A Novel Additive Internet of Things (IoT) Features and Convolutional Neural Network for Inter-Class Classification and Source Identification of IoT Devices

*Aamo Iorliam[1], Jessy Akaabo[2]
[1]Department of Mathematics/Computer Science, Benue State University, Makurdi, aamoiorliam@gmail.com
[2]Department of Mathematics/Computer Science, Benue State University, Makurdi, akaabojessy@yahoo.com
*Corresponding author: Aamo Iorliam

Abstract

The inter-class classification and source identification of IoT devices has attracted a great deal of attention due to the vast amount of available IoT devices and the huge amount of data these IoT devices generate almost every minute. As such there is every need to identify the source where the IoT data is generated and also separate an IoT device from the other based on the data they generate. This paper proposes a novel additive IoT features with the CNN system for the purpose of IoT source identification and classification. Experimental results shows that indeed the proposed method is very effective achieving an overall classification and source identification accuracy of 99.67 %. This result has a practical application to forensics purposes due to the fact that accurately identifying and classifying the source of an IoT device via the generated data can link organisations/persons to the activities they perform over the network. As such ensuring accountability and responsibility by IoT device users.

Key words: Internet of Things (IoT), Additive IoT Features**,** Inter-class classification, Source identification.

1. Introduction

The concept of inter-class classification was described by Iorliam, Ho, Waller, and Zhao [1] to mean the separability of biometric images that are not closely related and are generated by different devices. This concept is extended to the Internet of Things (IoTs) in order to perform the inter-class classification and source identification of IoT devices based on the data they generate.
IoT device source identification is concerned with determining which device has produced a particular IoT data. Source identification of devices is very important because it has the tendency of identifying devices within an organization and also unauthorized devices that are connected to the network of such an organization [2,3].
The concept of "Additive IoT features" is motivated from the concept of flow size difference proposed by Iorliam [4] as a network traffic feature for network traffic analysis. Flow size difference took into consideration the absolute values achieved by subtracting two adjacent flows. For the fact that subtraction and addition is associative, this paper extends this concept into the Additive IoT features, where two adjacent IoT values of a feature are added for the purpose of classification and source identification for the first time.
Convolutional Neural Network (CNN) is a powerful machine learning technique that has applications in images, network traffic analysis, document analysis, internet of things, amongst several other applications. Based on its huge capabilities, it is adapted for usage in this paper.
In this paper, the novel use of Additive IoT features and CNN for inter-class classification and source identification of IoT devices based on the benign data they generate is proposed.
Studies such as Bai *et al.* [5], Cvitić, Peraković, Periša, and Gupta [6], Zahid *et al.* [7], Zarzoor, Al-Jamali, and Al-Saedi [8], and Koball *et al.* [3] have proposed methods aimed at classifying IoT, however, our novel approach proposes a novel "Additive IoT features" and achieves an accuracy that is similar or greater than the existing state-of-the-art proposed methods.
This paper contributes to the area of IoT device classification and source identification as follows:

i. To the best of the researchers knowledge, this is the first time Additive IoT features are proposed as a stable IoT metric that could be utilized for classification purposes.
ii. This paper proposes the novel device classification and source identification of IoT devices based on Additive IoT features and CNN.

iii. The novel proposed approach is very simple and free from the overhead of feature engineering.

2. Literature Review

Classification and source identification of IoT devices has attracted huge attention recently. Most of the literature is focused on identifying and proposing new features that can effectively be used for classification and source identification purposes. While some literature is focused on developing/utilizing machine learning approaches in performing classification and source identification of IoT devices.

In this paper, a detail review is performed based on two areas, namely: feature extraction approaches for classification and source identification of IoT devices and Machine learning approaches for classification and source identification of IoT devices.

Bai *et al.* [5] used the flows from 15 devices categorized into 4 classes for the purpose of classifying seen and unseen IoT devices. They used the LSTM-CNN technique for the classification of IoT devices and achieved an accuracy of 74.8%.

