Journal of
Marine Science and Engineering

MDPI

Article

A Novel Framework of Real-Time Regional Collision Risk Prediction Based on the RNN Approach

Dapei Liu, Xin Wang *, Yao Cai, Zihao Liu and Zheng-Jiang Liu

College of Navigation, Dalian Maritime University, Dalian 116026, China; 18340878086@163.com (D.L.); caiyao@dlmu.edu.cn (Y.C.); zihaoliu0407@gmail.com (Z.L.); liuzhengjiang@dlmu.edu.cn (Z.-J.L.)
* Correspondence: xin.wang@dlmu.edu.cn

Received: 4 March 2020; Accepted: 20 March 2020; Published: 22 March 2020

Abstract: Regional collision risk identification and prediction is important for traffic surveillance in maritime transportation. This study proposes a framework of real-time prediction for regional collision risk by combining Density-Based Spatial Clustering of Applications with Noise (DBSCAN) technique, Shapley value method and Recurrent Neural Network (RNN). Firstly, the DBSCAN technique is applied to cluster vessels in specific sea area. Then the regional collision risk is quantified by calculating the contribution of each vessel and each cluster with Shapley value method. Afterwards, the optimized RNN method is employed to predict the regional collision risk of specific seas in short time. As a result, the framework is able to determine and forecast the regional collision risk precisely. At last, a case study is carried out with actual Automatic Identification System (AIS) data, the results show that the proposed framework is an effective tool for regional collision risk identification and prediction.

Keywords: regional collision risk; risk identification; risk prediction; Shapely value; Recurrent Neural Network

1. Introduction

Maritime transport is the backbone of international trade and the global economy. In recent decades, the rapid development and great volume of marine transportation [1] lead to higher marine traffic density and complexity which trigger vessel collision accidents easily. In general, vessel collision may cause great loss of human lives and property, as well as severe environment pollution [2]. Specifically, the collision risk is a major indicator for navigators and surveillance operators to judge the collision danger between meeting vessels [3], as well as the surveillance on shore plays an important role in preventing vessel collision accidents [4]. In practice, the operators of Vessel Traffic Service (VTS) system get access to real-time vessel data from modern navigational equipment [5], such as Automatic Radar Plotting Aids (ARPA), Automatic Identification System (AIS), Electronic Chart Display and Information System (ECDIS), etc. With the shipping increasing rapidly, the burden of VTS surveillance operators get heavier. Hence, an advanced marine surveillance framework, which could identify and predict collision risk between meeting ships precisely, has been emerged as an effective tool to lighten the burden of VTS surveillance operators.

Compared with a large amount of research in the field of collision risk identification for single or multiple vessels, the research on regional collision risk identification refers to determine the vessel collision risk in a certain water area, and only a few of studies have been carried out. First of all, based on the historical statistical data, the number of collision accidents per unit time in a certain water area was first considered to describe the regional collision risk by researchers. For example, the Formal Safety Assessment (FSA) concept and Bayesian network method were used to evaluate the collision risk of vessels in Yangtze River waters in China with real accident data [6]. In addition,

the accident data from 1995 to 2015 in the southern port area of Shenzhen were selected and analyzed to build a regional risk assessment model of port waterway [7].

Afterwards, risk modeling and risk analysis based on the probability and consequences of accidents is usually used to identify risk areas with high incidence of accidents in the selected water areas, and used to predict the consequences of accidents to further control the impact of accidents sometimes. Hu [8] considered five types of detailed accident characteristic information and used the information in the risk classification and quantitative modeling process of the pilot port safety assessment of Shanghai Port. Debnath and Chin [9] applied the vessel conflict theory to the quantitative measurement of the waterway collision risk in Singapore harbors, using two proximity indicators in time and space to quantify the collision risk in the waterway. Montewka et al [10] proposed a new geometric collision probability calculation method based on previous experience extracted from a large amount of vessel data using Monte Carlo and genetic algorithms. Le et al. [11] also uses the vessel collision probability of statistics from the Norwegian Classification Society to evaluate the risk factor of vessel collision with the platform.

Meanwhile, the study based on analysis of non-accident data is also a significant approach in the field of regional collision risk identification. In general, researchers have applied vessel factors such as vessel speed, Distance at Closet Point to Approach (DCPA), Time to Closet Point of Approach (TCPA), and ship domains to the study of regional collision risk in selected water waters. Qu et al. [12] used the three indicators of vessel speed dispersion, vessel acceleration, and number of fuzzy vessel domains to evaluate the regional collision risk of vessels in the Singapore strait. Bukhari et al. [5], in order to quantify the collision risk between all vessels in the area, combined DCPA, TCPA, bearing position, and change of Vessel's Compass Degree (VCD) to build a regional collision risk assessment framework based on fuzzy logic.

However, the shortcomings of the analysis based on historical case data are that the number of collisions and collision rate are based on historical data, which describing the results, and cannot indicate the real-time collision risk and danger zone of vessels in the water areas. As an estimated value based on statistical distribution and prediction, the collision probability is still different from the actual collision risk in actual water areas at that time, and there are certain limitations in the real-time nature of regional collision risk. Recently, Liu et al. [13] proposed a regional collision risk calculation model based on AIS data. Cooperative game theory is used to model the collision risk of regional collision risk in the framework, which stars from the information of single vessel's DCPA, TCPA, etc. Compared with the traditional model, the framework can obtain more accurate results of instantaneous regional collision risk, and avoids the influence of traffic flow in the water area. Therefore, in this paper, DCPA, TCPA and Ship Domain Overlapping Index (SDOI) are used as the vessel parameters for measuring regional collision risk of selected water area.

Moreover, the regional collision risk prediction in selected water areas can help the operators of maritime surveillance grasp the trend and value of the regional collision risk more accurately. In recent years, research on regional collision risk prediction has also been carried out and certain results were achieved. Nivoliantou et al. [14] selected the major accidents in the Aegean waters from 2008 to 2012 as historical data and carried out a Bayesian network-based forecasting study for the vessel's navigational environment risk in the Aegean waters. Fan et al. [15] extracted the influencing factors from the data of 218 vessel accidents in the research water area in 2013 and combined these with the Bayesian network to build a prediction model of collision accident levels in the Yangtze River. Kim et al. [16] used a deep neural network called ship traffic extraction network, which consisted with convolutional neural network and a large amount of historical AIS data to make mid-term and long-term predictions of vessel traffic in congested port waters. Okazaki et al. [17] used the Support Vector Machine (SVM) method in the study of collision risk prediction of vessels on the exit of the sea route. Fukuto and Imazu [18] combined the vessel's course prediction work and the Obstacle Zone by Target (OZT) method to determine the collision probability area.

Although the regional collision risk prediction has been studied by various methods in previous studies, certain limitations still exist. Some models require a large amount of historical data of vessel accidents to build a database for predicting the risk of regional vessel collisions, which has a random effect on the real-time quantification and short-term prediction of vessel collision risks in water areas. In addition, the existing models have good effects on mid-term and long-term predictions of regional collision risk, but there is no corresponding research on real-time and short-term predictions of regional collision risk. The real-time and short-term regional collision risk prediction in water areas with limited data is a shortcoming of current research.

