Using Machine Learning to Identify Outliers in Indoor Localization and the IoT

Kiran Kumain,
Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Abstract
Millions of gadgets are connected to create the "internet of things" (IoT), which offers intelligent services. One of the most worrying aspects of smart cities, the internet of things, including wireless sensor networks is indoor localisation. By combining supervised, unsupervised, along with ensemble machine learning techniques, one may analyse RSSs to understand the Wi-Fi indoor localization environment. Wi-Fi indoor localization is the format of the input dataset. The dataset is in.csv file format. In order to better represent data in the form of rows and columns, data has been moved from text files to.csv files during the pre-processing stage.Seven columns in the.csv file, labelled rss1 through rss7, contained RSS data. The class name 'y' was used for the header for these columns. In this study, supervised learning is carried out using Naives Bayes, SVM, and K-Nearest Neighbour. We have employed the IForest (Isolation Forest) method for unsupervised learning. The accuracy, recall, and f1-score are demonstrated by the experimental findings. The technique has demonstrated great efficacy and accuracy. The major goal is to use machine learning to examine the accuracy of indoor Wi-Fi localisation. In this case, supervised learning is done using Naives Bayes, SVMs, and K-Nearest Neighbour. We have employed the IForest (Isolation Forest) method for unsupervised learning. We can draw precise visual representations of such outlier detection using machine learning algorithms.

Keywords: Internet of Things (IoT), Outlier Detection, Machine Learning, K-Nearest Neighbour, Indoor Localization.
DOI Number: 10.48047/nq.2021.19.6.NQ21094 NeuroQuantology2021;19(6): 242-247

1. INTRODUCTION

Using smart devices, indoor localization is a technique by which nodes of a network determine their position in interior environments, including smart homes. Outlier, usually spelt anomaly, is a phrase that originated in the statistical community. Outliers can arise as a result of environmental deviance, human mistake, machine error, mechanical issues, changes in system behaviour, and more. In contrast, outliers are irregular and distinctive RSS values that are brought on by Wi-Fi interior localization situations. For Wi-Fi-based indoor localization environments, the "iF_Ensemble" outlier identification algorithm combines supervised, unsupervised, as well as ensemble learning techniques.

This study suggested using an unsupervised learning technique called isolation forest (I Forest) the data will initially be split into two groups, normal and abnormal, in order to identify outliers in the RSS data. Support vector machine also K-Nearest Neighbour machine learning classifiers are then used. As a way to conduct indoor localization testing, the dataset was gathered by analysing signal levels of seven Wi-Fi routers utilizing a smartphone along with recorded in a text file. In the literature, the procedure of determining the physical location of nodes in a wireless network has been referred to as localization, placement, geolocation, as well as self-regulating. Since localization is the most often used phrase, it is used here.

We may divide the nodes in a wireless network into three groups: anchor, un-localized, and localised. The initial set of nodes, often known as the anchors, are aware of their location or coordinates. The second group of nodes are referred to as un-localized since they are unaware of their location. The nodes in the third group, which are referred to as localised, were once members of the second group but had their locations later assessed. The position of nodes in a wireless network are capable of tracking for a variety of beneficial purposes, including cluster creation, load along with traffic

management, node lifespan control, as well as routing enhancement. The localization issue has a wide range of facets, including when and how frequently localization should be done.

All nodes should be initially localised when the network starts up. However, if there are movable nodes in the network, for instance, this could need to be performed on a regular basis. We now rely heavily on location-based services in our daily lives. However, in order to preserve the limited mobile phone battery supply, such services need ongoing user surveillance. a system that offers precise and cost-effective outdoor localisation using common mobile phone sensors. A WIRELESS sensor network is composed of several cheap, tiny sensor nodes that are dispersed across a vast area, and one or maybe more powerful sink nodes that get information via the sensor nodes. The sensor nodes include built-in computing, wireless connectivity, and sensing capabilities.