Cvitić, Peraković, Periša, and Gupta [6] used 13 network traffic features to perform the classification of IoT devices. These devices were classified into 4 major classes using their proposed multiclass classification model and achieved an accuracy of 99.79%.

Kotak, and Elovici [2] used grayscale snapshots of payloads of TCP sessions that are exchanged between IoT devices as features. The authors used the deep learning technique to identify known IoT devices and unknown IoT devices. They achieved an accuracy of 99% for identifying known devices and 99% for detecting unknown devices using the proposed deep learning technique.

Zahid *et al.* [7] achieved optimal features by performing recursive feature elimination and utilized features such as MAC address of the source, Port number of source, MAC address of destination, Destination port number, ID of network flow, Protocols for communication (6 and 17), Total forward packets, Total duration of network flow, Total number of packets backward, Length of all backward packets, Length of all forward packets, Max length of a forward packet from forward packets, Min length of a forward packet from forward packets, Mean of forward packets, STD of forward packets length, Greater length of a backward packet from all backward packets, Mean (average) length of backward packets, STD of backward packets length, Traffic flow in bytes/second, Packets flow/second, Length of header of forward packets, and Length of header of backward packets. They used the hierarchical deep neural networks with the features stated earlier and achieved a classification accuracy of 91% for the classification of IoT devices from non-IoT devices, and a classification accuracy of 91.33% for the classification of only IoT devices within a heterogeneous network.

Zarzoor, Al-Jamali, and Al-Saedi [8] utilized features such as packet intermediate time among two sequential packets receptions, packet length, IP source address, IP destination address, protocol utilized by the flow, source port number, destination port number, window size, source MAC address and heights number of hop that required for each packet to reach destination. The authors proposed a spike neural network to classify IoT devices. They showed that the proposed model consumed less energy and was able to perform IoT classification with a Precision of 0.98, Recall value of 0.97, and F1-score of 0.98.

Koball *et al.* [3] used 242 features from 8 IoT devices and achieved the highest classification accuracy of 96.5% using unsupervised machine learning techniques.

From the above reviewed literature, this is the first time additive IoT features will be proposed and fed as inputs into CNN to perform inter-class classification and source identification of IoT devices.

3.0 Methodology

This section first describes the dataset used and the preprocessing performed on the dataset. It further vividly describes the proposed Additive IoT Features for IoT device classification and source identification (AIFID). Furthermore, it explains the proposed model architecture and the evaluation metrics used in this paper.

3.1 Dataset and Dataset Pre-Processing

First, the study utilised the N-BaIoT dataset that is made of 9 IoT devices. The 9 devices include; Danmini Doorbell, Ennio Doorbell, Ecobee Thermostat, Philips B120N/10 Baby Monitor, Provision PT-737E Security Camera, Provision PT-838 Security Camera, SimpleHome XCS7-1002-WHT Security Camera, SimpleHome XCS7-1003-WHT Security Camera, and Samsung SNH 1011 N Webcam that produced benign data to include: 49548, 3910, 13113, 17524, 62154, 98514, 46585, 1952, and 52150 instances, respectively [9]. This traffic data collected using 9 IoT devices has also infected dataset with Mirai and BASHLITE. The benign N-BaIoT dataset is suitable for experimenting the proposed AIFID model because it can aid us to perform IoT device classification and source identification.

For the pre-processing, all NULL values were dropped using the Python "dropna" method.

The "MinMaxScaler()" Python command is used to scale each element of the features used in this experiment. The pre-processed dataset is split into 70 % train and 30% test sets. The train set is used to train the CNN learning model. While, the test set serves as input to test the performance of the model. The performance outcome of the model is then evaluated and the results presented as confusion matrix, F1-score, accuracy, precision and recall.

3.2 Additive IoT Features for IoT Device Classification and Source Identification

The additive IoT features are defined as the numeric sum of two consecutive adjacent IoT generated data as illustrated in Table 1.

