Anomaly Detection In Wireless Sensor Network

Aditya Tiwari[*],Siddharth Nahar[**]

[*] CS516: Ad-hoc Networking

Abstract- Consider a scenario where multiple IoT sensory devices collect data and send it to a sink IoT device which then later sends it for processing for required output. In recent scenarios it is very likely for the data to be sensed be faulty or be an outlier. Hence we need to detect if the data due to a node is an outlier or not. Current methods have drawbacks as computationally expensive, not dynamic to increase in detectors and etc.Hence we need a Machine Learning based approach that has a good accuracy and is fast.

I. Introduction

Consider the scenario of advance agriculture environment which is automated in terms of irrigation ,harvesting etc. In such a case, it is obvious that the whole system works on the basis of information collected from the environment. For this one deploys sensory IoT devices across field. Collects the data in real time, processes and takes action. For this system to work properly it is critical that the data collected from the nodes are correct and fault-free. But for uncontrollable reasons , these data may be outliers or faulty. The data collected may be an outlier mainly due to three reasons :

1. *Resource limitation*: Sensor nodes have constraints as battery or storage and computational capacity. Consider if the power of the node is exhausted, the probability of erroneous data will grow rapidly.
2. *Harsh environment*: Often nodes deployed are randomly to not have a bias in deployment. However due to harsh environment and potential damage to the sensory device, the data may be incorrect.
3. *Malicious attacks*: If the task at hand is costly , there may be parties that want to destroy the system, and hence an attacker may hijack a node and produce erroneous data.

As most of the data collected based on time series data, it forms a famous problem of anomaly detection in sensor networks. So we have applied Machine learning algorithms and design to detect outliers in data. Our approach was basically to learn logic how to predict anomaly based on data distribution so we need not need to keep tuning model after every time period. Global Learner will learn and will transmit this information to rest of sensor nodes.Our approach doesn't need much time , predicting outliers is local So energy is saved and results show approach works considerably good if data distribution doesn't change drastically.

II. Related Work

The data produced has some pattern and an outlier does not fit the pattern. Hence the task is to find the data points that do not fit the pattern of the data. There are several methods proposed in *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, 2nd Quart., 2010.*

- Statistical-Based

  The model generates a regression ML for the next fixed interval and then predicts the next point and compares the given point with the predicted point. If the difference is high, it is classified as an outlier. The model is two level where first level is at each node level and the second is on higher comparing every node.

- Clustering-Based

  Consider the scenario as unsupervised learning and each node as a cluster. Consider all the data generated by a node , merge them using an algorithm and report it to sink, if the inter-cluster distance is greater than some standard deviation, then the node is classified as producing outlier

- Classification-Based

  It first uses Statistical-Based. Then it uses SVM with a kernel function to map the data points into high dimension space using one class. This reduces the complexity in training and testing phase.However if the training data itself is tainted with outliers, the performance is poor. Hence another research paper proposed a two layer model , first layer was a Bayesian classifier which predicted the probability of a node producing outlier based on previous knowledge of the node.

- Spectral Decomposition Based

  In a hierarchical anomaly detection model in distributed large-scale sensor networks is proposed. It exploits principal component analysis (PCA) to deal with outliers in data generated by the faulty sensor.

- Paper’s Approach

Their framework contains a local detector and a global detector. The local detector is constructed based on the iNNE. The global detector is constructed by combining the local detectors of a node and its neighboring nodes, where the formation of a neighborhood is based on the spatial correlation among sensor nodes.The final detection of measurements is executed by a global detector of a sensor node. The main contributions of this paper are as follows:

1) An isolation-based distributed outlier detection framework using nearest neighbor ensembles is proposed.