In particular, Recurrent Neural Networks (RNN) have always been an interesting and important part of neural networks. Researchers have studied and applied RNNs since the 1990s [19–23]. As an important method in the field of machine learning and artificial intelligence networks, RNNs can realize the prediction of sequence-labeled data with time. Up to now, RNNs have been applied to a variety of problems, especially those involving ordered data processing. In terms of data prediction based on existing databases, RNN has shown great success, such as in image processing [24,25], language processing [26], and prediction problems in different fields [27–31]. RNNs contain recurrent connections that make them more powerful than traditional neural networks to model such sequence data. In the application of speech recognition, RNNs contain network activations that are recurrent from the previous time step as input networks to influence the prediction of the current time step. These activations are stored in the internal state of the network which can save long-term temporal context information in principle. This mechanism allows RNNs to take advantage of changing the context according to the history of input sequence dynamically instead of using static context such as traditional neural networks. Besides, RNNs do not need large memory storage in applications. Such advantages give RNNs better performance in prediction and dealing with sequence-labeled issues compared with traditional forward neural networks.

Based on above observation, to identify and predict regional collision risk precisely, a prediction framework for regional collision risk is proposed in this article. This framework uses limited non-accident data from selected water area to achieve the prediction of regional collision risk more accurately and effectively. Vessel parameters include speed, position, course, length, and the number of vessels entered the selected water area through the traffic lane in a specific period of time, are all extracted from AIS data. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, improved Shapely value method and RNN divide the framework into two steps. As a result, the application of the framework can display the distribution of regional collision risk in future clearly, and mark special areas that deserve attention.

The following content was arranged as follows: In Section 2, the regional collision risk identification is introduced. Section 3 describes the optimized RNN approach for regional collision risk prediction in detail. In Section 4, a case study to verify the feasibility and effectiveness of the framework is developed. Finally, conclusions, discussions and future perspectives are presented in Section 5.

2. The Prediction Framework and Regional Collision Risk Identification

2.1. Procedure of Regional Collision Risk Prediction

Overall, the procedure of proposed prediction framework for regional collision risk as follows: Firstly, by combining DBSCAN spatial clustering algorithm and improved Shapely value method, the real-time regional collision risk in selected water area is obtained. Then, history data consist of regional collision risk and the number of vessels entered the selected water area through traffic lane during the corresponding time period are used to build the RNN training data set. Based on the training data set, the prediction model was trained with proposed RNN algorithm in Tensorflow on a personal CPU system (Core i7 with 8 GB RAM). Furthermore, for certain time point, the regional collision risk can be predicted by trained RNN approach, which helps maritime traffic surveillance

department to grasp the regional collision risk in the future and allocate reasonable surveillance force more reasonable.

2.2. The Regional Collision Risk Identification

As an approach of expressing the collision risk of vessels in a certain area, the regional collision risk is the overall collision risk formed by all vessels in a certain area. Therefore, the contribution of all vessels in the area to the collision risk is used in the measurement of regional collision risk.

In a certain busy water area that meets the research requirements, plenty of vessels will appear at the same time. To calculate the regional collision risk in the water area at that moment, a clustering algorithm based on spatial density is used for data processing. To improve the calculation efficiency of regional collision risk and reduce the computational complexity, this study uses the spatial density clustering algorithm DBSCAN to cluster vessels in selected water areas.

Clustering belongs to an unsupervised learning method and is a branch of data mining. So far, many clustering algorithms have been proposed [32–37], such as K-Means clustering [32] and DBSCAN [34]. Among them, the DBSCAN algorithm is a density clustering method proposed by Ester [34], as a classic density-based clustering algorithm, the main idea of the DBSCAN algorithm is to cluster a given set of objects in space, group densely distributed points into a clustering cluster, and leave the remaining sparse individual points apart marked as noise points. The DBSCAN algorithm can identify noise among cluster objects, find clusters of any shape and size that reaches the target density. The clustering results of DBSCAN algorithm are relatively more accurate when there is no specified number of clustering results, and DBSCAN can find clusters of any shape in the data with noisy points. The DBSCAN algorithm has been widely used in various fields presently, such as medicine, biology [38], analytical chemistry [39], marine transportation [3,4], etc. This study used the DBSCAN clustering algorithm in the process of quantifying the regional collision risk in selected water area, which improved the efficiency and visualization of the quantization process, while reducing the amount of calculation and the complexity of the calculation. After clustering with DBSCAN, several clusters will be obtained in the area as shown in Figure 1.

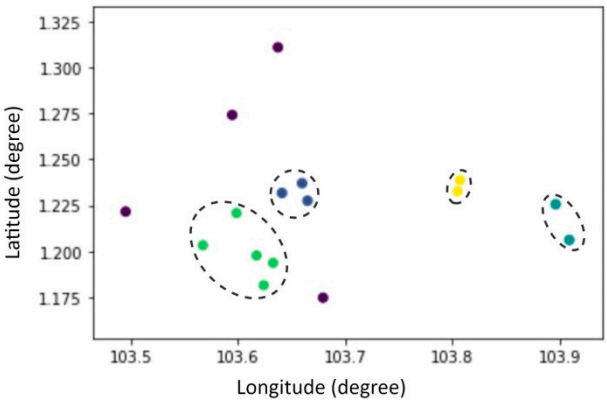

Figure 1. Vessel clusters.

The vessels in this area are divided into several clusters. Therefore, the cluster's collision risk is obtained by calculating the collision risk of every single vessel based on Equation (1), and then the regional collision risk is calculated from the cluster's collision risk based on Equation (2).

$$CRC = \sum_{i=1}^{n} CR_i * W_i, \quad (1)$$

where *CRC* refers to collision risk of each cluster, CR_i refers to collision risk of every single vessel and W_i refers to every single vessel's contribution.

$$RCR = \sum_{j=1}^{m} CRC_j \times W_j, \tag{2}$$

where *RCR* refers to regional collision risk, CRC_j refers to collision risk of every single cluster and W_j refers to every single cluster's contribution.

There are n vessels in each cluster, and any vessel in the cluster and other vessels form n vessel pairs. The collision risk of each single vessel in the cluster can be determined by summing all collision risk of corresponding vessel pairs. The collision risk of all vessel pairs of every single vessel are summed and expressed as:

$$CR_i = \frac{1}{n-1}(CR_{i1} + CR_{i2} + CR_{i3} + \cdots + CR_{in}), \tag{3}$$

where CR_i is the collision risk of Vessel i, CR_{ij}, $j = 1, 2, \ldots, n$ is the collision risk of vessel pairs which include Vessel i.

In this study, the analytical method was used to calculate the collision risk of the vessel pair. As a method to determine the collision risk through the vessel dynamic elements directly, the analytical method is more objective than the fuzzy logic method based on marine experts. Besides, it is simpler than the method based on artificial intelligence and the calculation results will unaffected by the previous training set. To overcome these limitations, this paper uses DCPA, TCPA, and SDOI to calculate collision risk to improve the calculation accuracy.

Considering the simplicity and practical application of the model, this study directly uses the AIS data to obtain the essential parameters of vessels, including speed, course, position (longitude and latitude) and length are obtained by decoding and extracting AIS data.

In every vessel pair, a vessel was assigned as own vessel, the other one was target vessel. The vessel parameters including speed, course, longitude, latitude and length of own vessel and target vessel can be expressed as (v_0, c_0, x_0, y_0, l_0) and (v_t, c_t, x_t, y_t, l_t).