Table 1: Sample Data For Additive IoT Features

S/ No	Additive_ MI_dir_L5_weight	MI_dir_L5_weight	Additive_ MI_dir_L5_mean	MI_dir_L5_ mean	Additive_ MI_dir_L5_variance	MI_dir_L5_variance
1.	2	1	414	60	0	0
2.	2.857878541	1	714.4589798	354	35.78933753	0
3.	2.857878541	1.857878541	697.4589798	360.4589798	35.78933753	35.78933753
4.	2.680222861	1	509.1409171	337	18487.44875	0
5.	4.284299211	1.680222861	306.2104017	172.1409171	32200.47372	18487.44875
6.	5.707896692	2.60407635	253.9405041	134.0694846	23432.06087	13713.02497

Additive IoT features have a background from the flow size difference proposed by Iorliam [4]. It has been proven from the literature that the flow size difference (flow subtraction) is a stable feature for network traffic classification and intrusion detection [4,10]. In our study, the "additive IoT features" is introduced for the first time for IoT classification and source identification purposes. This is inspired by the fact that addition and subtraction both share a closure property.

For that reason, if Iorliam [4] and Sethi *et al.* [11] used the flow size difference as features for network traffic analysis and intrusion detection purposes, and achieved their targeted goal of intrusion detection, then our proposed additive IoT features for IoT device classification and source identification would be very efficient and effective.

3.3 CNN for IoT Device Classification and Source Identification

Generally, CNN in terms of performance is very efficient in solving machine learning tasks [12].
CNN has proven over the years to be very effective in classification tasks especially when the datasets is huge. In our study we chose the CNN due to the fact that it is has the tendency to automatically select the best features in a particular dataset and has proven to achieve high accuracies.
The steps used for the novel proposed classification and source identification of IoT devices are as follows:

i. Get ALL the 115 statistical features from the IoT device dataset
ii. Calculate the IoT features addition (additive IoT features)
iii. Feed values from (ii) into the CNN classifier
iv. Perform classification

The 9-class classification and source identification is performed by merging the 115 benign IoT features for all the 9 IoT devices and labeling them from 0 to 8 as class labels. These features are fed into the CNN as shown below:

i. 70% of the IoT dataset is used for training, while 30% of the dataset is used for testing.
ii. The first layer used in this experiment is the sequential model "sequential ( )" which allows the network to be build layer by layer and it's well suited for our experiment.
iii. 480 neurons were used in the first hidden layer with 115 input parameters. The rectified linear activation function (ReLu) is first chosen due to its ability to achieve higher performance and again it is non-linear.
iv. Other two dense layers were added which had 240 and 120 neurons, accordingly.
v. The model ended with 9 dense layers, no activation, and a sigmoid activation function.
vi. The model is compiled using the binary crossentropy as loss, the adam as an optimizer, and accuracy as the metrics.

vii. 1000 epochs were used in this experiment with a batch size of 128.

3.4 Model Architecture

The proposed Additive IoT Features for IoT Device Classification and Source Identification (AIFID) is presented in Figure 1.

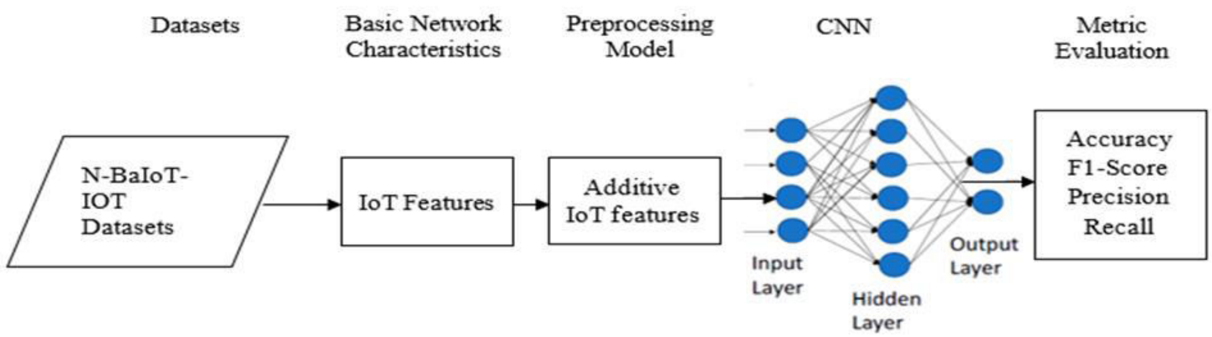

Figure 1: IoT-Based Additive Features for Classification and Source Identification Architecture