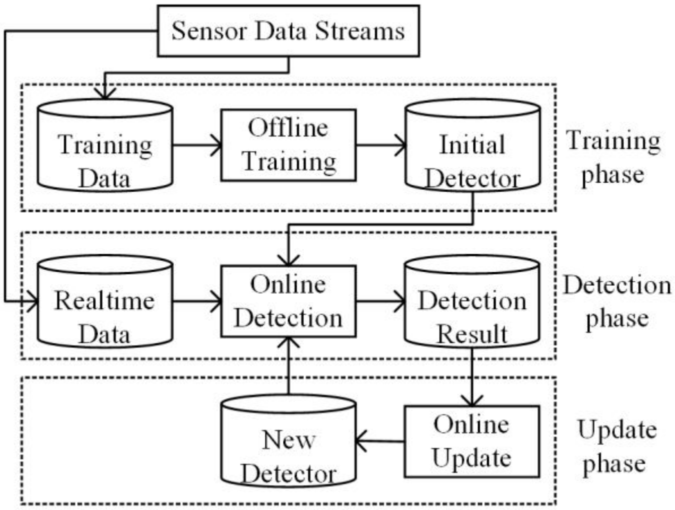
The framework combines the advantages of the isolation-based frameworks and the nearest neighbor-based frameworks. Our framework utilizes the idea of subset ensembles in the isolation-based frameworks to reduce computation burden and memory requirement.Moreover, we calculate outlier scores to identify outliers based on the distances between measurements which is the idea of the nearest neighbor-based frameworks.

2) A new combination method for local detectors based on the weighted voting is introduced. The actual distances among sensor nodes are used as the weights.The benefit of our combination method is the broadcast of the information of local detectors in WSNs is avoided. Only the measurements and the positions of sensor nodes are exchanged within a neighborhood.

3) A self-adaptive algorithm is developed to update the proposed framework. The key idea of the algorithm is a sliding window that keeps track of the dynamic changes of sensor data. Especially, the algorithm eliminates the influence of the biased values caused by dynamic changes in sensor data. Thus, the accuracy of detection is improved.

4) Side figure represents framework used by the paper we followed. It has three phases Training , Detection and Update Phase.

- Training Phase is local training mode, Sensor detect based on its history of outliers.
- Detection phase takes consideration of local neighbourhood data and detects outliers.
- Update Phase update detectors via current distribution of data.



Above Figure Presents the framework.

III. Dataset

We’ll be using ISSNIP dataset

ISSNIP: The Intelligent Sensors, Sensor Networks and Information Processing dataset is a real humidity temperature sensor data is collected using TelsoB motes in a single hop WSNs. This dataset has controlled outliers and all the data are labeled. There are a total of four sensor nodes and the data consists of temperature and humidity measurements over a period of 6 hour.

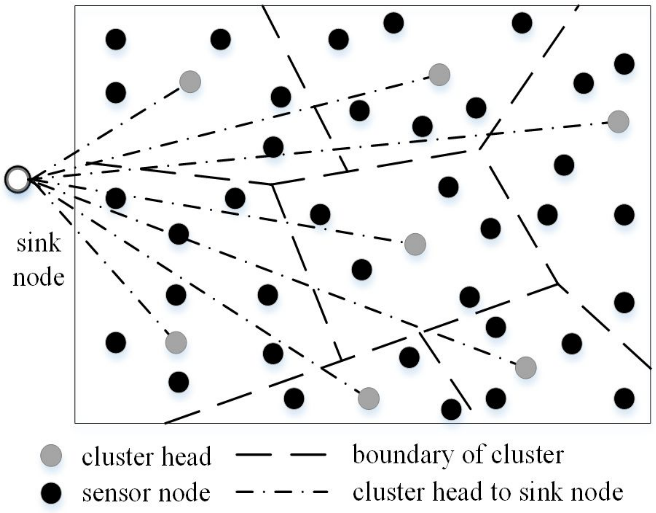
Figures at the end show visualization of the data

There are two datasets- indoor and outdoor both of them are highly biased i.e lesser number of negative data.

DataSet	Total Data	Positive Data	Negative Data	Bias Ratio
Indoor	4417	4300	117	37
Outdoor	5039	5007	32	156

IV. Problem Statement

A wireless sensor network typically consists of a large number of sensor nodes scattered over a region of interest to monitor specific physical phenomena. Sensor nodes may be arranged to various kinds of topologies for different applications.A typical network topology is shown below there are totally seven clusters in the network and clusters are separated



node by the dashed line. There is only one sink node, which is denoted by a black circle. A communication link between a cluster head
node and the sink node is denoted by a dotted line.

A subnetwork is one of the clusters in the whole network, in which nodes can directly communicate with each other. Each node in the sub-network does the same work such as data collection, communication,and outlier detection. We will try to propose a new framework based on the above sub-network to detect outliers for WSNs.

