 doi:10.20944/preprints202211.0130.v1

Article

Machine Learning Applied To LoRaWAN Network for Improving Fingerprint Localization Accuracy in Dense Urban Areas

Andrea Piroddi [1,†,‡], Maurizio Torregiani [2,‡]

[1] Università di Bologna; andrea.piroddi@unibo.it
[2] Telebit S.p.A.;
* Correspondence: andrea.piroddi@unibo.it;
† Current address: Università di Bologna.
‡ These authors contributed equally to this work.

Abstract: In the field of low power wireless networks, one of the techniques on which many researchers are putting their efforts is related to positioning methodologies such as fingerprinting in dense urban areas. This paper presents an experimental study aimed at quantifying the mean location estimation error in densely urbanized areas.Using a dataset made available by the University of Antwerp, a neural network was implemented with the aim of providing the position of the end-devices. In this way it was possible to measure the mean location estimation error in an area with high urban density. The results obtained show an accuracy in the localization of the end-device of less than 150 meters.This result would make it possible to use the fingerprint instead of alternative, energy consuming, methodologies such as GPS in IoT (Internet of Things) applications where battery life is the primary requirement to be met.

Keywords: IoT; localization; LoRaWAN; Deep Learning

1. Introduction

The growing interest of the telecommunications market for IoT technologies is driving research to develop different Low Power Wide Area Network (LPWAN) standards. Colloquially speaking, an LPWAN is supposed to be to the IoT what WiFi was to consumer networking: offering radio coverage over a (very) large area by way of base stations and adapting transmission rates, transmission power, modulation, duty cycles, etc., such that end-devices incur a very low energy consumption due to their being connected [13]. Ultra-low power consumption as well as ubiquitous outdoor and indoor connectivity are fundamental aspects to ensure that the network of IoT devices is reliable over the years. To ensure smooth operations on IoT network, it is necessary to take into account various elements such as network topology, modulation techniques, complexity of the hardware, the use of the radio spectrum and regulations [10]. From this point of view it follows that context-awareness is a key element in IoT applications. In order to set-up context-awareness, the location of the device must be identified with minimal location error. The simplest way to achieve this would be to use the GPS tracker. Unfortunately, however, the GPS receiver can consume up to 50 mA when detecting the position [6]. Added to this is the fact that once the position has been obtained it is necessary to transmit it to the gateway and this further step produces an additional energy consumption. A further element to consider is the high accuracy of a GPS measurement, an accuracy that is often not required in an IoT sensor network. So in the face of a high energy consumption we would have an excessively detailed measure in the context under analysis. A particularly interesting technique has been developed in [12], in which a simplified implementation of interferometry is presented obtaining high accuracies. This technique does not require additional hardware, but it cannot be implemented with all communication devices as it strongly depends on the

 doi:10.20944/preprints202211.0130.v1

modulation mechanisms used and the freedoms left to users. Hence, wireless localization based on LPWAN communication is an interesting alternative for localization in low power networks. These techniques estimate the position of a transceiver by analyzing the physical properties of the transmission link such as the value of the received signal strength (RSS), information on the packet time of arrival, etc ... This paper aims to demonstrate that by applying deep learning methodologies to fingerprint techniques it is possible to obtain interesting results in terms of minimizing the localization error. The performance of the fingerprint-based methods depends on the number of the reference points (RPs) in unit space. However, as RSS measurement is time-consuming and laborious, an increase in the number of RPs will increase positioning costs [8]. The remainder of the paper is structured as follows. Section 2 describes the LoRaWan standard used to collect the dataset. Section 3 shows how the dataset has been built. Section 4 illustrates the machine learning approach that we used to estimate the location of end-devices. Section 5 does show the results of our technique. In Section 6 we discuss those results. Finally, Section 7 shows the conclusions and the intended future work.

2. LoRaWAN Standard

LoRaWAN technology provides a two-way communication, but the transmission from node (also known as motes) to gateway (also known as concentrator or base station) or Uplink message is the most frequent one compared to that from gateway to node or Downlink, since usually the purpose of the nodes is to collect data and then send them to the Network Server and then to the Application Server.

