Deep Learning Framework for Locating Physical Internet Hubs using Latitude and Longitude Classification

El-Sayed Orabi Helmi[1], Osama Emam[2], Mohamed Abdel-Salam[3]
Dept. Business Information Systems, Faculty of Commerce and BA, Helwan University, Cairo, Egypt[1, 3]
Dept. of Computer Science, Faculty of Computing and AI, Helwan University, Cairo, Egypt[2]

***Abstract*—This article proposes framework for determining the optimal or near optimal locations of physical internet hubs using data mining and deep learning algorithms. The framework extracts latitude and longitude coordinates from various data types as data acquisition phase. These coordinates has been extracted from RIFD, online maps, GPS, and GSM data. These coordinates has been class labeled according to decision maker's preferences using k-mean, density based algorithm (DB Scan and hierarchical clustering analysis algorithms. The proposed algorithm uses haversine distance matrix to calculate the distance between each coordinates rather than the Euclidian distance matrix. The haversine matrix provides more accurate distance surface of a sphere. The framework uses the class labeled data after the clustering phase as input for the classification phase. The classification has been performed using decision tree, random forest, Bayesian, gradient decent, neural network, convolutional neural network and recurrent neural network. The classified coordinates has been evaluated for each algorithms. It has been found that CNN, RNN outperformed the other classification algorithms with accuracy 97.6% and 97.9% respectively.**

***Keywords*—Physical internet hubs (π hubs); deep learning; convolutional neural network (CNN); recurrent neural network (RNN); latitude and Longitude classification**

I. INTRODUCTION

Choosing the location of storage warehouses is one of the most important steps facing supply chain officials, if not the most important step. This step has gained importance because of its impact on the entire supply chain. It also affects transportation operations and determines the speed of response to customer requests. Therefore, it was necessary for many researchers and specialists in supply chains to use all available means and techniques to determine the best location for these warehouses. With the rapid development of communication technologies and their overlap with the Internet of Things, large amounts of data became available for analysis and to give accurate mathematical alternatives to solve the problem of choosing the locations of those repositories. There are devices to track the movements of cars and their various effects along the supply chain. These devices also show the time of movement and waiting for these cars in an accurate and round-the-clock manner. There were many ways to analyze this data to reach the optimal location for these repositories. Classification processes are one of the most used methods to reach the most accurate solutions. The researchers also used other methods, such as relying on the global positioning system technology to track trucks in various places. From this standpoint, the research team decided to use the available data from that technology, taking into account the waiting hours, to determine the best places for these warehouses. But before proceeding in detail in this research, some concepts related to the topic of research must be presented, including first the different data collection techniques and the reason for choosing some of them over the other. Second, the methods of data analysis and the algorithms used to analysis that data. Third, the research team will address the new phenomenon in the field of supply chains, namely the physical Internet (PI), and the extent of its expected impact in the near future, in context of what has been published on this subject in some international scientific journals. Fourth, a detailed overview of deep learning techniques in the context of the research topic will be illustrated.

First vehicles tracking [1] Global Positioning System, The Global Positioning System (GPS), formerly known as Navstar GPS, is a satellite-based radio navigation system owned by the US government and maintained by the US Space Force. The GPS does not require the user to send any data, and it works independently of telephonic or Internet reception, however both technologies can improve the accuracy of GPS positioning data. Military, civic, and commercial users all across the world rely on the GPS for crucial positioning. The US government designed, maintains, and administers the system, which is publicly accessible to anybody with a GPS device. Based on data received from several GPS satellites, the GPS receiver estimates its own four-dimensional position in space-time. Each satellite keeps a precise record of its position and time, which it sends to the receiver which in our case is attached to trucks.

On other hand, Radio Frequency Identification (RFID) is a passive wireless technology that allows an item or person to be tracked or identified. Tags and readers are the two main components of the system. Radio Frequency Identification (RFID) relies on a small electrical device, generally a microchip, to store data. These devices are usually quite small, about the size of a grain of rice, and can store a lot of information. Some contain a stored power source or batteries, even if they don't always emit electricity. The scanners that are used to read these devices can also offer enough power to read the microchip. The technology has a variety of applications, but it is most typically used to track objects.