V. Our Approach and Experiments

A. Data Distribution

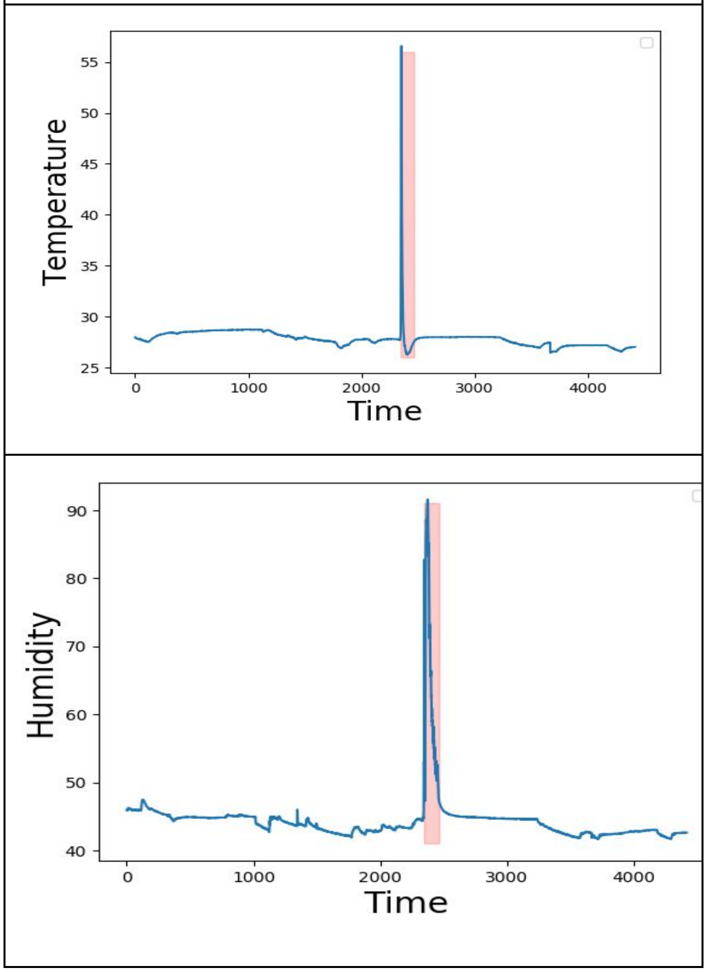
We plot the data to choose the algorithms to classify outlier and valid data. Understanding Data distribution really help what design to choose and how to tune our model. Below we show the distribution of data for both indoor and outdoor datasets. Challenges in Dataset, As we can see Red zone represents
Outlier data, Variance is high and classes are highly biased towards positive labels.

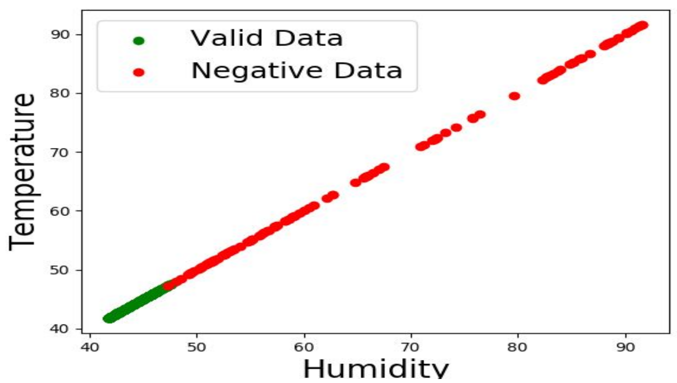
B. SUPPORT VECTOR MACHINES

SVM is a Machine Learning Model that does supervised classification on data. This tries to find a hyper-plane in data’s dimension that has least classification error. We experimented SVM classifier with different kernel.To balance out biased dataset, loss functions for false negatives were weighted with bias Ratio.Results are summarized in Table 1Now we considered what if our model is overfitting i.e It is specific to data. To test the same, we subtracted our indoor dataset of temperature and humidity by say 10, 20 ,50, 100 and Tested again. Since all of the data points are shifted down, the irregularity of notch remains same. We tested the same for each of the kernel for outdoor dataset.Results are summarized in Table 3.

Note that all the models trained on indoor dataset give higher error percentage on outdoor dataset. This is due to higher variance in outdoor data that is outdoor temperature and humidity varies more compared to indoor.