Then the relative distance (r), relative bearing (c_b), relative speed (v_r) and relative course (c_r) can be calculated as follows:

$$r = \sqrt{(x_t - x_0)^2 + (y_t - y_0)^2}, \tag{4}$$

$$c_b = \begin{cases} arctan\frac{x_t - x_0}{y_t - y_t} & if\ y_t > y_0 \\ arctan\frac{x_t - x_0}{y_t - y_0} + \pi & if\ y_t \le y_0 \end{cases}, \tag{5}$$

$$v_r = \sqrt{v_0^2 + v_t^2 - 2v_0 v_t \cos(c_t - c_0)}, \tag{6}$$

$$c_r = \begin{cases} c_0 + p + \arccos\frac{v_0^2 + v_r^2 - v_t^2}{2v_0 v_r}, & if\ \sin(c_0 + p - c_t) < 0 \\ c_0 + p - \arccos\frac{v_0^2 + v_r^2 - v_t^2}{2v_0 v_r}, & if\ \sin(c_0 + p - c_t) \ge 0 \end{cases}, \tag{7}$$

Here in, *DCPA*, *TCPA* and *SDOI* can be calculated as follows:

$$DCPA = r \times \left|\sin(c_r - c_b - \pi)\right|, \tag{8}$$

$$TCPA = r \times \cos(c_r - c_b - \pi)/v_r, \tag{9}$$

$$SDOI = \frac{\sqrt{\left(x_t' - x_0'\right)^2 + \left(y_t' - y_0'\right)^2}}{10(l_0 + l_t)}, \tag{10}$$

The calculation of collision risk of vessel pair by DCPA, TCPA [3] and SDOI [13] can be expressed in the form of negative exponential equations:

$$CR_{pair} = a_{DCPA}\alpha_{DCPA}\exp(-\beta_{DCPA}) + a_{TCPA}\alpha_{TCPA}\exp(-\beta_{TCPA}) + a_{SDOI}\alpha_{SDOI}\exp(-\beta_{SDOI}) \quad (11)$$

where the sum of a_{DCPA}, a_{TCPA} and a_{SDOI} is 1 and can be set according to the actual situation of selected water area for better accuracy.

After the collision risk calculation of vessel pairs in the cluster is quantified, in order to obtain the collision risk of each cluster more accurately, this paper uses improved Shapely value to determine the summing weight of each vessel pair in every single cluster.

$$s_i = \sum_{\substack{G \subseteq N \\ i \in G}} \frac{(g-1)!(n-g)!}{n!}[A(G) - A(G - \{i\})], \quad (12)$$

$$p_{fi} = CR_{fi} / \sum_{i=1}^{n} CR_{fi}, \quad (13)$$

$$s'_i = \sum {}_f s_f p_{ji} s_i / \left(\sum_{i=1}^{n} p_{fi} \times s_i \right), \quad (14)$$

where S_i refers to the summing weight determined by Shapely value of vessel i, G is the group formed according to vessel i, g represents the number of vessels in group G, N represents the group of all vessels, n refers the vessel number of group N, $A(G)$ refers to the total vessel number of group G and $A(G - \{i\})$ refers to the total vessel number of group G without vessel i. S'_i means the summing weight determined by Improved Shapley value of vessel i, f refers to the influencing factor of collision risk, CR_{fi} refers to the collision risk of the factor f of the vessel i, p_{fi} refers to the weight of the factor for vessel i, and σ_f refers to the influence coefficient which can be determined by maritime experts

The improved Shapely value method applied here can also be used in the quantification of the summing weight of clusters in Equation (2), therefore, regional collision risk can be quantified more precise. After calculating regional collision risk at different time points, a series of regional collision risk at different time points in selected water area is obtained. This study uses a RNN to predict regional collision risk of selected water area in the following content as the second step of the framework.

3. The RNN Based Prediction Framework

3.1. Recurrent Neural Network

In recent decades, the excellent prediction ability of RNN has studied widely. In the application of prediction, Emad et al. [40] pointed RNN has a better outcome in stock trend prediction compared with Time-Delay Neural Networks and Probabilistic Neural Network. Tian and Pan [41] used Long Short-Term Memory Recurrent Neural Network capture the nonlinearity and randomness in short-term traffic flow prediction more effectively. Maher and Biswajeet [42] applied an RNN model to predict the injury severity of traffic accidents, the RNN model outperformed the Multilayer Perceptron and Bayesian Logistic Regression models in comparative analyses. Xu et al. [43] carried out risk predictions from Electronic Health Records with the proposed RNN approach, the experimental results shown that RNN prediction model has good prediction performance. The application of RNN model in various fields' prediction motivated the evaluation of proposed RNN model for real-time prediction in regional collision risk.

A simple RNN consists of an input layer, a hidden layer, and an output layer, where $\boldsymbol{X_t}$ is the input, $\boldsymbol{Y_t}$ is the output, $\boldsymbol{H_t}$ is the hidden layer, U_t is the weight of the input value $\boldsymbol{X_t}$, and V_t is the weight

of the hidden layer $\boldsymbol{H_t}$. In a RNN, the input information continuously loops through the structure. The simple structure of RNN is shown in Figure 2.

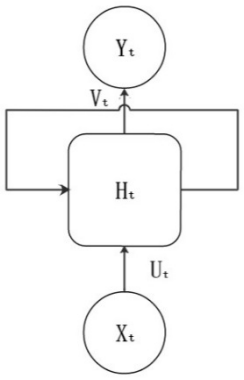

Figure 2. A simple structure of RNN.

The structure of the unfolded RNN is shown in Figure 3.

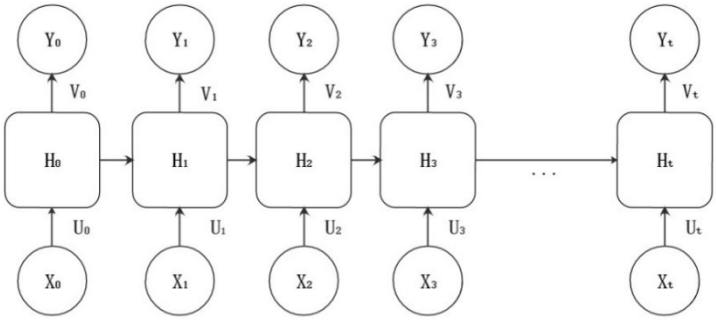

Figure 3. Unfolded structure of a simple RNN.

In Figure 3, $\boldsymbol{X_1}, \ldots, \boldsymbol{X_t}$ are inputs at different times, $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_t}$ are outputs at different times, $\boldsymbol{H_1}, \ldots, \boldsymbol{H_t}$ are hidden layers at different times in the network, and $U_1, \ldots, U_t$ are weights of input values at different times. $V_1, \ldots, V_t$ are the weights of the hidden layers at different times, $W_1, \ldots, W_{t-1}$ are the transfer weights of the hidden layers at different times. It can be seen that the inputs of the RNN include not only the current input data, but also the previous input information. The determination of the RNN at previous time steps will affect the subsequent determination of following time steps. These continuous massages are saved in the hidden state of the recurrent network. This hidden state spans multiple time steps and is passed forward layer by layer, which has always affected the network's processing of each new example. At the same time, the hidden state has been constantly corrected. Therefore, the real-time input data and the lasted input data become two input sources of the RNN. The combination of them will determine how the RNN processes new data. In traditional feed-forward networks, neurons directly forward information, and the transmitted information will not contact the nodes that have passed through it again, while the recurrent network uses historical information to update the weights in the network. The value of the hidden layer $\boldsymbol{H_t}$ of the RNN depends not only on the current input $\boldsymbol{X_t}$, but also on the value of the previous hidden layer $\boldsymbol{H_{t-1}}$. The weight matrix W_{t-1} is the weight collected by the previous value of the hidden layer and used in the hidden layer of this time. To realize the prediction of the regional collision risk in selected water area, RNN was used in this study.

$$\boldsymbol{Y_t} = gV_t\boldsymbol{H_t}, \tag{15}$$

$$\boldsymbol{H_t} = f(U_t\boldsymbol{X_t} + W_{t-1}\boldsymbol{H_{t-1}}), \tag{16}$$

3.2. RNN Model for Regional Collision Risk Prediction

Because the RNN has the advantage of updating the system weights based on historical input data, this study applies the RNN to the regional collision risk prediction at continuous time points. In this study, an RNN with one hidden layer structure is used as a prediction model with two sets of matrices as inputs and one matrix as output. A linear function is used as activation function of hidden layer in proposed network. The inputs include the number of vessels entered the selected water area through the traffic lane in different time periods and the regional collision risk value of the selected water area at different times. Among them, the number of vessels passing through the traffic separation lane in two consecutive time periods is used as a set of 2×2 input matrices, and the regional collision risk value in selected water area at two consecutive time points is used as another set of 2×2 size input matrix. A regional collision risk value at different time points in selected water area is used as the output data in this model, and the RNN is trained with a certain data sample. The learning results of the RNN on the samples are used as the basis for predicting the regional collision risk in selected water area.