Global System for Mobile Communication (GSM) based tracking system is presented that uses a mobile phone text message system to keep track of a vehicle's location and speed. The technology may send text notifications for speed and location in real time. One of the disadvantages of this technology is that it bears an additional cost and weak or no signals in some places during the supply chain.

On other hand, researcher used computer vision approach for latitude and Longitude coordinates extraction [2]. A lot of data is required for computer vision. It repeats data analysis until it detects distinctions and, eventually, recognizes images. To teach a computer to recognize automotive tires, for example, it must be fed a large number of tire photos and tire-related materials in order for it to understand the differences and recognize a tire, particularly one with no faults. Computer vision enables procedures to be automated, saving time and reducing the pressure on a limited labor supply. Computer vision has been used to unload inventory trailers with increased efficiency. Vehicle departure times must be carefully synchronized with the loading periods of other trucks, trains, or planes.

The motive that motivated the research team to conduct such a study is the lack of a single framework that can deal with different types of data, whether classified or unclassified data. The team believes that by creating a single framework that integrates many data sources and provides different mining and machine learning algorithms and techniques, it can provide an addition in supply chain and physical internet operations, especially in the current and future period.

This article is divided into several parts. The first part provides solid overview of the main concepts of physical internet and deep learning techniques. It also, provides a brief summary of some related researches in the logistics industry. The second part discusses in detail the proposed framework. The last part demonstrates the experiments and results.

II. Theoretical Background

The following section discusses briefly the main concept of physical internet and deep learning. Some related studies will be discussed in section.

A. Physical Internet

The physical internet (PI) was developed in response to the inefficiencies and unsustainable nature of current logistics and supply chain management strategies. It was proposed to address issues that make current logistics practices unsustainable, such as limited space utilization for road, rail, sea, and air transportation, empty travel being the norm rather than the exception; poor working conditions for truck drivers, products sitting idle, inefficiency of product distribution, inefficient use of production and storage facilities, mediocre coordination within distribution networks, and high inefficiency of multimodal transportation [3].

The primary goal of the PI is to change "the way physical goods are handled, moved, stored, realized, supplied, and used, with the goal of improving global logistics efficiency and sustainability." The PI intends to coordinate physical commodities transit in the same way as data packets are carried via the digital Internet. The movement of commodities will be optimized in terms of cost, speed, efficiency, and sustainability by pooling resources such as vehicles and data and constructing transit centers that enable smooth interoperability. The PI establishes common and generally agreed-upon standards and protocols to promote horizontal and vertical cooperation amongst businesses in order to achieve this optimization.

The PI does not directly manipulate physical items, but rather manipulates and maintains the transportation containers that house them, just as digital Internet packets store embedded data. Many factors must be coordinated for the PI to become fully functional, including tangible items such as PI modular containers or PI transit centers, as well as more abstract notions such as legislation and business models. Previous research, which concentrated on performing simulations with a small number of participants in specific industries, has demonstrated that the application of the PI, even if it is partial, can result in significant advantages [4].

As shown in Fig. 1, according to ALICE [5] The Physical Internet contemplates the transformation of Logistics Nodes into Physical Internet nodes with standardized service definition and operations. Services at PI nodes are visible and digitally accessible to businesses, and they cover planning, booking/transactions, execution, and information exchange. The Physical Internet is based on a comprehensive and systemic consolidation of flow and network of networks principles. The Physical Internet promotes full consolidation of logistics flows from independent shippers in logistics networks (e.g., extended pooling). The Physical Internet promotes pooling resources and assets in open, connected, and shared networks (i.e. connecting existing (business) networks, capabilities, and resources) so that network users and partners can use them easily. It is predicted that by pooling demand and resources to meet that demand, resource utilization will be more efficient. Transport, storage, and physical handling procedures of load units such as containers, swap-bodies, pallets, boxes, and so on are included in the Physical Internet, as is any other resource required for a freight transport and logistics operation [5].

