sustainability

MDPI

Article

Enhanced Deep Neural Networks for Traffic Speed Forecasting Regarding Sustainable Traffic Management Using Probe Data from Registered Transport Vehicles on Multilane Roads

Van Manh Do [1], Quang Hoc Tran [1,*], Khanh Giang Le [1], Xuan Can Vuong [2] and Van Truong Vu [3]

[1] Faculty of Civil Engineering, University of Transport and Communications, No. 3 Cau Giay Street, Lang Thuong Ward, Dong Da District, Hanoi 100000, Vietnam; manhdv@utc.edu.vn (V.M.D.); gianglk@utc.edu.vn (K.G.L.)
[2] Faculty of Transport Safety and Environment, University of Transport and Communications, No. 3 Cau Giay Street, Lang Thuong Ward, Dong Da District, Hanoi 100000, Vietnam; vuongcan@utc.edu.vn
[3] Institute of Techniques for Special Engineering, Le Quy Don Technical University, 236 Hoang Quoc Viet Rd., Co Nhue, Bac Tu Liem, Hanoi 100000, Vietnam; truongvv@lqdtu.edu.vn
* Correspondence: hoctq@utc.edu.vn

Abstract: Early forecasting of vehicle flow speeds is crucial for sustainable traffic development and establishing Traffic Speed Forecasting (TSF) systems for each country. While online mapping services offer significant benefits, dependence on them hampers the development of domestic alternative platforms, impeding sustainable traffic management and posing security risks. There is an urgent need for research to explore sustainable solutions, such as leveraging Global Positioning System (GPS) probe data, to support transportation management in urban areas effectively. Despite their vast potential, GPS probe data often present challenges, particularly in urban areas, including interference signals and missing data. This paper addresses these challenges by proposing a process for handling anomalous and missing GPS signals from probe vehicles on parallel multilane roads in Vietnam. Additionally, the paper investigates the effectiveness of techniques such as Particle Swarm Optimization Long Short-Term Memory (PSO-LSTM) and Genetic Algorithm Long Short-Term Memory (GA-LSTM) in enhancing LSTM networks for TSF using GPS data. Through empirical analysis, this paper demonstrates the efficacy of PSO-LSTM and GA-LSTM compared to existing methods and the state-of-the-art LSTM approach. Performance metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Median Absolute Error (MDAE) validate the proposed models, providing insights into their forecasting accuracy. The paper also offers a comprehensive process for handling GPS outlier data and applying GA and PSO algorithms to enhance LSTM network quality in TSF, enabling researchers to streamline calculations and improve supposed model efficiency in similar contexts.

Keywords: deep learning approach; PSO-LSTM; GA-LSTM; short-term traffic speed forecasting; urban traffic management; sustainability

Citation: Do, V.M.; Tran, Q.H.; Le, K.G.; Vuong, X.C.; Vu, V.T. Enhanced Deep Neural Networks for Traffic Speed Forecasting Regarding Sustainable Traffic Management Using Probe Data from Registered Transport Vehicles on Multilane Roads. *Sustainability* **2024**, *16*, 2453. https://doi.org/10.3390/su16062453

Academic Editor: Armando Cartenì

Received: 16 January 2024
Revised: 8 March 2024
Accepted: 11 March 2024
Published: 15 March 2024

1. Introduction

Early forecasting of vehicle flow speeds, from vehicle mobility data, plays a vital role in sustainable traffic development and supports countries in establishing their Traffic Speed Forecasting systems or applications [1]. Currently, countries worldwide mainly use online map services such as Google Maps, Apple Maps, Bing Maps, TomTom Maps, etc., to check travel routes and traffic status on roads [2,3]. This reliance unintentionally limits the development of alternative forecasting platforms, fails to promote sustainable development of domestic traffic management, and poses potential risks to information security. Therefore, researching more sustainable solutions to generate traffic self-warning systems in urban areas is urgent. Probe data, such as Global Positioning System (GPS) data,

can be used to forecast average vehicle speed in early warning systems for sustainable development in transportation management, which requires more in-depth research to analyze these data to support forecasting vehicle speeds and traffic status in urban areas.

Currently, traffic congestion frequently occurs, creating obstacles and difficulties for traffic participants in large urban areas [4,5]. Although there are many city government efforts to determine solutions to limit congestion, few synchronous methods meet the current reality. The optimal solution is to optimize the current traffic signal timings of traffic networks and provide a predicted road map for commuters [6–8]. Therefore, Traffic Speed Forecasting (TSF) plays a vital role in improving efficiency and reducing traffic congestion, and there is more and more quality scientific research on this issue. Taking advantage of GPS satellite data from vehicle tracking devices reveals new perspectives in reducing traffic congestion and managing traffic in urban areas [9,10].

There are many challenges when using GPS data in forecasting traffic speed regarding sustainable traffic management, such as data quality and accessibility. Data collected from GPS devices mounted on vehicles moving on the road are not always highly accurate due to the influence of environmental factors and signal noise [11]. Data collection capabilities may be limited in areas with underdeveloped telecommunications infrastructure and data management laws in each country. The second concern regards the transparency and information security of users [12]. Third, the cost of training, deploying, and maintaining the deep learning system requires large investment costs, from model development to data updates and staff training, putting financial pressure on the government [13]. The fourth is the ability to adapt to realistic conditions. The deep learning models are not always flexible and accurate to variations in traffic, especially in special situations, or not suitable for different weather conditions [14]. Fifth is consensus and cooperation; achieving consensus and cooperation from stakeholders is very important to ensure the effectiveness and sustainability of the system in traffic management [15]. Nevertheless, GPS data of vehicles participating in traffic hold value, enabling each country to establish its forecasting system and develop domestic information technology platforms, thereby ensuring sustainable development for the country.

The valuable data source from journey monitoring data collected by GPS signals mounted on vehicles only serves the purpose of managing or handling penalty situations due to commuter violations of traffic laws. Additionally, it can be exploited by a third party, such as Google Maps and other mobile applications, for different purposes [16]. This limits the development of new domestic application platforms to serve traffic optimization in each country and further concerns information security issues [17]. Therefore, processing and using this type of data source to assess traffic conditions for traffic networks is necessary. However, the data source received from the trip monitoring device mounted on vehicles participating in traffic requires a pre-processing step to remove abnormal or missing data. This is necessary before it can be used as input parameters for the Traffic Speed Forecasting (TSF) models and assessing traffic conditions, especially for roads with parallel multiple lanes in urban areas [18].

Recently, many promising methods have been used to forecast traffic conditions. Specifically, several traffic flow forecasting methods were applied, including linear regression models, artificial neural networks, and deep learning solutions [19]. However, these methods still have some limitations. These include inaccurate forecasting ability, failure to exploit complex relationships in big data, and difficulty in optimizing model parameters.

Furthermore, the LSTM (Long Short-Term Memory) method in TSF has advantages, such as the ability to process data series and solve the phenomenon of long-term dependent chain prediction. However, this method also has some disadvantages, such as difficulty in handling nonlinear data and the ability to optimize model parameters to achieve high performance. Hence, applying two deep learning neural networks, Long Short-Term Memory Optimized by Particle Swarm Optimization (PSO-LSTM) and Long Short-Term Memory Optimized by Genetic Algorithms (GA-LSTM), solved the current limitations of the LSTM method [20]. Using PSO and GA optimized hyperparameters such as window

size, epochs, neurons, and learning rate of the LSTM network more efficiently. This paper provides a solution for processing GPS data of journey monitoring devices mounted on commercial vehicles in Vietnam. By the combination of deep learning neural networks and optimization algorithms, the paper proposes a potential solution to significantly improve the ability to forecast future traffic speed, making a potential contribution to the application of artificial intelligence to solve traffic problems.

This scientific paper is the result of expanding, clarifying, and overcoming limitations of previously published research according to comments from readers and scientific researchers around the world. Specifically, this paper provides an entire proposed model to establish the traffic status for multilane roads. Additionally, it clearly outlines the process of processing GPS data to obtain revised data sources for TSF, which was not clearly stated in our previous study. Furthermore, this research proves that solutions to improve LSTM network efficiency using GA and PSO are necessary to enhance the quality of TSF models for transport management of urban areas.

The state-of-the-art solution (called LSTM*) is a contribution to further improving previously published research [7]. In a previous publication, the authors optimized three parameters of the LSTM network, including window size, the number of epochs, and the number of neurons. These parameters were not determined simultaneously but were instead determined in order of priority. First was window size (called bestWs), then the number of epochs (called bestEp), and finally the number of neurons (called bestNe). Although this approach found a set of three initialization parameters that were better than random initialization for the LSTM network, it did not guarantee that these parameters represented the optimal set. For example, when determining an optimal window size value number 1 (Ws1), this value was not the most optimal, but Ws1 combined with the set of the other two remaining parameters included the number of epochs 1 and the number of neurons 1 (Ep1 and Ne1) to form the set (Ws1, Ep1, and Ne1), which provided better results than the selected parameter set. In this paper, the authors proposed two enhanced LSTM algorithms, GA-LSTM and PSO-LSTM. Compared to the method used in other research [7] and the existing methods (parametric method, non-parametric method, and deep learning solution), these two algorithms demonstrate a major improvement in simultaneously optimizing the three parameters of the LSTM network using PSO and GA algorithms. The details of these two algorithms will be presented in the subsequent sections.