The structure of the RNN used in the prediction model is shown in Figure 4.

$$\boldsymbol{H_t} = W_{t-1}\boldsymbol{H_{t-1}} + U_{t1}\boldsymbol{X_{t,1}} + U_{t2}\boldsymbol{X_{t,2}} + b_H, \tag{17}$$

$$\boldsymbol{H_{t-1}} = W_{t-2}\boldsymbol{H_{t-2}} + U_{(t-1)1}\boldsymbol{X_{t-1,1}} + U_{(t-1)2}\boldsymbol{X_{t-1,2}} + b_H, \tag{18}$$

$$\hat{\boldsymbol{Y}}_{\boldsymbol{t-1}} = W_{2t-1}\boldsymbol{H_{2t}} + bY, \tag{19}$$

$$F = \left(\hat{\boldsymbol{Y}} - \boldsymbol{Y}\right)^2/2, \tag{20}$$

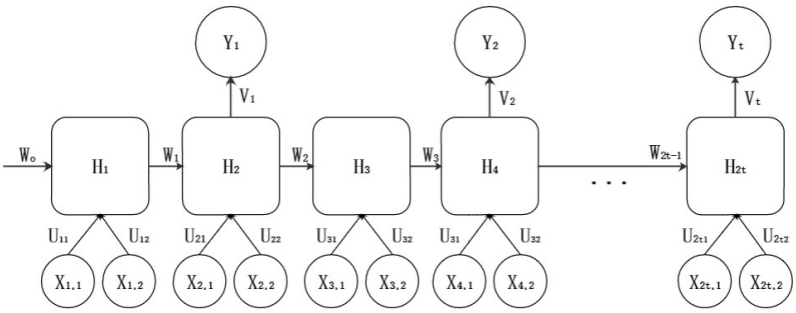

Figure 4. The structure of RNN prediction model.

Among them, $\boldsymbol{X_{1-2t,1}}$, $\boldsymbol{X_{1-2t,2}}$ are input matrices, $\boldsymbol{Y_{1-t}}$ are 1×1 output matrix, $\boldsymbol{H_{1-2t}}$ are hidden layers, b_H is bias vector of hidden layer, b_Y is bias vector of output, $U_{1-2t,1}$, $U_{1-2t,2}$, are the weights of different input matrices, and V_{1-t} are the weights of the hidden layer $\boldsymbol{H_{2-2t}}$. Therefore, we can get the formula, and the regression equation about the output value. The difference between the regression equation of the output value and the output value is used to obtain a new equation shown as Equation (20). When the new equation approaches to 0, it indicates that the RNN has achieved good results in learning the samples.

Construct the historical data obtained from Section 2 as a training database for RNN to learn. By calculating the regional collision risk at different time points, the time distribution of regional collision risk in selected water area can be obtained as inputs with the number of entered vessels at different time periods, therefore the RNN is used to predict regional collision risk of selected water area at following subsequent time point. The flow chart of this second-step framework is shown in Figure 5.

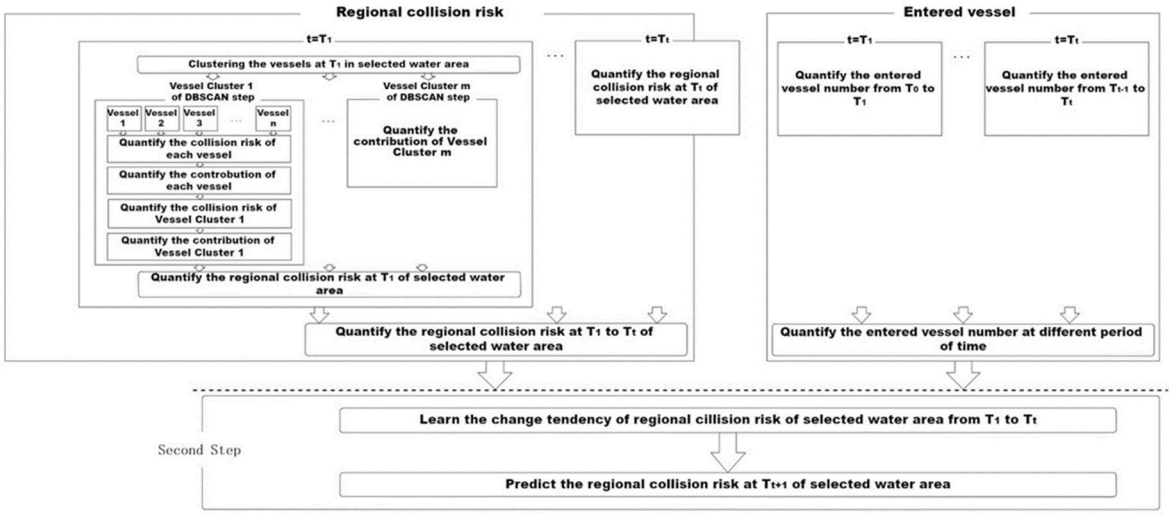

Figure 5. The flow chart of RNN prediction framework.

4. Case Study

4.1. Data Selection

To verify the validity of the prediction framework, a case study was performed at the western entrance of the Malacca Strait, which is the busiest water area of Singapore. The water area in this study is selected between 103.4° E to 103.6° E in longitude and 1.1° N to 1.2° N in latitude with the data from maritime safety research [44,45], as shown in Figures 6 and 7.

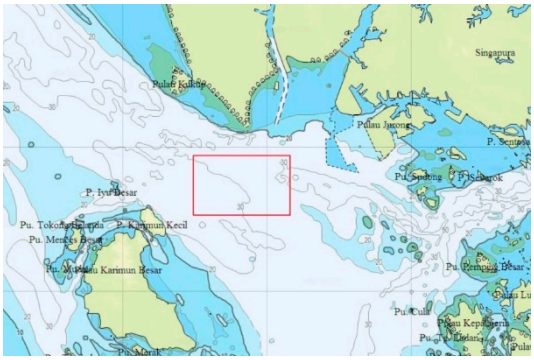

Figure 6. Selected water area-1 [44,45].

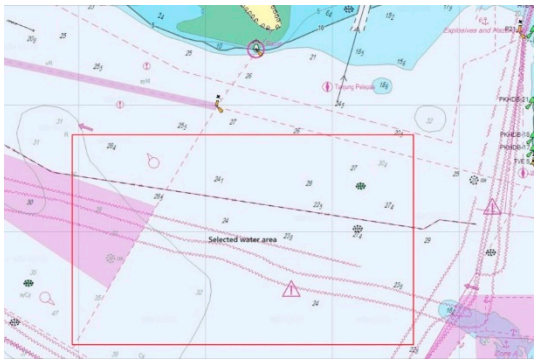

Figure 7. Selected water area-2 [44,45].