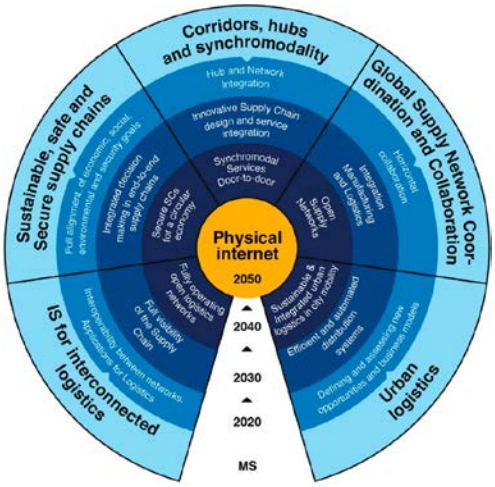

Fig. 1. Physical Internet according to the European Technology Platform ALICE [5].

B. Deep Learning

Deep learning is a machine learning technique that trains computers to do what people do instinctively. Deep learning is a major technology underpinning a lot of applications such as automated driving, medical research, voice control, text, inventory management and other critical fields [6]. Deep learning has received a lot of attention recently, and for good reason. It is attaining results that were previously unthinkable. Deep learning models can attain cutting-edge accuracy, sometimes outperforming humans. Models are trained utilizing a huge quantity of labeled data and multi-layered neural network architectures [7] Fig. 2 demonstrates architecture of deep learning network layers. Deep learning has several known algorithms such as convolution neural network (CNN), recurrent neural network (RNN), long short term memory networks (LSTM) and generative adversarial networks (GANs). CNN and RNN are the main concern in this research. CNNs are multilayer perceptron that have been regularized. Multilayer perceptron are typically completely connected networks, in which each neuron in one layer is linked to all neurons in the following layer. Because of their complete connectedness, these networks are subject to data over fitting. Regularization, or preventing over fitting, is commonly accomplished by punishing parameters during training (such as weight decay) or reducing connectivity (skipped connections, dropout, etc.) CNNs use the hierarchical pattern in data to assemble patterns of increasing complexity utilizing smaller and simpler patterns imprinted in their filters. As a result, CNNs are at the lowest end of the connectivity and complexity spectrum [8]. RNNs are a form of neural network capable of modeling sequence data. RNNs, which are generated from feed forward networks, behave similarly to human brains. Simply said, recurrent neural networks are capable of anticipating sequential data in ways that other algorithms are not. RNNs feature a Memory that retains all calculation information. It uses the same parameters for each input because doing the same task on all inputs or hidden layers yields the same result.

Deep learning has several known algorithms such as convolution neural network (CNN), recurrent neural network (RNN), long short term memory networks (LSTM) and generative adversarial networks (GANs). CNN and RNN are the main concern in this research. CNNs are multilayer perceptron that have been regularized. Multilayer perceptron are typically completely connected networks, in which each neuron in one layer is linked to all neurons in the following layer. Because of their complete connectedness, these networks are subject to data over fitting. Regularization, or preventing over fitting, is commonly accomplished by punishing parameters during training (such as weight decay) or reducing connectivity (skipped connections, dropout, etc.) CNNs use the hierarchical pattern in data to assemble patterns of increasing complexity utilizing smaller and simpler patterns imprinted in their filters. As a result, CNNs are at the lowest end of the connectivity and complexity spectrum [8]. Fig. 3 illustrates CNN network architecture.

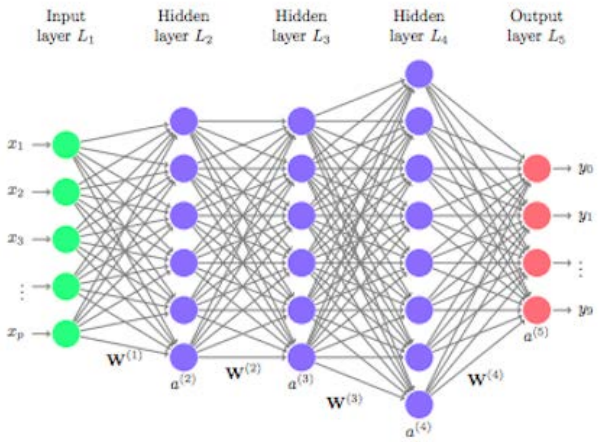

Fig. 2. Deep Learning Network Architecture [7].