The nodes send Uplink messages to the gateways in radio frequency through LoRaWAN modulation. Gateways forward messages to the Network Server by adding information regarding the quality of communication through an IP connection routed over Ethernet, Wi-Fi or 3/4/5G.

The nodes send messages in Uplink to all gateways in their transmission range in broadcast mode, the Network Server takes care of the management of duplicate Uplink messages and the selection of the best gateway to use if a Downlink message is to be sent to the node .

The Network Server also manages the transmission speed of the nodes through the ADR (Adaptive Data Rate) mechanism to maximize the network capacity and extend the battery life of the node. For example, the TTN Network Server uses the 20 most recent Uplink messages, starting from the moment the ADR bit is set, to determine the optimal Data Rate, these measurements contain the frame counter, the signal-to-noise ratio (SNR) and the number of gateways that received each Uplink message.

The Application Server instead takes care of receiving and analyzing the data sent by the nodes and determining the actions that must be performed by the nodes.

LoRaWAN is based on Chirp Spread Spectrum (CSS) technology, where chirps (also known as symbols) are the data carrier. The spread factor controls the frequency of the chirp and thus controls the data transmission rate. Lower spreading factors mean faster chirps and therefore higher data rates. For each increase in the spreading factor, the sweep rate of the chirp is halved and thus the data transmission rate is halved. [2]. LoRaWAN uses the ISM (Industrial, Scientific and Medical) frequency bands reserved for non-commercial radio communication applications, but for industrial, scientific and medical use. In particular, depending on the geographic area and related regulations, the two most common frequencies are 868 MHz in Europe and 915 MHz in North America. Figure 1 shows a picture of the LoRaWAN network architecture.

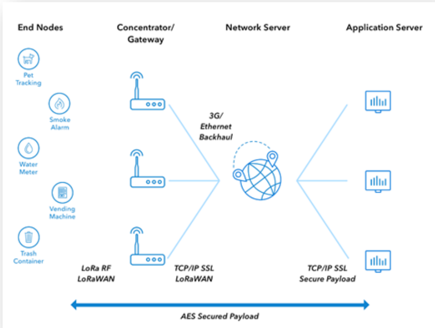

Figure 1. LoRaWAN network architecture

3. Dataset Analysis

In the period between November 2017 and February 2018, the dataset on which our study is based was collected by the Faculty of Applied Engineering of the University of Antwerp [1]. Hardware consisting of a GPS receiver and LoRaWAN end-device was installed on about twenty cars from the Antwerp postal service. While the 20 cars drove around in the city center, the location information was sent in a LoRaWAN message. On the LoRaWAN backend server a callback function was configured to forward the payload of each message, with additional network information attached, to the local data server. In the dense urban area explored, 72 LoRaWAN gateways were detected. Figure 2 shows a picture of the Antwerp Urban Area and message locations.

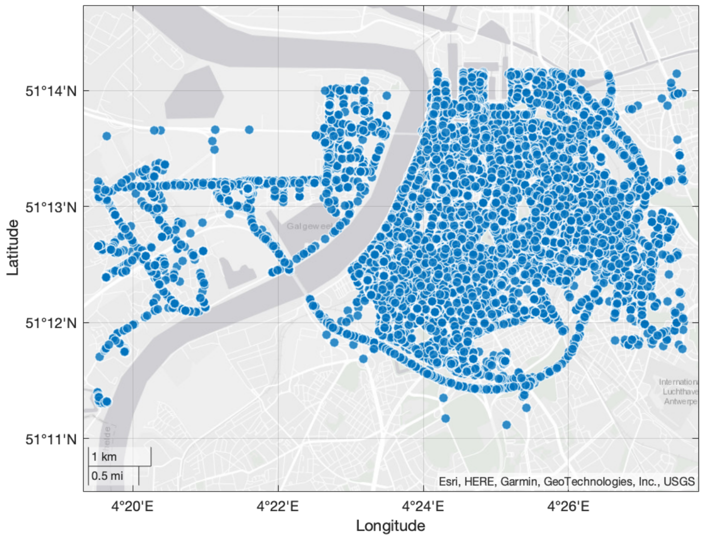

Figure 2. LoRaWAN Dataset collected in the City Center of Antwerp

Table 1. The Structure of the urban LoRaWAN dataset. Each row is a LoRaWAN message showing its receiving base stations (BS) with the RSSI value, the receiving time of the message (RX Time), the LoRa Spreading Factor, latitude and longitude of the transmitter at transmission time.