This area is located at the end of a traffic separation scheme. In this area, the vessel enters the port and anchorage after passing through the traffic lane, or enters the traffic lane from the port and anchorage. Without traffic separation scheme's control, the increase in traffic density and more vessel intersections will lead to greater collision risks. The selected area does not contain anchorages, narrow channels, and shallow water areas.

The Automatic Identification System (AIS) refers to a navigation assistance system applied to maritime safety and communication, which is applied between vessels and vessels, vessels, and shores. The AIS system can automatically exchange important information such as position, speed, heading, vessel name, call sign, Maritime Mobile Service Identify (MMSI), etc. According to the amendments to the International Convention for the Safety of Life at Sea adopted by the International Maritime Organization: All international navigation vessels over 300 t, non-international vessels of less than 500 t, and all passenger vessels must be compulsorily installed with AIS equipment, so that maritime traffic surveillance departments can obtain vessel data. As an important means to obtain vessel motion information data, the position information of the AIS system is derived from the Global Satellite Positioning System (civil GPS). Its positioning accuracy can already guarantee in 10 m, which meets the positioning requirements for vessels in maritime transportation surveillance. Based on GPS data that meets the accuracy requirements, the AIS system combines vessel dynamic information such as vessel position, vessel speed, changing heading rate, and heading. As well as vessel static information such as vessel name, call sign, draught, and dangerous goods. Such information is broadcasting from VHF channels to nearby vessel and shore station. The dynamic and static information enables neighboring vessel and shore stations to grasp the information of all vessels in the vicinity timely, which affords a great help to ensure the safety of maritime transportation.

In this study, MMSI, longitude, latitude, speed, heading, and vessel length information were selected from the 27 kinds of dynamic and static information contained in AIS data. After processing the AIS data, they were used to calculate the real-time regional collision risk of selected water area. The selected AIS data is the AIS vessel data received in the selected water area from 1800 to 1900 on 3 January 2014.

4.2. Data Optimizing

4.2.1. AIS Data Screening

The obtained AIS data is decoded and stored in the database, then carry out the work of pre-processing and cleaning to the data, so as to obtain valid AIS data. The main work of the pre-processing is to filter the AIS data on longitude and latitude according to the selected water area position information. Delete the data with MMSI of 0, and delete the AIS data where the position, speed or course exceeds a reasonable value.

4.2.2. AIS Data Processing

AIS information is sent discontinuously by different vessels at different time intervals. Because this study needs to calculate the regional collision risk at a specific time in selected water area, interpolation algorithm processing is performed on the filtered AIS data to obtain the different vessels', characteristic information at a specific time.

By collecting AIS data 3 minutes before and after 1800, 1810, ..., 1900 time points, the distribution of vessels in selected water area at the time point was optimized and applied in the following study.

4.3. Prediction Model Application

After the processed AIS information database was established according to the above process, Spatial clustering work was carried out to selected water area at different time point according to database, the results are shown in Figures 8–13.

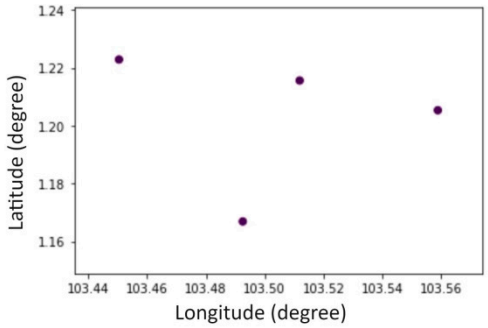

Figure 8. Result of DBSCAN at 1800.

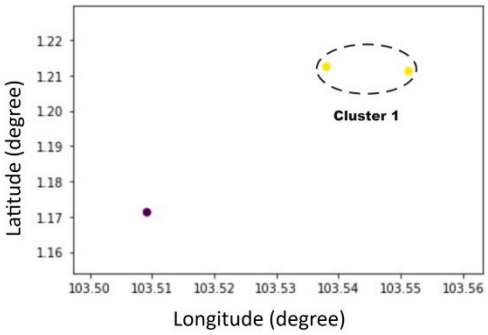

Figure 9. Result of DBSCAN at 1810.

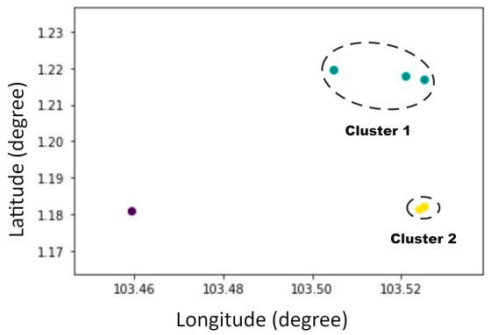

Figure 10. Result of DBSCAN at 1820.

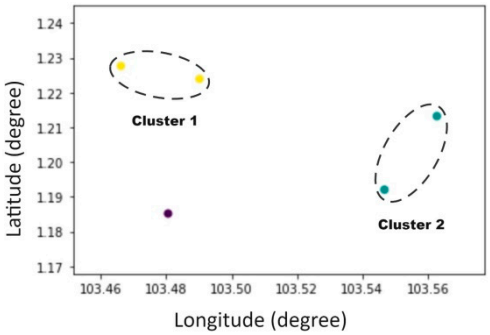

Figure 11. Result of DBSCAN at 1830.

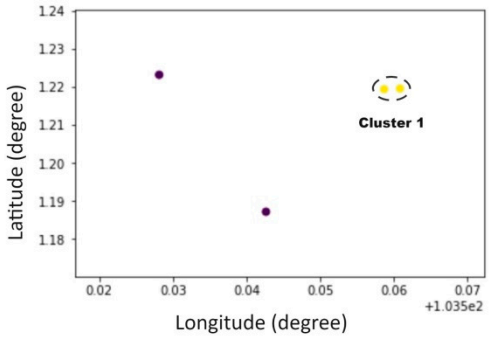

Figure 12. Result of DBSCAN at 1850.

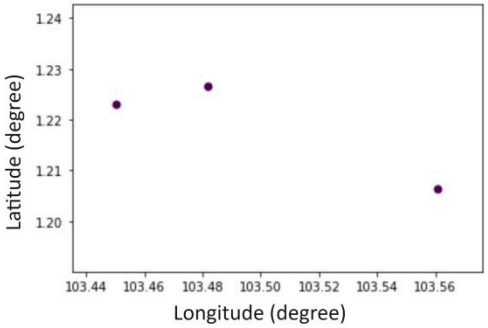

Figure 13. Result of DBSCAN at 1900.

The regional collision risk at the time of 1800, 1810, . . . , 1900 in the selected water area was calculated by using the improved Shapely value after clustering analysis at each time point, the regional collision values were given in Table 1. In addition, the number of entered vessels through traffic lane to selected water area is shown in Table 2.

Table 1. Regional collision risk of selected water area.

Moment	**1800**	**1810**	**1820**	**1830**
Regional Collision Risk	0	0.93935	0.61104	0.49891
Moment	**1840**	**1850**	**1900**	-
Regional Collision Risk	0	0.83017	0	-

Table 2. The Number of Entered Vessels.

Period	**1800–1810**	**1810–1820**	**1820–1830**
Entered Vessel Number	4	2	1
Period	**1830–1840**	**1840–1850**	**1850–1900**
Entered Vessel Number	0	3	0

In the next steps, the regional collision risk at different time points and the number of vessels entering the selected water area from the traffic lane in different time periods (10 minutes before each time point) were used as one input parameters to construct an RNN training data set, see Table 3.