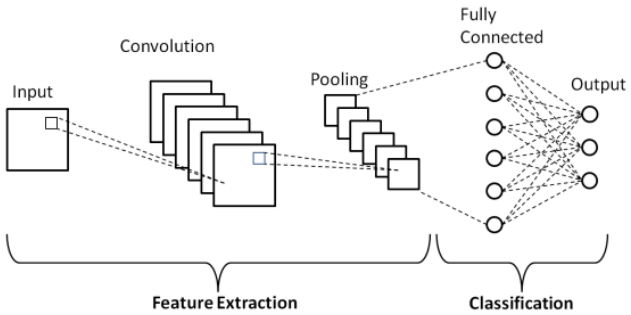

Fig. 3. Convolution Network Architecture [8].

RNNs are a form of neural network capable of modeling sequence data. RNNs, which are generated from feed forward networks, behave similarly to human brains. Simply said, recurrent neural networks are capable of anticipating sequential data in ways that other algorithms are not. RNNs feature a Memory that retains all calculation information. It uses the same parameters for each input because doing the same task on all inputs or hidden layers yields the same result.

C. Related Work

Several studies have been published to determine the best locations for supply chain warehouses using a number of different methods. Some of these researches used mathematical models to determine the locations and others used data mining techniques. In this section, the results of some of these studies, pointing out their advantages and disadvantages will be reviewed.

In [9], the researchers proposed an integrated Multi-Objective Hybrid Harmony Search-Simulated Annealing (MOHS-SA) algorithm to find the trade-off between the total cost of operating facilities, and the cost of CO2 emissions. The research presented a number of alternatives to warehouse locations, but it lacked any possibility of self-learning or determining the optimal location for these warehouses.

On other hand, in [10] the researchers examined a case study at a company that works in the Fast Moving Consumer Goods (FMCG) area and was undertaking a market test for its new alternative tobacco products using the B2C distribution system. The goal of this research is to identify both the number and location of warehouses that provide the lowest total logistics costs for the B2C process, while the company is still employing company-owned B2B warehouses. The number and location of these warehouses are determined using Agglomerative Hierarchical Clustering with Evolutionary Solver, where clustering is based on the shortest distance. One of the important effects that resulted from this research is that he was able to determine the number of warehouses based on the minimum costs, whether operating or transportation costs, in line with customer requests. The weak point of this research is that it locates warehouses based on cost only, without looking at other dimensions such as the environmental dimension, increased demand, or the introduction of other types of products. On other words this research is not the optimal solution for general case study.

In 2020 [11] proposed a mathematical model to allocate the warehouse centers in Far East of Russia based on population, trade turnover and volume of goods manufacturing. This research divided the warehouses to three categories federal, regional, and local centers. These logistics centers have been located in different cities regardless of the exact geographic information of these centers.

Using Lingo solver software [12] the researchers proposed a solution to locate warehouse based on linear programming. The decision in this research has been taken on cost of moving material, frequency trips of material, and the distance of location. This research suffered from the lack of the possibility of self-learning, as it relied entirely on linear mathematical equations without exposure to the non-linear nature of this problem.

In [13], the researchers proposed a solution to determine the optimal location of supply chain distribution centers using k-near neighbor clustering algorithm. They used spatial dataset in their case study. They have demonstrated the ability of this algorithm to divide geographical areas and locate warehouses. The defect of this algorithm was the necessity to determine the number of warehouses before starting the process of clustering the spatial data.

The authors of this paper [14] proposed a two-step clustering approach. In the first stage, they extract the frequent GPS trajectory halt locations. The second stage is to use a density-based clustering method to determine the nearest locations. Because it starts clustering from extracted points, the suggested approach saves time.

In [15], the authors proposed a clustering model using automatic identification system and DBSCAN algorithm to spot the location of ships using the navigation GPS latitude and longitude. The residence point of each ship is identified according to the ship speed and course change. This research uses only GPS data and lacks of self-learning possibility.

III. Proposed Framework

In this part, the proposed framework for solving the problem of locating supply chain repositories in context of the physical internet phenomenon will be presented. The proposed framework consists of four phases, data acquisition, data labeling, classification and testing. During data acquisition and selection phase, data from many sources is combined into a single data repository, resulting in a target dataset containing intriguing variables or data samples for discovery.