In Figure 1, the framework consists of five phases which include: (i) Selecting the suitable dataset (N-BaIoT Datasets) for the experiments; (ii) Utilising the basic network characteristics (IoT Features) for experimentation; (iii) Proposed preprocessing model (Additive IoT Features) from the IoT features; (iv) Adapt the CNN Model for IoT classification and source identification; and (v) Metric Evaluation (Accuracy, F1-Score, Precision and Recall). These phases are carefully followed to implement the proposed AIFID model.

3.5 Evaluation Measures

This study leverages on the strengths of Accuracy, F1 score, Precision, and Recall metrics to evaluate the effectiveness of the proposed Additive IoT Features for IoT device classification and source identification (AIFID).

4.0 Results/Discussions

This section of the study presents and discusses the experimental outcomes of the proposed IoT device classification and source identification model. The CNN model was trained and tested using the N-BaIoT dataset. These results are presented with clear discussion in two perspectives as shown below.

4.1 Performance Results of the CNN Classifier

First, Figure 2 depicts the training loss vs Epochs for the CNN classification and identification of IoT devices. It could be observed that epochs after 200 achieved relatively low loss values. When the loss values becomes very low, it means our proposed model learned properly.

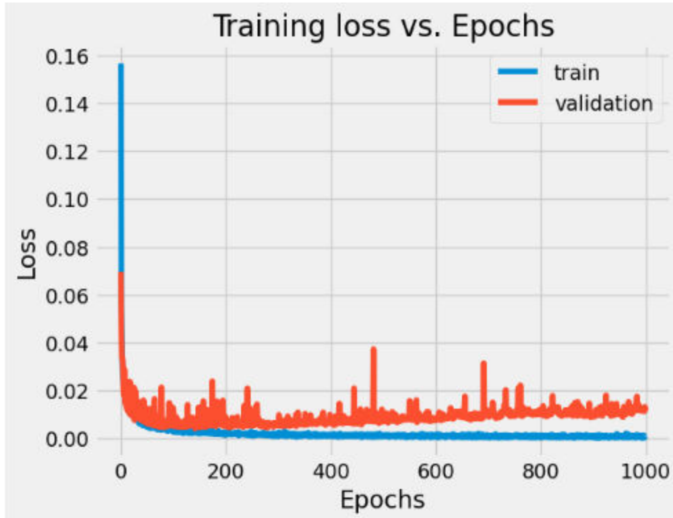

Figure 2: The Training Loss Vs Epochs of the CNN Model on N-BaIoT Datasets.

In Figure 3, as the Epochs increases especially after 200, the training and validation accuracy increased closely to 1.00 (100%). An accuracy very close to 100% shows that the proposed model was correctly trained.