The above data was applied to the RNN used in this study, and the parameters', weights were gradually modified, set the number of learning times to 3000 rounds. The output set obtained by RNN approach is shown in Table 4 and the training process of RNN is shown in Figure 14.

It can be illustrated from Figure 14 that the values of the parameter *mae* and parameter loss have changed from large to small, and have gradually stabilized.

Table 3. Input Set of RNN.

Input Set	Set 1.1		Set 1.2		Set 2.1		Set 2.2		Set 3.1		Set 3.2		Set 4.1		Set 4.2	
	0	0.939	4	2	0.939	0.611	2	1	0.611	0.499	1	0	0.499	0	0	3

Table 4. Output Set of RNN.

Output Set	Set 1	Set 2	Set 3	Set 4
	0.611	0.49891	0	0.83017

The meaning of parameter loss in RNN is shown in Equation (20), the meaning of *mae* here can be expressed as follow

$$mae = \frac{1}{n}\sum_{i=2}^{t}|Y_i - Y_{i-1}|, \tag{21}$$

The lower the value of *mae*, the better the goodness of fit. The lower the value of loss, the better the predictive ability of a model. Finally, the regional collision risk value at 1840 and 1850 of selected water area, the number of entered vessel in 1840–1850 and 1850–1900 were applied in the previous trained prediction model, therefore, a value of 1900 was calculated. The value of regional risk value at 1900 is 0.11747.

5. Validation of Prediction Framework

To verify the validity of the proposed framework, the actual regional collision risk of selected waters calculated based on AIS data was compared with the results obtained from the RNN prediction framework.

This study uses the RNN regional collision risk prediction model to predict the regional collision risk at different time points in selected water area from 1800 to 1900 on 3 January 2014. The results obtained from prediction framework and actual value of regional collision risk obtained from historical AIS data are shown in Table 5 and Figure 15.

In the following study, the data set was constructed from AIS data of selected water area at different time points from 1800 to 1900 on 4 and 5 January 2014, and the number of vessels entered the selected waters through traffic lane in different time periods. The RNN prediction framework proposed in this study was used to predict the regional collision risk respectively.

The data set constructed based on the AIS historical data on 4 and 5 January 2014 is shown in Table 6. The RNN prediction framework proposed in this study was trained based on the data set, and the factor weights were gradually modified, set the number of learning times as 3000 rounds and 7000 rounds. The prediction result is shown in Table 7, the training process diagram of RNN is shown in Figures 16 and 17, the parameter *mae* and parameter loss in each figure have changed from large to small, and have gradually stabilized.

Table 5. 03-Jan-14 Regional Collision Risk.

	03-Jan-14				
Regional Collision Risk	1820	1830	1840	1850	1900
Actual Value	0.611	0.498	0	0.830	0
Prediction Value	0.634	0.339	0.081	0.886	0.117

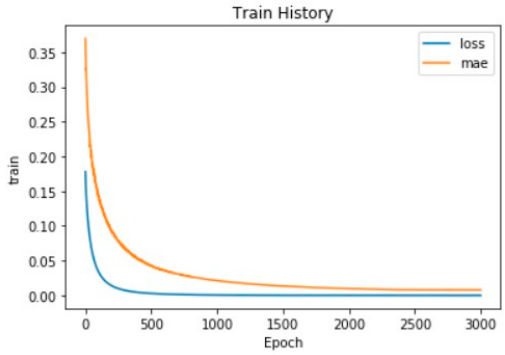

Figure 14. The training process of RNN.

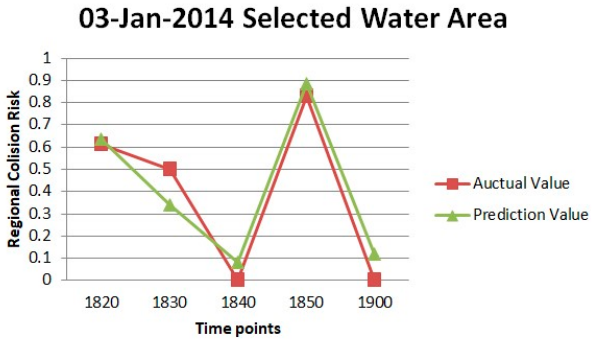

Figure 15. 03-Jan-2014 Regional collision risk.

Table 6. 04-Jan-14 Input Set.

Input Set	Date	Set 1.1		Set1.2		Set 2.1		Set 2.2		Set 3.1		Set 3.2		Set 4.1		Set 4.2	
	04-Jan-14	0	0.135	2	1	0.135	0	1	3	0	0	3	0	0	0.544	0	3
	05-Jan-14	0	0.614	1	1	0.614	0	1	2	0	1	2	0	1	0	0	1

Table 7. 04-Jan-14 Output Set.

Output Set	Date	Set1	Set2	Set3	Set4
	04-Jan-14	0	0	0.54425	0
	05-Jan-14	0	1	0	0

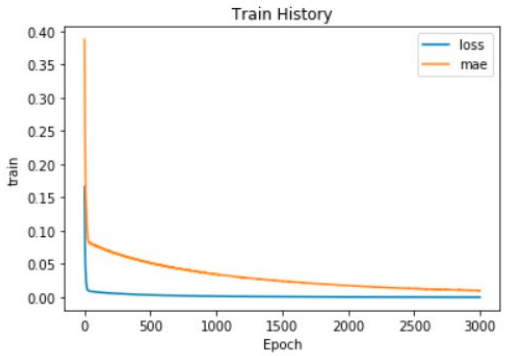

Figure 16. 04-Jan-2014 The training process of RNN.

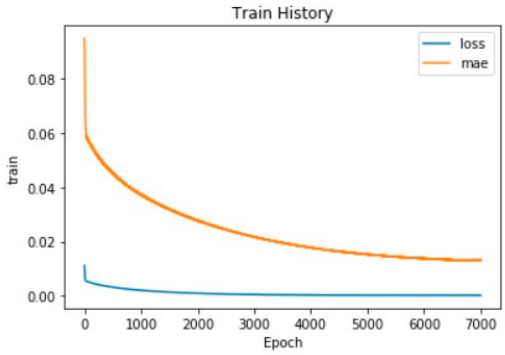

Figure 17. 05-Jan-2014 The training process of RNN.

In the next step, the prediction models obtained by this approach was used to predict the regional collision risk at 1900 on 4 and 5 January 2014. The predicted and true values for regional collision risk values are shown in Tables 8 and 9, and Figures 18 and 19.

Table 8. 04-Jan-14 Regional Collision Risk.

Regional Collision Risk	1820	1830	1840	1850	1900
Actual Value	0	0	0.544	0	0
Prediction Value	0.014	0	0.648	0.079	0

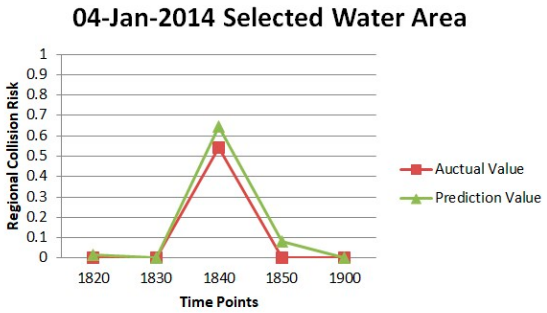

Figure 18. 04-Jan-2014 Regional collision risk.