According to the published results, the provided model, along with the optimal parameters, including optimal GA operators and PSO factors, also contribute. Then, scientists and researchers could apply the same GPS data sources directly to shorten calculation time. The paper is divided into the following three main parts. Section 2 outlines the methodology to process GPS data of commercial vehicles and presents solutions to improve machine learning efficiency using PSO and GA in TSF. Section 3 is the application of the methodology in Section 2 to the experimental arterial road in Vietnam. Section 4 comprises conclusions, lessons learned, limitations, and future work based on this research.

2. Materials and Methods

2.1. Model Development

Determining the average speed of vehicles on each road section is detailed in the model proposed below (Figure 1) to forecast and provide early traffic conditions to traffic participants. This includes information on specific routes and time frames for each day of the week, enabling participants to choose the most optimal time and journey.

Figure 1 illustrates our weekday traffic conditions forecasting framework. Initially, GPS tracking data and traffic geometry are gathered. These inputs then undergo data preprocessing, including road segmentation, map matching, velocity determination, and anomaly handling. Following this, data processing involves partitioning and normalization. Subsequently, TSF models, such as enhanced LSTM and other suggested models, are created. Finally, integrating traffic geometry and TSF models enables the generation of accurate traffic forecasting about future traffic conditions.

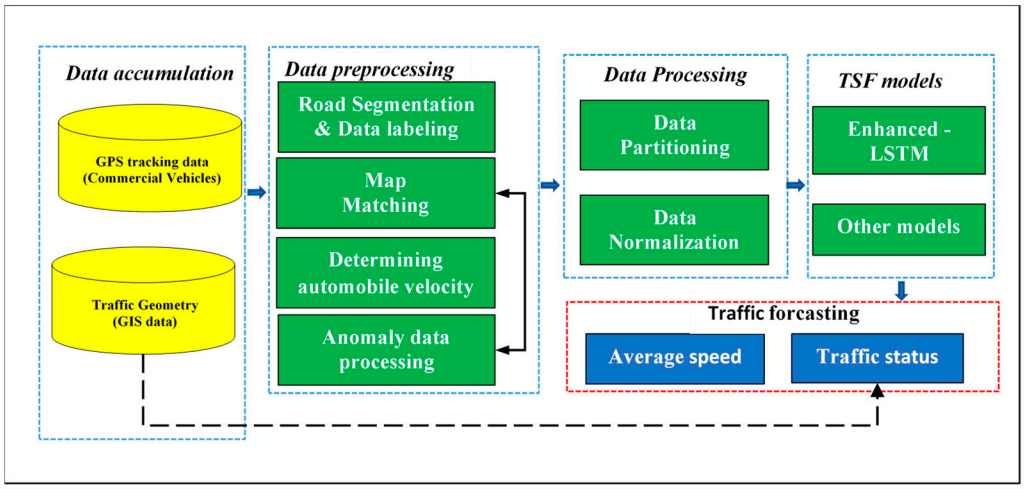


Figure 1. Weekday traffic conditions forecasting framework model: GPS tracking data and traffic geometry serve as primary inputs (shown in yellow), undergoing data preprocessing, data processing, and TSF model creation stages (shown in green), culminating in accurate traffic forecasting information (shown in blue).

However, this paper mainly emphasized the processing of abnormal or missing data from probe vehicle data and evaluated the quality of enhanced LSTM algorithms (PSO-LSTM and GA-LSMT) in generating a model to forecast the average speed of vehicles in each segment for arterial roads in Vietnam.

2.2. Data Processing

As mentioned earlier, this paper focused on processing GPS data to have accurate GPS data. Then, those data were used as input data for forecasting models on each road segment. Therefore, this section details how to process GPS data from trip monitoring devices mounted on vehicles registered for transport business in Vietnam, including handling anomalous signals and missing data. Processed data were standardized before inputting into the forecasting model. The remaining aspects, including the process of generating an online traffic forecast map from GPS data, were presented in detail in different publications.

Two current issues need to be addressed with GPS signals obtained from vehicle-mounted cruise monitoring devices. Firstly, it is necessary to determine the correct position of the vehicle on each specific road segment of multilane roads accurately. Secondly, GPS signals have anomalous values or missing values that appear due to reasons from the GPS signal transceiver. These are some challenges that this research must address.

The accuracy of the vehicle's location through GPS signals directly affects the results of calculating the average speed of each segment of the multilane road. Therefore, in addition to using accurate traffic geometry data from the 1/2000 scale map during the map matching stage, it is also vital to process which lane the GPS signal is located on. We described this incident in Figure 2 shown above. A registered vehicle moving on road segment 2 (Road Seg.2) transmits 15 GPS signals, which are red dots numbered from 1 to 15, in 15 min to the signal recording and processing center. However, 5 of the 15 received signals are not on Road Seg.2. There are 2 signals in Road Seg.1 (signals 5 and 6) and 3 on Road Seg.3 (signals 10,11). Road Seg.1 and Road Seg.3 could belong to the opposite direction in multilane road. This greatly affects the results of calculating the average speed of road segments and incorrectly determines the traffic status of each road segment and road network.

To solve the above issue, this research provided a process in pseudocode to describe a process for processing data from GPS signals to identify and remove outliers from vehicle trajectories. Algorithm 1 is overview of the GPS Data Filtering and Route Optimization Algorithm (GDF-ROA). The GDF-ROA algorithm refines GPS data by filtering points within the current time frame and assigning road segments to vehicles based on their codes.

It then removes outliers from each road segment by iteratively evaluating distances to road segments. Each point is assigned to the nearest segment, and segment densities are calculated. Outliers are identified and removed based on segment density. The algorithm ensures the accuracy of the vehicle's location through GPS signals while optimizing them for further analysis or navigation purposes. A detailed explanation of the process follows some steps below.

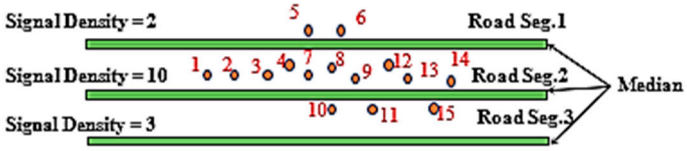


Figure 2. Illustration of signal confusion among lanes: there is a vehicle running on lane 2 (road segment 2), with 15 GPS signals received during the movement. Due to GPS signal transmission errors and small lanes, some confusion occurred; signals 5 and 6 were mistaken for lane 1 (road segment 1), and signals 10, 11, and 15 were mistaken for lane 3 (road segment 3).

Algorithm 1 GPS Data Filtering and Route Optimization Algorithm (GDF-ROA) for enhancing route accuracy and removing outliers from GPS data

```
Input:
    TmF: Current time frame
    SegmentList: List of segments (including 16 experimental segments)
    PointList: List of points (GPS signals)
Output:
    RouteList (containing routes of each vehicle after removing outliers)
//Step 1: Filter points within the current time frame
NewPointList ← FilterByTime(PointList, TmF);
//Step 2: Determine routes (set of points) for each vehicle based on vehicle code
RouteList ← FilterByVehicleCode(NewPointList);
//Step 3: Remove outliers for each route
For Each Route in RouteList
    ListSegmentDensity = new ArrayList<Integer>()
    RouteWithoutOutliers = new Route()
    For Each Point in Route
        MinDistance = +∞
        MinSegmentIndex = 0
        For Each Segment in SegmentList
            CurrDistance ← CalculateDistance(Point, Segment);
            If (CurrDistance < MinDistance) Then
                MinDistance = CurrDistance
                MinSegmentIndex = Segment.Id
        End If
        End For
        Point.SegmentID = MinSegmentIndex
       //Increase statistics for this road segment by 1.
        ListSegmentDensity[MinSegmentIndex] += 1
       //Check if the Point is not an outlier, add it to RouteWithoutOutliers
        If (Point.SegmentID = GetMaxDensityFrom(ListSegmentDensity)) Then
            RouteWithoutOutliers.Add(Point)
        End If
    End For
    Set Route to RouteWithoutOutliers
End For
//Final result: RouteList after removing outliers
Return RouteList;
```

Step 1 is to filter the points in the considered time frame. First, the GPS points in the current time frame (TmF) were filtered out of the PointList and saved to NewPointList. The original GPS big data source was reduced to filter out the data that needed to be processed within the corresponding time frame.

Step 2 is to determine the route of each vehicle through the vehicle code. Next, the GPS points in NewPointList were divided into separate routes based on their vehicle code. These routes were saved to RouteList.

Step 3 is to remove outliers for each road in RouteList. The algorithm determined the road segment for each point on the vehicle's motion trajectory. For each point in the road, this code compared the distance between that point and the road segments in the SegmentList. The lane closest to the GPS point was identified and assigned.

Specifically, this code uses the CalculateDistance (Point, Segment) function to calculate the distance between the point and the line segment (lane) and select the line segment with the smallest distance (MinDistance) for that point according to the formula below [21]:

$$\mathbf{D} = \frac{|(\mathbf{y_2} - \mathbf{y_1})\mathbf{x} - (\mathbf{x_2} - \mathbf{x_1})\mathbf{y} + \mathbf{x_2}\mathbf{y_1} - \mathbf{x_1}\mathbf{y_2}|}{\sqrt{(\mathbf{y_2} - \mathbf{y_1})^2 + (\mathbf{x_2} - \mathbf{x_1})^2}} \quad (1)$$

where (x, y) are the vehicle's GPS coordinates obtained from the vehicle's onboard trip monitoring device, (x_1, y_1) and (x_2, y_2) are the coordinates of the two corresponding road segment ending points.