INDOOR DATA



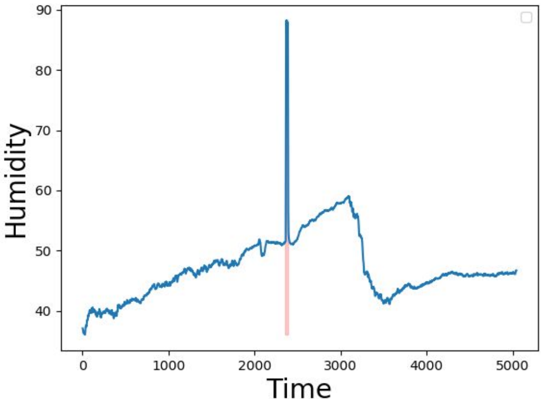


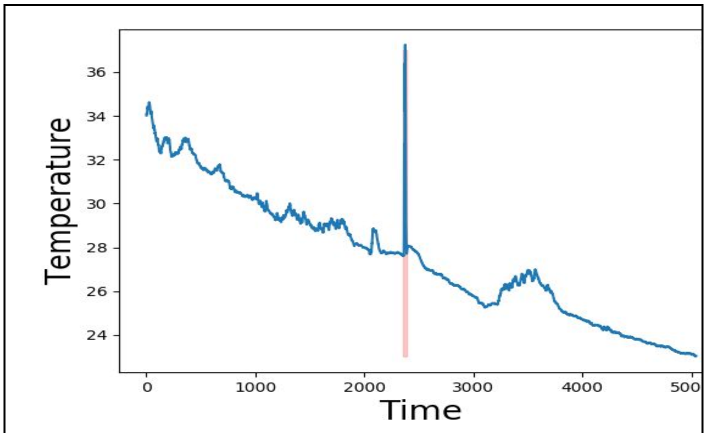
Linear kernel has 0 error for no shift and jumps to 100% with a little shift, which clearly shows overfitting i.e the model hasn’t learnt the logic but has learnt somehow to reduce the error if data is similar to the data it was trained on.

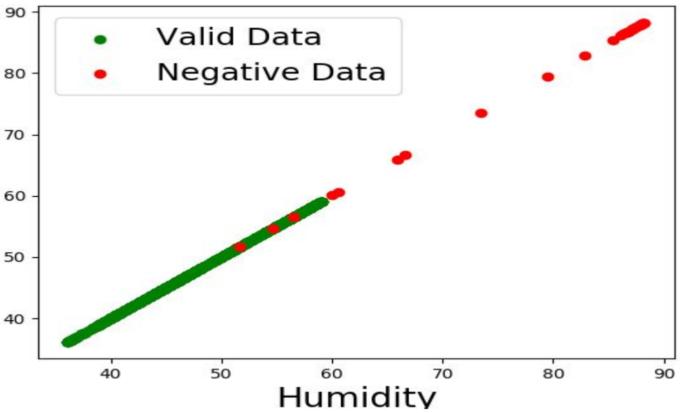
If you notice the graphs of variation of humidity and temperature with time, it has a sudden change in invalid range. So we trained gradient change on outdoor data and checked the error on indoor data and outdoor data. The results are given below.

Linear and Polynomial kernels have outperformed all other kernels with very less error.

OUTDOOR DATA







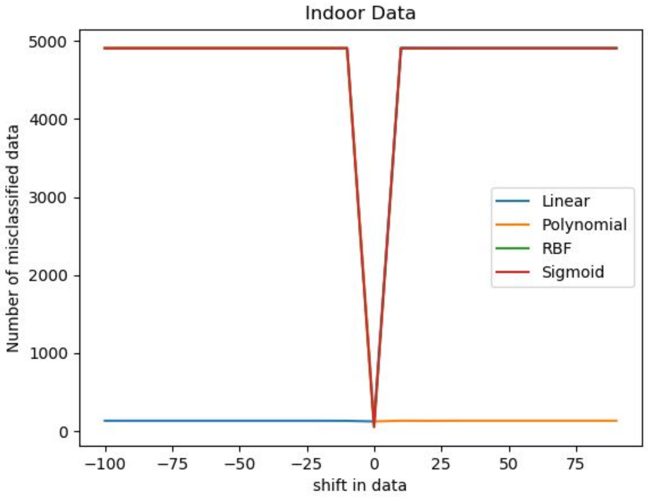
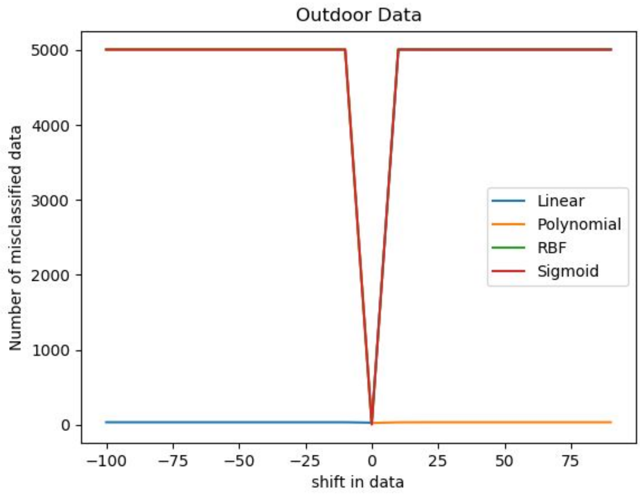
Notice the High Variance in Negative Data