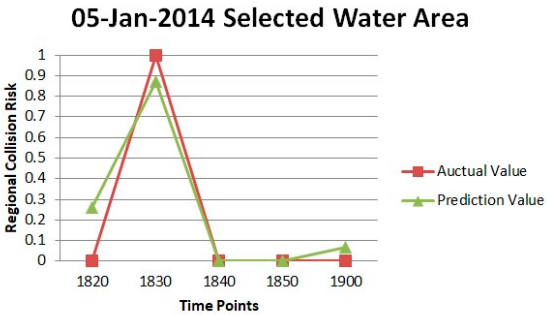

Figure 19. 05-Jan-2014 Regional collision risk.

Table 9. 05-Jan-14 Regional Collision Risk.

Regional Collision Risk	1820	1830	1840	1850	1900
Actual Value	0	1	0	0	0
Prediction Value	0.262	0.872	0	0	0.069

Based on above observation, at 1900 on 3, 4 and 5 January 2014, the prediction results of regional collision risk obtained through the RNN prediction framework is close to the actual value obtained based on AIS historical data. In addition, the prediction results of the prediction framework for the other time points are also closer to the actual values. Comparing the predicted values of regional collision risk with actual historical data, it can be seen from Figures 15, 18 and 19 that the predicted values obtained from RNN prediction framework are close to the actual values of regional collision risk. The change tendency of predicted regional collision risk value and the actual regional collision risk value have reach a good agreement. The results of previous application and validation all shown that the RNN prediction framework proposed in this study can effectively predict the regional collision risk in specific water area.

6. Conclusions

A regional collision risk prediction framework based on real-time AIS data is proposed in this paper. To improve the efficiency and reduce the computational complexity of the quantification part of the regional collision risk in the model, the DBSCAN spatial clustering algorithm was used to obtain the cluster distribution of vessels in selected water area, and several clusters each including several vessels in selected water area are obtained. Then, an improved Shapely value method is used to define every single vessel's contribution in each cluster and risk contribution of each cluster in selected water area to obtain the regional collision risk at a specific time point of selected water area. As following, the regional collision risk at a series of time points in selected water area obtained through the above steps, and the number of vessels entered the selected water area through traffic lane during the corresponding time period are used to build data set for training the RNN prediction model. In the second step of the RNN prediction framework, the RNN is trained by previous data sets, and finally the regional collision risk at future time point of selected water area is obtained. To validate the proposed prediction framework, a case study was carried out in a water area of Malacca, Singapore, and the results shown that the RNN prediction framework proposed in this study can effectively predict the regional collision risk in specific water area.

In contrast to other models, this prediction framework starts from the calculation of the collision risk of a single vessel, uses real-time AIS data in selected water area and other information to predict the regional collision risk in future time point. Moreover, the proposed framework can achieve more accurate real-time prediction of regional collision risk, without the limitation of construction of large database or large index set. The application of the RNN prediction framework in regional collision risk can help surveillance operators achieve better risk monitoring about the future trend of regional collision risk. In future, the framework can be applied to other water areas, such as port waters and waterways which have a need for regional collision risk monitoring.

During AIS data processing, interpolation is used to process different times' AIS data to obtain AIS information at the same time. To obtain more accurate vessel position information, consider more dynamic vessel parameters such as Speed Over Ground, Course Over Ground, Change of Speed, and Rate of Turn in future study. When calculating the collision risk of every single vessel, ship domain is considered to be circular in the SDOI parameters used in this model, while the more advanced geometric concepts of ship domain such as ellipse should be used to improve the accuracy of the model. In addition, the prediction framework involves cluster analysis of vessel distribution at multiple moments in selected water area, each cluster analysis needs to determine appropriate parameters based on real-time conditions. Therefore, an optimized DBSCAN spatial clustering algorithm should be used in future research. The limited size of training data and simple structure of RNN structure applied here can affect the predictions. To achieve more accurate prediction results, considering more factors, with more advanced structures and constructing larger training data, optimized RNN algorithms will also be applied in the future work.

Author Contributions: Conceptualization, X.W., Y.C. and Z.-J.L.; methodology, D.L. and Z.L.; software, D.L.; validation, D.L.; formal analysis, D.L.; investigation, D.L.; resources, X.W.; data curation, X.W.; writing—original

draft preparation, D.L. and X.W.; writing—review and editing, D.L. and X.W.; visualization, D.L.; supervision, Y.C. and Z.-J.L.; project administration, X.W.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported in part by the National Natural Science Foundation of China (Grant number 51909022, 61976033), the Natural Science Foundation of Liaoning Provence (Grant number 2019-BS-024), the Key Scientific Research Project of Ministry of Transport of China (Grant number 2019-ZD7-042), and the Fundamental Research Funds for the Central Universities (Grant number 3132019347).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

AIS	Automatic Identification System
ARPA	Automatic Radar Plotting Aids
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCPA	Distance at Closet Point to Approach
ECDIS	Electronic Chart Display and Information System
FSA	Formal Safety Assessment
GPS	Global Satellite Positioning System
MMSI	Maritime Mobile Service Identify
OZT	Obstacle Zone by Target
RNN	Recurrent Neural Network
SDOI	Ship Domain Overlapping Index
SVM	Support Vector Machine
TCPA	Time to Closet Point of Approach
VCD	Vessel's Compass Degree
VTS	Vessel Traffic Service

References

1. UNCTAD Review of Maritime Transport 2018. Available online: https://unctad.org/en/pages/PublicationWebflyer.aspx?publicationid=2245 (accessed on 27 November 2019).
2. Zhang, S.; Villavicencioa, R.; Zhu, L.; Pedersen, P.T. Ship collision damage assessment and validation with experiments and numerical simulations. *Mar. Struct.* **2019**, *63*, 239–256. [CrossRef]
3. Zhen, R.; Riveiro, M.; Jin, Y. A novel analytic framework of real-time multi-vessel collision risk assessment for maritime traffic surveillance. *Ocean Eng.* **2017**, *145*, 492–501. [CrossRef]
4. Liu, Z.H.; Wu, Z.L.; Zheng, Z.Y. A cooperative game approach for assessing the collision risk in multi-vessel encountering. *Ocean Eng.* **2019**, *187*, 106–175. [CrossRef]
5. Bukhari, A.C.; Tusseyeva, I.; Kim, Y.G. An intelligent real-time multi-vessel collision risk assessment system from VTS view point based on fuzzy inference system. *Expert Syst. Appl.* **2013**, *40*, 1220–1230. [CrossRef]
6. Zhang, D.; Yan, X.P.; Yang, Z.L.; Wall, A.; Wang, J. Incorporation of formal safety assessment and bayesian network in navigational risk estimation of the Yangtze river. Reliability. *Eng. Syst. Saf.* **2017**, *118*, 93–105. [CrossRef]
7. Mou, J.; Chen, P.; He, Y.; Yip, T.L.; Li, W.; Tang, J.; Zhang, H. Vessel traffic safety in busy waterways: A case study of accidents in western shenzhen port. *Accid. Anal. Prev.* **2016**, *123*, 461–468. [CrossRef]
8. Hu, S.; Fang, Q.; Xia, H.; Xi, Y. Formalsafety assessment based on relative risks model in ship navigation. *Reliab. Eng. Syst. Saf.* **2007**, *92*, 369–377. [CrossRef]
9. Debnath, A.K.; Chin, H.C. Navigational traffic conflict technique: A proactive approach to quantitative measurement of collision risks in port waters. *J. Navig.* **2010**, *63*, 137–152. [CrossRef]
10. Montewka, J.; Hinz, T.; Kujala, P.; Matusiak, J. Probability modelling of vessel collisions. *Reliab. Eng. Syst. Saf.* **2010**, *95*, 573–589. [CrossRef]
11. Le, C.H.; Ding, H.Y.; Dong, G.H.; Zhang, P.Y. Risk Assessment of Offshore Platform due to Ship Collision. In Proceedings of the International Conference on Electric Technology and Civil Engineering, Lushan, China, 22–24 April 2011.
12. Qu, X.; Meng, Q.; Li, S.Y. Ship collision risk assessment for the Singapore strait. *Accid. Anal. Prev.* **2011**, *43*, 2030–2036. [CrossRef]