At the same time, this code tracked the density (number of occurrences) of each line segment in ListSegmentDensity. The algorithm then found the road segment with maximum density and removed outliers. After determining the density of each road segment, this code found the road segment with the maximum density (maxIndex). Next, points on the route that did not belong to the road segment with maximum density were removed from the route. This eliminated outliers and retained only the points located on the most dense line segment. Finally, after completing step 3 for all routes in RouteList, the list of RouteList after outlier removal was returned as the final result of this data processing.

The second problem is described by partial data extraction in the table below. Specifically, after filtering the trip signal according to the specific time frame of each vehicle for each road segment as the algorithm model proposed above, the probe vehicle data still contained erroneous and missing GPS data. GPS signals were received by on-board unit (OBU) devices of commercial cars. Particularly, when the signal is lost or there is a signal error (NaN), the speed values return the value "0 km/h" in our system. Using the value 0 (km/h) to calculate the average speed of a road segment in a multilane road led to calculation errors on the experimental urban road because the Le Hong Phong experimental road has a maximum cycle length of 120 (s) for traffic light systems and stopping prohibition. Therefore, the speed returning "0 km/h" every 180 s can only be due to signal reception error or missing data. As shown in Table 1, the GPS data are still recorded (signals and density), but the speed is returned to 0 km/h.

Table 1. Anomalous or missing values by the GPS signal transceiver.

ID	Speed (Km/h)	Total Signals	Density	Status
2020-02-06T04:45:00.000Z	38.00	2	1	Normal
2020-02-06T05:48:00.000Z	43.50	4	1	Normal
2020-02-06T05:51:00.000Z	0.00	1	1	Missing

Table 1. *Cont.*

ID	Speed (Km/h)	Total Signals	Density	Status
2020-02-06T05:54:00.000Z	52.00	1	1	Normal
2020-02-06T05:57:00.000Z	0.00	2	1	Missing
2020-02-06T06:00:00.000Z	0.00	1	1	Missing
2020-02-06T06:03:00.000Z	0.00	1	1	Missing
2020-02-06T06:06:00.000Z	22.00	3	2	Normal
2020-02-06T06:09:00.000Z	0.00	1	1	Missing
2020-02-06T06:12:00.000Z	27.00	2	2	Normal
2020-02-06T06:15:00.000Z	44.08	8	4	Normal
2020-02-06T06:18:00.000Z	0.00	1	1	Missing
2020-02-06T06:21:00.000Z	35.00	4	2	Normal

As shown in the Table 1, column 1 describes the ID for vehicles with different identification codes, column 2 is the recorded vehicle's speed, column 3 is the number of GPS signals received, column 4 describes the density, and column 5 describes the current status of the received test signal. The red highlight colors are anomalous or missing values that need to be handled.

To handle the second problem, this paper proposed an algorithm called Interpolated Time Series Generator (ITST) as detailed in the pseudocode section below.

Algorithm 2 is the Interpolated Time Series Generator (ITSG). The ITSG algorithm aims to generate interpolated time series data from average velocities of segments within a specified time period. It begins by calculating the average velocity across all time frames for the segment under consideration. Then, the process identifies and interpolates outlier data points by checking for neighboring signals and computing averages. Finally, it generates time series data containing pairs of values {TimeFrame, Velocity} for further analysis or visualization. This method ensures the completeness and accuracy of time series data by addressing missing or outlier values through interpolation. The ITSG solved the problem of missing values in the road segment velocity data by the following steps.

Step 1: The model calculated the average velocity of each segment during the time frame by inputting the array V containing the average velocities and specifying the period under investigation.

Step 2: This process handled anomalous vehicle speed (zero velocity values) by interpolating them from neighboring values that are not abnormal. Specifically, for each time frame within the specified period, the process checks the velocity value (V_i) equals 0. If so, the process determines if all four neighboring signals are also outliers. If they are, replace the $V_i = 0$ with the calculated average velocity; otherwise, interpolate the velocity using the velocities of the neighboring non-outlier data points.

Algorithm 2 Interpolated Time Series Generator (ITSG): a process for generating interpolated time series data from average velocities of segments, enhancing accuracy and completeness

```
1. Input:
2.    V: An array containing the average velocities of the segment under consideration for all time frames.
3.    TimePeriod: The time period under investigation.
4. Output:
5.    TimeSeries (containing pairs of values {TimeFrame, Velocity})
6. //Step 1: Calculate the average velocity of the current segment across all time frames.
7. AverageVelocity = CalculateAverage(V)
8. //Step 2: Compute interpolated values for outlier data points.
9. For i = 2 To timeFrameLength Do
10.    //Check for cases where there is no signal in this time frame (i.e., V = 0).
11.    If (V_i = 0) Then
12.        //Check if all 4 neighboring signals are also outliers.
13.        If (V_i-2 = 0 & V_i-1 = 0 & V_i+1 = 0 & V_i+2 = 0) Then
14.            V_i = AverageVelocity.
15.        Else
16.            V_i = CalculateAverage (V_i-2, V_i-1, V_i+1, V_i+2)
17.        End If
18.    End If
19. //Step 3: Generate time series data.
20. For Each TimeFrame In TimePeriod
21.    TimeSeries_TimeFrame ← {TimeFrame, V_TimeFrame }
22. End For
23. Return TimeSeries;
```

Step 3: The provided method generates time series data by iterating over each time frame within the specified period and creating pairs of values {TimeFrame, Velocity}, returning the resulting time series containing these pairs of values.

The result of this algorithm is an improved time series, containing pairs of time and velocity values, ready to be used in speed analysis and Traffic Speed Forecasting applications on the corresponding road segment. These processing steps help us to remove anomalous GPS data and produce accurate and reliable data for decision support and analysis in the field of traffic monitoring and forecasting. Verification of the accuracy of the suggested model will be presented in the experimental data processing section below.

After the data preprocessing step, the revised data include three pieces of information: time frame (or time period), average travel velocity (km/h), and segment index, as shown in Algorithm 2.

To validate the accuracy of the ITSG model, we compared this model with the following popular data processing methods:

Method 1 (Me1) involved eliminating all abnormal or missing signals. This paper uses dropna function to eliminate all abnormal or missing data. Subsequently, the groupby function was employed to calculate the average speed for each road segment based on the filtered data. This approach mitigates the impact of incomplete data on empirical analysis [22,23].

Method 2 (Me2) utilized the average of all vehicles. The fillna function was used to fill in missing values using the average speed value of all registered cars. Then, the groupby function calculated the average speed for each road segment. This method stabilizes data and maintains accuracy during calculations [24].

Method 3 (Me3) involved averaging over time. The fillna function filled the missing value with the average speed value from all vehicles over the same period. Then, the groupby function averaged the speeds for 16 main road segments to reduce the impact of missing values and increase data uniformity [25].

Data normalization: after processing to determine the exact location of vehicles in specific segments and processing anomalous data, processed data were divided from the

data of all time frames into three parts: training set, validation set, and testing set. Then, the enhanced LSTMs (PSO-LSTM and GA-LSTM) are used to generate the TSF model.

2.3. TSF Models

As mentioned above, many studies demonstrated that the efficiency of the LSTM model increases significantly after applying GA-LSTM and PSO-LSTM [20,26]. However, the experiments on the GPS data to forecast the average velocity of segments of the arterial roads having parallel multiple lanes are limited. The PSO or GA enhanced the LSTM network to improve the efficiency of this machine learning model.

Normally, PSO-LSTM and GA-LSTM optimized the following effect parameters to make the model more efficient, such as the window size, the number of epochs, the number of neurons, and the learning rate. In the introduction section, in the previously published research [7], we only optimized three effective parameters: the window size, the number of epochs, and the number of neurons, respectively, similar to applying the LSMT network to optimize forecasting models in several studies [27–29]. This causes a disadvantage for the forecasting model because the effective parameters are not optimized simultaneously and the calculation time is often longer. Hence, we proposed the LSTM network optimization model by GA and PSO algorithms according to the information shown in Figure 3.