BS 1	BS 2	...	BS 72	RX Time	SF	Latitude	Longitude
-200	-200	-200	-200	"2019-01"	8	51.23399...	4.42610...
-200	-118	-200	-97	"2019-01"	7	51.20718...	4.40368...
...	...	...	...	"..."	...	...	...

 doi:10.20944/preprints202211.0130.v1

The urban dataset contains 130,429 messages and is available at [2]. All rows represent one of the 130,429 messages in the urban dataset, the last three columns represent the receiving time, latitude and longitude of a message. The columns before indicate which of the 72 base stations in the urban area have received the message. If a base station has not received the message, an RSSI of -200 dB is inserted in the cell. Received Signal Strength Indicator (RSSI) is basically a measurement of the power present in a received radio signal. This does not require additional bandwidth, energy or hardware. These features of RSS measurements make it relatively inexpensive, simple to implement and make this technique appealing [14]. In previous works basic kNN fingerprinting localization technique was used. A parameter sweep has been analyzed, varying k from 1 to 15. Considering that the optimal value of k was the one which produces the lowest mean location error, it emerged that optimal value for k was 11 nearest neighbors. Applying this value of k the LoRaWAN dataset returned a mean error of 398.4 m and a median error of 273.03 m.

By reproducing the same approach, it is possible to find the matrix of the centroids locations. Plotting the centroids on a map, Figure 3, it emerges that the resolution achievable with this technique is exactly the one indicated above. Calculating the distance of the centroids in the most densely urbanized area, it can be seen, in fact, that the values are around 796 meters. This result confirms what we have seen so far, namely that in this densely urban context, using the kNN resolution technique, resolution does not drop below 398 meters.

Is it possible to obtain a mean location error lower than that obtained with the kNN technique, using the dataset as input in a neural network for deep learning feature data classification? With the aim of answering this question, we started with a detailed analysis of the densely urbanized area. This area, in the specific case of Antwerp, is contained in a rectangle whose vertices have coordinates between [4° 20′ East, 4° 27′ East] and [51° 11′ North, 51° 15′ North]. We have therefore divided the area subtended by the rectangle into sub-areas. Within each of these sub-areas we can place a subset of the original dataset.

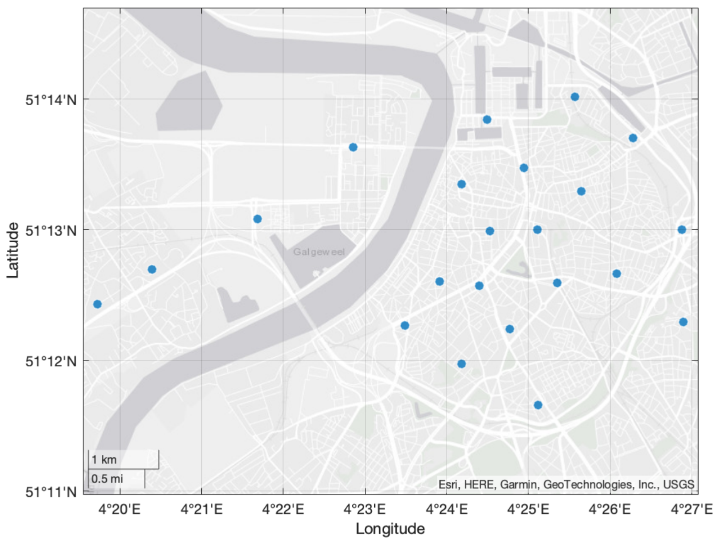

Figure 3. Geometrical Medians of Clusters obtained with kNN Technique

This allows us to reconstruct the dataset with a further column whose information is related to the sub-area which our message belongs to. For each of these sub-areas we have used the function for distance-based clustering of a set of XY coordinates [7]. This function finds clusters in a set of spatial points expressed in XY coordinates. The clustering is based on the distance between the points and it does not require the number of clusters to be known beforehand. Each point is clustered with the closest neighbouring point if the distance between the two points is shorter than the user-defined threshold, that we fixed to