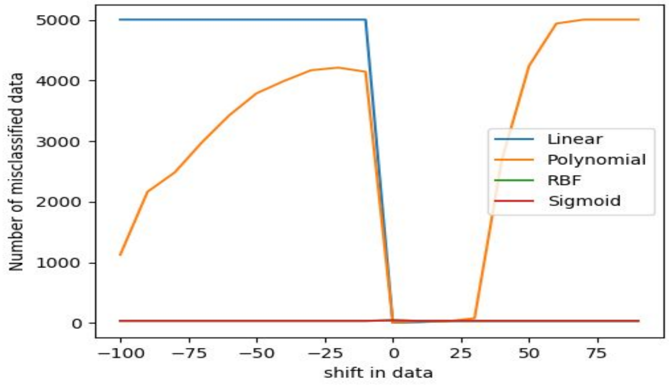
Table 1.

Kernel	**Trained On**	**Indoor Error (%)**	**Outdoor Error(%)**
Linear	Indoor	0.022639800769753225	32.34768803333995
	Outdoor	1.0414308354086486	0.8533439174439373
Polynomial	Indoor	**0.0**	34.29251835681683
	Outdoor	1.6527054561919854	0.13891645167692002
Sigmoid	Indoor	0.022639800769753225	86.62433022425084
	Outdoor	1.8338238623500112	0.8533439174439373
Radial	Indoor	0.022639800769753225	86.62433022425084

Basis Function	Outdoor	1.8338238623500112	0.8533439174439373

Results are as discussed above. Number of misclassified data graph will be shown below.

Results





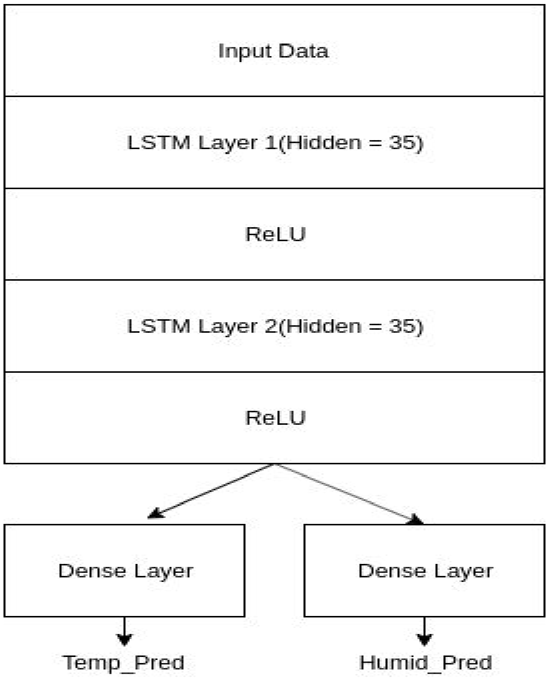
SVM on gradient trained on outdoor data

RBF and LINEAR performs best on data, but SVM doesn’t learns the logic just learns on current data. So we need to keep training on current data distribution. So we switched to LSTM model as discussed below.

C. LONG SHORT TERM MEMORY NETWORK

LSTM is supervised learning algorithm used more often for time-series data. Major change in the idea of using LSTM is to predict future Temp, Humid instead of errorizing on labels. For sensor network outlier detection we have three steps :

1. STEP 1



In this step we need train above model to locally detect outliers. For this step we train our model for Indoor moteid 2 dataset which doesn’t have any anomaly. For LSTM we divided dataset into sequence length of 10 and predicted length of 3. Batch size of 32. I trained for 1000 epochs and saved the best model.