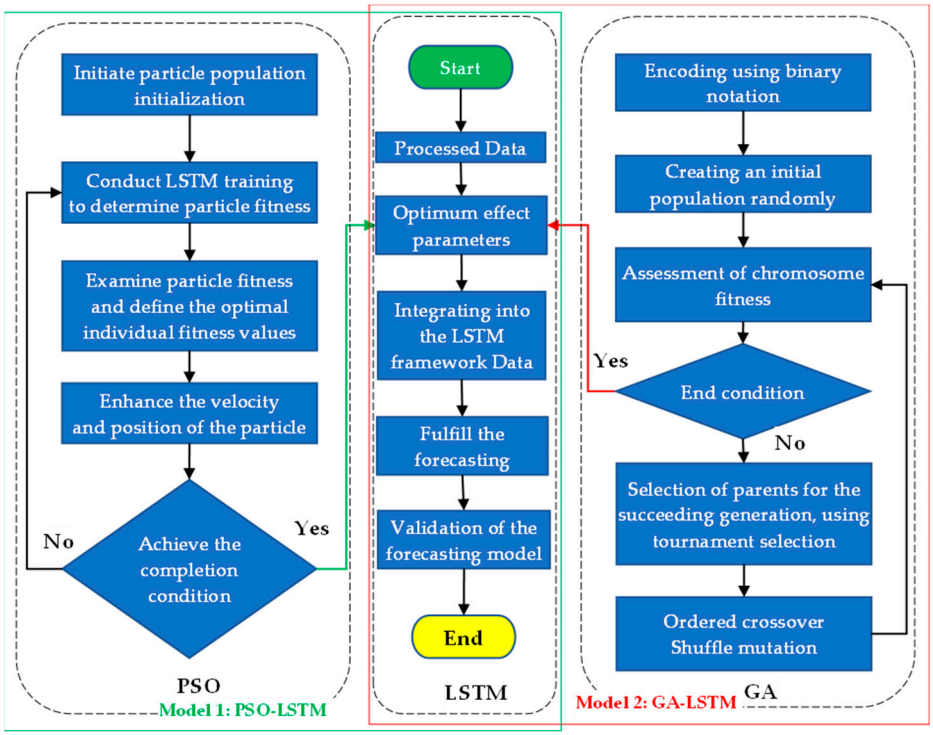


Figure 3. Enhanced Forecasting Models PSO-LSTM and GA-LSTM leverage optimization techniques to enhance the performance and accuracy of LSTM models by enhancing effect parameters.

PSO-enhanced LSTM employs binary encoding, weighting vectors, velocity limits, and MSE as fitness functions to efficiently optimize parameters. The GA-enhanced LSTM utilizes binary encoding, a population size of 20, tournament selection, ordered crossover, and shuffle mutation to enhance model evolution and diversity within the population. These innovative approaches collectively improve forecasting capabilities by optimizing parameters and ensuring robustness in handling complex data. Details of the two models are described as follows.

Model 1: The PSO-enhanced LSTM model utilized several techniques to improve its performance. Firstly, binary encoding was employed to represent the LSTM model's parameters, enhancing the size of the parameter vector and reducing the search space of

feasible solutions. This encoding scheme facilitated more efficient optimization. Secondly, weighting vectors C1 and C2 were introduced to provide different weights to each gene, promoting population diversification and enhancing individuals' search capabilities in the solution space. Thirdly, specific limits and velocity update formulas were designed to maintain stability and prevent velocity values from exceeding predefined limits during optimization. Fourthly, the Mean Squared Error (MSE) on the validation set served as the fitness function to evaluate the performance of each individual, guiding the optimization process effectively. Lastly, speed and position control mechanisms were integrated to prevent excessive or slow convergence, ensuring efficient optimization.

Model 2: For GA techniques, binary encoding was initially utilized to reduce the size of the search space, making the search process more efficient and alleviating computational pressure. Subsequently, each individual was encoded using 12 bits, facilitating the definition of the gene structure and providing sufficient information for the LSTM model. Additionally, a population size of 20 was chosen to ensure diversity within the population, allowing ample time for the GA algorithm to evolve and enhance the population effectively. Tournament selection was employed to impartially select the best individuals in the population, increasing the likelihood of retaining genetic diversity. Furthermore, ordered crossover was implemented to ensure that the newly formed model maintained similarity to the parent model, preserving beneficial characteristics. Lastly, shuffle mutation was introduced to enhance randomness in the genetic mutation process, preventing premature convergence and enabling exploration of a wider search space.

By providing these additional implementation details, this process aims to enhance the reproducibility of the LSTM model enhancements using PSO and GA optimization.

Through a long-term experimental process, we propose parameters that can be used to optimize the LSTM network for probe vehicle data on parallel multilane roads as the following Table 2. Scientists could verify and use directly provided optimal factors with the same GPS data type, thereby reducing processing time as well as improving the accuracy of the providing model.

Table 2. Optimal parameters for LSTM network.

No.	GA		PSO	
	Parameters (Operators)	**Techniques Used**	**Parameters (Factors)**	**Techniques Used**
1	Pre-initialization of the population	Binary encoding	Pre-initialization of the population	Binary encoding
2	Gene length	12	Gene length	12
3	Population size	20	PSO swarm size	20
4	Number of generation	20	Number of PSO Iterations	20
5	Selection Operator	Tournament selection	C_1, C_2 (Cognitive and social coefficients)	Use two vectors C1 and C2 with lengths equal to the number of individuals, with random values in the range [0, 1]
6	Crossover Operator	Ordered	W (Inertia weight)	Apply the weighting process by multiplying these two vectors (C1 and C2) together, giving different weights to each gene.
7	Mutation Operator	Shuffle mutation	Velocity and position control	Limiting velocity and position values to the range [0, 1] keeps the algorithm stable and avoids bursting or over-demanding
8	Evaluate the effectiveness of the fitness value	Mean Squared Error (MSE)	Evaluate the effectiveness of the fitness value	Mean Squared Error (MSE)

2.4. Performance Validation

The three indicators Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Median Absolute Error (MDAE) are common measurements to assess the accuracy of a predictive model. Each indicator offers its unique advantages [30,31].

RMSE (Root Mean Square Error) calculates the average of the square of the error between the prediction and the actual value. It evaluates the larger difference between false predictions and actual values, thereby identifying outliers or false predictions.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{\mathrm{real}} - \hat{x}_{\mathrm{Predicted}}\right)^2} \quad (2)$$

where n represents the count of observations, x_{real} denotes the actual value, and $\hat{x}_{\mathrm{Predicted}}$ stands for the predicted forecast value.

MAE (Mean Absolute Error) averages the absolute value of the error between prediction and reality. Since it does not square the error, MAE helps model formulation to determine the mean deviation without being affected by large errors.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|x_{\mathrm{real}} - \hat{x}_{\mathrm{Predicted}}\right| \quad (3)$$

MDAE (Median Absolute Error) calculates the median of the absolute value of the error between prediction and reality. The median eliminates the effect of outliers and represents a degree of stability and ability to deal with asymmetric data.

$$\mathrm{MDAE} = \mathrm{median}\left|x_{\mathrm{real}} - \hat{x}_{\mathrm{Predicted}}\right| \quad (4)$$

By combining all three indicators, the suggested models have a more comprehensive view of the accuracy of the forecasting issues. RMSE focuses on measuring large errors. MAE helps to determine the mean error. MDAE represents the median error [32]. This combination helps proposed models to evaluate the predictive model as a whole and consider all the important aspects.

3. Experimental Analysis

3.1. Data Type

This paper focuses on forecasting traffic speed on Le Hong Phong road, an arterial road in Hai Phong City, connecting the Hai Phong City center with Cat Bi International Airport. The starting point is located in Ngo Quyen district, Hai Phong City, with latitude 20°50′38.80″ N and longitude 106°42′34.36″ E (Figure 4). The ending point is in Hai An district, Hai Phong City, with latitude 20°49′27.00″ N and longitude 106°43′37.58″ E. This experimental road section has a total length of 2.33 km and is divided into four parts, starting at Ngo Quyen intersection (Intersection No. 1) and ending at Thanh To intersection.

Figure 4. Le Hong Phong experimental road, direction from Cat Bi airport to the city center.

The photograph captures Le Hong Phong experimental road, illustrating the urban road leading from Cat Bi airport towards the city center. This road serves as a critical artery facilitating transportation, bridging the gap between the airport and urban hub.

This figure showcases the distribution of commercial cars' GPS signals along the roadway. This research analyzes 16 road sections along Le Hong Phong road, along with traffic light systems, offering valuable insights into traffic patterns and vehicle movement dynamics on this crucial road. Each part of the road is divided into many lanes, including two central lanes 10 m wide, divided into three lanes for cars (four-wheel vehicles), and two side lanes 7.5 m wide, divided into two lanes for both cars and motorbikes. In total, this research examines 16 road sections on the Le Hong Phong road (Figure 5).

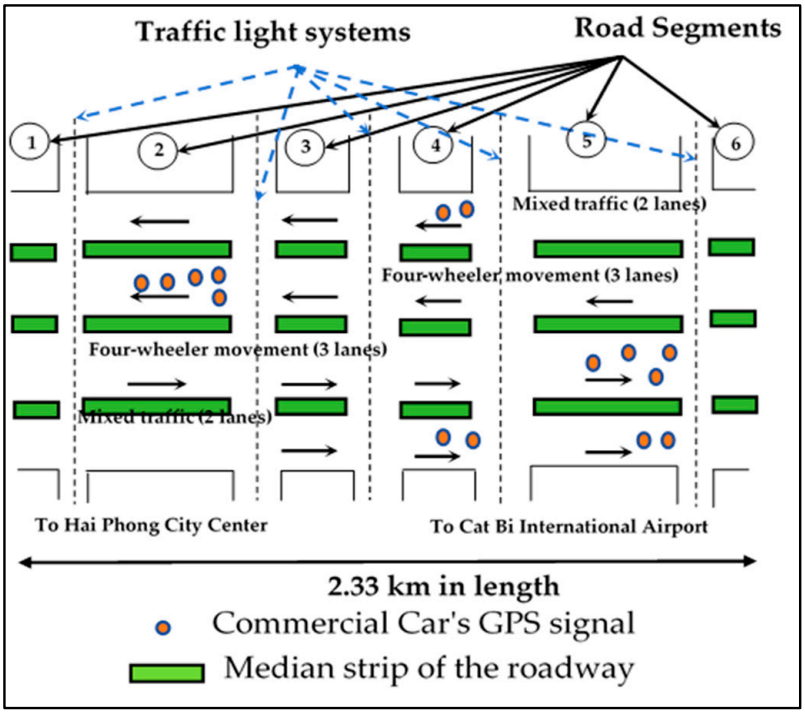


Figure 5. GPS signals of commercial cars on Le Hong Phong road. The red circles denote the GPS signals of commercial cars, while the green median strips delineate the separation between lanes.