150 meters. The function outputs basic summary statistics (number of clusters, minimum, maximum and average cluster sizes, etc.) on the screen and figures showing the clusters, the centroid points and the geometric median points of each clusters. It also creates two text files that contain the coordinates of all centroid points and geometric median points. The output variables return, for every cluster, the XY coordinates of the centroid and of the geometric median point, as well as the XY coordinates of every point that form the cluster. See an example in Tab 2

Table 2. Output of the distance-based clustering function in the central area of Antwerp Lat $\in [4°23'00'', 4°24'00''] - Long \in [51°12'00'', 51°13'00'']$.

Number of clusters	Size of smallest cluster	Size of largest cluster	Mean cluster size	Median cluster size	Number of points not part of any cluster
13	14	880	212	78	10

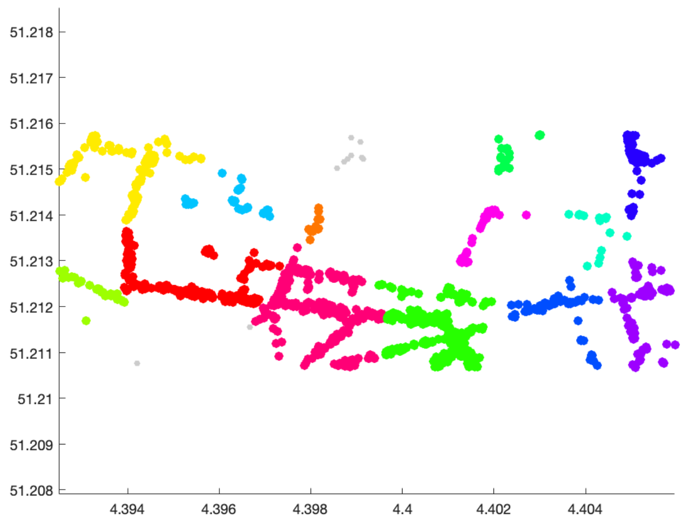

Figure 4. Number of clusters in the central Area of Antwerp Lat $\in [4°23'00'', 4°24'00''] - Long \in [51°12'00'', 51°13'00'']$

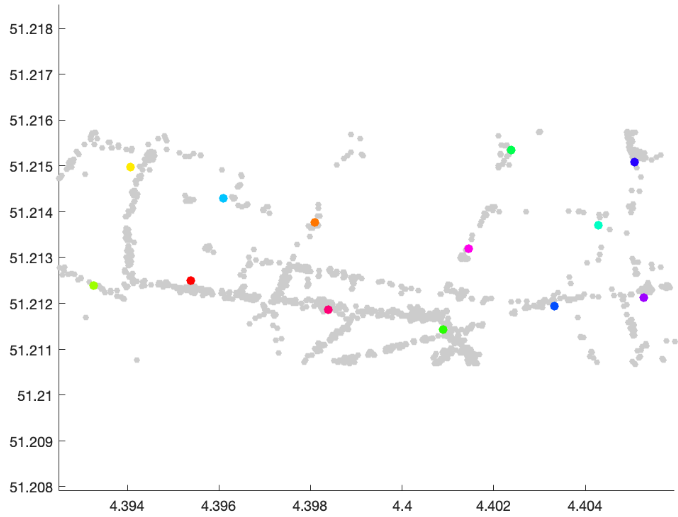

Figure 5. Centroids of clusters in the central Area of Antwerp Lat $\in [4°23'00'', 4°24'00''] - Long \in [51°12'00'', 51°13'00'']$