The typical configuration for each node includes a wireless radio transceiver, a tiny microprocessor, a power supply, and many types of sensors, including those for temperature, moisture, heat, pressure, sound, vibration, and others. The WSN is used to both detect time-sensitive events and deliver info that is precise and current regarding the physical environment. WSNs have applications throughout a broad spectrum in the personal, industrial, business, as well as military domains, including monitoring of the environment and habitats, tracking of objects also inventories, monitoring of one's health and well-being, Among other things, industrial safety and control as well as battlefield observation. Real-time data mining of sensor data to quickly reach wise conclusions is crucial in many of these applications.

In Wireless Sensor Networks (WSNs), outlier identification has been extensively studied and used in a wide range of applications. It provides data dependability, event reporting, and network security. Outlier detection regulates the accuracy of measured data, enhances the resilience of data processing in the face of noise and malfunctioning sensors, and offers a quick and effective technique to look for values that deviate from the typical sensor data pattern. GPS, RFID, Wi-Fi, and UWB are examples of radio positioning, whereas video cameras, infrared, ultrasound, and inertial systems are examples of non-radio location technologies. The fundamental issue with each sensor's signal dependability needs to be solved in an effort to advance the overall reliability and accuracy of localisation.

The article suggests an outlier identification method to deal with shaky measurement data. An innovative outlier identification approach is presented to deal with outliers in the localization process after an investigation of the localization algorithm utilised in the indoor localization system. Database management uses quality control together using the suggested outlier identification method. In Wireless Sensor Networks (WSNs), outlier identification has been extensively studied and used in a wide range of applications. It provides data dependability, event reporting, and network security. Outlier detection regulates the accuracy of measured data, enhances the resilience of data processing in the face of noise and malfunctioning sensors, and offers a quick and effective technique to look for values that deviate from the typical sensor data pattern. GPS, RFID, Wi-Fi, and UWB are examples of radio positioning, whereas video cameras, infrared, ultrasound, and inertial systems are examples of non-radio location technologies. The fundamental issue with each sensor's signal dependability needs to be solvedin an effort to advance the overall reliability and accuracy of localisation. The article suggests an outlier identification method to deal with shaky measurement data.

2. LITERATURE SURVEY

Measurements that considerably depart from the typical pattern of sensed data are referred to as outliers in the field of wireless sensor networks. Events, malicious network assaults, noise and mistakes are a few potential drivers of outliers.Because of how sensor data is, as well as the unique specifications &constraints of wireless sensor networks, conventional outlier identification approaches are not directly relevant to these systems. This report offers a thorough overview of current outlier identification methods created especially for wireless sensor networks. It also provides a method-based taxonomy as well as a comparison table that may serve as a reference for choosing a methodology appropriate for the given application predicated upon factors like data type, outlier type, outlier identification, as well as outlier degree.Sensor data's spatiotemporal correlations increase the outlier identification process's accuracy and resilience. Data distribution and distributional factors like standard deviation must be understood, which a severe drawback is [1].

Localization of nodes is frequently used in wireless networks. For instance, it is applied to strengthen security also to advance routing. Two categories of localisation algorithms exist: range-free and range-based. Range-based methods calculate the distance between two nodes using location metrics like ToA, TDoA, RSS, alsoAoA. The foundation of range-free algorithms is often proximity detection between nodes. Range-based algorithms are not

only more advanced nevertheless, more accurate, so there is a trade-off. However, localisation precision is crucial in applications like target tracking. In this research, we present a novel range-based approach that is depending upon the data mining technique known as density-based outlier detection (DBOD). The K-nearest neighbours (KNN) must be chosen. Each point utilised in the location estimate receives a density value from DBOD.These densities are averaged, and the spots with densities greater than the mean are preserved as candidate locations. We compare our method to the linear least squares as well as weighted linear least squares according to singular value decomposition algorithms using various performance metrics. It has been demonstrated that the suggested technique outperforms conventional algorithms even in cases when the anchor geometry surrounding an unlocalized node is subpar [2].

The work proposes an outlier identification approach carrying out quality control of the RSS database as well as data filtering in real-time localization with the goal to account for the complicated RF propagation effects with restricted resources in WSN-based indoor localization. The method is demonstrated to be reliable and efficient when the management of data that are prone to abnormalities. According to the experimental findings, adding an outlier detection system can increase localization accuracy by 13–30%. Thus, the localization system and outlier detection technique can open up a variety of WSN applications for automated surveillance, exploration, as well as context awareness. The ease with which a wireless network may be expanded or altered is one of its greatest advantages. It might increase the localization's precision [3].