2. STEP 2

Fit gaussian distribution on error of test set. Suppose x_pred and x_true be two vectors.

$$error = x\_true - x\_pred$$
$$\mu = (x\_true - x\_pred)/N$$
$$\sigma = ((error - \mu)^T * (error - \mu))/N$$

σ represent covariance matrix, μ represents mean of loss and will be used in next step.

3. STEP 3

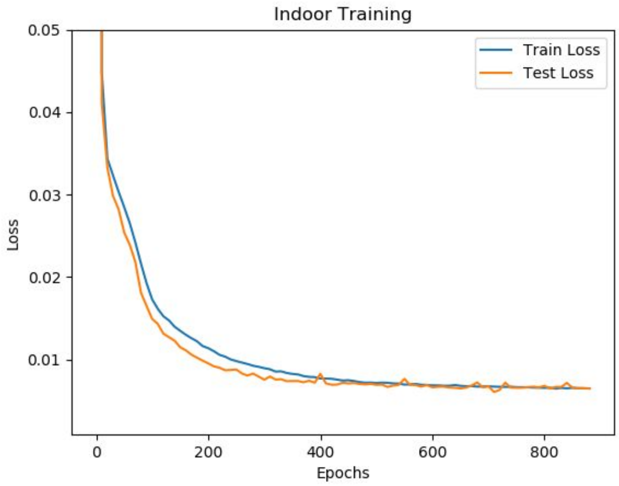
Mahalanobis Distance, We can measure the rarity of the event with the location in the distribution. The Mahalanobis' distance is statistics representing an anomaly score.

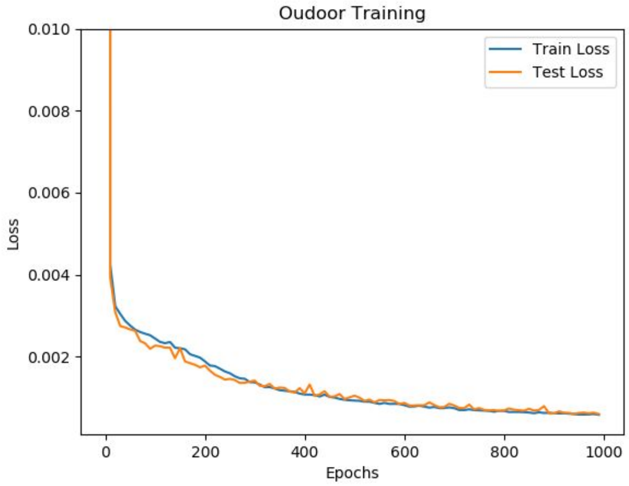
$$p(x/Data) = N(x|\mu, \Sigma)$$

$$a(x) = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

a(x) represents rarity of event known as Mahalanobis distance. Overall Algorithm is a series of above steps.

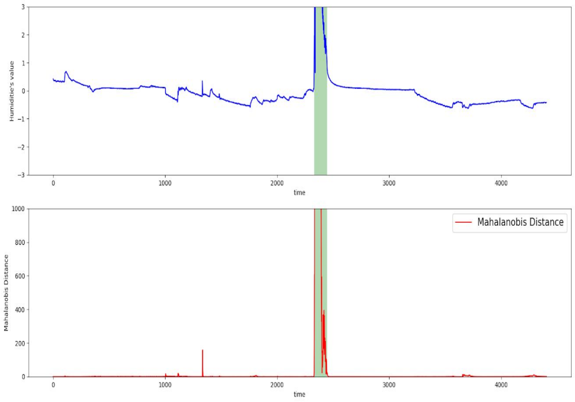
Training LSTM model has been stable and we can see loss function over epochs as below :

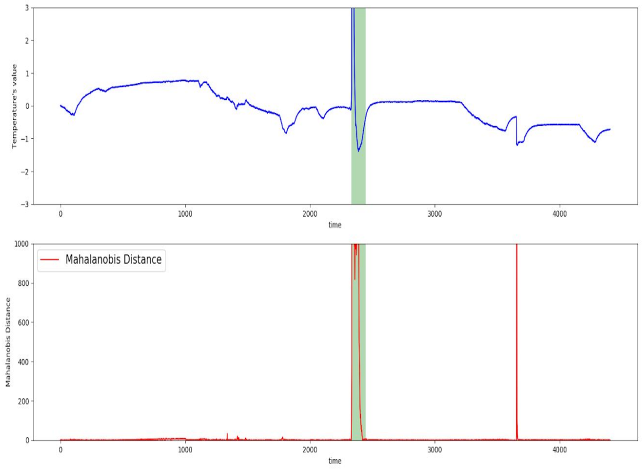




EXPERIMENTS of LSTM model have been very interesting. Training on dataset of moteid 2 and Testing on dataset of moteid 1 We get a very high precision and accuracy. This Experiments concludes LSTM model is learning logic and learning data distribution of Indoor Doors.