4. Machine Learning Approach

Deep neural networks rely on massive and high-quality data to gain satisfying performance. When training a large and complex architecture, data volume and quality are very important, as deeper model usually has a huge set of parameters to be learned and configured. This issue remains true in mobile network applications [11]. Using the MATLAB suite, we then configured, a neural network that would take as input the RSSI data received from each of the 72 base stations, the values of the coordinates (Latitude and Longitude) of the message and the categorical value of the assigned sub-area. This operation has been recursively applied to all the sub-areas. To train a network using categorical features, we had to first convert the categorical features to numeric using the ***convertvars*** function by specifying a string array containing the names of all the categorical input variables. The data set has been partitioned into training, validation, and test set using 15% of the data for validation, and 15% for testing. The neural network was defined with a feature input layer (BS1, BS2,..., BS72, lat, long, subarea), normalizing the data using Z-score procedure. Next, a fully connected layer was included with output size 70 followed by a batch normalization layer and a ReLU layer. For classification, another fully connected layer was specified with output size corresponding to the number of classes, followed by a softmax layer and a classification layer.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2022 doi:10.20944/preprints202211.0130.v1

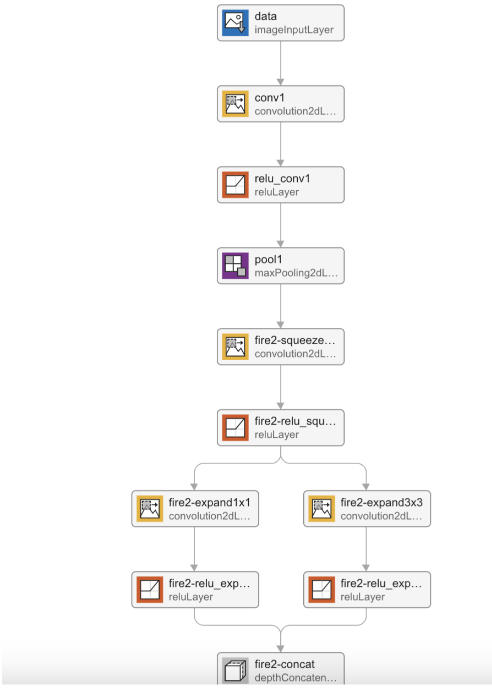

Figure 6. Architecture of the Neural Network

In Figure 6 is showed a piece of the architecture of the Neural Network used in this work.

Table 3. Neural Network Architecture and Training Options.

```
layers = [
    featureInputLayer(numFeatures,'Normalization', 'zscore')
    fullyConnectedLayer(70)
    batchNormalizationLayer
    reluLayer
    fullyConnectedLayer(numClasses)
    softmaxLayer
    classificationLayer];

miniBatchSize = 20;
options = trainingOptions('adam', ...
    'InitialLearnRate',0.0007,...
    'MiniBatchSize',miniBatchSize, ...
    'Shuffle','every-epoch', ...
    'ValidationData',tblValidation, ...
    'Plots','training-progress', ...
    'Verbose',false);
```

The software trained the network on the training data and calculated the accuracy on the validation data at regular intervals during training.

5. Results

The results obtained do show that it is possible to reach an average localization error lower than that obtained with the kNN technique. In fact, we got an error of less than 150 meters. The keystone is linked to the ratio between the number of clusters and the mean cluster size.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2022 doi:10.20944/preprints202211.0130.v1

$$\lambda = \frac{number\_of\_clusters}{mean\_cluster\_size} \tag{1}$$

If this ratio, named as "crowding index", is greater than 10%, the accuracy is not guaranteed because the number of clusters is too high compared to the mean cluster size. On the other hand, when the value is less than 10% we find a very high accuracy. Accuracy is the ratio of correct predictions and the total number of classes. That is:

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalseNegative + FalsePositive} \tag{2}$$

5.1. Low Accuracy

When λ , the crowding index, is greater than 15% the poor accuracy in the localization measurement emerges. The results are distributed as the following example:

1. Number of clusters: 4
2. Size of smallest cluster: 13
3. Size of largest cluster: 123
4. Mean cluster size: 50.500000
5. Median cluster size: 33
6. Number of points that are not part of any cluster: 9

