6,2 arasındaki *MAPE* değerleri ile en başarılı yaklaşımlar olmuştur. Ayrıca, bu model üzerinden Marmara Denizi verileri ağırlıklandırma yapmadan saf halleri ile kullanımının sonucunda hataların ancak %9,4'e kadar indiği gözlemlenmiştir.

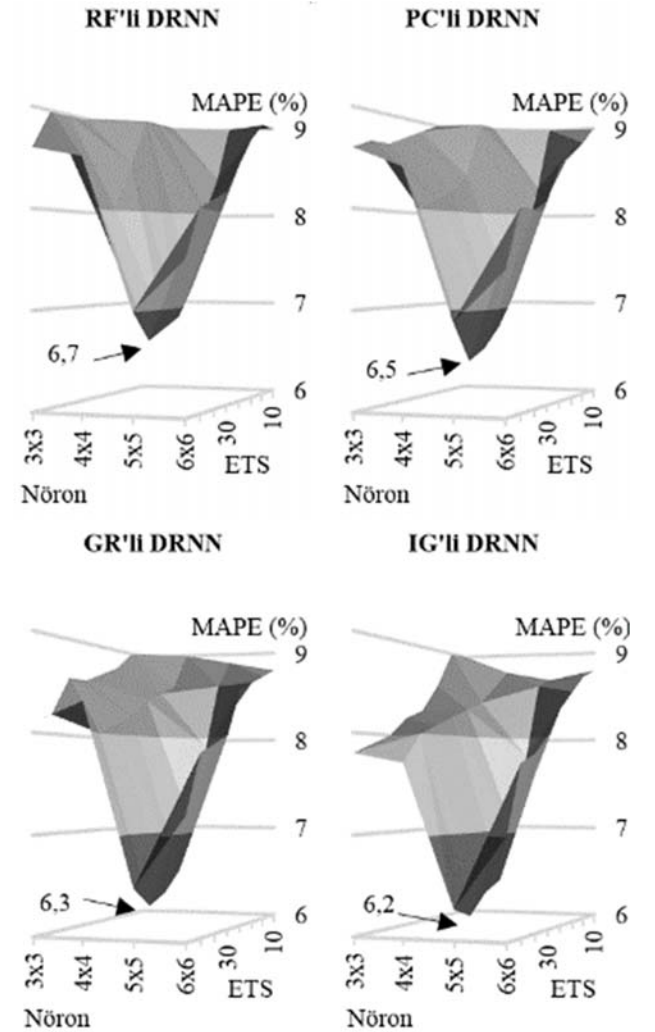
3.2. Çarpışma Tahminleme Sonuçları (Collision Estimation Results)

Test verileri gerçek veriler olduğu için hiçbir kaza içermeyen verilerden oluşmaktadır. Çarpışma tahminlemesi yapabilmek için Çanakkale Boğazı ve Marmara Denizi

verilerinde iki geminin aynı anda aynı konumda olabilecekleri şekilde 100'er durum simule edilmiş ve bu veriler üzerinden çarpışma tahminlemesi testleri gerçekleştirilmiştir. Bu testlerde sadece bu 100 durumu yakalamak tek başına doğruluk (A_c) için yeterli değildir. Kaza durumu olmamasına rağmen kaza olarak yakalanan durumlar da sistemin başarısızlığını göstereceğinden F-skor (F_s) değerlerine de bakılmıştır. A_c formülü Eş. 39'da ve F_s formülü Eş. 40'da verilmiştir.

$$A_c = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \quad (39)$$

$$F_s = 2 \times \left(\left(\frac{T_p}{T_p + F_p} \times \frac{T_p}{T_p + T_n} \right) / \left(\frac{T_p}{T_p + F_p} \times \frac{T_p}{T_p + T_n} \right) \right) \quad (40)$$



Şekil 10. DRNN modellerinin Marmara Denizi verileri için *MAPE* sonuçları grafikleri
(*MAPE* results graphs of Deep Recurrent Neural Network models for Marmara Sea data)

Burada T_p , doğru onaylanmış; T_n , doğru reddedilmiş; F_p , hatalı onaylanmış ve F_n , hatalı reddedilmiş değerlerini simgelemektedir. Testler'de Çanakkale Boğazı verileri için