Particularly, traffic data were collected by using the GPS transceiver installed in business vehicles. Collection occurred from 6:00 a.m. to 10:00 p.m. between 1 February and 26 February 2020. Data were sampled daily. They included traffic information from other vehicles. It is worth noting that the variability in traffic flows and mixed traffic over time make Traffic Speed Forecasting complex and require the use of effective methods in data processing. After processing to determine the exact location of vehicles in specific segments and processing anomalous and missing data, the processed data from the 26-day period (1 February to 26) were divided into three parts: training set, validation set, and testing set. The data from the first eighteen days were used as the training dataset. The next four days were used for validation, and the remaining four days were used as the test dataset. The proposed model was developed based on the training dataset and validation set.

Subsequently, aiming to improve the traffic conditions and optimize Traffic Speed Forecasting on the Le Hong Phong route, this study used two advanced forecasting algorithms: GA-LSTM and PSO-LSTM. These algorithms have the potential to contribute improvements in TSF and support traffic flow management efficiently on this important road of Hai Phong City.

3.2. Data Processing and Analysis

Anomaly data from the experimental road, as presented in Section 2.2, were processed to produce improved datasets for TSF on each road segment. There are many solutions to handle this problem. As outlined in the ITSG algorithm, the proposed model identified anomalous values by averaging two GPS signals before and two GPS signals after the anomaly value and then changing in accordance with abnormal or missing data. In case all four GPS signals are abnormal, the average value of all vehicles on the road segment is calculated to change the anomalous GPS data. To ensure the effectiveness of this proposed method, the model was compared with three other processing methods mentioned above (Me1, Me2, and Me3).

To ensure effectiveness, the datasets for all four cases were divided into a ratio of 70:15:15, corresponding to the training, validation, and testing datasets. The four methods were compared based on the Root Mean Square Error (RMSE) measure with the experimental dataset.

Based on the detailed analysis of the methods in Table 3, the following main observations and conclusions are presented. The ITSG demonstrated the highest stability and accuracy in evaluating the data, with the best mean value of 9.81 and low variability. This method exhibits uniformity in the distribution of the data, illustrated by the closeness of the 25th and 75th percentiles (8.33 and 11.52), indicating a degree of stability in the data. On the other hand, Me1, despite having the highest mean value (10.64), has the greatest variability, with a standard deviation of 2.44. This indicates instability in the evaluation of the data, possibly due to abnormal values causing large fluctuations. Me2 and Me3, the other two methods, show comparable performance with stable mean values (10.13 and 10.12, respectively) and low variability (standard deviation 2.13 and 2.09). Both methods showed stability and accuracy in evaluating the data, being less affected by outliers. Overall, ITSG showed effective and stable performance in data evaluation, while Me1 showed instability and ineffectiveness compared to other methods.

Table 3. Comparative analysis of parameters across four methods for abnormal data processing.

Factors	ITSG	Me1	Me2	Me3
Count	16	16	16	16
Mean	9.81	10.64	10.13	10.12
Std	1.99	2.44	2.134	2.094
Min	6.81	7.00	6.91	6.91
25%	8.33	8.86	8.64	8.64
50%	9.04	10.213	9.813	9.90
75%	11.52	12.81	11.74	11.87
Max	13.25	14.22	13.89	13.46

Figure 6 illustrates the data post-processing of abnormal GPS tracking, segmented into time frames, velocity (km/h), and segment numbers. This structured representation allows for detailed analysis, providing insights into vehicle speed variations and segment-specific information after processing abnormal GPS data. The speed forecasting results are displayed more clearly in the figure below.

Furthermore, Figure 7 depicts speed values plotted against time. This visualization provides insights into how vehicle speeds fluctuate over time, aiding in the analysis of traffic patterns and dynamics. After processing the data, no unusual velocity values appear.

GPS-monitored data after pre-processing

Time Frame	Speed (Km/h)	Segment
2/1/2020 6:00	32.053	16
2/1/2020 6:40	44.580	1
2/1/2020 6:50	39.547	6
2/1/2020 7:00	33.149	7
2/1/2020 7:10	45.372	4
2/1/2020 7:20	36.292	2
2/1/2020 7:30	56.800	11
2/1/2020 7:50	22.367	14
2/1/2020 8:00	32.693	13
2/1/2020 8:10	38.112	9
2/1/2020 8:40	41.763	5
2/1/2020 8:50	32.665	10
2/1/2020 9:00	29.130	3
2/1/2020 9:30	32.693	15
2/1/2020 10:50	33.168	8
2/1/2020 11:20	25.878	12
...	...	...

Figure 6. Data after processing abnormal GPS tracking data: segmented into time frames, velocity (km/h), and segment numbers.

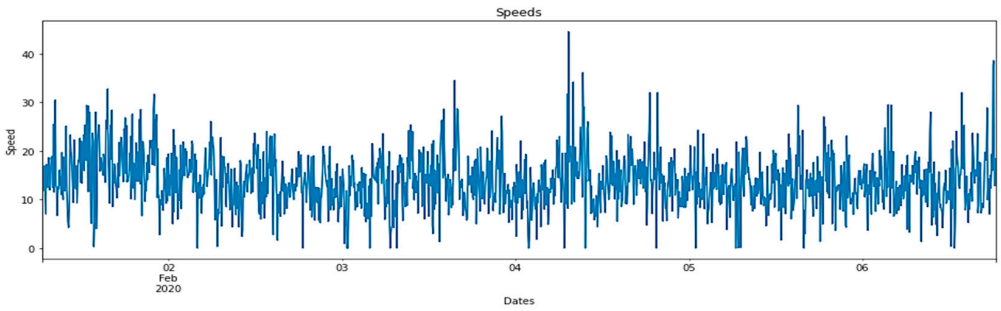


Figure 7. Time series data obtained after initial raw data processing: speed values plotted against time.

3.3. PSO-LSTM and GA-LSTM in Traffic Speed Forecasting

As mentioned above, there are three types of existing solutions to generate a model to forecast the average speed of parallel multilane segments of roads in urban areas: the parametric method, non-parametric method, and machine learning solution. We proposed the following performance comparisons to effectively validate the suggested model by enhancing the LSTM network utilizing GA and PSO.

In the first performance comparison, the research compared proposed methods PSO-LSTM and GA-LSTM with other machine learning methods, LSTM*, CNN, and MLP, based on their effectiveness and applicability in forecasting the average speed of multi-parallel road segments in urban areas.

In particular, LSTM* represented the state-of-the-art method that we previously published by optimizing the parameters window size, number of epochs, and number of neurons sequentially to improve the accuracy of the original LSTM network [7]. LSTM was chosen because of its ability to process data sequences and to learn sample buffers in data time. This is consistent with the time series nature of the GPS data that we are studying.

Convolutional Neural Network (CNN) is considered an effective forecasted method when simultaneously processing many spatial data, such as image space, to generate TSF

models [33,34]. CNN was chosen because of its ability to handle spatial data and to learn special buffers from data images. Even though the road is not an actual target of the image, CNN can still learn specific roads in the road dataset, helping to improve the traffic forecasting model.

Multilayer Perceptron (MLP) is a classic neural network structure; it is highly appreciated for its ability to learn abstract representations of input data to make predictions through layers (input, hidden, and output) [35,36]. MLP is used as a reference model because of its ability to learn symbolic objects of data input in time series prediction and develop a predictive framework [37,38]. Although MLP lacks the sequence processing capabilities of LSTM and CNN, it can still provide a reliable comparison basis for more complex models.

Based on the statistics from Table 4 and other statistical analyses, a violin chart was presented for a more objective comparison. The distribution of enhanced LSTM such as GA-LSTM and PSO-LSTM on the violin graph was uniform, narrow, and lower than graphs of other machine learning methods. Particularly, the median value of PSO-LSTM was always the smallest at 8.15, 6.49, and 5.10, with standard deviations (std) of 1.12, 0.99, and 0.75 for the RMSE, MAE, and MDAE indices.

Table 4. Comparing PSO-LSTM and GA-LSTM with alternative machine learning approaches.