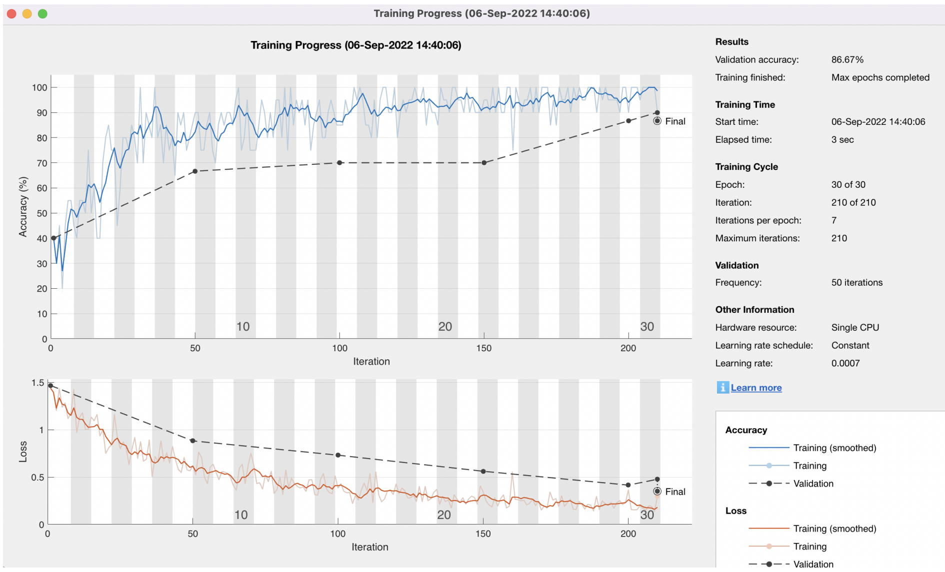

Figure 7. Training Process Low Accuracy

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2022 doi:10.20944/preprints202211.0130.v1

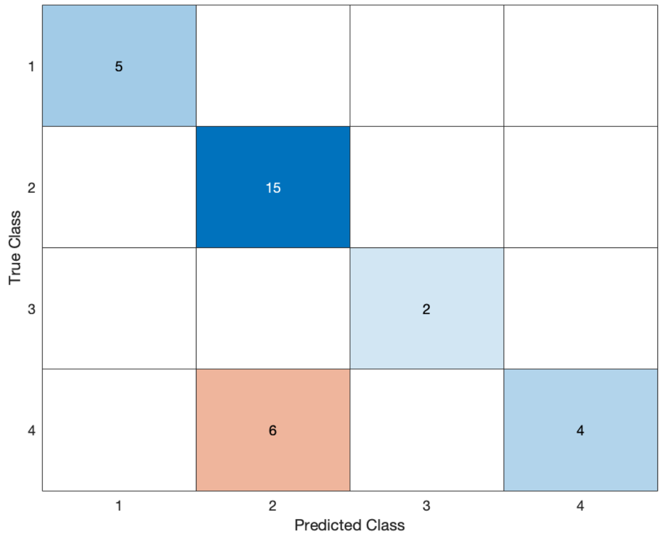

Figure 8. Confusion Matrix Low Accuracy

In Figure 7 and Figure 8 we can see respectively the behavior of the training process and the confusion matrix in case of $\lambda = 20\%$. The Confusion Matrix Table briefly describes the predicted outcome for the classification problem. In this case it shows that it predicts 86% of data correctly and 14% of data was miss labeled in the validation data set.