Actual environment to gather priceless raw felt data. Getting high-level information out of this massive amount of data is the difficult part. However, recognising outliers can help researchers learn important and practical information. Any measurement that differs from the norm is an outlier the typical behaviour of sensed data in the context of wireless sensor networks. Over the past 10 years, several outlier identification strategies in WSNs have been thoroughly investigated, with a particular emphasis on traditional based algorithms. These methods locate anomalies in the dataset of real transactions. The goal of this work is to present a well-organized and thorough review of the current research on classification-based outlier identification methods as they relate to WSNs.Furthermore, this work aims to offer a clearer and more accessible explanation of classification-based strategies.

It was compared to the OCSVM and RE-based methods, which both accurately identify outliers. The test examples being susceptible to classification-based procedures when a significant outlier score is sought for 689 [4].

The proposed approach consists of in two stages: training along with testing. In the training phase, the authors used the BBNs to model the normal behavior of the network and used a maximum likelihood estimation algorithm to learn the parameters of the BBNs. In the testing phase, the authors used the learned BBNs to identify outliers in the sensor data. The outcomes demonstrated that the suggested strategy functioned better existing outlier detection approaches with regard to precision &false positive rate. The suggested approach has shown promising results and has the potential to be applied in other WSN applications [5].

3. PROPOSED SYSTEM

For Wi-Fi indoor localization environments, the priorresearch has created an outlier identification tool called iF_Ensemble by analysing RSSs use a mixture of supervised, unsupervised, as well as ensemble machine learning algorithms. A technique for unsupervised learning is isolation forest (iForest). Support vector machine, K-nearest neighbour, along with random forest classifiers with stacking are examples of supervised learning techniques. Accuracy, precision, recall, F-score, and ROC-AUC curve are utilised for assessment purposes. After removing outliers, the evaluation of the machine learning approach utilised yields a high accuracy of 97.8% using the suggested outlier identification techniques with an increase in accuracy of the localization process in indoor environments of about 2%.

Wi-Fi indoor localization is the format of the input dataset. The dataset is in.csv file format. The pre-processing of data has changed the format of the data from text file to.csv. Seven columns in the.csv file holding RSS data were given the titles rss1 through rss7, and the class label 'y' was used for the heading for the seventh column. In this study, supervised learning is carried out using Naives Bayes, SVM, and K-Nearest Neighbour. We have employed the IForest (Isolation Forest) method for unsupervised learning. Unsupervised learning divides data into normal and pathological categories. The accuracy, recall, and f1-score are demonstrated by the experimental findings. The technique has demonstrated great efficacy and accuracy. The comparison graph between supervised and unsupervised learning is then shown. The following are a few of the suggested system's benefits:

- When compared to existing methods, both supervised and unsupervised learning perform with high accuracy.
- Additionally, it shows visual graphs.

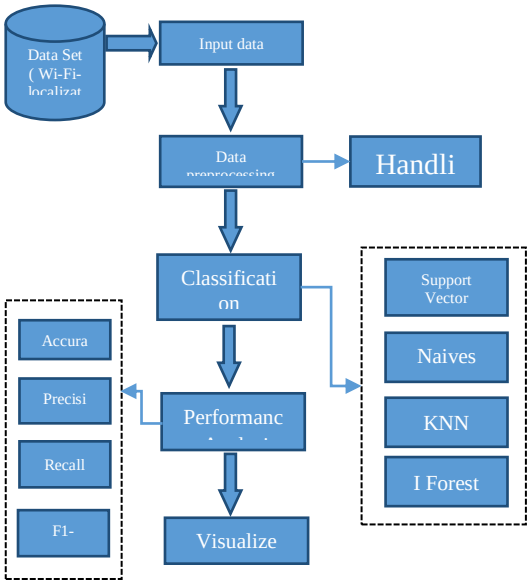

Fig 1: System Architecture

The next section provides an explanation of the many phases that are involved in putting the suggested technique into practise:

1. Data Selection and Loading

The supplied dataset's format is Wi-Fi indoor localization. A text file in the.txt format serves as the dataset. The information was moved from a text file to a.csv file. In order to conduct indoor localization testing, the dataset was gathered by analysing signal levels of seven Wi-Fi routers employing a smartphone that is recorded in a text file. Seven columns and 2000 rows make up the dataset. The eighth column has a class designation of 1 to 4 that designates numerous interior locations.