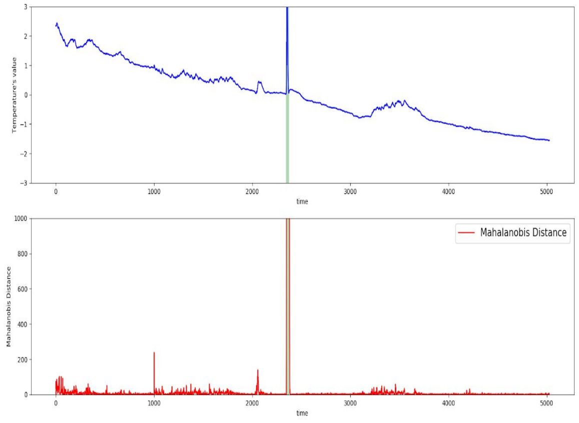
Indoor Test LSTM Results, Red graph shows predicted and green zone sees anomaly.

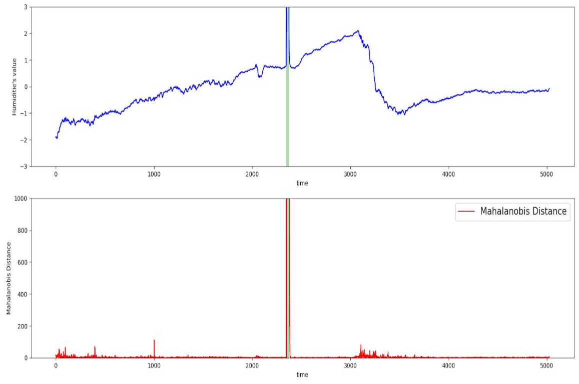




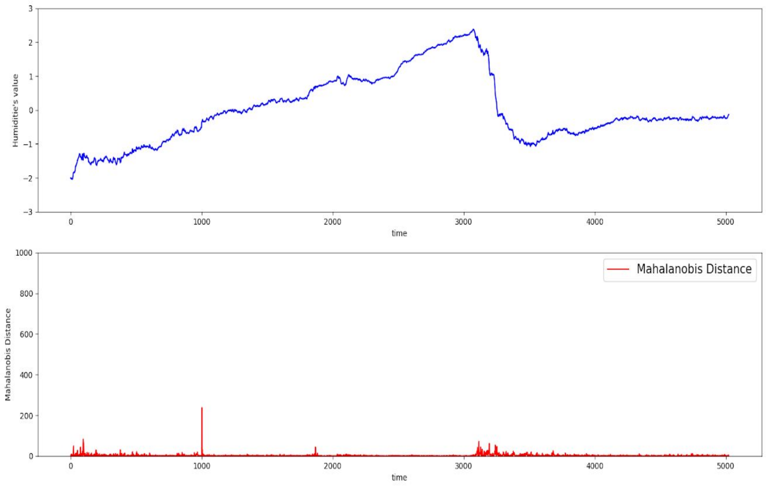
Similar Results are observed for Outdoor data. Precision and Accuracy is very high. Training is done on Outdoor moted id 4 and Tested on mote id 3 dataset.

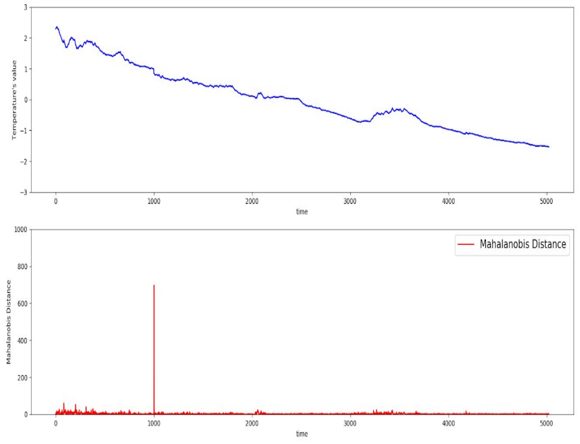
Outdoor Test LSTM Results, Red Graph shows predicted and Green zone shows anomaly.