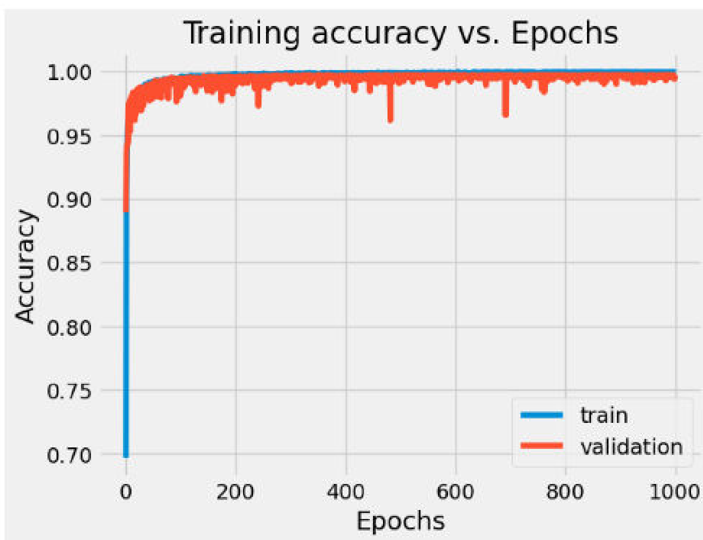

Figure 3: The Training Accuracy Vs Epochs of the CNN Model

The performance of the proposed model in terms of the confusion matrix is depicted in Figure 4. The CNN algorithm was fed with the 115 features of the N-BaIoT dataset for experimental purposes. The 9-class confusion matrix comprising of 9 IoT devices confirms that the CNN model achieved excellent identification and classification results for the IoT devices with overall accuracy of 99.67 %.

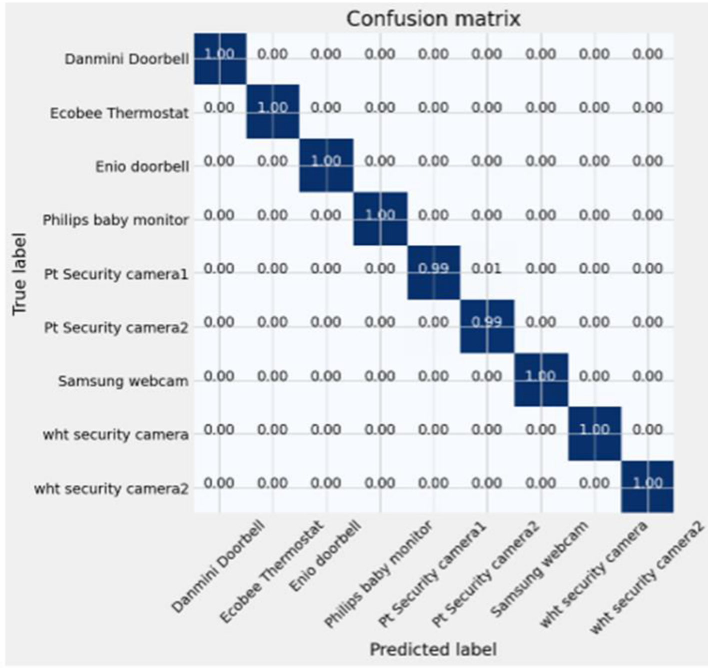

Figure 4: Confusion Matrix for the CNN Model

From Figure 4, it is clear that devices such as Danmini Doorbell (d1), Ecobee Thermostat (d2), Enio doorbell (d3), Philips baby monitor (d4), Samsung webcam (d7), wht security camera (d8), and wht security camera2 (d9) were all correctly identified and classified at an accuracy of 100 percent. Whereas Pt Security camera1 (d5), and Pt Security camera2 (d6) were all identified and classified at an accuracy of 99.0 percent.

Furthermore, Table 2 provides a comprehensive summary of Precision, Recall and F1-score results for the various IoT devices considered in this study. These results clearly demonstrate that the CNN model unambiguously understood the N-BaIoT dataset utilized in the study and accurately identified and classified them.

Comparatively, the proposed AIFID with the overall accuracy of 99.67 % performs at par with existing state-of-the-art models such as Cvitić, Peraković, Periša, and Gupta [6] where they achieved the highest IoT device classification accuracy of 99.79%. This analysis illustrates that the proposed model understands the N-BaIoT dataset and it can effectively and efficiently perform IoT device classification and source identification.