2. Data Preprocessing

Pre-processing data involves deleting unnecessary information from the dataset. Missing values and Nan values are replaced by 0 throughout this procedure to remove missing data. Data was cleared of any errors and missing values as well as duplicates.

3. Splitting Dataset into Train and Test Data

The process of breaking accessible data into two pieces, Data splitting is frequently used for cross-validator requirements.

A component of the data is used to develop a prediction model, and a different portion of the data is used to determine the model's efficacy. A crucial step in reviewing data mining algorithms is dividing the data into training and testing sets. The majority of the data is often used for training, while a smaller piece of the data is utilised for testing when you divide a data set into a training set and testing set.

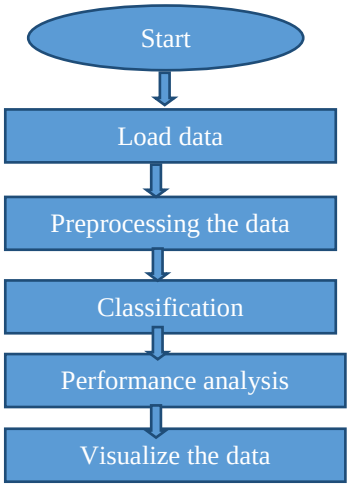

Fig 2; Flow Diagram

4. Classification

Support vector machine (SVM) models are multidimensional hyperplane representations of many classes. K nearest neighbours is a straightforward algorithm that sorts incoming instances according to a similarity metric after storing all of the existing examples. Nave Bayes is a classification method that uses the Bayes Theorem and the premise of predictor independence. An unsupervised learning system for anomaly identification called IForest (Isolation forest) operates on the idea of separating abnormalities. By randomly picking an attribute and then a split value for that property, amid the minimum and maximum values permitted for that attribute, it builds partitions on the sample in a recursive manner.

5. Prediction

This method involves inferring the analysis from the dataset. It refers to the methodical identification, extraction, quantification, as well asresearch of emotional states and subjective data via natural language processing, text analysis, also biometrics. By optimising the total prediction outcomes, this project will successfully forecast the data from the dataset.

4. RESULTS

Applying machine learning techniques as Naives Bayes, Support vector machines, k-nearest

neighbours, and the IForest (Isolation Forest) algorithm, this research intends to examine the accuracy based on the Wi-Fi-indoor localization scenario. The data pre-processing method involves moving the data from a text file to a.csv file when the input dataset is in the form of a.csv file. The following screenshots illustrate the experimental findings showing the procedure is very accurate and efficient. The accuracy outcomes of the user's indoor localization have also been computed before, and outlier identification and outlier reduction are noted. It is argued that this technique for outlier identification in indoor localization as well as IoT can be reliable and helpful.

```
dtype: int64
......SUPPORT VECTOR MACHINE.......

acuracy: 98.08823529411764
.....Naivies Bayes.....

 accuracy: 97.5

......K-NEAREST NEIGHBOUR......

acuracy: 98.38235294117646

......IForest......

Accuracy: 96.0
```