Due to the lack of classified data in many cases in actual supply chain applications and to maximize the effectiveness of the proposed framework, the research team performed a preliminary clustering of the collected data as data labeling phase. The clustering stage will be done using three known algorithms K-mean, density based algorithm (DB Scan) and hierarchical cluster analysis (HCA) to satisfy most of the decision-makers requirements. For example, when using k-mean algorithm, the decision maker can determine the number of PI hubs according to the available resources and future plans of the organization such equipment, transportation and workers. Decision makers can use the DB Scan algorithm to cluster the data without specifying a specific number of Hubs, due to the nature of the work of this algorithm. It divides the data into any number of groups so that the differences between the points of the same group are reduced to a minimum. If it is expected that there will be groups with arbitrary shapes, the decision maker can use DB Scan algorithm. The decision makers can use HCA to cluster the latitude and longitude coordinates according to regions or cities. The most significant aspect in hierarchical clustering is the linkage mechanism, which determines how the distances between clusters will be calculated. It has a significant impact on not just the clustering quality but also the algorithm's efficiency. The k-centroid link method has been chosen in our implementation. The output of this phase is labeled training dataset.

The clustering distance matrix used in the proposed framework is haversine formula. The haversine formula is a very accurate method of estimating distances between two places on the surface of a sphere given the two points' latitude and longitude. The haversine formula is a re-formulation of the spherical law of cosines; however the haversine formulation is more useful for tiny angles and distances. The central angle (θ) between any two points on a sphere is:

$$\theta = \frac{d}{r}$$

where:

d is the distance between the two points along a great circle of the sphere.

r is the radius of the sphere.

The haversine formula allows the haversine of θ (that is, hav(θ)) to be computed directly from the latitude (represented by φ) and longitude (represented by λ) of the two points:

$$hav(\theta) = hav(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\, hav(\lambda_2 - \lambda_1)$$

where:

φ1, φ2 are the latitude of point 1 and latitude of point 2,

λ1, λ2 are the longitude of point 1 and longitude of point 2.

Finally, the haversine function hav(θ), applied above to both the central angle θ and the differences in latitude and longitude, is

$$hav(\theta) = sin^2\left(\frac{\theta}{2}\right) = \frac{1-\cos(\theta)}{2}$$

In the third phase, the proposed framework presents the possibility of using a number of well-known classification algorithms to classify data, such as: decision tree, Naïve Bayes, random forest, gradient descent, k-nearest neighbors, support vector machine (SVM), CNN, and RNN. Although some of these algorithms are lazy learning algorithms, they classify data accurately in many cases. The proposed framework uses CNN and RNN deep learning algorithms to add self-learning nature.