Count	**16 Segments**														
PI	**RMSE Testing**					**MAE Testing**					**MDAE Testing**				
Solutions	**PSO-LSTM**	**GA-LSTM**	**LSTM***	**CNN**	**MLP**	**PSO-LSTM**	**GA-LSTM**	**LSTM***	**CNN**	**MLP**	**PSO-LSTM**	**GA-LSTM**	**LSTM***	**CNN**	**MLP**
Mean	8.23	8.53	9.01	10.38	10.01	6.55	6.74	6.99	8.14	7.88	5.10	5.44	5.74	6.74	6.67
Std	1.12	1.24	1.65	3.23	2.11	0.99	1.07	1.33	2.65	1.75	0.75	0.84	1.17	2.27	1.59
Min Value	6.14	6.18	6.29	6.40	6.63	4.86	4.99	4.80	4.87	5.23	3.55	3.69	4.11	4.16	4.30
Q1	7.31	7.92	8.05	8.16	8.58	6.10	6.14	6.32	6.44	6.72	4.74	4.88	5.12	4.98	5.36
Median	8.15	8.53	8.70	9.50	9.93	6.49	6.69	6.70	7.39	7.63	5.10	5.53	5.52	6.46	6.63
Q3	8.98	9.10	10.06	11.87	11.64	7.30	7.48	7.87	9.26	8.98	5.67	5.98	6.56	7.43	7.60
Max Value	9.89	10.87	12.05	17.12	13.21	8.04	8.25	9.40	13.36	10.78	6.29	6.56	7.81	10.98	10.21
IQR	3.76	4.69	5.76	10.72	6.57	3.18	3.26	4.60	8.49	5.55	2.73	2.88	3.70	6.82	5.91

Additionally, Figure 8 is the violin chart comparison of the PSO-LSTM, GA-LSTM, LSTM, CNN, and MLP methods. The violin charts illustrate the performance of different machine learning methods for vehicle speed forecasting by RMSE, MAE, and MDAE values for 16 road segments.

In the chart above, the CNN method was the least effective method when compared to the other methods through relatively high IQR indices of 10.72 (RMSE test), 8.49 (MAE test), and 6.82 (MDAE test), respectively. These observations are also objectively shown through the width of the violin plot of the CNN method. This could explain why the CNN machine learning method is a more suitable forecasting solution for spatial data or image data as some studies have mentioned.

On the other hand, the GA-LSTM and PSO-LSTM methods appear to be more powerful than the LSTM* method in terms of RMSE, MAE, and MDAE indicators. The two proposed methods also had shorter calculation times than the LSTM* method by optimizing the LSTM network indices simultaneously. In total, PSO-LSTM exhibited superiority, with fewer errors and stable predictability, while GA-LSTM had relative advantages, with low median values and stable predictability. Depending on the specific requirements of the problem, PSO-LSTM or GA-LSTM could be a suitable solution to optimize vehicle speed forecasting performance.

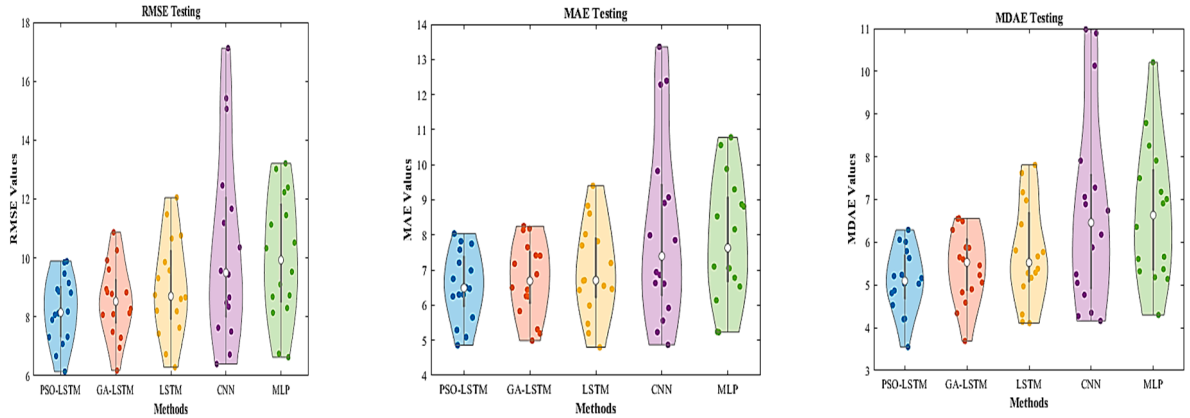


Figure 8. Violin chart comparison of PSO-LSTM, GA-LSTM, LSTM, CNN, and MLP methods based on statistical analyses.

More specifically, PSO-LSTM was the most effective machine learning method to handle our processed probe data type of registered vehicles on Le Hong Phong Road.

To further confirm the effectiveness of our proposed model, this research compared the best mentioned machine learning method above, PSO-LSTM, with other popular methods, such as ARIMA (representing the parametric method [39]) and PROPHET (representing the non-parametric method [40]).

Based on the statistical tables (Table 5) and other analyses, the violin plot was presented according to accuracy assessment indices RMSE, MAE, and MDAE for three TSF methods (PSO-LSTM, ARIMA, and PROPHET) for 16 road segments of Le Hong Phong's experiment road (Figure 9).

Table 5. TSF model comparison (PSO-LSTM, ARIMA, and PROPHET).

Count	PI	RMSE Testing			MAE Testing			MDAE Testing		
	Solutions	PSO-LSTM	ARIMA	PROPHET	PSO-LSTM	ARIMA	PROPHET	PSO-LSTM	ARIMA	PROPHET
16 (Segments)	Mean	8.23	9.57	9.59	6.55	7.45	7.47	5.10	6.17	6.23
	Std	1.12	1.72	1.75	0.99	1.50	1.50	0.75	1.41	1.42
	Min Value	6.14	7.02	6.98	4.86	5.35	5.26	3.55	4.24	4.29
	Q1	7.31	8.21	8.35	6.10	6.27	6.36	4.74	5.10	5.06
	Median	8.15	9.37	9.42	6.49	7.25	7.32	5.10	6.13	6.25
	Q3	8.98	10.66	10.68	7.30	8.53	8.38	5.67	7.09	7.00
	Max Value	9.89	12.20	12.24	8.04	9.96	10.09	6.29	9.12	9.34
	IQR	3.76	5.18	5.26	3.18	4.61	4.83	2.73	4.88	5.05

Figure 9 presents violin charts comparing the PSO-LSTM, ARIMA, and PROPHET forecasting models, highlighting the RMSE, MAE, and MDAE distributions.

The experimental distributions of RMSE, MAE, and MDAE of PSO-LSTM are narrower than the other two representative methods that stand for non-parametric methods and parametric methods. Specifically, the IQR values of PSO-LSTM were the smallest at 3.76, 3.18, and 2.73 regarding RMSE, MAE, and MDAE, respectively.

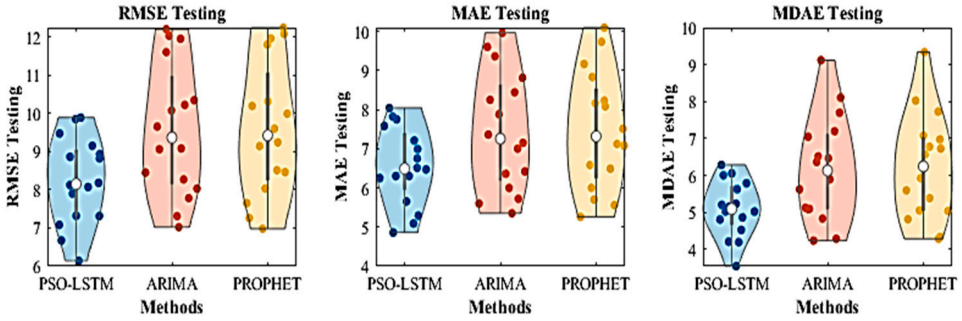
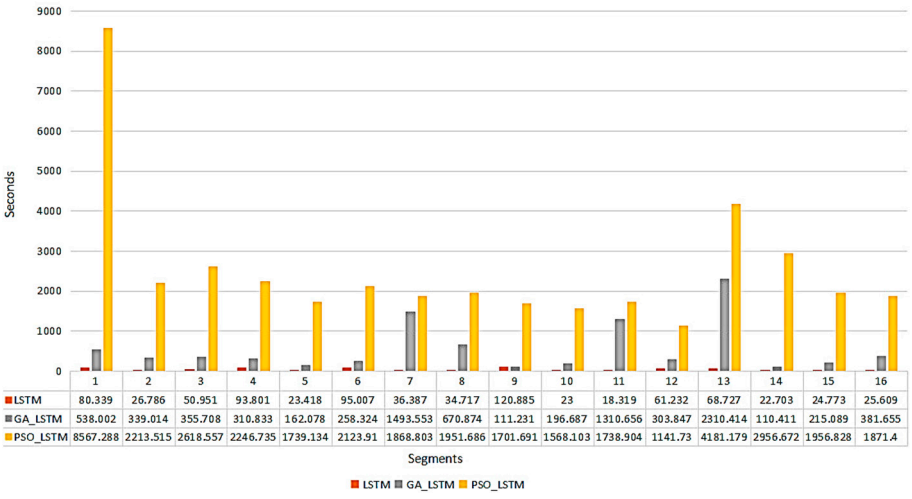


Figure 9. Violin chart comparison of PSO-LSTM, ARIMA, and PROPHET forecasting models.

The violin plot of the RMSE test for ARIMA displays a wider distribution than PSO-LSTM and PROPHET. The graph shows that PROPHET's MDAE test has a wider and higher distribution than the other two methods, indicating significant bias in the forecasting model. Based on the RMSE test, MAE test, MDAE test, and statistical analysis in Table 5 above, PSO-LSTM showed more stability and accuracy in TSF with this type of data.