We can see wrong labeled dataset is correctly classified but what about correct data We show some graphs for both indoor and outdoor dataset for no anomaly datasets.

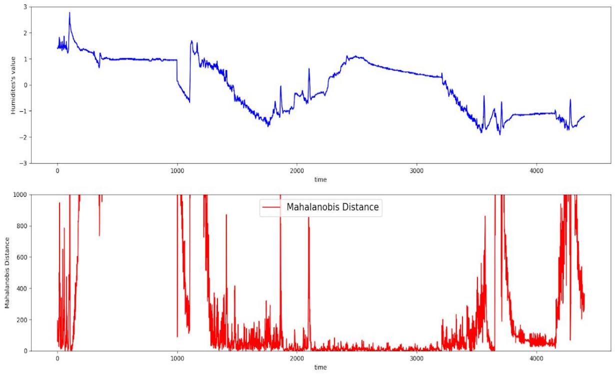


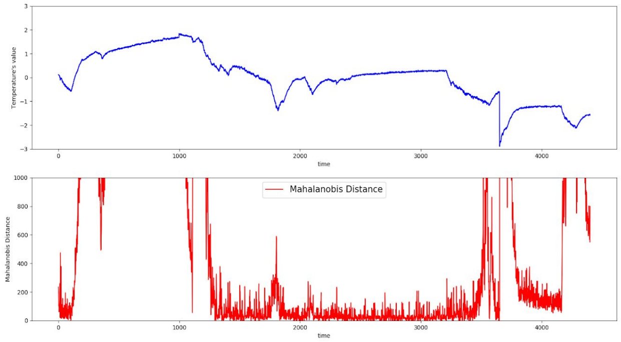


We can see Graphs with red peaks shows anomaly. We can clearly see even if data has high variance our model has predicted them accurately.

We also tried experimenting training model on indoor dataset and predict results for outdoor dataset. But we see results are not good because data are from completely different data distributions. Below we show results for Temp Predicted from training outdoor and testing on indoor.

Results are not good as data distributions vary differently.





VI. CONCLUSIONS

Our results clearly show that we can model data distributions. Our approach learns logic how to predict outliers , no need to keep updating model with current data distribution. If we train model with good variant distribution Results are much better.

Model doesn't predict accurately if data distribution varies widely as shown in the last experiment. So in future work we can try to create model independent of data distribution.

VII. REFERENCES

1. S. Suthaharan, M. Alzahrani, S. Rajasegarar, C. Leckie, and M. Palaniswami, ‘‘Labelled data collection for anomaly detection in wireless sensor networks,`` in Proc. 6th Int. Conf. Intell. Sensors, Sensor Netw. Inf. Process. (ISSNIP), Dec. 2010, pp. 269–274.
2. Y. Zhang, N. A. S. Hamm, N. Meratnia, A. Stein, M. van de Voort, and P. J. M. Havinga, ‘‘Statistics-based outlier detection for wireless sensor networks,’’ Int. J. Geograph. Inf. Sci., vol. 26, no. 8, pp. 1373–1392, 2012 Y. Zhang, N. A. S. Hamm, N. Meratnia, A. Stein, M. van de Voort, and P. J. M. Havinga, ‘‘Statistics-based outlier detection for wireless sensor networks,’’ Int. J. Geograph. Inf. Sci., vol. 26, no. 8, pp. 1373–1392, 2012
3. H. Feng, L. Liang, and H. Lei, ‘‘Distributed outlier detection algorithm based on credibility feedback in wireless sensor networks,’’ IET Commun., vol. 11, no. 8, pp. 1291–1296, Jun. 2017.
4. Z. Wang, G. Sang, C. Gao ,”An Isolation-Based Distributed Outlier Detection Framework Using Nearest Neighbor Ensembles for Wireless Sensor Networks” IEEE Vol 7, 2019.

AUTHORS

First Author – Siddharth Nahar, B.Tech CSE, IIT Ropar, 2016csb1043@iitrpr.ac.in
Second Author – Aditya Tiwari , B. Tech CSE, IIT Ropar, 2016csb1029@iitrpr.ac.in

Supervisor – Sujata Pal, sujata@iitrpr.ac.in, Assistant Professor IIT Ropar