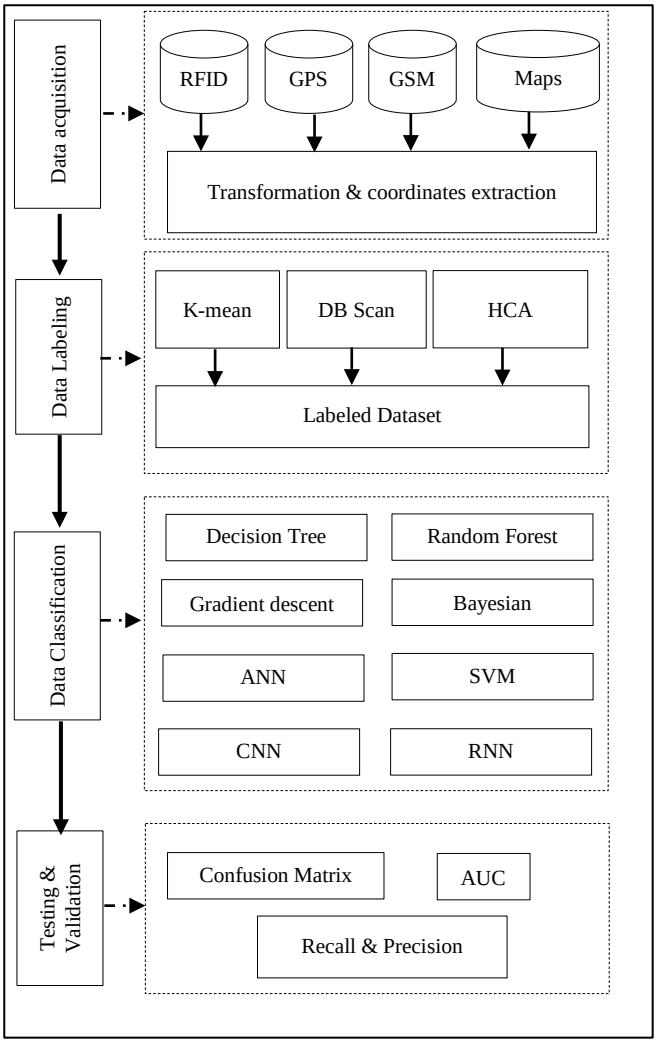

Fig. 4. Physical Internet Hubs Locating using Latitude and Longitude Classification.

Testing and validation is the fourth phase. At this phase, the results of all classification algorithms that have been used in the proposed framework are evaluated to find out the optimal outcomes within standardized testing criteria. In this phase the area under curve (AUC), confusion matrix and precision have been used to calculate the framework accuracy.

As in Fig. 4 the framework extracts latitude and longitude coordinates from four major resources (RFID, GPS, GSM, and maps) at data acquisition phase the framework has been developed to handle all of these data types. It is not necessary to use all data types in real case scenarios; any user can customize this step according to his application.

IV. Experiments and Results

This section discusses in detail the performed experiments. All experiments were carried out using Spain's coordinate's dataset on the Kaggle website [16]. The frame work has been implemented with python3 environment. Table I illustrates the Spain's coordinate's dataset main features.

All null features (city, district, & region) and missing instances have been dropped as preprocessing phase. The reset 975000 instances will be divided as 70% for training and 30% for testing and validation for the classification purposes after the preparation phase.

The dataset was divided into 15 class labeled groups using k-mean algorithm as data preparation phase. The number of groups was chosen to reduce distances to a minimum in order to reduce the time and cost of transporting goods between points which achieves sustainability. Then this labeled dataset has been fed to the proposed framework for classification as 3rd phase. The architecture implementation of CNN algorithms was as the following: 1 Conv. layer, 5 dense layers and the used activation function was Relu. The RNN implementation also as the following; 1 RNN layer, 4 dense function using tanh activation function. Table II shows the evaluating comparison of the used classification algorithms.

As shown in Table II, the Bayesian algorithm classification accuracy was 86.2%. The Random forest algorithm achieved 90.2% accuracy. CNN and RNN algorithms made the classification with 97.6%, and 97.9% respectively. On other hand SVM classified the dataset with accuracy 95.1%. The previous results have been calculated with confidence level 95%. The results showed also, that RNN outperformed the classification rather than other algorithms.

TABLE I. Spain Open Address Dataset Features

#	Features	Unique Instances
1	Latitude	977080
2	Longitude	977070
3	Street name	470045
4	City	Null
5	District	Null
6	Region	Null
7	Postal code	976000

TABLE II. THE PROPOSED FRAMEWORK ACCURACY COMPARISON

	AUC	CA	Precision	Recall
Bayesian	0.8956	0.862	0.885	0.890
Random Forest	0.9145	0.902	0.903	0.912
CNN	0.9806	0.976	0.971	0.978
RNN	0.9812	0.979	0.982	0.989
SVM	0.9342	0.928	0.913	0.922
ANN	0.9564	0.951	0.934	0.946
Gradient descent	0.9012	0.899	0.891	0.898
Decision tree	0.9102	0.909	0.906	0.904

V. CONCLUSION

The proposed framework consists of four major phases. The first phase is data acquisition. In this phase the latitude and longitude data extracted from different data sources such as: RFID, GSM, GPS, and maps. This data has been class label using one of three clustering algorithms (k-mean, DB scan, & HCA) as second phase. The third phase is classification phase. In this phase, the decision makers can choose from one or more classification algorithms such as: Bayesian, decision tree, random forest, SVM, ANN, CNN and RNN. The performed experiments showed that RNN classified the data better than other algorithms with accuracy 97.9%. According testing results, the research team believes that it is preferable to locate the physical internet hubs by one of deep learning algorithms especially CNN or RNN rather than other classification techniques. This study can be extended in the future by adding new deep learning algorithms or using different activation functions.