Figure 10 shows the runtime of three algorithms on sixteen experimental road segments. Upon observing the chart, it becomes evident that the runtime of the optimal algorithm LSTM using PSO and GA takes much longer than LSTM. This is understandable because, while LSTM fits directly with randomly initialized parameters on the training datasets, GA-LSTM and PSO-LSTM have to undergo an evolutionary process of GA and a movement iteration process of particles in PSO to find the optimal parameter set for LSTM. Another noticeable point is that the runtime of PSO significantly surpasses GA. Because of the GA technique, only a few offsprings need to calculate fitness values at each generation. Meanwhile, in PSO, at each movement iteration, all the particles need to update their positions and velocities, thus all having to recalculate fitness values. This leads to PSO having the longest runtime. This tradeoff can be acceptable because both PSO-LSTM and GA-LSTM yield better results than LSTM alone, and PSO-LSTM provides the best result.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
LSTM	80.339	26.786	50.951	93.801	23.418	95.007	36.387	34.717	120.885	23	18.319	61.232	68.727	22.703	24.773	25.609
GA_LSTM	538.002	339.014	355.708	310.833	162.078	258.324	1493.553	670.874	111.231	196.687	1310.656	303.847	2310.414	110.411	215.089	381.655
PSO_LSTM	8567.288	2213.515	2618.557	2246.735	1739.134	2123.91	1868.803	1951.686	1701.691	1568.103	1738.904	1141.73	4181.179	2956.672	1956.828	1871.4

Figure 10. Computational performance comparison: LSTM, PSO-LSTM, and GA-LSTM.

After processing the data and running the TSF model by PSO-LSTM, the forecasting results for both the training and testing datasets of all the road segments are presented. The forecast results of Road Seg.1 shown in Figure 11 below are representative results for 16 experimental road sections of the Le Hong Phong road.

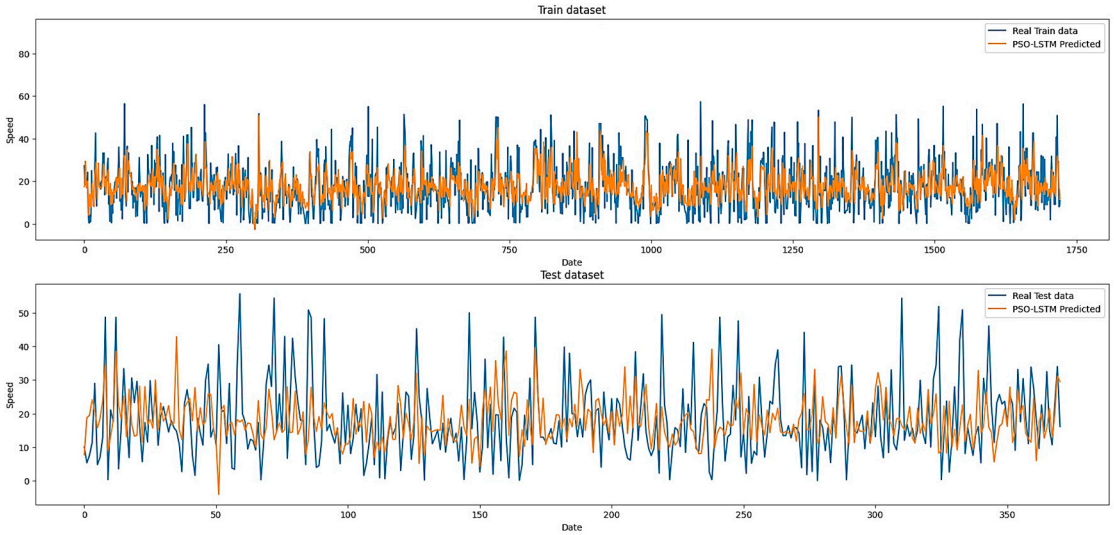


Figure 11. TSF results of PSO-LSTM with training and testing dataset on Seg.1 of the Le Hong Phong road. Comparing real train and test data with PSO-LSTM forecasting model.

Figure 11 shows the TSF results of PSO-LSTM on Seg.1 of Le Hong Phong Street, depicting real train data and test data alongside the PSO-LSTM forecasting values. The above image showcases real train data (blue zigzag line) and PSO-LSTM predictions (orange zigzag line) for Seg.1 of Le Hong Phong Street. Similarly, the below image displays real test data and PSO-LSTM forecasting values. These visualizations offer insights into the accuracy of PSO-LSTM in forecasting vehicle speeds on this road segment.

4. Discussion

This paper focused on processing big data sources from GPS tracking devices of vehicles registered for transport business in Vietnam. The paper demonstrated that abnormal or missing values always exist in data sources obtained from GPS tracking devices and contributed methods for handling these abnormal or missing GPS values. Furthermore, this research developed a process to optimize TSF on parallel multilane roads in Hai Phong City using two advanced methods, PSO-LSTM and GA-LSTM. Through objective comparisons, the paper proposed a model for handling abnormal or missing signals of GPS tracking devices through two suggested models by pseudocode above.

Specifically, this paper proposed an effective approach to handling anomalous data through using the ITST (Interpolated Time Series Generator) algorithm to handle missing and heterogeneous data values. In addition, the paper also provided optimal values as well as usage techniques of PSO and GA that can be directly applied to further research to reduce calculation time and improve the accuracy of the proposed forecasting model in the field of TSF. From there, scientists in each country can directly apply or develop other research to serve the establishment of domestic traffic forecast models or applications towards smart, sustainable urban development. Comparisons between PSO-LSTM, GA-LSTM, and existing forecasting models demonstrated that PSO-LSTM was superior in constructing the TSF. Moreover, this research showed that LSTM networks were much more effective when using optimization algorithms such as PSO and GA to optimize the crucial parameters of the model, such as the number of hidden neurons, window size, number of iterations (epochs), and learning rate (learning rate).

Through detailed probe data processing models and performed comparisons, this study demonstrated the superiority, expanded, and overcame the limits of the previously published research paper.

The proposed model needs to be further improved to adapt to real-world applications, which hinges on its adaptability, generalizability, and capacity to provide sustainable benefits in dynamic transportation environments. By applying the contributions of this

paper and addressing these factors, researchers can develop a robust and versatile TSF model that supports decisionmaking and enhances transportation systems' resilience and efficiency towards sustainable traffic management.

5. Limitations and Future

Although the research results are quite promising, this research is limited to the scope of the experimental road named Le Hong Phong Road, Hai Phong City, Vietnam. Therefore, applying the results of this paper to other proposed model formulations and in the traffic network is necessary to verify the accuracy of the suggested model. Additionally, the data collected are mainly from commercial cars. Hence, data collection is necessary on many different vehicles of mixed traffic flows and different traffic conditions to suit each country and territory better.

While the research results are feasible, several limitations require attention. Reliance on specific GPS data may limit generalizability, and potential feedback loops necessitate ongoing monitoring.

The results of this research apply to practical traffic management systems in Vietnam. They also reveal prospects for integrating artificial intelligence into improving traffic management capacity and solving the challenges regarding the transportation industry. In the future, the proposed model will be tested on many different types of roads under different conditions and for mixed traffic flows to improve the accuracy of the TSF model. The research will aim to generate an entire traffic forecast model for Hai Phong City and develop it in other cities in Vietnam, which will promote the development of domestic application platforms that contribute to sustainable urban development in the future.

6. Conclusions

This paper once again demonstrates that PSO and GA improved the accuracy of LSTM networks in establishing traffic forecast models. However, for each different data type, the algorithms need to be customizable to fit and find the most optimal model. Research needs to be continuously improved to enhance the accuracy of forecast models to handle traffic congestion in urban areas.

Furthermore, this paper presents a comprehensive methodology for traffic forecasting. Despite the inherent challenges associated with portability and dynamic traffic patterns, the proposed model needs to address these concerns. This can be achieved through adaptability, generalizability, and continuous refinement to ensure the model's relevance and applicability in diverse geographic and infrastructural contexts.

Author Contributions: Conceptualization, V.M.D. and Q.H.T.; Data curation, Q.H.T., K.G.L., and V.T.V.; Formal analysis, Q.H.T.; Methodology, V.M.D., X.C.V., V.T.V. and K.G.L.; Supervision, V.M.D.; Validation, X.C.V. and Q.H.T.; Visualization, Q.H.T. and X.C.V.; Writing—original draft, Q.H.T.; Writing—review and editing, V.M.D., K.G.L., V.T.V. and X.C.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by University of Transport and Communications (UTC) under grant number T2022-CT-004TD.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this paper are available on request from the corresponding author.