13. Liu, Z.H.; Wu, Z.L.; Zheng, Z.Y. A novel framework for regional collision risk identification based on AIS data. *Appl. Ocean Res.* **2019**, *89*, 261–272. [CrossRef]
14. Nivoliantou, Z.S.; Koromila, I.A.; Giannakopoulos, T. Bayesian Network to Predict Environmental Risk of a Possible Ship Accident. *Int. J. Risk Assess. Manag.* **2016**, *19*, 1–5. [CrossRef]
15. Fan, S.; Sang, L.; Mao, Z. The prediction of the collision incident level in the lower reaches of the Yangtze River based on the mutual information. In Proceedings of the 2017 4th International Conference on Transportation Information and Safety (ICTIS), Banff, AB, Canada, 8–10 Auguest 2017.
16. Kim, K.; Lee, K.M. Deep Learning-Based Caution Area Traffic Prediction with Automatic Identification System Sensor Data. *Sensors* **2018**, *18*, 3172. [CrossRef] [PubMed]
17. Okazaki, T.; Terayama, M.; Nishizaki, C. Feasibility Study for Predicting Collision Possibility Sea Area for Each Ship by Using Support Vector Machine. In Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics, Miyazaki, Japan, 17 January 2019.
18. Fukuto, J.; Imazu, H. Application of "Obstacle Zone by Target (OZT)" Algorithm for Collision Alarm. *J. Jpn. Inst. Navig.* **2013**, *128*, 49–54. [CrossRef]
19. Costa, M.; Pasero, E.; Piglione, F.; Radasanu, D. Short term load forecasting using a synchronously operated recurrent neural network. In Proceedings of the International Joint Conference on Neural Networks, Washington, DC, USA, 10–16 July 1999.
20. Coulibaly, P.; Anctil, F. Real-time short-term water inflows forecasting using recurrent neural networks. In Proceedings of the International Joint Conference on Neural Networks, Washington, DC, USA, 10–16 July 1999.
21. Li, S.; Wunsch, D.C.; O'Hair, E.; Giesselmann, M.G. Wind turbine power estimation by neural networks with Kalman filter training on a SIMD parallel machine. In Proceedings of the International Joint Conference on Neural Networks, Washington, DC, USA, 10–16 July 1999.
22. Liang, S.F.; Su, A.W.Y.; Lin, C.T. A new recurrent-network-based music synthesis method for Chinese plucked-string instruments-pipa and qin. In Proceedings of the International Joint Conference on Neural Networks, Washington, DC, USA, 10–16 July 1999.
23. Giles, C.L.; Lawrence, S.; Tsoi, A.C. Rule inference for financial prediction using recurrent neural networks. In Proceedings of the IEEE Conference on Computational Intelligence for Financial Engineering, New York City, NY, USA, 24–25 March 1997.
24. Karpathy, A.; Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 3128–3137. [CrossRef] [PubMed]
25. Mou, L.; Ghamisi, P.Z. Deep Recurrent Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
26. Kolbæk, M.; Yu, D.; Tan, Z.H.; Jensen, J. Multitalker Speech Separation with Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1901–1913. [CrossRef]
27. Ma, X.; Dai, Z.; He, Z.; Ma, J.; Wang, Y. Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction. *Sensors* **2017**, *17*, 818. [CrossRef]
28. Guo, L.; Li, N.; Jia, F.; Lei, Y.; Lin, J. A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing* **2017**, *240*, 98–109. [CrossRef]
29. Shi, Z.; Xu, M.; Pan, Q.; Yan, B.; Zhang, H. LSTM-based Flight Trajectory Prediction. In Proceedings of the 2018 International Joint Conference on Neural Networks, Rio de Janeiro, Brazil, 8–13 July 2018.
30. Gao, M.; Shi, G.; Li, S. Online Prediction of Ship Behavior with Automatic Identification System Sensor Data Using Bidirectional Long Short-Term Memory Recurrent Neural Network. *Sensors* **2018**, *18*, 4211. [CrossRef]
31. Kong, W.; Dong, Z.Y.; Jia, Y.; Hill, D.J.; Xu, Y.; Zhang, Y. Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network. *IEEE Trans. Smart Grid* **2019**, *10*, 841–851. [CrossRef]
32. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 1990.
33. Ester, M.; Kriegel, H.P.; Xu, X. Knowledge discovery in large SPATIAL database: Focusing techniques for efficient class identification. In *International Symposium on Spatial Databases*; Springer: Berlin/Heidelberg, Germany, 1995; pp. 67–82.

34. Ester, M.; Kriegel, H.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland, OR, USA, 2–4 August 1996; 1996; pp. 226–231.
35. Guha, S.; Rastogi, R.; Shim, K. CURE: An efficient clustering algorithm for large databases. *Inf. Syst.* **2001**, *26*, 35–38. [CrossRef]
36. Agrawal, R.; Gehrke, J.; Gunopolos, D.; Raghavan, P. Automatic subspace clustering of high dimensional data for data mining application. In Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, Seattle, WA, USA, 1–4 June 1998; pp. 94–105.
37. Nanopoulos, A.; Theodoridis, Y.; Manolopoulos, Y. C2P: Clustering based on closest pairs. In Proceedings of the 27th International Conference on Very Large Databases, Roma, Italy, 11–14 September 2001; pp. 331–340.
38. Filipovych, R.; Resnick, S.M.; Davatzikos, C. Semi-supervised cluster analysis of imaging data. *NeuroImage* **2011**, *54*, 2185–2197. [CrossRef] [PubMed]
39. Huth, R.; Beck, C.; Philipp, A.; Demuzere, M.; Ustrnul, Z.; Cahynová, M.; Kyselý, J.; Tveito, O.E. Classifications of atmospheric circulation patterns. *Ann. N. Y. Acad. Sci.* **2008**, *1146*, 105–152. [CrossRef]
40. Emad, W.S.; Danil, V.; Donald, C. Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks. *IEEE Trans. Neural Netw.* **1998**, *9*, 1456–1470.
41. Tian, Y.; Pan, L. Predicting Short-Term Traffic Flow by Long Short-Term Memory Recurrent Neural Network. In Proceedings of the 2015 IEEE International Conference on Smart City, Chengdu, China, 19–21 December 2015.
42. Maher, I.S.; Biswajeet, P. Severity Prediction of Traffic Accidents with Recurrent Neural Networks. *Appl. Sci.* **2017**, *7*, 476.
43. Xu, E.; Zhao, S.; Mei, J.; Xia, E.; Yu, Y.; Huang, S.F. Multiple MACE Risk Prediction using Multi-Task Recurrent Neural Network with Attention. In Proceedings of the 2019 IEEE International Conference on Healthcare Informatics, Xi'an, China, 10–13 June 2019.
44. Zaman, M.B.; Kobayashi, E.; Wakabayashi, N.; Khanfir, S.; Pitana, T.; Maimun, A. Fuzzy FMEA model for risk evaluation of ship collisions in the Malacca Strait: Based on AIS data. *J. Simul.* **2014**, *8*, 91–104. [CrossRef]
45. Zaman, M.B.; Santoso, A. Formal Safety Assessment (FSA) for Analysis of Ship Collision Using AIS Data. *Int. J. Mar. Navig. Saf. Sea Transp.* **2015**, *9*, 67–72. [CrossRef]