REFERENCES

[1] Sumit S. Dukare, Dattatray A. Patil, Kantilal P. Rane, Vehicle Tracking, Monitoring and Alerting System: A Review, International Journal of Computer Applications, Vol. 119, pp. 39-40, 2015.

[2] Bo Yang, Mingyue Tang, Shaohui Chen, Gang Wang, Yan Tan & Bijun Li., A vehicle tracking algorithm combining detector and tracker, Journal on Image and Video Processing, Vol. 17, pp. 10-17, 2020.

[3] Horst Treiblmaier, Kristijan Mirkovski, Paul Benjamin Lowry, Zach G. Zacharia., the physical internet as a new supply chain paradigm: a systematic literature review and a comprehensive framework, The International Journal of Logistics Management, Vol. 31, pp. 240-280, 2020.

[4] Sarraj, R., Ballot, E., Pan, S., Hakimi, D. and Montreuil, B, Interconnected logistic networks and protocols: simulation-based efficiency assessment, International Journal of Production Research, Vol. 52 , pp. 3185-3208, 2014.

[5] Klumpp, Matthias, Automation and artificial intelligence in business logistics systems: human reactions and collaboration requirements, International Journal of Logistics, Vol. 21, pp. 224- 242, 2017.

[6] Vega-Márquez B,Nepomuceno-Chamorro I,Jurado-Campos N and Rubio-Escudero C , Deep Learning Techniques to Improve the Performance of Olive Oil Classification, Frontiers Chemistry, Vol. 7, pp. 1-10, 2020.

[7] Vankara, J., Krishna, M.M., Dasari, S, Classification of Brain Tumors Using Deep Learning-Based Neural Networks, , Smart Technologies in Data Science and Communication, Singapore, Springer , pp. 33-40, 2021.

[8] Avilov, Oleksii; Rimbert, Sebastien; Popov, Anton; Bougrain, Laurent. Montreal , Deep Learning Techniques to Improve Intraoperative Awareness Detection from Electroencephalographic Signal,2nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 142–145, 2020.

[9] F Misni, L S Lee, N I Jaini, Multi-objective hybrid harmony search-simulated annealing for location-inventory-routing problem in supply chain network design of reverse logistics with CO2 emission, Journal of Physics: Conference Series, pp. 1-16, 2021

[10] Nyoman Sutapa, Magdalena Wullur, Tania Nano Cahyono, Determining the Number and Location of Warehouses to Minimize Logistics Costs of Business to Consumer (B2C) Distribution, SHS Web of Conferences, pp. 1-8, 2020.

[11] A Bardal , M Sigitova, Localization of Transport and Logistics Centers in the Region, IOP Conf. Series: Materials Science and Engineering, pp. 1-6, 2020.

[12] Ade Irman, Y Muharni, Andri Yusuf. Bali, Design of warehouse model with dedicated policy to minimize total travel costs: a case study in a construction workshop: International Conference on Advanced Mechanical and Industrial engineering, Indonesia pp. 1-7, 2020.

[13] Sumit S. Dukare, Dattatray A. Patil, Kantilal P. Rane., Vehicle Tracking, Monitoring and Alerting System: A Review, International Journal of Computer Applications, Vol. 119, pp. 29-40, 2015.

[14] Bo Yang, Mingyue Tang, Shaohui Chen, Gang Wang, Yan Tan & Bijun Li, A vehicle tracking algorithm combining detector and tracker.. s.l. : J Image Video Proc., Vol. 17, pp. 10-20, 2020.

[15] Horst Treiblmaier, Kristijan Mirkovski, Paul Benjamin Lowry, Zach G. Zacharia., the physical internet as a new supply chain paradigm: a systematic literature review and a comprehensive framework, The International Journal of Logistics Management, Vol. 31, pp. 240-280, 2020.

[16] Open address Europe. www.Kaggle.com. [Online] May 10, 2017. [Cited: J https://www.kaggle.com/datasets/openaddresses/openaddresses-europe?select=spain.csv, last access: 5 june 2022..