Table 2: Results Summary of the other Evaluation Metrics

IoT Device ID	IoT Devices	Precision	Recall	F1-Score
d1	Danmini Doorbell	1.00	1.00	1.00
d2	Ecobee Thermostat	1.00	1.00	1.00
d3	Enio doorbell	1.00	1.00	1.00
d4	Philips baby monitor	1.00	1.00	1.00
d5	Pt Security camera1	0.99	0.99	0.99

d6	Pt Security camera2	0.99	0.99	0.99
d7	Samsung webcam	1.00	1.00	1.00
d8	wht security camera	1.00	1.00	1.00
d9	wht security camera2	1.00	1.00	1.00

4.0 Conclusion

In this study, a novel Additive IoT Features for IoT device classification and source identification (AIFID) is presented. This model leveraged on the features of the N-BaIoT dataset. The dataset was fed to the CNN learning model. Usually, evaluation metrics are used to assess the effectiveness of a model. Thus, the study employed the Accuracy, F1-Measure, and Precision including Recall to measure the efficiency of the proposed CNN model. The performance results of the proposed AIFID was presented and discussed and as well compared to the state-of-the-art IoT device classification technique proposed by Cvitić, Peraković, Periša, and Gupta [6]. The experimental performance results of the AIFID model performs favourably well with existing models. This study has shown that the Additive IoT Features for IoT device classification and source identification is very effective. The study addresses the rarity of a model to classify and identify devices sources. In the future, the researchers hope to experiment and get the best features for IoT device classification and improve on the performance accuracies as well.

References

1. Iorliam, A., Ho, A. T., Waller, A., & Zhao, X. (2017). Using benford's law divergence and neural networks for classification and source identification of biometric images. In *Digital Forensics and Watermarking: 15th International Workshop, IWDW 2016, Beijing, China, September 17-19, 2016, Revised Selected Papers 15* (pp. 88-105). Springer International Publishing.
2. Kotak, J., & Elovici, Y. (2021). IoT device identification using deep learning. In *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) 12* (pp. 76-86). Springer International Publishing.
3. Koball, C., Rimal, B. P., Wang, Y., Salmen, T., & Ford, C. (2023). IoT Device Identification Using Unsupervised Machine Learning. *Information*, *14*(6), 320.
4. Iorliam, A. (2016). *Application of power laws to biometrics, forensics and network traffic analysis*. University of Surrey (United Kingdom).
5. Bai, L., Yao, L., Kanhere, S. S., Wang, X., & Yang, Z. (2018, October). Automatic device classification from network traffic streams of internet of things. In *2018 IEEE 43rd conference on local computer networks (LCN)* (pp. 1-9). IEEE.
6. Cvitić, I., Peraković, D., Periša, M., & Gupta, B. (2021). Ensemble machine learning approach for classification of IoT devices in smart home. *International Journal of Machine Learning and Cybernetics*, *12*(11), 3179-3202.
7. Zahid, H. M., Saleem, Y., Hayat, F., Khan, F. Z., Alroobaea, R., Almansour, F., ... & Ali, I. (2022). A framework for identification and classification of IoT devices for security analysis in heterogeneous network. *Wireless Communications and Mobile Computing*, *2022*.
8. Zarzoor, A. R., Al-Jamali, N. A. S., & Al-Saedi, I. R. (2023). Traffic Classification of IoT Devices by Utilizing Spike Neural Network Learning Approach. *Mathematical Modelling of Engineering Problems*, *10*(2).
9. Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Shabtai, A., Breitenbacher, D., & Elovici, Y. (2018). N-BaIoT—network-based detection of IoT botnet attacks using deep autoencoders. *IEEE Pervasive Computing*, *17*(3), 12-22.
10. Iorliam, A., Tirunagari, S., Ho, A. T., Li, S., Waller, A., & Poh, N. (2016). " Flow Size Difference" Can Make a Difference: Detecting Malicious TCP Network Flows Based on Benford's Law. *arXiv preprint arXiv:1609.04214*.
11. Sethi, K., Sai Rupesh, E., Kumar, R., Bera, P., & Venu Madhav, Y. (2020). A context-aware robust intrusion detection system: a reinforcement learning-based approach. *International Journal of Information Security*, *19*, 657-678.
12. Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)* (pp. 1-6). IEEE.