Fig 3: Performance Measure

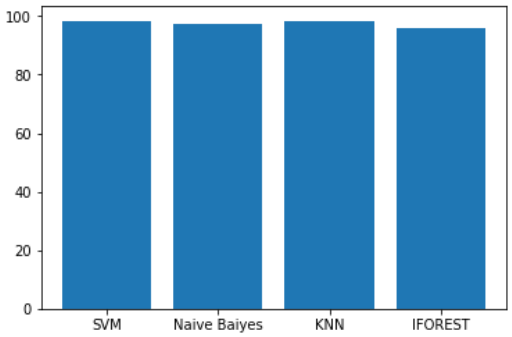

Fig 4: Performance Analysis

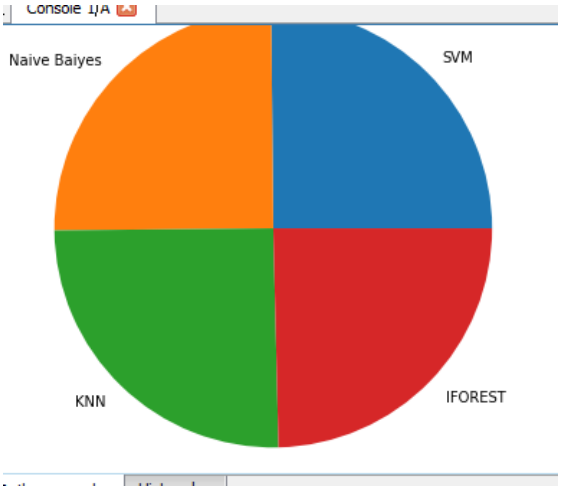

Fig 5: Comparative Analysis

5. CONCLUSION

Indoor localization and IoT location data are crucial for accurate localization in this research. However, the radio signal limitations in a Wi-Fi interior setting alter RSS values and lead to irregularities in RSSs. a Wi-Fi indoor localization environment with the use of machine learning techniques to analyse RSSs. Results of the suggested approach's evaluation utilising accuracy, precision, recall, and F-score. The testing showed that it was highly accurate and effective.The superior The ROC curves displayed for the evaluation also support the outcomes of the proposed technique. Additionally, the accuracy outcomes of the user's indoor localisation have been computedin this study before outlier identification as well as outlier reduction. The localisation accuracy of indoor users has been seen to increase.We may thus draw the conclusion that this can be a practical and trustworthy method for outlier identification in indoor localization and IoT.

6. FUTURE ENHANCEMENT

In localization contexts and wireless sensor networks generally, there aren't many solutions for outlier identification. Future research can look at how artificial intelligence (AI) approaches can be used to discover outliers in localization and WSNs.

REFERENCE

[1] K. K. Almuzaini and A. Gulliver, “Range-based localization in wireless networks using density-based outlier detection,” Wireless Sensor Network, vol. 2, no. 11, p. 807, Nov. 2010.
[2] H. Aly, A. Basalamah, and M. Youssef, “Accurate and energy-efficient GPS-less outdoor localization,” ACM Trans. Spatial Algorithms Syst., vol, 3, no. 2, pp. 1–31, July 2017.
[3] Y. Zhang, N. Meratnia, and P. Havinga, “Outlier detection techniques for wireless sensor

networks: A survey," IEEE Commun. Surveys &Tuts., vol, 12. no. 2, pp. 159–170, 2010.
[4] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," J. Network Computer Applicat., vol. 60, pp. 19–31, 2016.
[5] D. Janakiram, V. A. Reddy, and A. P. Kumar, "Outlier detection in wireless sensor networks using Bayesian belief networks," in Proc. COMSWARE, 2006.
[6] K. Niu, F. Zhao, and X. Qiao, "An outlier detection algorithm in wireless sensor network based on clustering," in Proc. IEEE ICCT, 2013.
[7] X. Wei, L. Wa ng and J. Wan, "A New Localization Technique Based on Network TDOA Information," *Proceedings of IEEE International Conference onITS Telecommunications*, Chengdu, China, June 2006, pp. 127-130
[8] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: a survey," IEEE Communication Surveys & Tutorials, vol. 12, no. 2, 2010.
[9]Y. Shang, W. Ruml, Y. Zhang, and M. Fromherz, "Localization from connectivity in sensor networks," IEEE Transactions on Parallel and Distributed Systems, vol. 15, no. 11, pp. 961–974, 2004.
[10] A. Ayadi et al., "Performance of outlier detection techniques based classification in wireless sensor networks," in IEEE IWCMC, 2017.
[11] K. Niu, F. Zhao, and X. Qiao, "An outlier detection algorithm in wireless sensor network based on clustering," in Proc. IEEE ICCT, 2013.