DT ve ARIMA'nın dışında; bu veriler için rota tahminlemede en başarılı sonucu vermiş versiyonu olan *PC* ağırlıklandırma yöntemi, 12 adet gizli katman nöronlu ve 30 ETS içeren ÇKA modeli, ve *GR* ağırlıklandırma yöntemi, 4x4 adet gizli katman nöronlu ve 35 ETS içeren DRNN modeli kullanılmıştır. Marmara Denizi verileri için yine DT ve ARIMA'nın dışında; bu veriler için rota tahminlemede en başarılı sonucu vermiş versiyonu olan *IG* ağırlıklandırma yöntemi, 18 adet gizli katman nöronlu ve 40 ETS içeren ÇKA modeli ve *IG* ağırlıklandırma yöntemi, 5x5 adet gizli katman nöronlu ve 40 ETS içeren DRNN modeli kullanılmıştır. Sonuç karşılaştırmaları Tablo 4'te verilmiştir. Bu sonuçlar yorumlandığında, rota tahminleme ile hemen hemen paralel sonuçların olduğunu görmekteyiz. DRNN'nin Çanakkale Boğazı için %100 ve Marmara Denizi için %99 çarpışma durumlarını yakalayabilmesi ve Çanakkale Boğazı için %97,2 ve Marmara Denizi için %96,5 F_s değerleri ile en başarılı model olduğu söylenebilir.

Tablo 4. DT, ARIMA, ÇKA ve DRNN için A_c ve F_s değerleri

(Accuracy values for Linear Estimation Model, ARIMA, Multilayer Perceptron and Deep Recurrent Neural Network models)

	DT	ARIMA	ÇKA	DRNN
Çanakkale Boğazı - A_c (%)	92	94	100	100
Çanakkale Boğazı - F_s (%)	86,2	87,5	96,5	97,2
Marmara Denizi - A_c (%)	80	81	95	99
Marmara Denizi - F_s (%)	75,2	77,2	93,2	96,5

5. SONUÇLAR (CONCLUSIONS)

Bu çalışma ile deniz ulaşımı konusunda günümüzde tüm sektörlerde hızla yaygınlaşan yapay zekâ yaklaşımları ele alınmış ve yeni yöntemler ortaya konulmuştur. İki farklı karakteristik özelliği bulunan Çanakkale Denizi ve Marmara Denizi AIS verileri kullanılarak, gemi hareketlerinin gemi tiplerine göre değişimi gösterilmiş ve bu duyarlılığa dayanarak da her gemi tipi için ayrı bir rota ve çarpışma tahminleme modelleri sunulmuştur. Kullanılan geleneksel yöntemler olan DT ve ARIMA'nın yanında makine öğrenmesi algoritması ÇKA ve derin öğrenme algoritması DRNN kullanılarak bu veriler üzerinden testler gerçekleştirilmiştir. Veriler oldukları gibi değil ağırlıklandırma yapılarak hassasiyetleri artırılmış ve çok daha başarılı *lat* ve *lng* tahminlemelerinin yapılabileceği kanıtlanmıştır. Bu çalışmanın literatüre bir başka katkısı da AIS sisteminin tespit edilemeyip, tipi belirsiz olarak kaydedilen gemilerin, göz ardı edilmeden SOM++ yöntemiyle keşifleri yapılmış ve çarpışma ihtimallerinin değerlendirilmiş olmasıdır. Sonuçlar yorumlandığında, düzenli hareket yapısındaki Çanakkale Boğazı verilerinde tüm modellerin tatmin edici bir *MAPE*, A_c ve F_s değerlerini verdikleri gözlemlenirken; düzensiz hareket yapısına sahip Marmara Denizi verilerinde *IG* öznitelik seçim tekniğini kullanan ve LSTM içeren DRNN yönteminin %6,2 *MAPE*, %99 A_c ve %96,5 F_s değerleri ile öne çıkan yöntem olduğu

görülmüştür. Gelecek çalışmalar adına, bu önerilen derin öğrenme modeli başka deniz ve boğazlara da uyarlanıp kullanılabilir; hatta, bu model her ne kadar denizcilik verisi olan AIS verileri kullanılmış ve test edilmiş olsa da finans, sağlık, meteoroloji, vb. sektörlerde zaman serisi analizleri için yeni çalışmalara ışık tutabilir.