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

References

1. Yuan, H.; Li, G. A Survey of Traffic Prediction: From Spatio-Temporal Data to Intelligent Transportation. *Data Sci. Eng.* **2021**, *6*, 63–85. [CrossRef]
2. Mishra, S.; Bhattacharya, D.; Gupta, A. Congestion Adaptive Traffic Light Control and Notification Architecture Using Google Maps APIs. *Data* **2018**, *3*, 67. [CrossRef]
3. Barrington-Leigh, C.; Millard-Ball, A. The world's user-generated road map is more than 80% complete. *PLoS ONE* **2017**, *12*, e0180698. [CrossRef] [PubMed]
4. Gayialis, S.P.; Kechagias, E.P.; Konstantakopoulos, G.D. A city logistics system for freight transportation: Integrating information technology and operational research. *Oper. Res.* **2022**, *22*, 5953–5982. [CrossRef]
5. Agyapong, F.; Ojo, T.K. Managing traffic congestion in the Accra central market, Ghana. *J. Urban Manag.* **2018**, *7*, 85–96. [CrossRef]
6. Zheng, F.; van Zuylen, H.J.; Liu, X.; Le Vine, S. Reliability-Based Traffic Signal Control for Urban Arterial Roads. *IEEE Trans. Intell. Transp. Syst.* **2016**, *18*, 643–655. [CrossRef]
7. Tran, Q.H.; Fang, Y.-M.; Chou, T.-Y.; Hoang, T.-V.; Wang, C.-T.; Vu, V.T.; Ho, T.L.H.; Le, Q.; Chen, M.-H. Short-Term Traffic Speed Forecasting Model for a Parallel Multi-Lane Arterial Road Using GPS-Monitored Data Based on Deep Learning Approach. *Sustainability* **2022**, *14*, 6351. [CrossRef]
8. Tran, Q.H.; Do, V.M.; Dinh, T.H. Traffic signal timing optimization for isolated urban intersections considering environmental problems and non-motorized vehicles by using constrained optimization solutions. *Innov. Infrastruct. Solutions* **2022**, *7*, 299. [CrossRef]
9. Cvetek, D.; Muštra, M.; Jelušić, N.; Tišljarić, L. A Survey of Methods and Technologies for Congestion Estimation Based on Multisource Data Fusion. *Appl. Sci.* **2021**, *11*, 2306. [CrossRef]
10. Astarita, V.; Festa, D.C.; Giofrè, V.P. Mobile Systems applied to Traffic Management and Safety: A state of the art. *Procedia Comput. Sci.* **2018**, *134*, 407–414. [CrossRef]
11. Leduc, G. Road traffic data: Collection methods and applications. *Work. Pap. Energy Transp. Clim. Chang.* **2008**, *1*, 1–55.
12. Fernández, J.D.; Sabou, M.; Kirrane, S.; Kiesling, E.; Ekaputra, F.J.; Azzam, A.; Wenning, R. User consent modeling for ensuring transparency and compliance in smart cities. *Pers. Ubiquitous Comput.* **2020**, *24*, 465–486. [CrossRef]
13. Lee, I.; Shin, Y.J. Machine learning for enterprises: Applications, algorithm selection, and challenges. *Bus. Horizons* **2019**, *63*, 157–170. [CrossRef]
14. Bao, X.; Jiang, D.; Yang, X.; Wang, H. An improved deep belief network for traffic prediction considering weather factors. *Alex. Eng. J.* **2020**, *60*, 413–420. [CrossRef]
15. Yuan, T.; Neto, W.D.R.; Rothenberg, C.E.; Obraczka, K.; Barakat, C.; Turletti, T. Machine learning for next-generation intelligent transportation systems: A survey. *Trans. Emerg. Telecommun. Technol.* **2021**, *33*, e4427. [CrossRef]
16. Bradbury, M.; Taylor, P.; Atmaca, U.I.; Maple, C.; Griffiths, N. Privacy Challenges with Protecting Live Vehicular Location Context. *IEEE Access* **2020**, *8*, 207465–207484. [CrossRef]
17. Siuhi, S.; Mwakalonge, J. Opportunities and challenges of smart mobile applications in transportation. *J. Traffic Transp. Eng.* **2016**, *3*, 582–592. [CrossRef]
18. Klos, A.; Bogusz, J.; Figurski, M.; Kosek, W. On the Handling of Outliers in the GNSS Time Series by Means of the Noise and Probability Analysis. In *IAG 150 Years*; Springer: Cham, Switzerland, 2015; pp. 657–664.
19. Fang, W.; Zhuo, W.; Yan, J.; Song, Y.; Jiang, D.; Zhou, T. Attention meets long short-term memory: A deep learning network for traffic flow forecasting. *Phys. A Stat. Mech. Appl.* **2022**, *587*, 126485. [CrossRef]
20. Mahjoub, S.; Labdai, S.; Chrifi-Alaoui, L.; Marhic, B.; Delahoche, L. Short-Term Occupancy Forecasting for a Smart Home Using Optimized Weight Updates Based on GA and PSO Algorithms for an LSTM Network. *Energies* **2023**, *16*, 1641. [CrossRef]
21. Zhang, H.; Luo, Y.; Qin, F.; He, Y.; Liu, X. Elsd: Efficient Line Segment Detector and Descriptor. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021.
22. Pigott, T.D. Handling and Meta-analysis, Handling Missing Data. 2009. 2: pp. 399–416. Available online: https://books.google.com.vn/books?hl=en&lr=&id=LUGd6B9eyc4C&oi=fnd&pg=PA399&dq=Handling+missing+data.+2009&ots=5QyDuRZq4R&sig=P-3mJg97jEqurK7P-VKpEuYC1HI&redir_esc=y#v=onepage&q=Handling%20missing%20data.%202009&f=false (accessed on 12 February 2024).
23. Ilias, L.; Doukas, G.; Kontoulis, M.; Alexakis, K.; Michalitsi-Psarrou, A.; Ntanos, C.; Askounis, D. Overview of methods and available tools used in complex brain disorders. *Open Res. Eur.* **2023**, *3*, 152. [CrossRef]
24. Bloice, M.D.; Holzinger, A. A tutorial on machine learning and data science tools with python. *Mach. Learn. Health Inform. State Art Future Chall.* **2016**, *2016*, 435–480.
25. Makarov, A.; Namiot, D. Overview of data cleaning methods for machine learning. *Int. J. Open Inf. Technol.* **2023**, *11*, 70–78.
26. Eid, M.M.; El-Kenawy, E.-S.M.; Khodadadi, N.; Mirjalili, S.; Khodadadi, E.; Abotaleb, M.; Alharbi, A.H.; Abdelhamid, A.A.; Ibrahim, A.; Amer, G.M.; et al. Meta-Heuristic Optimization of LSTM-Based Deep Network for Boosting the Prediction of Monkeypox Cases. *Mathematics* **2022**, *10*, 3845. [CrossRef]
27. Meng, X.; Fu, H.; Peng, L.; Liu, G.; Yu, Y.; Wang, Z.; Chen, E. D-LSTM: Short-Term Road Traffic Speed Prediction Model Based on GPS Positioning Data. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 2021–2030. [CrossRef]
28. Hsueh, Y.-L.; Yang, Y.-R. A Short-term Traffic Speed Prediction Model Based on LSTM Networks. *Int. J. Intell. Transp. Syst. Res.* **2021**, *19*, 510–524. [CrossRef]

29. Gao, Y.; Zhou, C.; Rong, J.; Wang, Y.; Liu, S. Short-Term Traffic Speed Forecasting Using a Deep Learning Method Based on Multitemporal Traffic Flow Volume. *IEEE Access* **2022**, *10*, 82384–82395. [CrossRef]
30. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]
31. Karunasingha, D.S.K. Root mean square error or mean absolute error? Use their ratio as well. *Inf. Sci.* **2022**, *585*, 609–629. [CrossRef]
32. Botchkarev, A. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv* **2018**, arXiv:1809.03006.
33. Ma, X.; Dai, Z.; He, Z.; Ma, J.; Wang, Y.; Wang, Y. Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction. *Sensors* **2017**, *17*, 818. [CrossRef]
34. Jeong, J.; Kim, H. Multi-Site Photovoltaic Forecasting Exploiting Space-Time Convolutional Neural Network. *Energies* **2019**, *12*, 4490. [CrossRef]
35. Song, C.; Lee, H.; Kang, C.; Lee, W.; Kim, Y.B.; Cha, S.W. Traffic Speed Prediction under Weekday Using Convolutional Neural Networks Concepts. In Proceedings of the Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; pp. 1293–1298. [CrossRef]
36. Oliveira, D.D.; Rampinelli, M.; Tozatto, G.Z.; Andreão, R.V.; Müller, S.M.T. Forecasting vehicular traffic flow using MLP and LSTM. *Neural Comput. Appl.* **2021**, *33*, 17245–17256. [CrossRef]
37. Shiblee, M.; Kalra, P.K.; Chandra, B. Time Series Prediction with Multilayer Perceptron (MLP): A New Generalized Error Based Approach. In *International Conference on Neural Information Processing, Auckland, New Zealand, 2008*; Koppen, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 37–44.
38. Khan, R.U.; Almakdi, S.; Alshehri, M.; Kumar, R.; Ali, I.; Hussain, S.M.; Haq, A.U.; Khan, I.; Ullah, A.; Uddin, M.I. Probabilistic Approach to COVID-19 Data Analysis and Forecasting Future Outbreaks Using a Multi-Layer Perceptron Neural Network. *Diagnostics* **2022**, *12*, 2539. [CrossRef] [PubMed]
39. Shahriari, S.; Ghasri, M.; Sisson, S.A.; Rashidi, T. Ensemble of ARIMA: Combining parametric and bootstrapping technique for traffic flow prediction. *Transp. A: Transp. Sci.* **2020**, *16*, 1552–1573. [CrossRef]
40. Wu, Z.; Chen, X.; Gao, Z. Bayesian non-parametric method for decision support: Forecasting online product sales. *Decis. Support Syst.* **2023**, *174*, 114019. [CrossRef]