5.2. High Accuracy

In the case of high accuracy in the localization measurement the results are distributed as the following example:

1. Number of clusters: 15
2. Size of smallest cluster: 16
3. Size of largest cluster: 1175
4. Mean cluster size: 359
5. Median cluster size: 350
6. Number of points that are not part of any cluster: 6

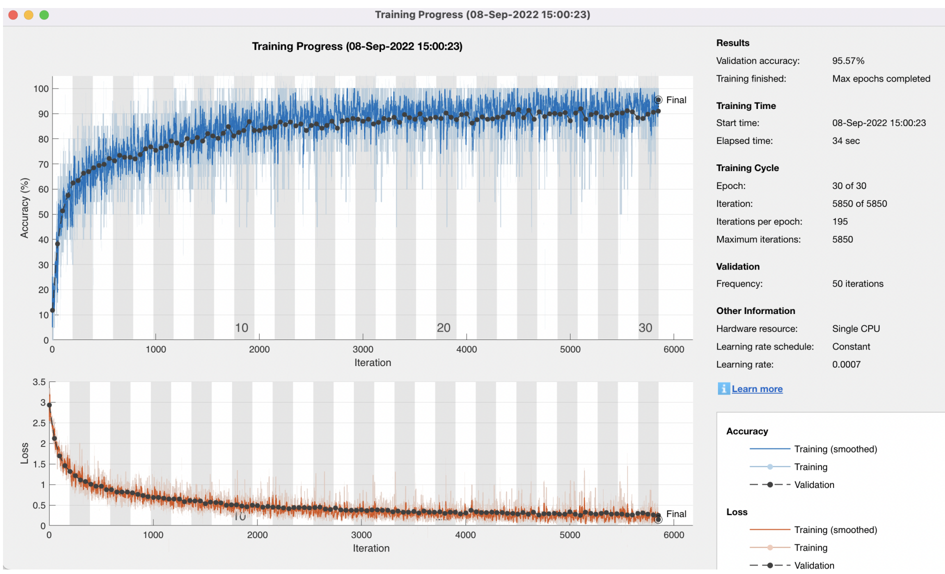

Figure 9. Training Process High Accuracy

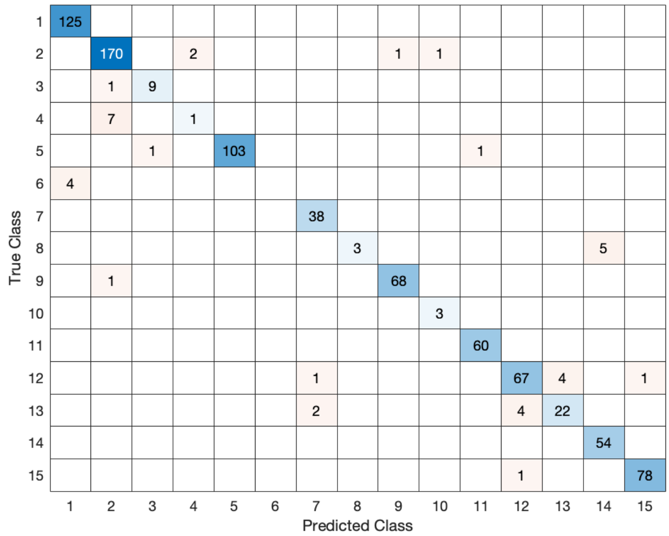

Figure 10. Confusion Matrix High Accuracy

In this case, Figure 9 and Figure 10 show that the neural network predicts 95% of data correctly and 5% of data was miss labeled in the validation data set.

6. Discussion

Literature did show that the first results of basic fingerprinting implementation indicate a mean location estimation error of 398.40 m for the urban LoRaWAN dataset using a standard kNN algorithm. The purpose of this work was to verify whether a lower location estimation error level could be achieved with a machine learning approach. The results of this work show that it is possible to achieve greater accuracy as long as the lambda ratio has values lower than 0.15, that is the ratio between the number of clusters and the mean cluster size (crowding index) is less that 15%. The average position estimation error obtained with the machine learning approach is less than 150 m, and this is an important milestone for the increasing relevance of the Internet of Things and location-based services.

7. Conclusions

Localization algorithms with an estimation error of this order could be suitable for many applications, e.g., a self-driving car company could classify if their vehicles are on their parking site or on their way. A better localization allows to guarantee safety and timeliness in case of need. Future work will be focused on verifying the degree of accuracy it is possible to achieve using the optimized neural network.

Author Contributions: "Conceptualization, A.P. and M.T.; methodology, A.P.; software, A.P.; validation, M.T.; formal analysis, A.P.; investigation, M.T.; writing—original draft preparation, A.P.; writing—review and editing, M.T. All authors have read and agreed to the published version of the manuscript."