TEŞEKKÜR (ACKNOWLEDGEMENT)

Bu çalışmanın testlerini yapabilmek için gerekli olan AIS verilerinin sağlanmasında yardımcı olan Dokuz Eylül Üniversitesi Denizcilik Fakültesi'nden Prof. Dr. Selçuk Nas'a teşekkür ederim.

KAYNAKLAR (REFERENCES)

1. UNCTAD, Review of Maritime Transport, 2016.
2. SOLAS, Safety of Life At Sea Consolidated Edition, 2014.
3. IMO, Revised Guidelines for the Onboard Operational Use of Shipborne Automatic Identification Systems (AIS), 2015.
4. Mustaffa, M., Ahmat, N. H., Ahmad, S., Mapping vessel path of marine traffic density of Port Klang, Malaysia using Automatic Identification System data, International Journal of Science and Research (IJSR), 4 (11), 245-248, 2015.
5. Cimino, G., Ancieri, G., Horn, S., Bryan, K., Sensor data management to achieve information superiority in maritime situational awareness, CMRE Formal Report, NATO Unclassified, 2014.
6. ITU, Technical characteristics for an automatic identification system using time division multiple access in the VHF maritime mobile frequency band, Recommendation ITU-R M.1371-5, 2014.
7. Aarsæther, K. G., Moan, T., Estimating navigation patterns from AIS, Journal of Navigation, 62 (4), 587-607, 2009.
8. Sang, L. Z., Yan, X. P., Wall, A., Wang, J., Mao, Z., CPA calculation method based on AIS position prediction, Journal of Navigation, 69 (6), 1409-1426, 2016.
9. Tang, Q., Gu, D., Day-ahead electricity prices forecasting using artificial neural networks, IEEE International Conference on Artificial Intelligence and Computational Intelligence, 2, 511-514, 2009.
10. Vahidinasab, V., Jadid, S., Kazemi, A., Day-ahead price forecasting in restructured power systems using artificial neural networks, Electric Power Systems Research, 78 (8), 1332-1342, 2008.
11. Zhang, Z. G., Yin, J. C., Wang, N. N., Hui, Z. G., Vessel traffic flow analysis and prediction by an improved PSO-BP mechanism based on AIS data, Evolving Systems, 10 (3), 397-407, 2009.
12. Århus, G. H., Salen, S. R., Predicting shipping freight rate movements using recurrent neural networks and ais data-on the tanker route between the Arabian Gulf and Singapore, Master's thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2018.

13. Nguyen, Q. V., Extreme weather disaster resilient port and waterway infrastructure for sustainable global supply chain, University of Mississippi, 2017.
14. Xiao, Z., Ponnambalam, L., Fu, X., Zhang, W., Maritime traffic probabilistic forecasting based on vessels' waterway patterns and motion behaviors, *IEEE Transactions on Intelligent Transportation Systems*, 18 (11), 3122-3134, 2017.
15. Pallotta, G., Vespe, M., Bryan, K., Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction, *Entropy*, 15 (6), 2218-2245, 2013.
16. Lei, B., A DBSCAN based algorithm for ship spot area detection in AIS trajectory data. *MATEC Web of Conferences*, EDP Sciences, 291, 2019.
17. Liang, M., Liu, R. W., Zhong, Q., Liu, J., Zhang, J., Neural network-based automatic reconstruction of missing vessel trajectory data, *IEEE 4th International Conference on Big Data Analytics*, 426-430, 2019.
18. Westerdijk, L., Classifying vessel types based on AIS data, MSc thesis, Vrije University, Amsterdam, Holland, 2019.
19. Zhou, Y., Daamen, W., Vellinga, T., Hoogendoorn, S. P., Ship classification based on ship behavior clustering from AIS data, *Ocean Engineering*, 175, 176-187, 2019.
20. Lei, P. R., Mining maritime traffic conflict trajectories from a massive AIS data, *Knowledge and Information Systems*. 1-27, 2019.
21. Hanyang, Z., Xin, S., Zhenguo, Y., Vessel sailing patterns analysis from s-ais data based on k-means clustering algorithm, *IEEE 4th International Conference on Big Data Analytics*, 10-13, 2019.
22. Mustafa, M., Ahmad, S., Ali, A. M. M., Ahmad, N., Jais, M., Hamidi, M., Data mining analysis on Ships collision risk and marine traffic characteristic of Port Klang Malaysia waterways from automatic identification system (AIS), *International MultiConference of Engineers and Computer Scientists*, 242-246, 2019.
23. Yang, D., Wu, L., Wang, S., Jia, H., Li, K. X., How big data enriches maritime research—a critical review of Automatic Identification System (AIS) data applications, *Transport Reviews*, 39 (6), 755-773, 2019.
24. Bye, R. J., Almklov, P. G., Normalization of maritime accident data using AIS, *Marine Policy*, 109, 103675, 2019.
25. Goerlandt, F., Kujala, P., Traffic simulation based ship collision probability modeling, *Reliability Engineering & System Safety*, 96 (1), 91-107, 2011.
26. Wang, K., Liang, M., Li, Y., Liu, J., Liu, R. W., Maritime traffic data visualization: a brief review, *IEEE 4th International Conference on Big Data Analytics*, 67-72, 2019.
27. Fujii, M., Hashimoto, H., Taniguchi, Y., Kobayashi, E., Statistical validation of a voyage simulation model for ocean-going ships using satellite AIS data, *Journal of Marine Science and Technology*, 1-11, 2019.
28. Liu, Y., Song, R., Bucknall, R., Intelligent tracking of moving ships in constrained maritime environments using ais, *Cybernetics and Systems*, 50 (6), 539-555, 2019.
29. Liu, Z., Wu, Z., Zheng, Z., A novel framework for regional collision risk identification based on AIS data, *Applied Ocean Research*, 89, 261-272, 2019.
30. Kilimci, Z.H., Financial sentiment analysis with Deep Ensemble Models (DEMs) for stock market prediction, *Journal of the Faculty of Engineering and Architecture of Gazi University*, 35 (2), 635-650, 2020.
31. Kaya A., Keçeli A. S., Can A.B., Examination of various classification strategies in classification of lung nodüle characteristics, *Journal of the Faculty of Engineering and Architecture of Gazi University*, 34 (2), 709-725, 2019.
32. Vadaine, R., Hajduch, G., Garello, R., Fablet, R., A Multi-task Deep Learning Architecture for Maritime Surveillance using AIS Data Streams, Cornell University, New York, USA, 2018.
33. Lechtenberg, S., Siqueira Braga, D. D., Hellingrath, B., Automatic identification system (AIS) data based ship-supply forecasting. In *Proceedings of the Hamburg International Conference of Logistics (HICL)*, 3-24, epubli GmbH, 2019.
34. Hoque, X., & Sharma, S. K., Ensembled deep learning approach for maritime anomaly detection system, *Springer ICETIT 2019*, 862-869, 2019.
35. García, S., Luengo, J., Herrera, F., Data preprocessing in data mining, New York: Springer, 59-139, 2015.
36. Packiam, R. M., Prakash, V. S. J., A novel integrated framework based NN modular optimization for efficient analytics on Twitter big data, *Springer Information and Communication Technology for Intelligent Systems*, 213-224, 2019.
37. Krouska, A., Troussas, C., Virvou, M., The effect of preprocessing techniques on Twitter sentiment analysis, *IEEE 7th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 1-5, 2016.
38. Hassler, A. P., Menasalvas, E., García-García, F. J., Rodríguez-Mañas, L., Holzinger, A., Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome, *BMC medical informatics and decision making*, 19 (1), 33, 2019.
39. Benhar, H., Idri, A., Fernández-Alemán, J. L., Data preprocessing for decision making in medical informatics: potential and analysis, *Springer World Conference on Information Systems and Technologies*, 1208-1218, 2018.
40. Tian, C., Hao, Y., Hu, J., A novel wind speed forecasting system based on hybrid data preprocessing and multi-objective optimization, *Applied Energy*, 231, 301-319, 2018.
41. Xiao, L., Wang, J., Yang, X., Xiao, L., A hybrid model based on data preprocessing for electrical power forecasting, *International Journal of Electrical Power & Energy Systems*, 64, 311-327, 2015.
42. Zhang, X., Peng, Y., Zhang, C., Wang, B., Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? Some