Funding: "This research received no external funding".

Institutional Review Board Statement: "Not applicable".

Data Availability Statement: In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Please refer to suggested Data Availability Statements in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics. If the study did not report any data, you might add "Not applicable" here.

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

 doi:10.20944/preprints202211.0130.v1

Sample Availability: The Mathlab Code is available at the following repository: https://github.com/apirodd/LoRaWAN .

References

1. Aernouts, M; Berkvens, R; Vlaenderen, K.V.; Weyn, M.; Sigfox and LoRaWAN Datasets for Fingerprint Localization in Large Urban and Rural Areas. *Data* **2018**, *15*, 1–15.
2. LoRaWAN Dataset Antwerp "https://zenodo.org/record/3342253/files/lorawan_antwerp_2019_dataset.csv?download=1"
3. Anagnostopoulos, G.G.; Kalousis, A. Can I Trust This Location Estimate? Reproducibly Benchmarking the Methods of Dynamic Accuracy Estimation of Localization. Sensors 2022, 22, 1088. https://doi.org/10.3390/s22031088
4. Margelis, G.; Piechocki, R.; Kaleshi, D.; Thomas, P. Low Throughput Networks for the IoT: Lessons learned from industrial implementations. In Proceedings of the IEEE World Forum on Internet of Things (WF-IoT), Milan, Italy, 14–16 December 2015; pp. 181–186.
5. LoRaWANTM 1.1 Specification. Available online: "https://lora-alliance.org/wp-content/uploads/2020/11/lorawantm_specification_-v1.1.pdf" (accessed on 26 Sept 2022).
6. Fargas, B.C.; Petersen, M.N. GPS-free geolocation using LoRa in low-power WANs. In Proceedings of the 2017 Global Internet of Things Summit (GIoTS), Geneva, Switzerland, 6–9 June 2017; pp. 1–6.
7. Yann Marcon (2022). Distance-based clustering of a set of XY coordinates ("https://www.mathworks.com/matlabcentral/fileexchange/56150-distance-based-clustering-of-a-set-of-xy-coordinates"), MATLAB Central File Exchange. Retrieved October 26, 2022.
8. Yu, Xm., Wang, Hq. Wu, Jq. A method of fingerprint indoor localization based on received signal strength difference by using compressive sensing. J Wireless Com Network 2020, 72 (2020). https://doi.org/10.1186/s13638-020-01683-8
9. Thrane, J., Sliwa, B., Wietfeld, C., amp; Christiansen, H. L. (2020). Deep learning-based signal strength prediction using geographical images and expert knowledge. GLOBECOM 2020 - 2020 IEEE Global Communications Conference. https://doi.org/10.1109/globecom42002.2020.9322089
10. Piroddi, A. L. Torregiani, M. (2021). Combining Q-Learning and Multi-Layer Perceptron Models on Wireless Channel Quality Prediction. American Journal of Engineering and Applied Sciences, 14(1), 139-151. https://doi.org/10.3844/ajeassp.2021.139.151
11. Zhang, C., Patras, P., amp; Haddadi, H. (2019). Deep Learning in Mobile and wireless networking: A survey. IEEE Communications Surveys amp; Tutorials, 21(3), 2224–2287. https://doi.org/10.1109/comst.2019.2904897
12. Maróti, Miklós, Péter Völgyesi, Sebestyen Dóra, Branislav Kusy, András Nádas, Ákos Lédeczi, Gyorgy Balogh and Károly Molnár. “Radio interferometric geolocation.” SenSys ’05 (2005).
13. Augustin, A., Yi, J., Clausen, T., amp; Townsley, W. (2016). A study of lora: Long range amp; low power networks for the internet of things. Sensors, 16(9), 1466. https://doi.org/10.3390/s16091466
14. Kannan Srinivasan and Philip Levis: RSSI Is Under-Appreciated, Proceedings of the Third Workshop on Embedded Networked Sensors (EmNets), 2006.