- experiment evidences, *Journal of Hydrology*, 530, 137-152, 2015.
43. Al Shalabi, L., Shaaban, Z., Kasasbeh, B., Data mining: A preprocessing engine, *Journal of Computer Science*, 2 (9), 735-739, 2006.
 44. Dash, M., Liu, H., Feature selection for classification. *Intelligent Data Analysis*, 1 (1-4), 131-156, 1997.
 45. Ben-David, A., Comparison of classification accuracy using Cohen's Weighted Kappa, *Expert Systems with Applications*, 34 (2), 825-832, 2008.
 46. Çelik, C., Bilge, H. Ş., Feature selection with weighted conditional mutual information, *Journal of the Faculty of Engineering and Architecture of Gazi University*, 30(4), 585-596, 2015.
 47. Babu, M. S., Vijayalakshmi, V., An effective approach for sub-acute Ischemic stroke lesion segmentation by adopting meta-heuristics feature selection technique along with hybrid Naive Bayes and sample-weighted random forest classification, *Sensing and Imaging*, 20 (1), 7, 2019.
 48. Alwidian, J., Hammo, B. H., Obeid, N., WCBA: Weighted classification based on association rules algorithm for breast cancer disease, *Applied Soft Computing*, 62, 536-549, 2018.
 49. Jindal, R., Taneja, S., A novel weighted classification approach using linguistic text mining, *Int J Comput Appl*, 180 (2), 9-15, 2017.
 50. Cao, X., Ge, Y., Li, R., Zhao, J., Jiao, L., Hyperspectral imagery classification with deep metric learning, *Neurocomputing*, 356, 217-227, 2019.
 51. Zhang, L., Chen, H., Hu, Y., Compressive tracking via weighted classification boosted by feature selection, *Springer Electronics, Communications and Networks V.*, 137-145, 2016.
 52. Fredstam, M., Johansson, G., Comparing Database Management Systems With Sqlalchemy: A Quantitative Study on Database Management Systems, 2019.
 53. De Silva, A. M., Leong, P. H. W., Grammar-Based Feature Generation for Time-Series Prediction, Springer, Berlin, Germany, 2015.
 54. Dogan, Y., Birant, D., Kut, A., SOM++: Integration of self-organizing map and k-means++ algorithms, *Springer International Workshop on Machine Learning and Data Mining in Pattern Recognition*, 246-259, 2013.
 55. İnce H., İmamoğlu S.Z., Keskin H., Comparing self organizing maps with k-means clustering: an application to customer profiling *Journal of the Faculty of Engineering and Architecture of Gazi University*, 28 (4), 723-731, 2013.
 56. Le-Tien, T., Phung-The, V., Routing and tracking system for mobile vehicles in large area. *IEEE 5th International Symposium on Electronic Design, Test & Applications*, 297-300, 2010.
 57. Montgomery, D. C., Jennings, C. L., Kulahci, M., Introduction to time series analysis and forecasting, John Wiley & Sons, 394-419, 2015.
 58. Eswaran, C., Logeswaran, R., An enhanced hybrid method for time series prediction using linear and neural network models, *Applied Intelligence*, 37 (4), 511-519, 2012.
 59. Zhang, Q., Yang, L. T., Chen, Z., Li, P., A survey on deep learning for big data, *Information Fusion*, 42, 146-157, 2018.
 60. Chamorro, J. A., Bermudez, J. D., Happ, P. N., Feitosa, R. Q., A many-to-many fully convolutional recurrent network for multitemporal crop recognition, *Isprs Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4, 2019.
 61. Abdel-Nasser, M., Mahmoud, K., Accurate photovoltaic power forecasting models using deep LSTM-RNN. *Neural Computing and Applications*, 31 (7), 2727-2740, 2019.
 62. Hochreiter, S., Schmidhuber, J., Long short-term memory, *Neural computation*, 9 (8), 1735-1780, 1997.

